Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Technically co-sponsored by

Financially co-sponsored by

Platinum Sponsor

Gold Sponsor

Silver Sponsor

Organized by

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

ISBN: 978-1-5090-0202-3

IEEE Catalog Number: CFP16A38-CDR

Proceedings of 2016 International Conference on Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

ISBN : 978-1-5090-0202-3

IEEE Catalog Number : CFP16A38-CDR

Contact Address:

Head of the Department
Department of Computer Science and Engineering
ECE Building, West Polashi
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh
Telephone: 880-2-9665612
PABX: 880-2-55167100, 55167228-57 Ext: 6432
Fax: 880-2-9665612
E-mail: headcse@cse.buet.ac.bd

Message from the Technical Program Committee (TPC) Co-Chairs

We take this opportunity to welcome you all to the 2016 International Conference on Networking Systems and Security (NSysS) in purview of exchanging the latest research ideas and results in networking systems and security. The theme of the conference includes recent advances in both theoretical and experimental research addressing the broader aspect of computer networks, networking systems, and security across academia and industry. To harvest the benefit of information technology, developing countries like Bangladesh will need outstanding researchers in the field of networking. This conference aims to bring the opportunity to get immensely benefited from knowing the trends of technological advancement and through assimilating it to develop our economy.

NSysS 2016, a three-day conference, entails a diverse program including keynote and invited speeches, industry talks, panel discussion, full and short original research paper presentations and student activities session. We strongly believe that this conference will have far reaching impact on research in this field by inciting lots of interaction and enthusiasm among its participants, both young and experienced researchers, through enhancing their current knowledge and understanding of the topics discussed in this conference.

NSysS 2016 invited full and short research papers from academia, industries and research centers around the world. A total of 43 submissions were received and reviewed. The technical program committee comprised of 32 members from 5 different countries including 9 external reviewers. After a rigorous review based on novelty and technical merit, only 14 full papers (out of 36 with the acceptance rate of about 39%) and 8 short papers have been selected for oral presentation. We also have received three invited papers out of which one has been accepted.

The conference features keynote speeches from three outstanding researchers: Dr. Mohammed Atiquzzaman (University of Oklahoma, USA), Dr. Ying-Dar Lin (National Chiao Tung University, Taiwan), and Dr. Vijay Raghunathan (Purdue University, USA). Invited speakers include Dr. Javed I. Khan (Kent State University, USA), Dr. Shahriar Nirjon (University of North Carolina, USA), Dr. Jelena Misic (Ryerson University, Canada) and Dr. Mahbubur Rahman Syed (Minnesota State University, USA).

NSysS 2016 could not be successful without sincere efforts exerted by its organizing members. We thank the members of the organizing committee. We also thank its program committee members and advisory members for their supports all the way down to this very day.

We would specially like to thank our technical co-sponsors, BUET ACM Chapter and IEEE Bangladesh Section.

We wish all the success of this conference.

Dr. A.K.M. Ashikur Rahman
Professor, Dept. of CSE, BUET,
&
TPC Co-Chair, NSysS 2016.

Dr. Mohammed Atiquzzaman
Professor, University of Oklahoma, USA,
&
TPC Co-Chair, NSysS 2016.

Message from the Organizing Committee Co-chairs

It gives us immense pleasure to welcome everyone to the 2016 International Conference on Networking Systems and Security (NSysS). We greet experts, academicians, researchers, entrepreneurs and students from diverse grounds who have come together to discuss the issues of concern to state of the art technology and contribute to the enrichment of human civilization. This conference will definitely give us the opportunity to know the trends of technological advancement in different areas of computer networks, networking systems, and security. NSysS 2016 gives researchers and practitioners a unique opportunity for sharing their perspectives with others interested in the various aspects of networking systems. Findings of this conference will open new areas of research and pave the way for facing upcoming challenges.

NSysS 2016 presents a series of important sessions. Our keynote speaker list includes three esteemed researchers: Dr. Mohammed Atiquzzaman (University of Oklahoma, USA), Dr. Ying-Dar Lin (National Chiao Tung University, Taiwan), and Dr. Vijay Raghunathan (Purdue University, USA). NSysS 2016 also includes three invited speeches featuring both academic and industry aspects of networking systems research. The invited speakers are Dr. Javed I. Khan (Kent State University, USA), Dr. Shahriar Nirjon (University of North Carolina, USA), Dr. Jelena Misic (Ryerson University, Canada) and Dr. Mahbubur Rahman Syed (Minnesota State University, USA). Besides, as an important part of our conference, full and short original research paper presentations provide the opportunity to researchers to exhibit their recent research outcomes.

It is our great pleasure to acknowledge that putting together all these events of NSysS 2016 was a great team effort. We first thank our keynote and invited speakers for participating in our conference. We are grateful to the authors for providing the content of the program. It gives us great pleasure to thank the technical program committee Co-chairs of NSysS 2016, whose contribution was immense to make this event successful. We also thank the reviewers for their hard work in timely reviewing papers and providing their valuable feedback for authors. Finally, we thank the Organizing Committee for their relentless effort.

We would specially like to thank our platinum sponsors, HUAWEI Technologies Co. Limited, TigerIT Bangladesh Limited, our gold sponsor, Dohatec and silver sponsor, REVE Systems. We would also like to thank our technical co-sponsors, BUET ACM Chapter and IEEE Bangladesh Section. Last but not the least, we thank the department of CSE, BUET for hosting the conference.

We have the very best wishes for the grand success of this conference.

Dr. Mohammad Mahfuzul Islam
Professor & Head, Dept. of CSE, BUET,
&
General Co-Chair, NSysS 2016.

Dr. M. Kaykobad
Professor, Dept. of CSE, BUET,
Dean, Faculty of EEE, BUET,
&
General Co-Chair, NSysS 2016.

NSysS 2016 Committees

International Advisory Committee

Mohammed Atiquzzaman, University of Oklahoma, USA
Raouf Boutaba, University of Waterloo, Canada
Ekram Hossain, University of Manitoba, Canada
Mohammad Ataul Karim, University of Massachusetts Dartmouth, USA
M. Kaykobad, BUET, Bangladesh
Manzur Murshed, Federation University, Australia
Suman Kumar Nath, Microsoft Research, USA
M. Sohel Rahman (Convener), BUET, Bangladesh

Technical Program Committee Co-Chairs

Mohammed Atiquzzaman, University of Oklahoma, USA
A.K.M. Ashikur Rahman, BUET, Bangladesh

Technical Program Committee Members

Nova Ahmed, North South University, Bangladesh
Khawza Iftekhar Uddin Ahmed, United International University, Bangladesh
S M Iftekharul Alam, Intel Labs, USA
Kazi Md. Rokibul Alam, Khulna University of Engineering & Technology, Bangladesh
Mohammed Eunus Ali, BUET, Bangladesh
Pulak Chowdhury, University of California, Davis, USA
S. M. Farhad, BUET, Bangladesh
Shaikh Anowarul Fattah, BUET, Bangladesh
Tanzima Hashem, BUET, Bangladesh
Mohammad A. Hoque, East Tennessee State University, USA
Md Shohrab Hossain, BUET, Bangladesh
A. B. M. Alim Al Islam, BUET, Bangladesh
Salekul Islam, United International University, Bangladesh
Tanzima Islam, Lawrence Livermore National Laboratory, USA
Joarder Kamruzzaman, Monash University, Australia
Maleq Khan, Virginia Tech, USA
Mohammad Mannan, Concordia University, Canada
Manzur Murshed, Federation University, Australia
Suman Kumar Nath, Microsoft Research, USA
Mahmuda Naznin, BUET, Bangladesh
Shahriar Nirjon, University of North Carolina at Chapel Hill, USA
Rajesh Palit, North South University, Bangladesh
Al-Sakib Khan Pathan, International Islamic University, Malaysia
Farzana Rahman, James Madison University, USA
Md. Saidur Rahman, BUET, Bangladesh

M. Sohel Rahman, BUET, Bangladesh
Mohammad Rashedur Rahman, North South University, Bangladesh
Md. Abdur Razzaque, University of Dhaka, Bangladesh
Pran Kanai Saha, BUET, Bangladesh
Md Yusuf Sarwar Uddin, BUET, Bangladesh
Mohammad Shorif Uddin, Jahangirnagar University, Bangladesh
Salmin Sultana, Intel Labs, USA
Yun Wang, Bradley University, USA
Mohammad Zulkernine, Queens University, Canada

Organizing Committee

General Chairs

Mohammad Mahfuzul Islam, BUET, Bangladesh
M. Kaykobad, BUET, Bangladesh

Publicity and Sponsorship Chairs

Md. Mostofa Akbar, BUET, Bangladesh
Md. Humayun Kabir, University of Victoria, Canada
Mohammed Eunus Ali, BUET, Bangladesh

Local Organizing Chairs

Abu Sayed Md. Latiful Hoque, BUET, Bangladesh
Md. Yusuf Sarwar Uddin, BUET, Bangladesh
Anindya Iqbal, BUET, Bangladesh

Panel Chairs

Mahmuda Naznin, BUET, Bangladesh
Sadia Sharmin, BUET, Bangladesh

Tutorial Chairs

Tanzima Hashem, BUET, Bangladesh
Md. Abdur Razzaque, University of Dhaka, Bangladesh

Keynote/Industry Speakers Chair

Md. Shohrab Hossain, BUET, Bangladesh

Student Activities Chair

S. M. Farhad, BUET, Bangladesh

Md. Mustafizur Rahman, University of Dhaka, Bangladesh

Poster and Demo Chair

Md. Monirul Islam, BUET, Bangladesh

Registration Chair

Abu Wasif, BUET, Bangladesh

Finance Chair

Khaled Mahmud Shahriar, BUET, Bangladesh

Web Chair

Rifat Shahriyar, BUET, Bangladesh

Publication Chair

Mohammad Saifur Rahman, BUET, Bangladesh

Conference Coordinator

A. B. M. Alim Al Islam, BUET, Bangladesh

Members

Mohammed Kaysar Abdullah, BUET, Bangladesh

Toufique Ahmed, BUET, Bangladesh

Abdus Salam Azad, BUET, Bangladesh

Md. Aashikur Rahman Azim, BUET, Bangladesh

Madhusudan Basak, BUET, Bangladesh

Siddhartha Shankar Das, BUET, Bangladesh

Tanvir Ahmed Khan, BUET, Bangladesh

Johra Muhammad Moosa, BUET, Bangladesh

Ishat E Rabban, BUET, Bangladesh

Radi Muhammad Reza, BUET, Bangladesh

Md. Iftekharul Islam Sakib, BUET, Bangladesh

Tanmoy Sen, BUET, Bangladesh

Fatema Tuz Zohora, BUET, Bangladesh

Table of Contents

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Invited Paper

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Comparative Study of Different Methods of Social Network Analysis and Visualization

Md. Marufur Rahman
Technical Lead, Samsung R&D Institute Bangladesh Ltd.
Dhaka, Bangladesh
maruf.rh@samsung.com

Dr. Md. Rezaul Karim
Professor, Department of Computer Science & Engineering
University of Dhaka, Dhaka, Bangladesh
rkarim@univdhaka.edu

***Abstract*—A Social Network is a social structure made of individuals or organizations that are linked by one or more specific types of interdependency, such as friends, kinship, terrorist relation, conflict, financial exchange, disease transmission (epidemiology), airline routes etc. Social Network Analysis is an approach to the study of human or organization social interactions. It can be used to investigate kinship patterns, community structure or the organization of other formal and informal social networks. Recently, email networks have been popular source of data for both analysis and visualization of social network of a person or organization. In this paper, three types of visualization technique to analyze social networks have been considered - Force directed layout, Spherical layout and Clustered layout. Each method was evaluated with various data sets from an organization. Force directed layout is used to view the total network structure (Overview). Spherical layout is a 3D method to reveal communication patterns and relationships between different groups. Clustered graph layout visualizes a large amount of data in an efficient and compact form. It gives a hierarchical view of the network. Among the three methods, Clustered layout is the best to handle large network and group relationship. Finally, the comparative study of these three methods has been given.**

Keyword— Social Network; Force directed layout; Spherical layout; Clustered graph layout.

I. INTRODUCTION

A *Social Network* [2, 3] is a map of all of the relevant ties (links or relationships) between the nodes (individuals or organizations) being studied. The network can also be used to determine the social capital of individual actors. Figure 1 illustrates the email communication of individuals where each circle represents a person and each link between two circles represents email exchange. Here, we can easily observe that Susan have more interaction than others.

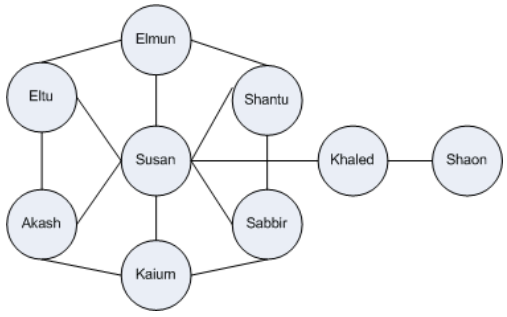


Figure 1: An example of social network

Social Network Analysis (SNA) [1, 3, 4] is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information knowledge processing entities. Social network analysis has now used to investigate the crime, terrorist attack, virus attack and overall the relationship of human and organization to reveal the communication patterns in social life. SNA provides both a visual and a mathematical analysis of complex human systems. Visualization [6, 7, 8] is a technique for creating images, diagrams, or animations to communicate a message. Visualization has been an effective way to communicate both abstract and concrete ideas.

There are various sources of data for social network. They are: email network, blog, portal, questionnaires, direct observation, written records and experiments. Recently, email networks have been very much popular for both analysis and visualization of social network. For example, analysis of email networks was used to identify the informal communication structure within an organization, to discover the shared interests between people and in relation to the spread of computer viruses. In this paper, we consider the email network to analyze and visualize the social network. An email network is derived from an email log file from the email server. In the email network, each node represents an email address and each edge between two nodes represents an email exchange between these two email addresses.

There are different methods for analysis and visualization of the social networks. Force directed layout is used to view the total network structure (Overview). It is simple, easy to implement and suitable for graphs of medium size (up to 50-100 vertices). Spherical layout is used to reveal communication patterns, relationships between different groups Clustered graph layout visualizes a large amount of data in an efficient and compact form. It gives hierarchical view and group relationships. Each method is application specific and each has merits and demerits. But there is no comparative study of these methods in social network analysis. So we cannot get an idea which method is suitable for which scenario. This makes us to study different methods of social network analysis and visualization.

The remainder of the paper is organized as follows. In Section 2, we give the analysis of a social network for an organization. Section 3, 4 and 5 provides three visualization methods of social network: Force directed layout, Spherical layout and Clustered graph layout. The comparative

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

performance analysis of these three methods has been presented in Section 6. Finally Section 7 concludes the paper.

II. STATISTICAL ANALYSIS OF AN EMAIL NETWORK

For our research, the email log file (one day mail log) was collected from the email server of Dhaka Stock Exchange Ltd. (DSE) on 25 November, 2007. That mail server uses postfix as mail transport agent (MTA). For analysis and visualization, we parse the necessary information from the mail log. On the 25 November, 2007 DSE email network has the following statistics:

Total email exchange : 905
Nodes: 369 Edges: 905
Highest Email Sender/Receiver: jramey1@bellsouth.net and email send/Receive: 206
Total email exchange inside DSE: 12
Nodes: 80 Edges: 12
Highest Email Sender/Receiver: ruhul@dsebd.org and email send/Receive: 49

	From	To
2	<kang@stealthpromotions.com>	<shamim@dsebd.org>
3	<randall@mercury-spb.com>	<rafiq@dsebd.org>
4	<juliana@e-business-associates.	<raquib@dsebd.org>
5	<kfnuosh@blsarchitects.com>	<nur@dsebd.org>
6	<gair@plastmarket.com>	<sheuli@dsebd.org>
7	<graham@valukargo.com>	<syed@dsebd.org>
8	<rupert@ltindia.com>	<sheuli@dsebd.org>
9	<cornelius@i-gts.com>	<iqbal@dsebd.org>
10	<cornelius@i-gts.com>	<iqbal@dsebd.org>
11	<cornelius@i-gts.com>	<iqbal@dsebd.org>
12	<sebastiansparrowhawk@cristalb	<rezaur@dsebd.org>
13	<linfmgmedicalmet@fmgmedical.	<shirin@dsebd.org>
14	<Cassidy@aurocon.ro>	<shamim@dsebd.org>
15	<konrad@olgafilippova.com>	<mahmud@dsebd.org>
16	<linccrmet@ccr.de>	<skzaman@dsebd.org>
17	<pablo.ruiz@tm.net.my>	<sheuli@dsebd.org>
18	<G2GIMRYRM3KLITEGPOX4GZE	<sheuli@dsebd.org>
19	<Pollanen@breadmaking.com>	<dseclearing@dsebd.org>
20	<Pollanen@breadmaking.com>	<dseclearing@dsebd.org>
21	<Mai.Gourlay@noegel.de>	<darain@dsebd.org>
22	<Mai.Gourlay@noegel.de>	<darain@dsebd.org>
23	<rzsg@tiscali.nl>	<webmaster@dsebd.org>
24	<dwsoftlabm@softlab.it>	<skzaman@dsebd.org>
25	<huntley@telesensventures.com>	<imam@dsebd.org>
26	<typhoon1302000@tcat.ac.uk>	<eleas@dsebd.org>
27	<nicole@dsecargonet.com>	<sydbhq@dsebd.org>
28	<dsearcy@dsearcy.com>	<anup_bd@dsebd.org>
29	<darain@dsebd.org>	<6fd41251@dsebd.org>
30	<darain@dsebd.org>	<6fd41251@dsebd.org>
31	<darain@dsebd.org>	<6fd41251@dsebd.org>
32	<eleas@dsebd.org>	<casad@dsebd.org>

(a) Sender and Receiver List

Id	Node	Total Email	Adjacent List
0	<kang@stealthpromotions.com>	1	1
1	<shamim@dsebd.org>	28	0 19 110 144 159 159 200 200 202 246 246 246 246 277 277 282 282 290 270 270 295 301 302
2	<randall@mercury-spb.com>	1	3
3	<rafiq@dsebd.org>	14	2 62 157 157 184 277 277 282 282 270 270 302 302 340
4	<juliana@e-business-associates.(	1	5
5	<raquib@dsebd.org>	16	4 54 156 157 157 187 222 252 259 277 277 270 270 302 302 358
6	<kfnuosh@blsarchitects.com>	1	7
7	<nur@dsebd.org>	18	6 89 102 104 123 157 157 253 277 282 282 291 291 291 270 270 302 302
8	<gair@plastmarket.com>	1	9
9	<sheuli@dsebd.org>	21	8 12 24 25 69 118 159 159 200 200 264 277 277 282 282 270 270 297 302 302 324
10	<graham@valukargo.com>	1	11
11	<syed@dsebd.org>	17	10 140 140 147 149 159 159 169 200 200 269 270 270 272 288 299 299
12	<rupert@ltindia.com>	1	9
13	<cornelius@i-gts.com>	3	14 14 14
14	<iqbal@dsebd.org>	11	13 13 13 29 103 157 157 300 336 368 368
15	<sebastiansparrowhawk@cristalb	1	16
16	<rezaur@dsebd.org>	12	15 159 159 200 200 255 277 277 270 270 302 302
17	<linfmgmedicalmet@fmgmedical.(	1	18
18	<shirin@dsebd.org>	23	17 57 67 121 155 159 159 200 200 216 267 277 277 282 282 270 270 302 302 330 364 360 362
19	<Cassidy@aurocon.ro>	1	1
20	<konrad@olgafilippova.com>	1	21
21	<mahmud@dsebd.org>	5	20 72 100 100 250
22	<linccrmet@ccr.de>	1	23
23	<skzaman@dsebd.org>	42	22 32 83 90 109 115 124 124 124 125 131 145 150 150 159 159 170 196 199 199 200 200 209
24	<pablo.ruiz@tm.net.my>	1	9
25	<G2GIMRYRM3KLITEGPOX4GZE	1	9
26	<Pollanen@breadmaking.com>	2	27 27
27	<dseclearing@dsebd.org>	22	26 26 45 50 171 171 201 201 247 265 271 271 285 285 285 285 287 287 325 327 328 335
28	<Mai.Gourlay@noegel.de>	2	29 29
29	<darain@dsebd.org>	34	28 28 41 41 41 44 14 171 171 194 201 201 223 224 224 245 271 271 285 285 285 285 287
30	<rzsg@tiscali.nl>	1	31
31	<webmaster@dsebd.org>	9	30 159 159 200 200 270 270 299 299
32	<dwsoftlabm@softlab.it>	1	23
33	<huntley@telesensventures.com>	1	34

(b) Adjacency list of all nodes

Figure 2: Analysis result of maillot parsing

The comma separated value (CSV) files in Figure 2 are the output of mail log parsing. Figure 2(a) illustrates the entire sender and receiver list. Figure 2(b) shows the adjacency list and total email exchange of each node (email id).

III. FORCE DIRECTED LAYOUT

Force directed layout [9, 10] is used to visualize the overview of a network structure. It is an algorithm for drawing graphs in an aesthetically pleasing way. They are good for achieving the following aesthetic criteria: uniform edge length, uniform vertex distribution and showing symmetry. Their purpose is to position the nodes of a graph in two dimensional or three dimensional spaces so that all the edges are of more or less equal length. The process flow of force directed layout has been given in Figure 3.

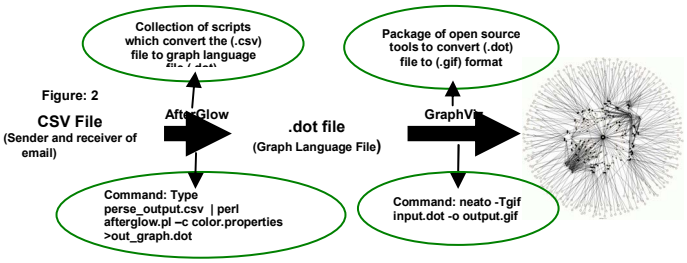


Figure 3: Process flow of force directed layout

From the process flow in Figure 3, the parsing output CSV (Comma Separated Value) file has been parsed using AfterGlow script and then visualizes using a package of open source tools (GraphViz). Figure 4 illustrates the sample output of force directed layout. Ellipse represents sender, rectangle represents receiver and circle is the central node that is connected to all receivers.

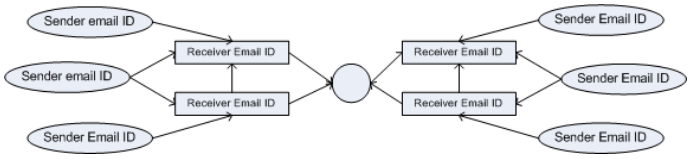


Figure 4: Visualization output of Force directed Method

All the nodes (labeling email id) with sender and receiver are illustrated in figure 5. It gives total overview of the email exchange of Dhaka Stock Exchange Ltd. (DSE) on November 25, 2007. But it is very difficult to get desired information.

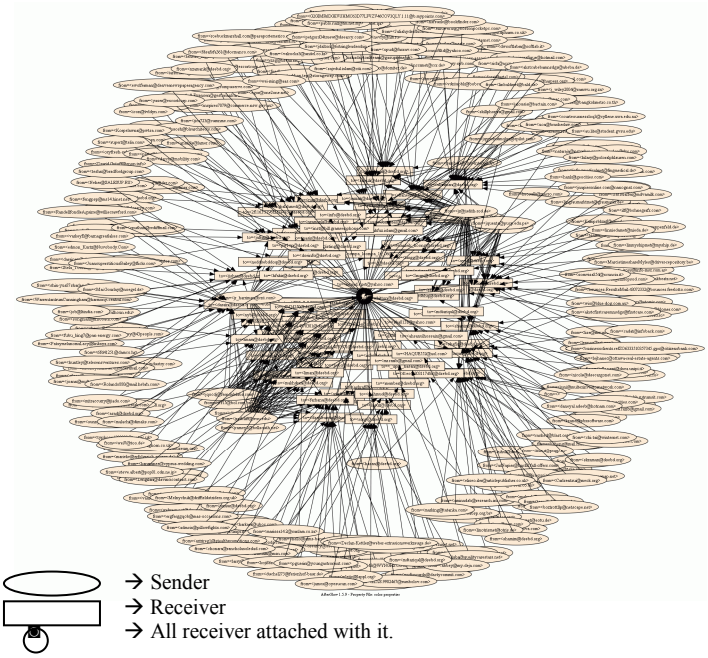


Figure 5: Force Directed Layout with label (DSE Email Network@25/11/2007)

If the label of email Id is omitted then Figure 5 looks like Figure 6 which is more understandable. Circle represents the sender and rectangle represents receiver in Figure 6. All receivers are connected to a central node.

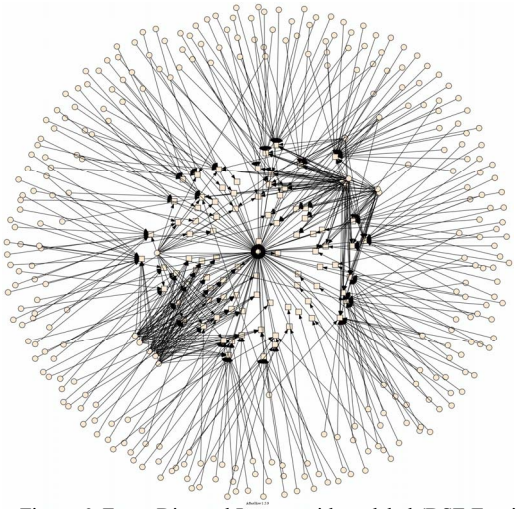


Figure 6: Force Directed Layout without label (DSE Email Network@25/11/2007)

○ → Sender
□ → Receiver
→ Mail exchange direction
→ All receiver attached with it.

Force directed layout is suitable for graphs of medium size (up to 50-100 vertices). While graph drawing is a difficult problem, force-directed algorithms usually require no special knowledge about graph theory such as planarity. It is simple and easy to implement. The running time of force-directed algorithm is O(n3) [9, 10], where n is the number of nodes of the input graph. Complexity per single iteration is O(n2). Classical force directed algorithms are unable to handle larger network due to the inherent n squared cost at each time step.

IV. SPHERICAL LAYOUT

Spherical layout [11] describes a 3D method to visualize an email network on the surface of a sphere. Visualization on the surface of a sphere [12] is used to reveal communication patterns, relationships between different groups. Spherical layout employs Self-Organizing Map (SOM) for this visualization. A self-organizing map (SOM) [13, 14] is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two dimensional) representation of the input space of the training samples. The map seeks to preserve the topological properties of the input space. This makes SOM useful for visualizing low-dimensional views of high-dimensional data.

The algorithm of the spherical layout has been given in [11]. The process flow of that algorithm has been illustrated in Figure 7.

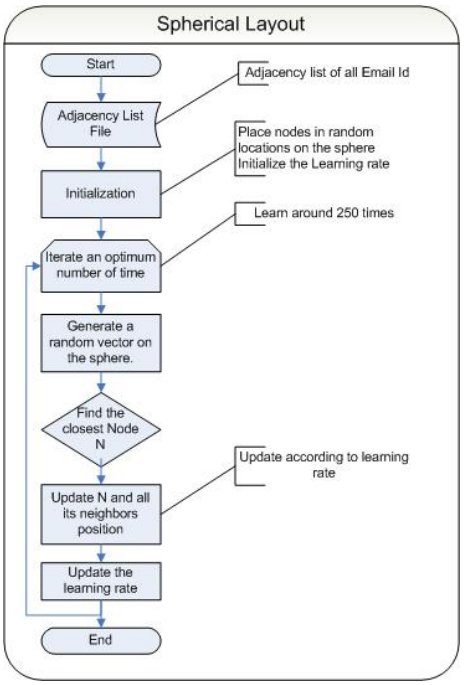


Figure 7: Process flow of Spherical layout.

The Spherical layout has been drawn using visual C++ and OpenGL.

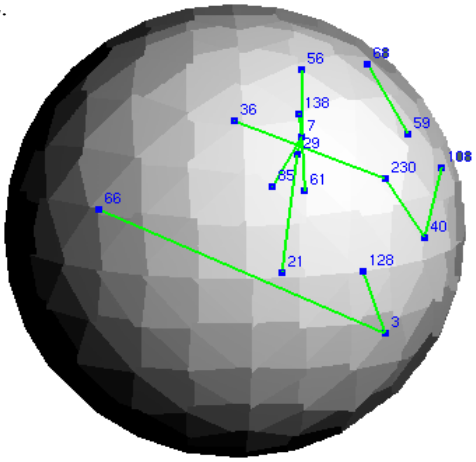


Figure 8: Spherical Layout of four groups of DSE internal email network

Figure 8 illustrates relationship between groups like Application support, Network, Hardware of ICT Division and administration department of Dhaka Stock Exchange Ltd. (DSE). Nodes of each department are given below.

Administration : {3, 66, 128}
Application Support: {7, 21, 29, 56, 61, 85, 138}
Network Dept. : {108, 40, 230, 36}
Hardware Dept. : {68, 59}

Here only edges representing intra-group communications are shown to reduce visual complexity. No inter group edge is shown. Inter-group communications can be observed by the closeness of the groups: the closer the groups, the more communication between them.

People in the ICT division (Application support and Network) have more interaction than that of administration department. This means that they often communicate and collaborate with each other in order to complete a task (see Figure 8). Application support department has some interaction with administration but there is no interaction between hardware and administration. We can easily observe it from the Figure 8. Network and application support department are mixed together to exhibit more communication. Hardware department has interaction with network and application support but not with administration.

SOM sphere layout shows communication patterns between groups more clearly compared to the force directed layout. The nodes are distributed more evenly on the surface of the sphere, instead of collapsing at the center. The spherical surface provides a natural fisheye effect which enlarges the focus point and shows other portions of the image with less detail. The main disadvantage of using the SOM for graph layout is the overlapping between nodes and edges. Another major problem with SOM sphere layout is that they are very computationally expensive.

V. CLUSTERED GRAPH LAYOUT

Clustered graph layout [15] gives a hierarchical view of the network. It shows the clustered structure of the network according to the clustering attribute such as advice relation, friendship relation, who reports to whom, age, department, level, education, gender, and job-ranking. A clustered graph has its vertices grouped into clusters in a hierarchical way via subset inclusion, thereby imposing a tree structure on the clustering relationship. In clustered layout, groups of related nodes are clustered into super-nodes. The user sees a summary of the graph: the super-nodes and super-edges between the super nodes. Some clusters may be shown in more detail than others. Graphs which arise in information visualization applications are typically very large: thousands, or perhaps millions of nodes. Recent graph drawing competitions have shown that visualization systems for classical graphs are limited to (at best) a few hundred nodes.

A clustered graph [15], C = (G, T) consists of an undirected graph G and a rooted tree T such that the leaves of T (1, 2, 3, 4 in Figure 9) are exactly the vertices of G. Each node of T (P, Q in Figure 9) represents a cluster V (v) of the vertices of G that are leaves of the sub tree rooted at v.

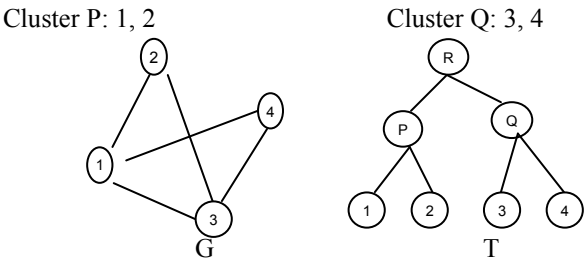


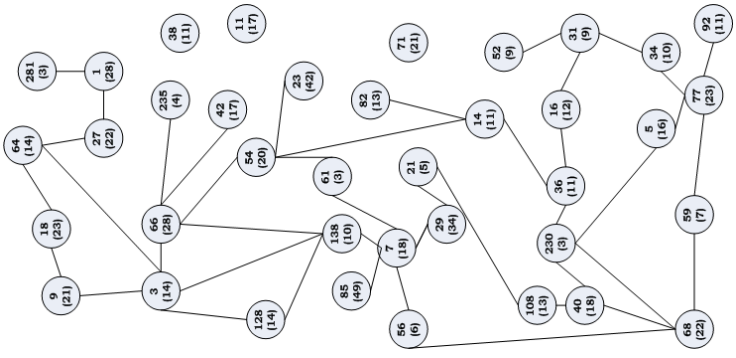
Figure 9 : Cluster graph C = (G, T)

The intra email network of DSE has been clustered according to department. The departments are combined to division (super node) – Information Technology (ICT) Division and Non-Technical Division. The clustering structure of DSE email network is described as follows:

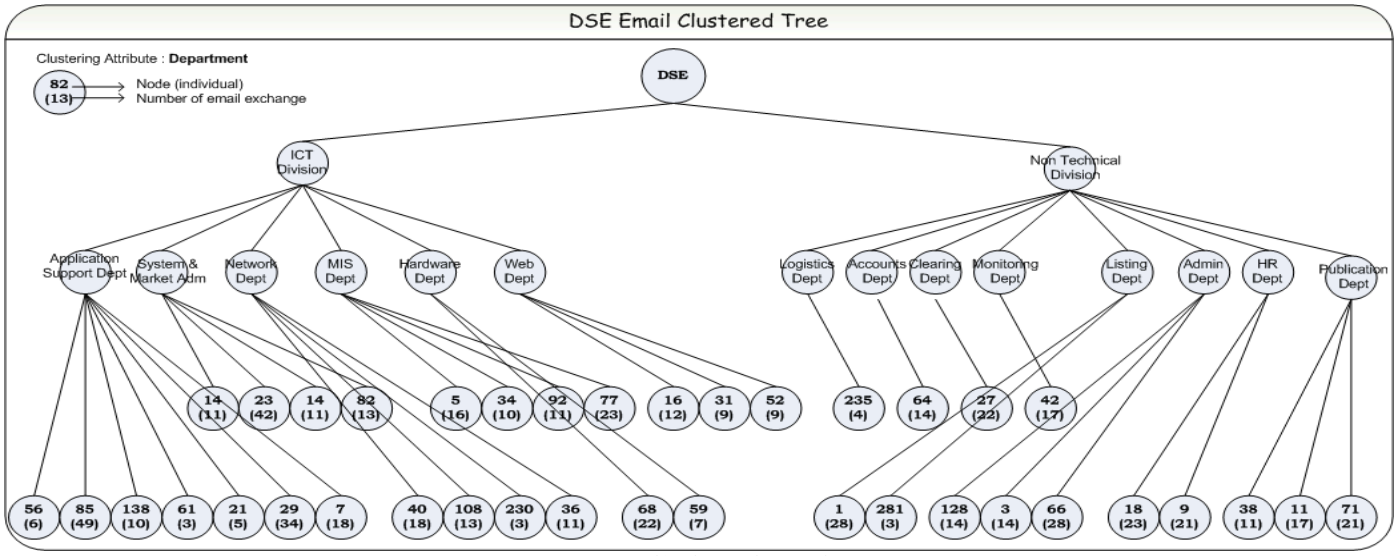
Clustering attribute: Department
Root: {P, Q}
P: {A, B, C, D, E, F, G, H}: Non-Technical Div.
Q: {I, J, K, L, M, N}: Information Technology (IT) Div.

A: {11, 38, 71} : Publication and R&D Dept.
B: {42} : Monitoring Dept.
C: {1, 281} : Listing Dept.
D: {235} : Logistics
E: {9, 18} : HRD
F: {64} : Accounts
G: {27} : DSE Clearing
H: {3, 66, 128} : Administration
I: {56, 7, 29, 21, 85, 138, 61}: Application Support
J: {82, 14, 23, 54}: System and Market Administration
K: {68, 59} : Hardware Dept
L: {5, 77, 34, 92}: MIS Dept
M: {108, 40, 230, 36}: Network Dept.
N: {16, 31, 52}: Web Dept.

Figure 10 illustrates clustered graph of Dhaka Stock Exchange Ltd. (DSE) on November 25, 2007 obtains from the adjacency list and sender and receiver list of Figure 2. Since email exchange inside DSE is few, we assumed some email exchange inside DSE to make this clustering more understandable.



a) Undirected Graph of DSE, G



b) Clustered Tree of DSE, T

Figure 10: Clustered graph C = (G, T) of DSE.

Clustered Graph Layout algorithm:

Clustered_Graph_Embed()

{

For each node v of T in clustered graph C = (G, T) [15],

- Compute upward embedding and the PQ-tree(s) [18, 19] returned by *Upward_Embed [16]*.
- If v is not the root of T then update the upward embedding [16, 17] according to ORD (v) (the circular ordering of edges incident to cluster v in embedding.
- Compute a complete planar embedding [16, 17] Hv of G (v) from the upward embedding using *Entire_Embed* [16].
- Call Procedure Formalize (Hv) [15] to modify Hv, such that every wheel graph [2] F of G(v) is embedded canonically and the vertices and edges of G-F are embedded in the external face (the rim face) of F.
- For each child u of v, find the circular order [15] ORD(u) of the edges incident to cluster u according to Hv.

}

This clustering is done by the algorithm Clustered_Graph_Embed()

- Input: A connected clustered graph C = (G, T) which is shown in figure 10.
- Output: An embedding of C which consists of a circular ordering of the edges incident to each cluster of C which is shows in Figure 11 and Figure 12.

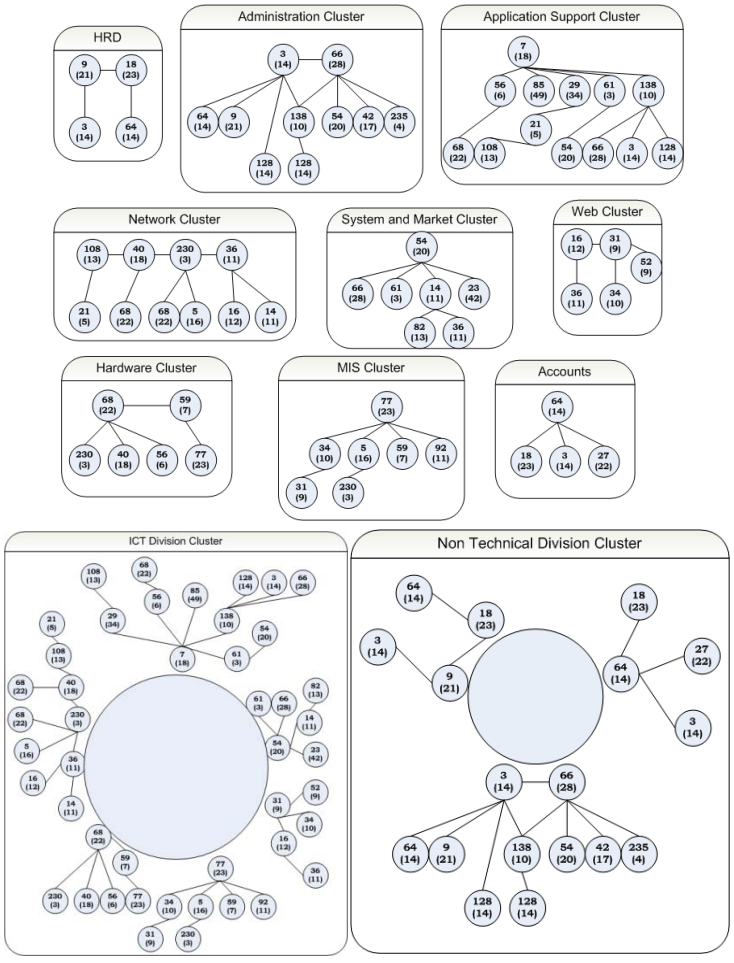


Figure 11: The clustering structure of DSE email network

Algorithm Clustered Graph Embed from [15] obtains an embedding recursively from top to bottom. The circular orderings of edges around the cluster in the embedding are as follows:

E (HRD) : (9, 3), (9, 18), (9, 64);

F (Accounts) : (64, 27), (64, 3), (64, 18);

H (Administration) :(66, 235), (66, 42), (66, 54), (66,138), (66, 3), (3, 138), (128, 138), (3, 128), (3, 9), (3, 64);

I (Application Support) : (7, 138), (138,128), (138, 3), (138, 66), (7, 61), (61, 54), (7.29), (29, 21), (21, 108), (7, 85), (7, 56), (56, 68);

J (System and Market) : (54, 23), (54, 14), (14, 36), (14, 82), (54, 61), (54, 66);

K (Hardware Dept) : (59, 77), (59, 68), (68, 56), (68,56), (68, 40), (68, 230);

L (MIS Dept) : (77, 92), (77, 59), (77, 5), (5,230), (77, 34), (34, 31);

M (Network Dept.) : (36, 14), (36, 16), (36, 230), (230, 5), (230, 68), (230, 40), (40, 68), (40, 108), (108, 21);

N (WEB) : (31, 52), (31, 34), (31, 16), (16, 36);

Q (ICT Division) : (77, 92), (77, 59), (77, 5), (5,230), (77, 34), (34, 31), (59, 77), (59, 68), (68, 56), (68,56), (68, 40), (68, 230), (36, 14), (36, 16), (36, 230), (230, 5), (230, 68), (230, 40), (40, 68), (40, 108), (108, 21), (7, 138), (138,128), (138,3), (138, 66), (7, 61), (61, 54), (7.29), (29, 21), (21, 108), (7, 85), (7, 56), (56, 68), (54, 23), (54, 14), (14, 36), (14, 82), (54, 61), (54, 66), (31, 52), (31, 34), (31, 16), (16,36);

P (Non-Technical Division): (66, 235), (66, 42), (66, 54), (66,138), (66, 3), (3, 138), (128, 138), (3, 128), (3, 9), (3, 64), (9,3), (9,18), (9,64), 64, 27), (64,3), (64,18);

We now draw the graph according to this circular ordering which is shown in Figure 11.

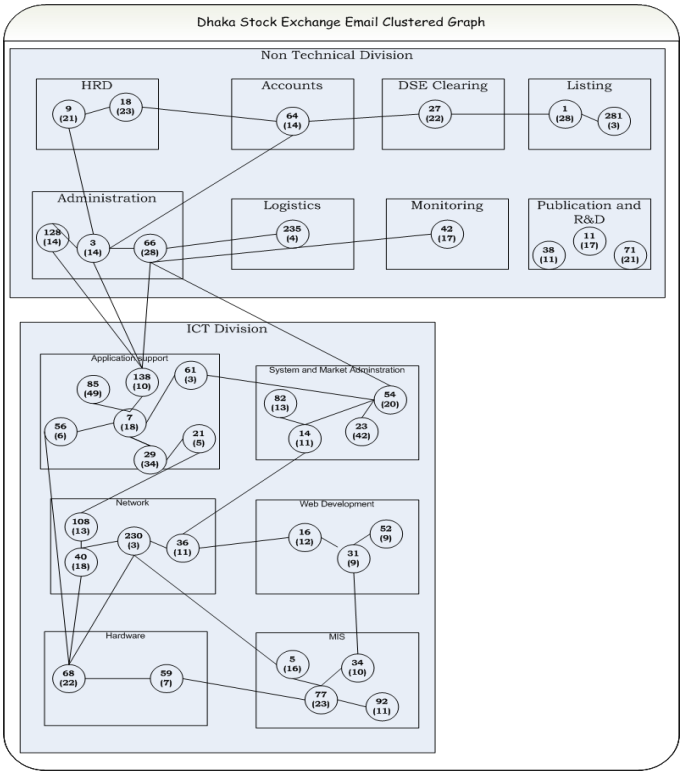


Figure 12: Final clustered layout of DSE internal email network

Clustered graph layout visualizes a large amount of data in an efficient and compact form. It can handle larger networks and group relationships. As the amount of information that we want to visualize becomes larger and the relations become more complex, the classical graph model tends to be inadequate. It gives the layout with minimum edge crossing. It uses planarity based methods for this layout. This algorithm is only for connected clustered graph. Non-connected graph with this algorithm does not work. Its running time is not satisfactory which is O(n2).where n is the number of nodes. Among the three layouts, clustered layout is the best to handle large network and group relationships.

VI. PERFORMANCE ANALYSIS

We can easily observe from figure 13(a) that clustered layout can handle more nodes relative to those of force directed and spherical layout. Force directed and spherical layout is suitable for small number of nodes but simple and easy to implement than clustered layout. The running time complexity of these three methods is shown in figure 13(b). Force directed layout has cubic complexity. Clustered and spherical layout time complexity is quadratic.

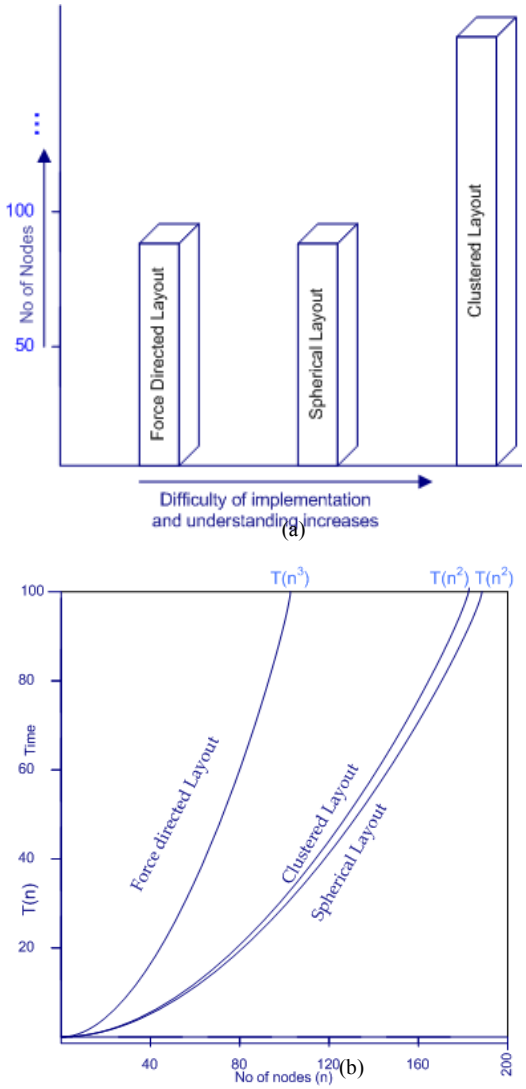


Figure 13: (a) difficulty level and nodes (*n*) handling by three methods (b) Time complexity of three methods.

Performance Analysis of three methods of visualization:

Criteria	Force Directed Layout	Spherical Layout	Clustered Graph Layout
Visualization Type	2D visualization on the plane.	3D visualization on the surface of a sphere.	Hierarchical view of the network.
Visualization Criteria	Gives the overview of the network	Gives the relationships between different groups of the network.	Shows the clustered structure of the network according to the attribute.
Output graph	Does not give clear information.	Gives group visualization clearly.	Best to visualize large network and group relationships.
Number of nodes handling	Suitable for graphs of medium size (up to 50-100 vertices).	Suitable for small number of groups.	Visualizes huge number of nodes in an efficient and compact form.
Complexity	$O(n^3)$	$O(n^2)$	$O(n^2)$

From the above performance analysis, we can conclude the following:

Force directed drawing is used to view the total network structure (Overview) and it is simple and easy to implement. It is suitable for graphs of medium size (up to 50-100 vertices). Force-directed algorithms usually require no special knowledge about graph theory such as planarity. Its running time is cubic.

Spherical layout is used to reveal communication patterns, relationships between different groups. The nodes are distributed more evenly on the surface of the sphere, instead of collapsing at the center. It is also suitable for graphs of medium size. Its running time is quadratic.

Clustered graph layout visualizes a large amount of data in an efficient and compact form. It gives hierarchical view and group relationships. So among the three layouts, clustered layout is the best to handle large network and group relationships. Its running time is quadratic.

VII. CONCLUSION

This paper presents various methods for visualization and analysis of social networks. For our test dataset, our analysis and comparison conveys new insights which would help anyone to select the appropriate method according to a specific data set. There is no explicit and details comparison of different methods of social network yet. Nevertheless, we needed to analyze a large dataset of some months and other visualization methods to obtain a full understanding of this analysis and visualization of social network. Our continuing goals are to gain deeper insights into the correlation of different methods of analysis and visualization of the interconnectivity of social groups.

REFERENCES

[1] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Ch. 3&4, Cambridge University Press, New York, 1995.

[2] L. A. Adamic and E.Adar. How to search a social network, Social Networks, Vol. 27(3), pp. 187-203, 2005.

[3] J. Scott. Social Network Analysis: A Handbook, SAGE. London, 2000

[4] T. Coffman and S. Marcus. Pattern classification in social network analysis: A case study. In Proceedings of the 2004 IEEE Aerospace Conference, Vol. 5, pp. 3162 – 3175, March 2004

[5] Diane J Cook, Lawrence B. Holder, Mining Graph Data, pp. 35-62, 443-468, Wiley-interscience, November,2006

[6] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. Graph Drawing - Algorithms for the Visualization of Graphs. Prentice-Hall, Englewood, 1999

[7] T. Kamada & S. Kawai. An algorithm for drawing general undirected graphs. Information Processing Letters, Vol. 31, pp. 7-15, 1989

[8] V. Batagelj and A. Mrvar. Pajek— Analysis and visualization of large networks. Graph Drawing Software, pp. 77–103. Springer, Berlin 2003.

[9] P. Eades. A Heuristic for Graph Drawing, Congressus Numerantium, Vol. 42, pp. 149-160, 1984

[10] T.M. Fruchterman & E.M Reingold. Graph Drawing by Force-Directed Placement. Software: Practice and Experience, Vol. 21(11), pp. 1129-1164, 1991

[11] X. Fu, S.-H. Hong, N.S. Nikolov, X. Shen, Y. Wu, K. Xu, Visualization and analysis of email networks, In Proceedings of 6th International Asia-Pacific Symposium on Visualization, apvis, pp.1-8, 2007

[12] S. Kettner, C. Madden, and R. Ziegler. Direct rotational interaction with a spherical projection in Proceedings of Interaction: Systems, Practice and Theory, ACM SIGCHI 2004.

[13] E. Bonabeau and F. Hnaux. Self-organizing maps for drawing large graphs. Information Processing Letters, Vol. 67, pp. 177 – 184, 1998.

[14] B. Meyer. Self-organizing graphs - a neural network perspective of graph layout. In Proceedings of the 6th International Symposium on Graph Drawing, pp. 246–262, Springer-Verlag, London, 1998.

[15] Q. Feng. Algorithms for drawing clustered graphs. PhD thesis, Department of Computer Science & Software Engineering, The University of Newcastle, Australia, April 1997.

[16] N.Chiba and T. Nishizeki, Planar Graphs. Theory and Algorithms. North-Holland, Amsterdam, 1988.

[17] N. Chiba, T. Nishizeki, S. Abe, and T. Ozawa. A linear algorithm for embedding planar graphs using PQ-trees. Journal of computer and System Sciences, Vol. 30(1), pp. 54-76, 1985

[18] K. Booth and G.Lueker. Testing for the consecutive ones property interval graphs and graph planarity using PQ-tree algorithms. Journal of Computer and System Sciences, Vol. 13, pp. 335-379, 1976

[19] Gorril Vollen, PQ-Trees and maximal planarization, PhD thesis. Department of Informatics. Universitas Osloensis, February, 1998

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Full Papers

Application Specific Communication

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

ShonaBondhu: A Cloud Based System to Handle Flash Flood

Nova Ahmed, A.K. Azad, Mahmudur Rahman Khan, Ahsan Habib, Shuvashish Ghosh, *Sabiha Shahid

Department of Electrical and Computer Engineering

**Department of Environmental Science and Management*

ECE, North South University

Dhaka, Bangladesh

nova@northsouth.edu

***Abstract*— Flash Flood is a natural disaster that floods away large area where there are dense presence of rivers. Bangladesh is one such country where people face this sudden flood problem and loses valuable assets using manual water level monitoring. The challenge lies in the sudden increase of water level once the flood water is in. We are proposing a distributed system using water level monitoring sensors named *Shonabondhu*. The sensing nodes are distributed all across the country and the servers that collect data from sensors are spread around various regions. The servers use a function of rainfall and current water level that indicate a particular gradient to that sensor. The gradient information among sensors are related using the water level and rainfall data over four years (2008 to 2011). This gradient information is updated and propagated when any kind of change is present near the source of the river. A communication abstraction is created to propagate sensitive information and periodic updates of current status. We have used actual sensors to monitor the water level in the river and have used emulated sensors to mimic the behavior in large distributed system. Our current system works as a proof of a concept system before the actual deployment of this system in collaboration of Water Development Board of Bangladesh.**

Keywords— Flash Flood; Middleware; Sensor based System; Solution for Natural Disaster; in Bangladesh

I. INTRODUCTION

Our advancement in technology is not reflected when we are going back to a third world country fighting natural disasters often. We have considered Bangladesh as our field of study and we consider the problem of Flash Flood. Flash Flood occurs when the river water level rises rapidly within a short period of time (often within 2 hours water level can rise as high as few meters) mainly caused by sudden heavy rainfall in locality. It is a major problem in many different countries like Bangladesh, India, China, Nepal, Malaysia, Philippines and many other countries [23, 27, 28].

There have been an array of research studies that consider water level sensing, monitoring and /or prediction and offer valid system support for deployment of such system. E. A. Bashar et al. [18] have worked on water monitoring system deployed at Honduras, a river in Massachusetts using a predictive sensor environment; Ivan Stonianov et al. [20] used sensor network to monitor Boston Sewer Commission Data, Jirapon Sunkpho et al. [19] uses real time flood monitoring system deployed at Thammarat a Southern province of Thailand; Danny Hughes et al. [23] uses sensor based system to study flood monitoring system on the River Ribble in North West England. The major difference lies in the problem scopes. In the cases of previously conducted studies – a single river and river bank is studied. When the geographic location is closely situated – a light weight solution like wireless sensor network based architecture can be an excellent solution as we have seen in many existing systems. In Bangladesh, however, we have a network of rivers that contribute to the flash flood problem.

Our water monitoring system must include various locations that are geographically distant. We have proposed a system named Shonabondhu, Flash Flood Forecaster that uses local wireless networks and a stable network based infrastructure to support sensor nodes placed around the rivers causing the disaster. Shonabondhu consists of servers which are assigned a gradient information as a function of its current state in terms of local water level, rainfall along with historical information for that region regarding that particular time of year named as *Gradient Servers* which collect information from the sensor nodes and process them. If for any environmental factor, the gradient information changes in a server – it will propagate that information to interested servers which may be affected by the change of that particular server affecting nearby servers to change gradient information in a cascading fashion accordingly. It can be the case where a gradient server near the source of the river experiences sudden rise of water level and as electrons travel faster than any physical entity, we can expect that server to propagate that water level rise information to other servers along the connected river system and if certain thresholds are crossed, adequate alarms can be generated. We have also considered a webserver that will act as the spokesperson for the entire system and will be responsible for doing long term data analysis unlike the local servers. In our infrastructure, we have considered the central server to be part of a cloud for resilient and reliable performance.

Our research work presents our study on the large scale system and sensors that must be done prior to actual deployment. We have studied the feasibility and performance of every single component of Shonabondhu. The learning

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

based gradient level is still work on progress which we hope to share in our future research studies.

The rest of the paper is organized as follows: we discuss the background information regarding flash flood in the next section, Section III discusses System Architecture followed by Implementation and Evaluation in Section IV. Section V illustrates Related work, finally, Conclusion and Future work.

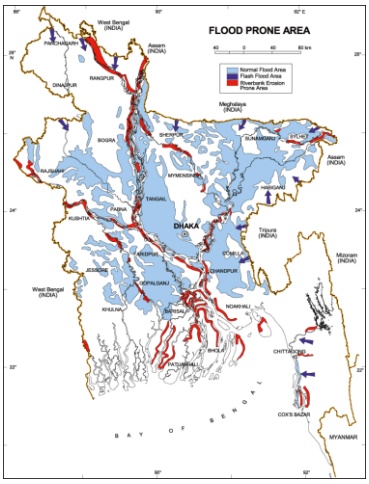


Figure 1. Flash Flood Prone Area in Bangladesh (Shown in Red)

II. The Problem of Flash Flood

Flash flood refers to a phenomenon where water level raises rapidly within a short period of time. It is caused by sudden incident of heavy rainfall. Flash flood is a major problem in many countries such as India, China, Nepal, Malaysia, Philippines and many other countries [23, 27, 28, 29]. We have focused our attention to flash flood problem in Bangladesh. There are many places that are prone to this problem as can be seen in Figure 1. However, the effect is severe in north eastern side of Bangladesh. There has been major flash flood incidents recorded in 1954, 1961, 1966, 1970, 1987, 1988, 1993, 1999, 2004, 2006, 2007 and 2010 [29].

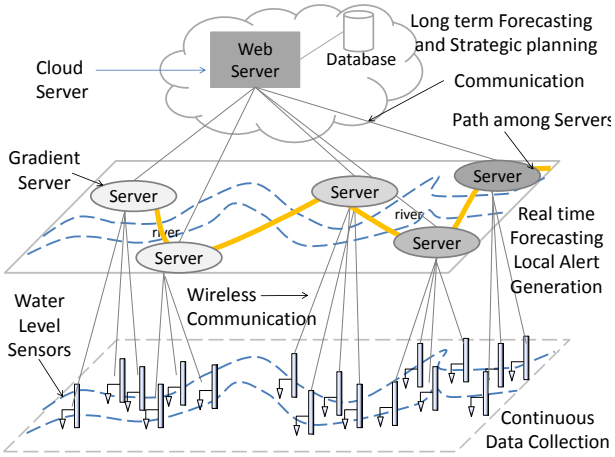


Figure 2. System Architecture

III. System Architecture

Shonabondhu System works in three different planes as can be seen in the schematic presented in Figure 2 – the first plane consist of Server/Servers residing in Cloud and is responsible for long term data forecasting which impacts strategic planning and making long term decisions regarding flood and flood affected areas; the second plane consists of the distributed local servers known as Gradient Servers which use real time decision making and forecasting at local points based on its gradient information calculated by local water level, rainfall level, time and expected water level; the third plane is responsible for real time monitoring and consists of water level sensors deployed at various localities.

A. Cloud Server

The cloud server is responsible for aggregated data collection and long term data analysis so that corresponding strategic decisions can be made. A cloud server is considered due to the dynamic nature of flash flood itself which occurs within a particular time frame (monsoon season comprising mainly rainy season) requires very high level of data processing and does not require intense data processing through the rest of the year.

B. Gradient Server

Gradient servers are distributed multithreaded servers placed at various physical locations. The servers communicate among themselves using their gradient information which is a function of its current water level, rainfall, time and expected water level. The servers are responsible for making local forecasting of data and work proactively to alert other servers along the physical vicinity so that advanced alert messages can be generated. It uses three levels of communication – a stable and secured communication mechanism to send summary data to cloud server periodically, a stable and gradient based communication system (described in the following text) to forecast flood locally and a wireless communication mechanism to collect, filter and process continuously generated sensor data.

C. Sensors

At the lowest plane of our architecture, we have abundance of sensors near the river banks. There are several kinds of water level sensors available under limited cost (we are focused to provide a lower cost solution considering the economic situation of our country). We have considered ultrasonic sensors which uses the reflection of a signal to find out the distance of an object and liquid level sensor using optical devices which uses optoschmitt trigger to detect the level of water [4, 5]. The details of our sensors are discussed in the implementation section.

IV. Implementation and Evaluation

Our system implementation is based on a proof of concept system deployed in a smaller region near a river bank which has been emulated in a larger distributed system setup. We discuss the implementation along with the system's performance evaluation in this section.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

A. Study On Sensors

The Gradient server wirelessly communicates to the ultrasonic sensor, optical flow sensor or emulated sensor. The ultrasonic sensor and liquid level sensors are attached to the arduino board [17] for its flexible programming interface.

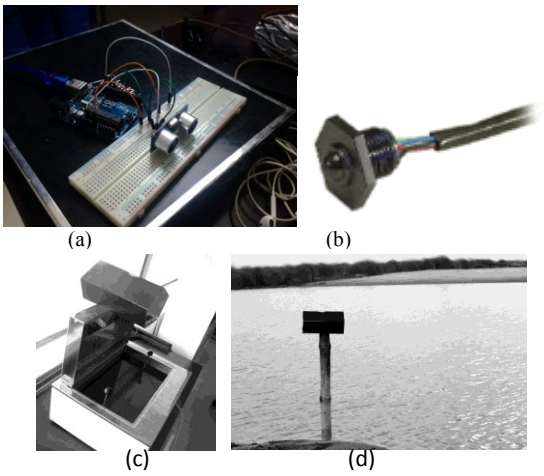


Figure 3. (a) Ultrasonic Sensor (b) Liquid Level Sensor (c) clean water in Laboratory (d) dirty water in river

*1) Ultrasonic Sensor:*The ultrasonic sensor of named HC SR04 [5] is used for a choice of sensor during our study shown in Figure 3-(a). The sensor performances are presented in the evaluation section. The major motivation behind using this particular sensor module was the availability of this product in Bangladesh. The major specifications are presented in Table I.

Table I. System Specifications for Ultrasonic Sensor

Specification	Value
Max Range	4000 cm
Min Range	2 cm
Measuring Angle	15 degree
Dimension	40 * 20 * 15 mm

**Test distance= (high level time x velocity of sound)/2
**The module automatically sends eight 40kHz and detects if there is a pulse back

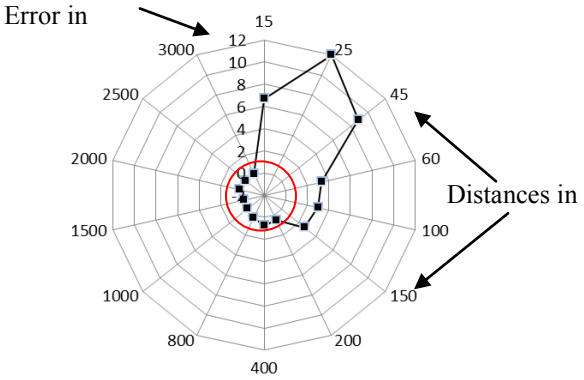


Figure 4. Error in Percentage using Ultra Sonic Sensors

We have studied the ultrasonic sensor of named HC SR04 [5] which shows error rate in percentage within the level of 0 to 1 when the sensor s are placed 200 cm to 3000 cm. Beyond the 3000 cm distance, the sensor fails to present valid data. On the other hand water level of 200 cm or less provides a higher error level ranging from 2% to as high as 12%. This behavior is summarized in our study shown in Figure 4. We have studied the sensor accuracy in clean water and in river water (has sediment and other dirt in it) – the ultrasound sensor has not shown any significant difference in detecting the water level which is evident from the way this sensor works using the reflection of signals. However, the sensor failed to read at all once it was covered by sediment which may be the case when the sensor is installed in the river bank.

*2) Liquid Level Sensor:*We have used the liquid level sensor of LL Series named LL103101from Honeywell [4] to study the performance of this particular sensor type. It has the advantage of being compatible to microcontroller modules, reasonable response time and being cost effective. The major specifications are presented in Table II and the sensor is presented in Figure 3-(b). We have conducted the experiment by attaching this sensor to a microcontroller based module.

Table II. System Specification for Optical Flow Sensor

Specification	Value
Operating Temperature Range	-40 ˚C to 125˚C
Storage Temperature Range	-50 ˚C to 150˚C
Response Time- Rising Liquid	50µs
Response Time- Falling Liquid	1s max.(in ethanol)

**Average maximum temperature in Bangladesh 30.6 ˚C
**Average minimum temperature in Bangladesh 21.5 ˚C
**Average temperature in Bangladesh 26.1 ˚C
*data source: Weather Base over a period of 30 years.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

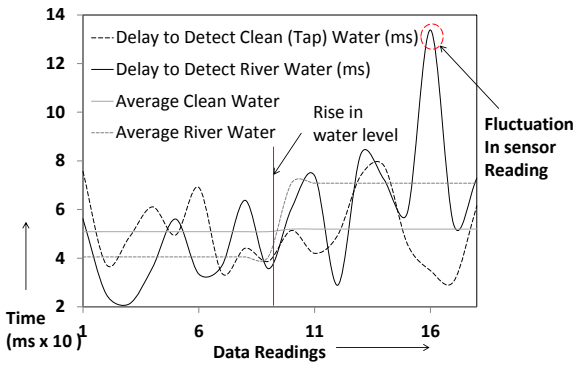


Figure 5. Comparison of Difference in Delay in Clean Water and River Water

The liquid level sensor of LL Series named LL103101from Honeywell [4] shows varying performance in clear water and river water as shown in Figure 5. The water level detection time is lower for river water compared to clean water possibly for the underlying optical technology it uses detects heavier liquid easily. The problem with this sensor lies in its input phase when the sensor disseminates invalid output. Our close study using four liquid level sensors placed at various water level shows that the sensor shows erroneous data over a short period of time which may cause the system to be unaware of an imminent threat for couple of milliseconds. Our system will be able to overcome this kind of entry using redundant sensors as can be seen in our study.

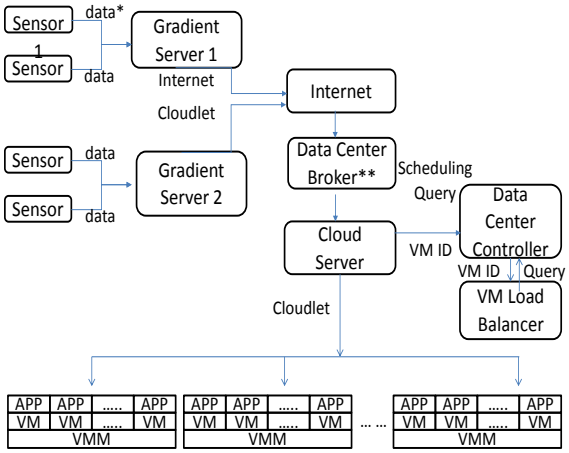


Figure 6. Gradient Server Cloud Server connection in Simulation

B. Study onGradient Server and Cloud Server

We have considered the Meghna basin in our current experiments. The Meghna system originates in the hills of Shilong and Meghalaya of India. Other rivers like Barak, Amalshid, Surma and Kusiara rivers join in the journey in this river system which are mostly highly flashy and steep rivers originated where the highest aveage waterfall is recorded (Cherapunji in Asam, 10,000 mm annual rainfall). The cloud and gradient server connection is shown in Figure 1.

We have used CloudSim [1] simulation tool to design and evaluate our cloud server. We have a simple version of the Virtualized data center where *Virtual Machines* (*VM*) are the basic computing block. Here the *Userbase* defines the overall behavior of the users in a geographic region of the world. In our simulation, for simplicity, we combine the user requests from huge number of users in a Userbase to group them to internet cloudlets. These internet cloudlets are sent to Data Center via internet. As can be seen in this simulation we used 100 bytes for user request length and for Brokering policy we used "Nearest Datacenter First" method.

We have considered two distinct scenarios in our study where the server is lightly loaded which reflects the current web server load located in Flood Forecasting Center. The other scenario of consideration the scenario when the server is heavily loaded considering a situation that may arise during a flood or when a flood is imminent and there are many hits to the central server. The server setup under various load are discussed below. Although the current web server has not incurred such traffic so far, we have designed our assumptions taking other Government web servers used to publish public data (e.g., Server that publishes Secondary School Exam results). The results are presented in evaluation section.

We have set 6 user bases which incorporate sensors in the architecture, with 10 requests per user per hour. data size per byte set to 100, peak hour starts at 3 and ends at 6 GMT, average peak users being 1000 and off peak users being 100. Data Centers have 5 brokers, number of virtual machine is set to 50, image size 10,000, memory size 2048 (MB). For data centers we also have considered x86 architecture, Linux Xen VMM and cost per VM to be 0.1 $/s, memory cost to be 0.05 $/s, data transfer 0.1 $/s for our cost analysis. We have used round robin scheduling for load balancing. Heavily loaded system considers similar parameters as the lightly loaded system parameters except for some changes. We have considered 60 requests per user per hour (as opposed to 10), having peak starting at 3 and ending at 5 GMT, average peak users being 10 times more than lightly loaded system having peak and off peak users to be 10,000 and 1000 respectively.

We have conducted number of experiments under the heavily loaded and lightly loaded situation in cloud architecture using CloudSim. We present the results in the following subsection.

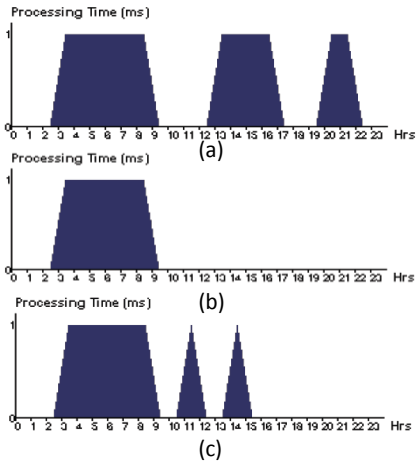


Figure 11. Lightly Loaded Nodes

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

1) Lightly Loaded Nodes

We have an average response time of 50.69 (ms) for lightly loaded cloud nodes where the minimum and maximum response time is 35.88 (ms) and 65.40 (ms) respectively. Individual response time at user bases remain close to 50 (ms) throughout the execution time. The data center service time is around 1 (ms) on average with as low as 0.07 (ms) and as high as 1.62(ms). The processing time varies over time dpending on the assigned load as can be seen in Figure 11 (a) to 11(c). Every data center of our consideration does not provde more than 10,000 requests per hour over the functioning time. In general, our simulation study has estimated a total virtual machine cost of 720.22 in dollars and data transfer cost of 8.98 dollars which is affordable considering the data security and reliability.

2) Heavily Loaded Nodes

Heavily loaded nodes show average response time of 53.13 (ms) which is close to lightly loaded nodes. In heavily loaded nodes, the variability is high having minimum response time as low as 33 (ms) and maximum being 74 (ms). Hourly load show similar behavior to user base 1 as shown in Figure 12(a). Data center processing time is 3.03 (ms) on average having minimum and maximum as 0.62 (ms) and 13.92 (ms). respectively. The data centers show high number of requests per hour during the peak load hours as shown in Figure 12(b). The general cost for virtual machine is the same 720.08 $ but data transfer cost increases being 157.72 $ as opposed to 8 dollars.

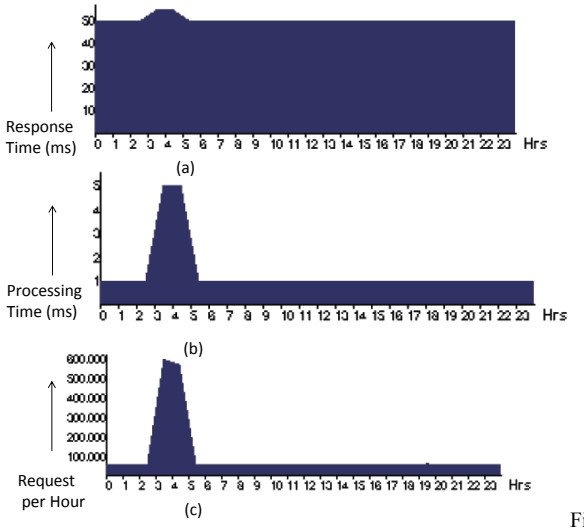


Figure 12. Behavior of Heavily loaded nodes

Our studies on cloud based system validates the feasibility in terms of response time and cost which are the major concerns being a country without a cloud center from international vendors locally.

Our study on Shonabondhu system provides an overview of the system along with its large scale deployment feasibility. Although most of our large scale experiments are studied using emulated sensors, we have seen tolerable latency of the system in terms of communication overhead among servers.

V. RELATED WORK

We have looked at works that are similar in concept with our research in many different ways: existing work on water level sensing, disaster sensing mostly uses wireless sensor network as infrastructure and we have consulted existing state of the art work, there are some work from different research studies that use the concept of gradient sensing works but in different context.

Elizabeth A. Basha et al. [18] has worked on a deployment that covers Honduras, a river in Massachusetts using a predictive sensor environment. It uses a two tier approach to minimize cost – some nodes work in long range using radio signal while other nodes work for shorter range.,It uses three sensor nodes for testing purpose to cover the particular river of interest. Jiranpon Sunkpho et al. [19] discusses about a real time flood monitoring system. It is installed to monitor Nakhon Si Thammarat, a southern province of Thailand where flooding is a recurrent event. It uses sensing monitoring in 15 sites using internet based real time monitoring. The middleware named VitualCom is in charge of communication among sensors and Application Server. Ivan Stonianov et al. [20] used sensor networks to monitor water supply and sewer system. This system uses wireless sensor network to cover the area of interest. It is used at Boston Water Sewer Commission and it collects water level information in real time. The remote monitoring system uses high data rate sampling which is 1000 samples per second. The entire system is based on Wireless Sensor Network (WSN). LooCI System [22] developed by Danny Hughes et al. focuses on a novel component binding model using even binding a wireless sensing system. It can be a good candidate to implement real time sensing system. Danny Hughes et al. [23] discusses Gridit middleware used for flood monitoring and warning scenario. It supports a WSN based flood monitoring on the River Ribble in North West England using GridStix hardware platform. It uses an overlay based framework in Gridit. Slim Rekhins et al. [24] uses a combination of WSN and RFID readers to collect water quality information. It does not discuss aobut any middleware explicitly but it works based on data collected from underlying sensor system and analyzes them in real time. The systems of consideration for water level monitoring provide great insight towards implementing a real time monitoring system using sensors. However, wireless sensor network may not be suitable for a country wide development. Wireless sensor network modules can be locally deployed but at the same time there must be a coordination among the local components. Shonabondhu also focuses on a server to server trigger based communication which actually mimics the flow of river and spreads out data in a similar fashion. This is where our system stands out from the rest of the system. Dan Chen et al. [25] discusses challenges in using WSN for monitoring natural disasters. WSN having quick response, low cost and being scalable suffer when there is a presence of large amount of data or requirement for low level processing at lower level. Chistian Seerag et al. [26] similarly approaches challenges deploying WSN in the wild and middleware centric issues. These insights provided us with valuable insights selecting a platform for a robust system.Chuljin Park et al. [21] worked on water quality monitoring for hypothetical river using

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

simulation based study. It should be able to provide a base line to compare our system theoretically.

The work developed by F. Ye et al [6] considers gradient broadcast as a method to deliver messages for large scale sensor network rather than consider the sensors to have gradient information. There has been gradient based routing [7] by J. Faruque et al. where the gradient information is used for routing which is similar to our concept while we use the gradient information more for data collection importance rather than involving that information for routing. Our concept is similar to the concept proposed by H. Lin et al. [8] where the gradient information is used among the sensors for sensing capabilities.The concept of path is used in many different contexts including fault tolerance [9], compiler optimization techniques [10], profiling distributed systems [12,13], and resource allocation [14]. Scout OS [8] defines a path abstraction to navigate through the layers of the network stack and the Ninja project [16] utilizes a path abstraction as a way to compose multiple services distributed on the Internet into a single logical unit. Our work is inspired by the use of paths in these various contexts. We have used the concept of path used to improve system reliability from RF^{2}ID[15].

VI. CONCLUSION AND FUTURE WORK

Large scale deployment of a sensor based system is a challenging task specifically when we are targeting an eminent natural disaster. We have presented Shonabondhu system which mainly focuses a distributed architecture using embedded sensor nodes deployed around a flood prone country Bangladesh. Our detailed study of sensor nodes, network and scalability study provides an overview of the system structure along with hope of coping a problem like flash flood with better preparation.

REFERENCES

[1] CloudSim, The CLOUDS Lab, www.cloudbus.org/cloudsim. (Last accessed on March, 2014)

[2] Bangladesh Water Development Board, Weblink: www.bwdb.gov.bd. (Last accessed on March, 2014)

[3] Flood Forecasting and Warning Center, Bangladesh. Weblink: www.ffwc.gov.bd (Last accessed on March, 2014)

[4] Honeywell, Liquid Level Sensor, Weblink: http://sensing.honeywell.com/index.php?ci_id=4128&la_id=1&Ntk=si_all_products&N=1235&Ntt=ll103101(Last accessed on March, 2014)

[5] Ultrasonic Rangile Module, Weblink: http://www.micropik.com/PDF/HCSR04.pdf (Last accessed on March, 2014)

[6] F. Ye et al., " Gradient Broadcast: A robust data delivery protocol for large scale sensor network,", Wireless Networks11, 285-298, 2005.

[7] J. Faruque et al., " Analysis of Gradient-based protocols in sensor networks,"DCOSS, 2005.

[8] H. Lin., "Composable Information Gradients in Wireless Sensor Networks", IPSN '08 Proceedings of the 7th international conference on Information processing in sensor networks, 121-132, 2008

[9] M. Chen et al., "Path-based Failure and Evolution Management," In Proceedings of the First Symposium on Networked Systems Design and Implementation (NSDI), 2004

[10] G. Ammons et al., "Improving data-flow analysis with path profiles," In Proceedings of the ACM SIGPLAN '98 Conference on Programming Language Design and Implementation, 1998

[11] P. Barham et al., "Magpie: real-time modelling and performance-aware systems," In Proceedings of the 9th Workshop on Hot Topics in Operating Systems, 2003.

[12] T. Gschwind et al., "WebMon: A Performance Profiler for Web Transactions," WECWIS : 171-176, 2002.

[13] M. Welsh et al., " SEDA: An Architecture for Well-

[14] Conditioned, Scalable Internet Services, " In Proceedings of the Symposium on Operating Systems Principles, 2001.

[15] N. Ahmed et al., "RF^{2}ID: A Reliable Framework for RFID Deployment," IPDPS, 2007

[16] S.D. Gribble et al, "The Ninja Architecture for Robust Internet-scale Systems and Services," Computer Networks35, 473-497, 2001.

[17] Arduino, Weblink: http://www.arduino.cc/ (Last accessed on March, 2014)

[18] Elizabeth A. Basha et. al., 2008. Model-based monitoring for early warning flood detection. In *Proceedings of the 6th ACM conference on Embedded network sensor systems* (SenSys '08). ACM, New York, NY, USA, 295-308.

[19] Jirapon Sunkpho et. al., 2011. Real-time flood monitoring and warning system. In *Songklanakarin Journal of Science and Technology (SJST)*.33(2), 227-235, Mar-Apr. 2011.

[20] Ivan Staoianov et al. Sensor Networks for monitoring water supply and sewer systems: Lessons from Boston. 2006. In *Eight Annual Water Distribution Systems Analysis Symposium (*WSDA), Ohio, USA.

[21] Chuljin Park et. al., 2010. Designing optimal water quality monitoring network for river systems and application to a hypothetical river. In *Proceedings of the Winter Simulation Conference* (WSC '10), Björn Johansson, Sanjay Jain, and Jairo Montoya-Torres (Eds.). Winter Simulation Conference 3506-3513.

[22] Danny Hughes, et. al.,. 2009. LooCI: a loosely-coupled component infrastructure for networked embedded systems. In *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia* (MoMM '09). ACM, New York, NY, USA, 195-203.

[23] Danny Hughes et al.. Exploiting extreme hetegeneity in a Flood Warning Scenario using the GridKit Middleware. In *Proceedings of the 5th International Workshop on Middleware Tools, Services and Run-Time Support for Sensor Networks,* NY, USA, 29-34.

[24] Slim Rekhis, Nourhene Ellouze, and Noureddine Boudriga. 2012. A wireless sensor network based water monitoring system. In *Proceedings of the 8h ACM symposium on QoS and security for wireless and mobile networks* (Q2SWinet '12). ACM, New York, NY, USA, 33-40.

[25] Dan Chen et. al. Natural Disaster Monitoring with Wireless Sensor Networks: A Case Study of Data-intensive Applications upon Low-Cost Scalable Systems. *Mob. Netw. Appl.* 18, 5, 651-663.

[26] Christian Seeger et. al. Wireless sensor networks in the wild: three practical issues after a middleware deployment. In *Proceedings of the 6th International Workshop on Middleware Tools, Services and Run-time Support for Networked Embedded Systems,* NY, USA, , Article 1 , 6 pages.

[27] List of Flash Flood, Web link: http://en.wikipedia.org/wiki/List_of_flash_floods (Last accessed on March, 2014)

[28] Nova Ahmed. Gradient Sensor Middleware System design for Flash Flood prone River banks in Bangladesh. Grace Hopper Celebration for Women in Computing, 2013.

[29] Abdul Awal Sarker et. al.,, Landslide and Flash flood in Bangladesh, Disaster Risk Reduction in Bangladesh, 2013, Chapter 8, Springer, pp 165 – 189

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

Using Adaptive Heartbeat rate on Long-lived TCP Connections

M Saifur Rahman, Md. Yusuf Sarwar Uddin, M Sohel Rahman and M Kaykobad
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Email: {mrahman, yusufsarwar, msrahman, kaykobad}@cse.buet.ac.bd, saifur80@gmail.com

***Abstract*—In this paper, we propose techniques for dynamically adjusting heartbeat or keep-alive interval of long-lived TCP connections, particularly the ones that are used in push notification service in mobile platforms. When a TCP connection between a client behind a NAT (or any other middle-box) and a server is idle for a long time, it may get torn down due to TCP binding timeout. In order to keep the connection alive, the client device needs to send keep-alive packets through the connection when it is otherwise idle. To reduce resource consumption, the keep-alive packet should preferably be sent at the farthest possible time within the NAT binding timeout. We refer to this interval as the *Optimal Keep-alive Interval*. Due to varied settings of different network equipments, optimal keep-alive interval will not be identical in different networks. Hence, the heartbeat rate used in different networks should be changed dynamically. We propose a set of iterative probing techniques, namely binary, exponential and composite search, that detect optimal keep-alive interval with varying degree of accuracy; and in the process, keeps improving the keep-alive interval used by the client device. We also analytically derive performance bounds of these techniques. To the best of our knowledge, ours is the first work that systematically studies several techniques to dynamically improve keep-alive interval. To this end, we run experiments in simulation as well as make a real implementation on Android to demonstrate the proof-of-concept of the proposed schemes.**

I. Introduction

Smart phones, tablet PCs and other customized PDAs try to provide the user with fresh data. This includes the user's emails, social news feed etc. Real time communications, such as voice and video calls, are also performed through such devices. Since the devices in question have limited battery life, they cannot frequently poll for information updates, incoming call notifications etc. Rather, they rely on change notifications being pushed by a notification server. Mobile platforms like iPhone, Android, Windows Phone, Black Berry etc. provide a *Notification Service* which, at a high level, can be modeled as shown in Figure 1. Rather than polling different services to check if data needs to be downloaded, the user's device only maintains a TCP connection to a notification server. This connection is called a *Notification Channel*. When the user's social network service wants to send recent activity feeds, it sends a new activity notification to the notification server. This is delivered to the device through the notification channel. Based on this notification, the device opens a new connection to the social network service, downloads the activity feed and closes the connection. Similar workflow is executed for downloading unread emails, receiving voice and video calls etc. Therefore, to enable these seemingly *always connected* scenarios, the device must keep the notification channel alive at all times, even during low power modes. Such a connection is quiet for the most part, with traffic flowing only when the server has an update to share. When the device is behind a Network Address Translator (NAT), firewall or any other stateful middle-box, the connection would be subjected to binding timeouts. That means, the connection state may be cleared by the middle-box if no data passes through it for some specified amount of time.

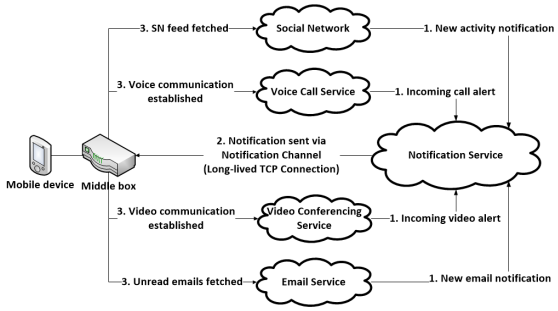


Fig. 1. A model of notification service used in different mobile platforms.

Studies have shown that the NAT binding timeouts vary dramatically amongst the commercially available home gateways [1]. To safeguard against this timeout, the device needs to periodically probe the other end of the connection when the connection is otherwise idle. This is called a keep-alive [2] or a heartbeat; with the interval being referred to as keep-alive interval or KA interval in short. To balance the battery life constraint with the necessity of keeping the connection alive, it is preferred that the keep-alive probe is made at the farthest possible time within the NAT binding timeout. This interval is termed as optimal keep-alive interval. Such an interval needs to be computed via experiment in the network environment that the device is currently in.

In this paper, we present several iterative probing approaches to dynamically improve the KA interval of a long-lived TCP connection. These include binary search, exponential search and composite search. Composite search combines

978-1-5090-0203-0/16/$31.00

different aspects of binary and exponential search techniques. We analyze the proposed techniques theoretically and validate them through simulation. We show that composite search technique results in the least number of KA messages sent during several study durations of different lengths.

The rest of the paper is organized as follows. Section II reviews some earlier work that is relevant to our research. Section III describes several techniques to dynamically change the keep-alive interval of TCP connection. Section IV provides analytical bounds on several properties of these techniques. Section V describes the simulation experiments and their results. Finally, Section VI concludes this paper.

II. Related Work

We organized the literature review in four sections. Firstly, we look at a study that suggests binding timeout of TCP connections varies widely amongst different networking equipments. This motivates the need for sending keep-alive packets periodically to retain the connection. Then we review use of KA packets in existing systems. We also review some literature on impact of KA interval on power consumption. This motivates the need for detecting and using longer KA intervals when possible. Finally, we review some earlier works that uses iterative probing for measuring different network parameters.

A. Study of NAT Binding Timeouts

In [1], Hätönen et al. experimentally analyzed the TCP binding timeouts of different home gateways. The shortest timeout observed was about 4 minutes; median was about an hour. While the Internet Engineering Task Force (IETF) recommends a timeout of 124 minutes [3], more than 50% of the devices did not comply.

On the other hand, some of the devices retained TCP bindings for considerably longer; they did not timeout the TCP connection even after 24 hours of idleness. As the binding timeout varies widely, a mobile device in real network environment may also need to perform testing to find longer keep-alive interval and send keep-alives using that period. While protocols such as NSIS [4] or SIMCO [5] do exist to explicitly negotiate the binding timeout with the middle-box, equipments currently used by mobile operators do not usually support these protocols.

B. Keep-alive in Existing Systems

The *Direct Push* feature of Exchange ActiveSync (EAS) protocol uses long-standing HTTPS requests to maintain a channel with the service [6]. Each request is 'parked' at the server for a time specified by the request, if there is no update on the server to share. After the time elapses, the server returns a '200 OK'. If, however, the underlying network has a smaller binding time out, then no response is received from the server. Then the client sends another HTTPS request with a lesser amount of wait time specified. Detailed discussion on the keep-alive interval adjustment, along with the configurable parameters, can be found in [7].

Android, iPhone, Windows Phone etc. platforms maintain a persistent connection between the client (device) and a notification service [8], [9], [10], [11]. It is expected that each of these eco-systems employ exchange of keep-alive messages periodically and has mechanism to tune the keep-alive interval to obtain good battery life performance. Patents granted to Microsoft Corporation indicate use of a test connection to dynamically detect an efficient keep-alive interval for communication between client and server via middle-box [12], [13]. However, the exact strategy used in choosing the test intervals is not called out in the patents.

The Heartbeat Extension [14] of Transport Layer Security (TLS) protocol [15] also uses keep-alive messages for ensuring liveness of peers.

C. Impact of KA Interval on Power Consumption

Haverinen et al. showed in [16] that battery life is significantly influenced by the frequency of keep-alive messages. They performed real power measurements in two different 3G networks, as well as, in 2G GPRS network.

A comprehensive survey on general solutions for energy efficiency on mobile devices, published between 1999 and May 2011, is available in [17]. Here the authors classify and provide a short summary of the various efforts on studying, modeling and reducing energy consumption in mobile devices. We discuss one of the power models, due to Balasubramanian et al. [18] here. In this energy model, the energy spent to download or upload x kilobytes of data over the cellular network consists of three components: ramp energy, transmission energy and tail energy. $R(x)$ denotes the sum of the ramp energy and the transfer energy to send x bytes of data. Tail energy is represented by E per second. For WiFi, there is no ramp energy. In this case, $R(x)$ denotes the sum of the transfer energy and the energy for scanning and association. Tail energy is 0 in this case. The total energy to transmit a packet further depends on the time the interface is on. The energy consumption to keep the interface on is represented using M, the maintenance energy per second. Finally, T denotes the tail-time. Since the keep-alives are sent with a period that is longer than the tail-time, each keep-alive incurs the overhead of the tail-time, if the device is not transferring any other data at that time. In that case, by reducing the number of keep-alives, we can reduce the overall power consumption.

D. Iterative Probing to Measure Network Parameters

Iterative probing has been widely used in the literature for measuring different network parameters like end-to-end available bandwidth (avail-bw), its variability, TCP congestion window size etc. The congestion window size in TCP doubles up in each round trip iteration in the slow start mode. In the subsequent additive increase mode, it is increased by one segment in every round trip iteration. Whenever a packet loss is detected, the slow start threshold is set to be half of the congestion window size and the entire process is restarted [19].

Jain et al. used iterative probing to measure end-to-end avail-bw. Their measurement methodology, Self-Loading

Periodic Streams (SLoPS), was implemented in a tool called pathload [20]. Other iterative techniques for avail-bw measurements include Bfind [21], PTR [22], TOPP [23], pathChirp [24] etc. The variability in the available bandwidth has also been measured by Jain et al. using iterative probing [25].

III. Techniques to Dynamically Improve Keep-alive Interval

In this section, we describe our proposed techniques for improving keep-alive interval of TCP connections. All schemes use an iterative probing that repeatedly send keep-alive packets at progressively longer intervals, until a bound is found beyond which the connection can't be kept alive. This bound is referred to as the optimal keep-alive interval. The schemes vary on how these test intervals are chosen one after another.

Figure 2 shows a flow chart of the steps taken in detecting the optimal keep-alive interval. The first step is to open a separate TCP connection with the target service. If the probes were done in the data connection (e.g. notification channel), the connection may suffer disruption, which may cause adverse effect on the ongoing service. Instead, we conduct the probing on a separate connection. We refer to this connection as *test* connection.

As the testing reveals improved intervals, the new intervals can be applied on the notification channel right away. Initially, however, a conservative keep-alive interval is used. This is the maximum keep-alive interval that is already known to work. The same interval is also used as the lower bound of the search range for searching a better interval. We also specify a higher bound on the search space. The higher bound is 1 minute less than the minimum keep-alive interval that is already known to not work. Note that we use intervals in minute boundaries only. In the figure, the search range has been represented as the interval $[low, high]$.

Starting from the lower bound, we try to improve by guessing new keep-alive intervals. Once a guess is made, the connection is kept silent for that much time. Afterwards a keep-alive is sent to check whether the connection is alive or not. If the connection is alive, it means our guessed KA interval is able to keep the connection alive. Now, a higher KA interval is guessed and tested. On the other hand, if the connection was dropped, then we need to lower the guess and perform the test again. This process is continued until the difference between 2 consecutive guesses is less than 1 minute.

We explore three different techniques for optimal KA interval detection. These techniques vary in how they select the next KA interval to test. The techniques are: binary search, exponential search and composite search. Intuitively, we first applied binary search to this problem. This takes the least amount of time to find the optimal KA interval. The problem with this approach, however, is that the first probe is made after a long period of waiting during which time no improvements can be made to the KA interval in the data connection. As such, we subsequently examined exponential search, where probes are made with shorter wait times initially; Improvements could immediately be made to KA interval in the data connection. However, exponential search takes very long time to detect the optimal KA interval. Therefore, we finally combined the two approaches into what we call the composite search and were able to get good results.

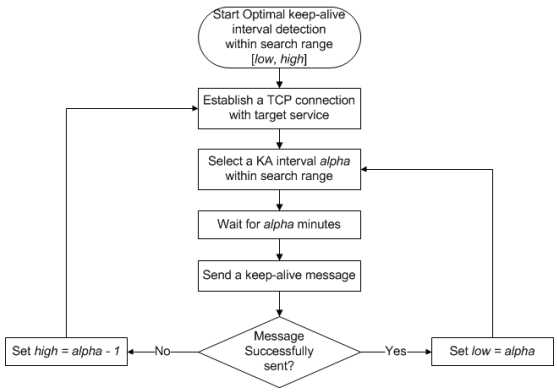


Fig. 2. Flow chart of KA interval detection technique.

For each of these techniques, we initialize low with 1 minute. This is a reasonable choice, since it is smaller than the smallest keep-alive interval in a home gateway, as found from the data presented in [1]. Exponential and composite search techniques do not need to specify a higher bound on the search range. In other words, we initialize $high = \infty$. For binary search, we initialize $high$ with 128 minutes. This is a power of 2 and is slightly higher than the IETF recommendation.

A. Binary Search

At each iteration, the mid-point of the search space is chosen as the next KA interval to be tested. After the test is completed, the search space gets halved. If the test was successful, then we search the second half of the initial search space to see if a better interval can be obtained. If, on the other hand, the test failed, then we continue searching in the first half of the initial search space. And this process is repeated. As an example, if the binding timeout of a NAT (or another middle-box) is 24 minutes, and the search range is $[1, 128]$, the sequence of intervals tested is:

$$ProbeSeq_{24}^{binary} = \{65^*, 33^*, 17, 25^*, 21, 23, 24\}$$

The probes that failed are annotated with '*' mark.

B. Exponential Search

In this approach, the difference between successive tested intervals follows geometric progression with a ratio of 2. So, the first few test intervals are $2, 4, 8, 16, 32$ minutes and so on. If the next interval to be tested goes beyond $high$, then there is no need to test that interval. Instead, low is increased to the last successfully tested interval; and the progression of the interval differences is restarted at 1. On the other hand, if the next tested interval overshoots the binding timeout, then the connection is lost. In this case, in addition to the updates

mentioned above, the $high$ is reduced to 1 minute less than the failed test interval. This process is repeated until the optimal KA interval is reached. For example, if the binding timeout is 24 minutes, the sequence of intervals tested would be

$$ProbeSeq_{24}^{exp} = \{2, 4, 8, 16, 32^*, 17, 19, 23, 31^*, 24, 26^*, 25^*\}$$

C. Composite Search

Composite search is a combination of binary and exponential searches. Initially, it acts like the exponential search. The difference between successive tested intervals follows geometric progression with ratio 2. When a tested interval overshoots the actual binding timeout, the test connection will get terminated. At that point, low is increased to the last successfully tested interval; and $high$ is reduced to 1 minute less than the failed test interval. Subsequently, binary search is performed in the new search range. As an example, if the network timeout is 24 minutes, the sequence of intervals tested would be:

$$ProbeSeq_{24}^{comp} = \{2, 4, 8, 16, 32^*, 24, 28^*, 26^*, 25^*\}$$

IV. Analytical Bounds of KA Schemes

We have analytically derived two metrics, namely the number of probes performed in the test connection to detect the optimal KA interval (Probe Count), and total time taken to do so (Convergence Time). Our schemes can be compared based on these.

Let α denote the optimal keep-alive interval. Let the search range be $[1, h]$. For binary search, $h = 2^k$ for some $k \geq 1$. For exponential and composite search, $h = \infty$. Let $N(\alpha)$ denote the number of test probes needed to detect the optimal KA interval. The best, worst and average case values of $N(\alpha)$ are represented respectively by N_{best}, N_{worst} and N_{avg}. Let $T(\alpha)$ denote the time it takes to detect the optimal KA interval α. This is called the convergence time. The best, worst and average case convergence time for the specified search range is represented by T_{best}, T_{worst} and T_{avg} respectively. In what follows, we obtain analytical bounds on these properties for the different detection techniques. Due to space limitation, we only show the derivations for binary search. For the other techniques we only record the results here. The reader is referred to [26] for details.

A. Probe Count

In binary search,

$$N_{best}^{binary} = N_{avg}^{binary} = N_{worst}^{binary} = \lg h \tag{1}$$

In exponential search, we have:

$$N_{best}^{exp} = 1 \tag{2}$$

$$N_{worst}^{exp} = \begin{cases} \lg h + 2 & \text{when } h \leq 4 \\ \frac{\lg h(\lg h+1)}{2} & \text{otherwise.} \end{cases} \tag{3}$$

$$N_{avg}^{exp} = \frac{\lg h(1 + \lg h)}{4} + 1 + \frac{1}{h} \tag{4}$$

In composite search, we have:

$$N_{best}^{comp} = 1 \tag{5}$$

$$N_{worst}^{comp} = 1 + 2\lg h \tag{6}$$

$$N_{avg}^{comp} = 2\lg h + \frac{2}{h}(2 + \lg h) - 3 \tag{7}$$

Probe count in different search techniques has been compared in Table I.

TABLE I
PROBE COUNT COMPARISON AMONG TECHNIQUES TO DYNAMICALLY IMPROVE KA INTERVAL.

	Best	**Average**	**Worst**
Binary search	$O(\lg h)$	$O(\lg h)$	$O(\lg h)$
Exponential search	1	$O(\lg^2 h)$	$O(\lg^2 h)$
Composite search	1	$O(\lg h)$	$O(\lg h)$

B. Convergence Time

In binary search, the first probe takes $1 + 2^{k-1}$ unit time. if $\alpha \geq 1 + 2^{k-1}$, then the second probe takes $1 + 2^{k-1} + 2^{k-2}$ unit time after completion of first probe. On the other hand, if $\alpha \leq 2^{k-1}$, then the second probe takes only $1 + 2^{k-2}$ unit time. Let $a_{k-1}a_{k-2}...a_1a_0$ denote the binary representation of $\alpha - 1$. Therefore, we can write the following expression:

$$T^{binary}(\alpha) = (h + \lg h - 1) + \sum_{i=1}^{\lg h - 1} a_i i 2^i \tag{8}$$

The best case convergence time happens when α is 1 or 2. And that value is: $2^k + k - 1$ unit time. Therefore,

$$T_{best}^{binary} = h + \lg h - 1 \tag{9}$$

The worst case convergence time occurs for $\alpha = 2^k$ or $\alpha = 2^k - 1$.

$$\begin{aligned} T_{worst}^{binary} &= (2^k + k - 1) + \sum_{i=1}^{k-1} i 2^i \\ &= h\lg h + \lg h - h + 1 \end{aligned} \tag{10}$$

The convergence time on average is:

$$\begin{aligned} T_{avg}^{binary} &= \frac{1}{2^k} \sum_{\alpha=1}^{2^k} T(\alpha) \\ &= (\frac{h}{2} + 1)\lg h \end{aligned} \tag{11}$$

In exponential search, we have:

$$T_{best}^{exp} = 2 \tag{12}$$

$$T_{worst}^{exp} = \begin{cases} 5h - 1 & \text{when } h \leq 8 \\ \frac{\lg^2 h - 3lgh + 12}{2} h - \frac{\lg^3 h + 11 \lg h + 36}{6} & \text{otherwise.} \end{cases} \tag{13}$$

$$T_{avg}^{exp} = \frac{\lg^2 h - 3\lg h + 24}{8} h - \frac{\lg^3 h + 23\lg h - 24}{24} \quad (14)$$

In composite search, we have:

$$T_{best}^{comp} = 2 \quad (15)$$

$$T_{worst}^{comp} = h\lg h + 5h - 3 \quad (16)$$

$$T_{avg}^{comp} = (\frac{h}{2} + 1)\lg h + \frac{2}{3}h + 3 - \frac{5}{3h} \quad (17)$$

TABLE II
CONVERGENCE TIME COMPARISON AMONG TECHNIQUES TO DYNAMICALLY IMPROVE KA INTERVAL.

	Best	**Average**	**Worst**
Binary search	$O(h)$	$O(h\lg h)$	$O(h\lg h)$
Exponential search	2	$O(h\lg^2 h)$	$O(h\lg^2 h)$
Composite search	2	$O(h\lg h)$	$O(h\lg h)$

Table II show the convergence time comparison of the search approaches.

V. EXPERIMENTS

We have implemented the different techniques to detect the optimal keep-alive interval on Omnet++ simulation platform [27]. We made an experimental setup with a client connected to a server through a single middle-box. Although in practice a connection may pass through a series of NAT boxes and firewalls with different timeout bindings, in terms of connection timeout the smallest interval along the path applies. In that, the series of middle-boxes can effectively be replaced by that particular middle-box with smallest binding timeout. The topology of the simulation setup is shown in Figure 3.

The delay from a node to the middle-box is set to 10ms. Between middle-box and the server, the delay was set to 100ms. These choices have been based on the observed round trip time (RTT) to the gateway and to prominent cloud services respectively, when connected to the network of the Department of CSE, BUET through different access points. As long as these delays are much less than 1 minute, our results will continue to hold.

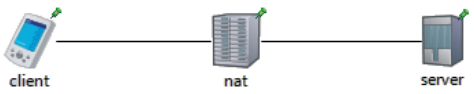


Fig. 3. Simulation topology.

A. Correctness of Analytical Bounds

Figure 4a plots the probe count of binary search, as found through simulation, against binding timeout. The theoretical curve is also put in the same plot. As can be seen, they are identical. This is expected, since the probe count in a given search range is independent of any specific binding timeout.

Figure 4b and 4c respectively plot the probe count of exponential and composite search techniques as a function of binding timeout. In both plots, observe that the experimental curve is off from the theoretical curve by 1 minute. That is, the observed probe count for any α is equal to the theoretical probe count for $\alpha-1$. This is due to network delay τ. While the middle-box times out a binding after α unit of time, if a node sends keep-alive after α unit time of silence, the KA packet reaches the middle-box after $\alpha + \tau$ unit time of quietness, where $0 < \tau \ll \alpha$. Therefore, the connection gets dropped. So, from the node's perspective, the binding timeout is $\alpha - 1$. For the same reason, the experimental convergence time curves (not shown for brevity) are also off from their theoretical counterparts by 1 unit along the Y-axis.

B. Keep-alive Interval with Tunable Accuracy

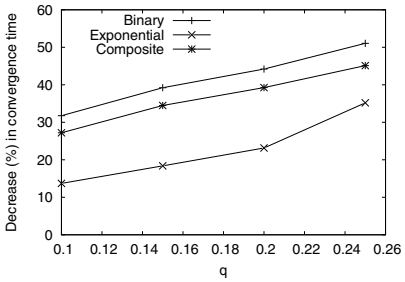
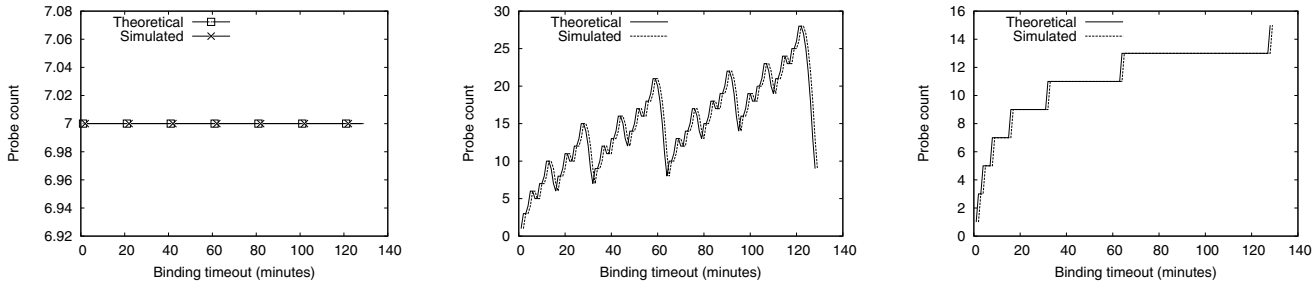


Fig. 5. Average decrease in convergence time due to relaxing accuracy of detected KA interval.

Long convergence time of the iterative probing techniques may result in increased load on the server. To mitigate that, we can sacrifice accuracy of the detected KA interval to reduce the convergence time. We introduce a tunable parameter q in our algorithms for this purpose. Let α be the optimal keep-alive interval. In effort to reduce the convergence time, let us settle on a sub-optimal keep-alive interval α'. We want to ensure that it admits an error that is no more than q fraction of α, for some $0 \leq q < 1$. That is: $\alpha' \geq (1-q)\alpha$. Let us continue testing so long as $\frac{high-low}{high} > q$. In that case, when the testing is completed, we can write:

$$\begin{aligned} \alpha' &= low \\ &\geq (1-q)high \\ &\geq (1-q)\alpha \end{aligned}$$

Thus with this simple modification, we can settle to a sub-optimal keep-alive interval with an assurance that it admits error no more than q fraction of the actual optimal keep-alive interval. We implemented this change and ran our algorithms with different values of q to see the impact on convergence time. As can be seen from Figure 5, the average decrease in



(a) Binary search. (b) Exponential search. (c) Composite search.

Fig. 4. Probe count of different search techniques in presence of network delay.

convergence time shows a linear relationship with q. Binary search has the most reduction in convergence time.

It is not necessary to hard code the value of q into the algorithm. Rather, it could be tuned dynamically. The server could send different values of q at different times to different nodes, if it needed to balance some load. The value could be embedded into the response to the KA message sent by the client.

C. Number of Keep-alives Sent

From the commencement of testing, we counted the total number of keep-alive packets sent through the data and test connection over a period of time. The smaller this number, the better it is. Because, each time a keep-alive is sent, there is the cost of bringing the radio to high power state, if the device was otherwise idle. We conducted this experiment for several different time durations that reflect some real world scenarios, where a device remains in the same network settings for some amount of time.

Firstly, we performed a 30 minute and a 1 hour run. These are typical durations of meetings in any organization.We also simulated a 2 hour run. This matches the duration of graduate classes, seminars or mini workshops. We plot the number of keep-alives sent over the data and the test connection against the binding timeout of middle-box. The resulting curves are shown in Figure 6. No testing is performed in case of the curve marked as 'Default'. One keep-alive packet is sent every minute in this case. On the other hand, 'Oracle' curve represents the behavior of a system that knows the optimal keep-alive interval apriori. Observe that the curve for binary search matches the Default curve in Figure 6a and Figure 6b. This is because the wait time for sending the first probe over the test connection is longer (65 minutes) than the duration of the scenario. As such, no improvements could be offered by the binary search technique. Figure 6c shows reduction in number of keep-alive sent using binary search technique. The duration of this run permitted one or more probes, which results in improved keep-alive interval.

Let us take a look at the behavior of composite and exponential search, on the contrary, in Figure 6. The curves for these approaches look identical. Both these techniques are able to reduce the number of keep-alive packets sent significantly. This is because, the initial improvements to the keep-alive intervals during the testing happens much earlier in these approaches. The curves are within 20 units (along Y-axis) of the Oracle curve in Figure 6c.

TABLE III
AVERAGE REDUCTION (%) OF NUMBER OF KEEP-ALIVES SENT WHEN TESTING IS PERFORMED TO IMPROVE THE KEEP-ALIVE INTERVAL FROM THE CONSERVATIVE DEFAULT VALUE.

Experiment duration	Binary search	Exponential search	Composite search
30 minutes	0	64.30	64.25
1 hour	0	77.79	77.83
2 hours	26.79	84.92	85.05
6 hours	71.62	90.88	91.06
8 hours	77.52	91.74	92.09
12 hours	83.47	92.84	93.13
24 hours	89.61	94.12	94.39

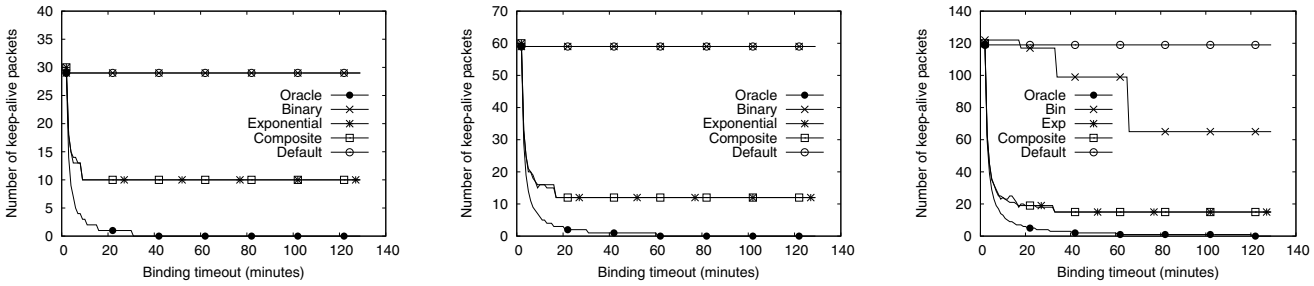
Next, we experimented with 8, 12 and 24 hour long runs. The number of keep-alives sent in all these cases, for the different search techniques are shown in Figure 7. In all these runs, composite search and exponential search curves are almost identical and they approach the 'Oracle' curve with increasing study durations.

For each different binding timeout, we calculated the percentage reduction in the number of KA packets sent using the different search techniques, compared to no testing. Then we averaged this over the entire range of binding timeouts used in the experiments: [1, 128]. This average percentage reduction in number of KA packets sent is listed in Table III. Based on this result, it is clear that exponential and composite search techniques should be preferred to binary search. And since composite search has the better convergence time of the two, it is the best choice.

D. Impact of Packet Failure

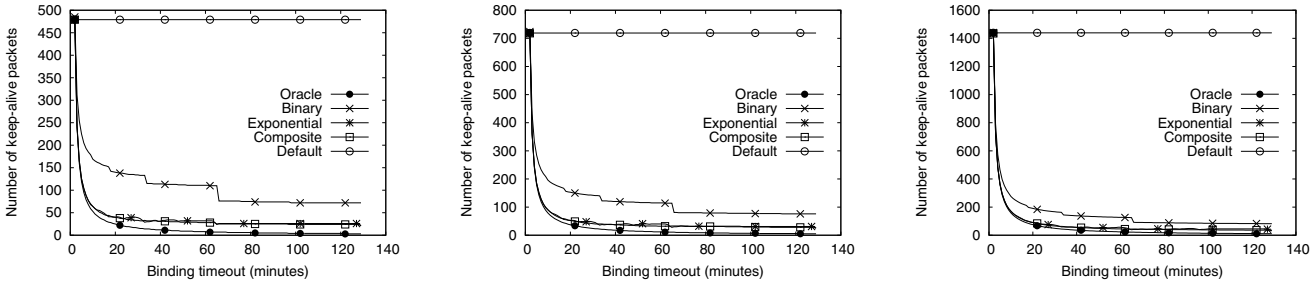
So far, when a keep-alive message fails over the test connection, we have assumed that the middle-box has dropped the connection due to too long an idleness. However, a packet could also be dropped for some transient issues in the network. The sender has no way to differentiate between the different causes of failure. As such, the technique we developed may not considerably improve KA interval.

In this experiment, we try to simulate transient network failures and observe the behavior of the different search tech-



(a) 30 minute run. (b) 1 hour run. (c) 2 hour run.

Fig. 6. Number of keep-alive packets sent during meeting or seminar scenario.



(a) 8 hour run. (b) 12 hour run. (c) 24 hour run.

Fig. 7. Number of keep-alive packets sent during longer durations.

niques. To model the transient failures, we define a parameter p that represents the probability that a keep-alive message will fail. The message failures are independent of each other. For different values of p, we observe the detected binding timeout. For each value of p, we repeated our experiment 20 times and averaged the results over those runs. Figure 8 plots the detected binding timeouts against the actual binding timeouts. The curve marked as 'No Error' represents the detected binding timeout when there is no error involved. From these curves, it is clear that, impact of packet failure cannot be neglected.

E. Retry on Packet Failure

We modified each algorithm as follows: When keep-alive message fails, we re-establish the TCP connection and test the same interval again. It is unlikely that both packets will encounter the transient failure. So, if the latter attempt fails too, we conclude that we have overshoot the optimal KA interval. With this modification, we re-ran the same experiments. The results are shown in Figure 9. For $p = 0.02$ and $p = 0.05$, the error in the detected binding timeout is very negligible.

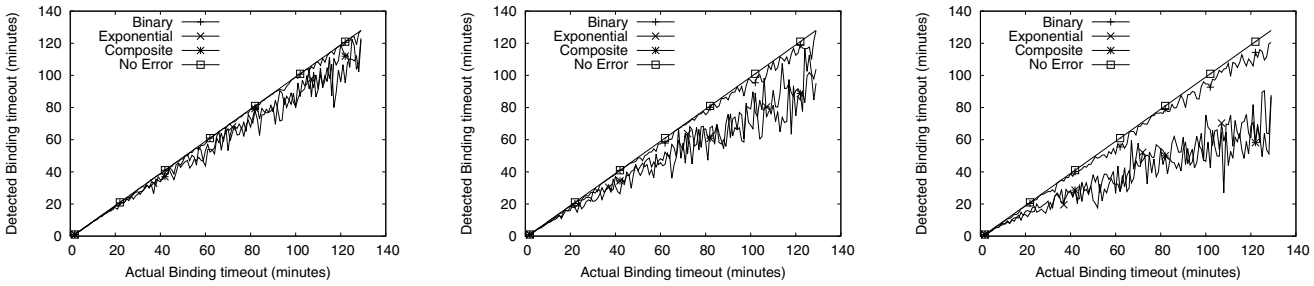
For $p = 0.10$, the retry scheme in binary search approach is quite successful in coping with transient network issues. On average, the error in detected binding timeout is less than 0.5%. In case of exponential search, the error remaining in detected binding timeout is around 5%, on average; and for composite search, it is around 3.5%. While the error is not negligible, we think this is within tolerable limit.

With single retry scheme, we encounter double the wait time for each valid keep-alive failure. Hence the convergence time grows significantly. Binary search encounters the most increase in convergence time: around 60% on average. Probe count in each technique is also increased due to the retry scheme. Since the error in detected binding timeout is in tolerable range and there is significant impact on convergence time and probe count for each number of additional retries, we would not implement more than single retry on packet failure.

With the retry scheme, we would now like to know how well do the different techniques perform in reducing the number of keep-alive packets sent over the data and the test connection combined. Like earlier, we conducted runs of different durations reflecting real world scenarios. The packet failure probability p was set to 0.02. q was set to 0.10. (Recall that q is the allowed error in the detected binding timeout, measured as a fraction of actual binding timeout.)

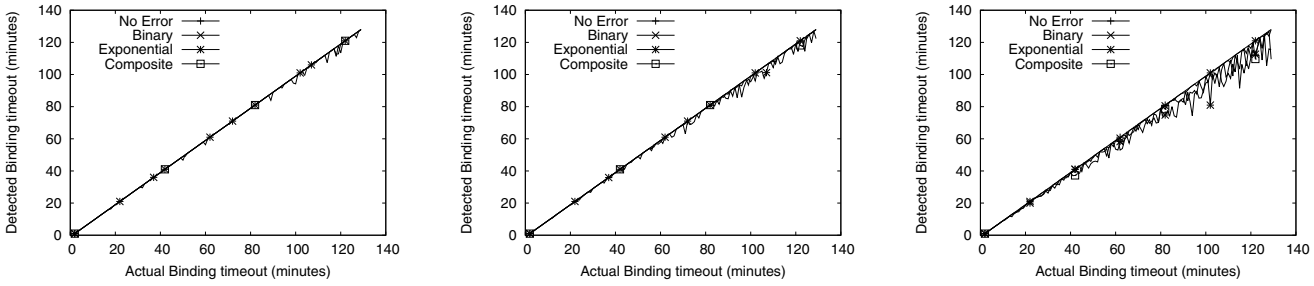
The result for each binding timeout was averaged over 20 repetitions. The results are shown in Figure 10 and Figure 11. These curves are very similar to the corresponding curves of Figure 6 and Figure 7. The average percentage reduction in number of KA packets sent is listed in Table IV. The values are comparable to the corresponding values in Table III.

Therefore, we can conclude that the retry scheme successfully coped with transient network failures and was able to reduce the number of keep-alive packets sent over the data and test connection considerably. Using the q parameter, some accuracy of detected binding timeout is sacrificed to reduce the convergence time. Even then, the reduction in number of keep-alive packets sent were very much comparable with earlier experiment. Overall, based on all the experiments, we conclude that composite search technique should be used to dynamically improve keep-alive interval.



(a) $p = 0.02$. (b) $p = 0.05$. (c) $p = 0.10$.

Fig. 8. Detected vs. actual binding timeout in presence of transient network errors.



(a) $p = 0.02$. (b) $p = 0.05$. (c) $p = 0.10$.

Fig. 9. Detected vs. actual binding timeout when retry scheme is applied in presence of transient network errors.

TABLE IV
AVERAGE REDUCTION (%) OF NUMBER OF KEEP-ALIVES SENT WHEN TESTING IS PERFORMED TO IMPROVE THE KEEP-ALIVE INTERVAL FROM THE CONSERVATIVE DEFAULT VALUE. HERE $p = 0.02$ AND $q = 0.10$.

Experiment duration	Binary search	Exponential search	Composite search
30 minutes	0	62.37	62.37
1 hour	0	75.44	75.58
2 hours	26.45	82.39	82.62
6 hours	71.59	88.19	88.56
8 hours	77.57	89.03	89.52
12 hours	83.59	90.10	90.56
24 hours	89.61	91.32	91.69

F. Proof of Concept Implementation

We implemented the composite search technique with no retry on an Andriod device. We deployed a server at kopottakha.cs.uiuc.edu on port 8080 and had the device connect to it. This proof-of-concept system only used a test connection to detect the optimal KA interval. Updating the KA interval on the data connection was not done - that would require changes in the OS code. We used our system in 2 different mobile operator's networks. The optimal KA interval detected were 9 and 10 minutes. The sequence of tests and the convergence time in the real implementation matched with our analytical bounds. We also ran the test over WiFi networks.

VI. CONCLUSION

In this research, we applied several iterative probing techniques to dynamically adapt the keep-alive interval of long-lived TCP connections. These include binary search, exponential search and composite search. We performed theoretical analysis as well as experiments on a simulation platform to compare these techniques. To the best of our knowledge, such analysis has not been done in any earlier work. We evaluated the performance of our techniques by varying different parameters and found composite search to be the best choice.

Other search techniques could be explored in future research work. In particular, searching with multiple test connections can reduce the convergence time significantly and also improve the keep-alive intervals quickly. However, care should be taken to not overload the server with too many connections. Occasionally, it is possible that due to changes in the network infrastructure, the binding timeout of the middle-box has decreased. In that case, the data connection will experience frequent disconnections. Experiments should be conducted to develop a strategy to bring the data connection out of this unstable state (frequent disconnections). Finally, work could be done to write a redistributable library that can be plugged into any device to improve the keep-alive interval. The API and protocol design for such a library remains to be investigated.

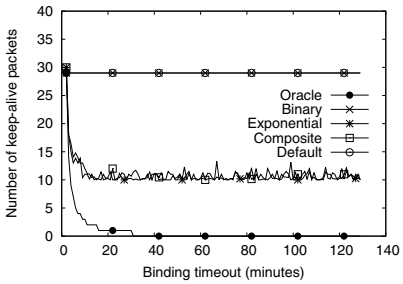
ACKNOWLEDGMENT

The authors would like to acknowledge all the anonymous referees for their valuable suggestions.

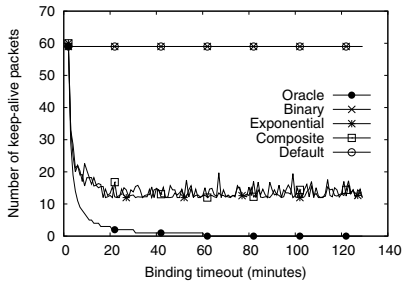
REFERENCES

[1] S. Hätönen, A. Nyrhinen, L. Eggert, S. Strowes, P. Sarolahti, and M. Kojo, "An experimental study of home gateway characteristics," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 260–266.

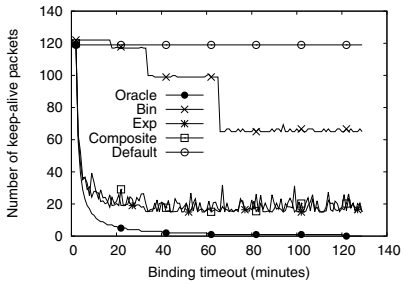
[2] R. Braden, "Requirements for Internet hosts-communication layers," October 1989, RFC 1122 (Standard).



(a) 30 minute run.

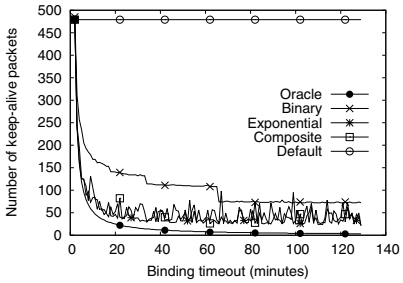


(b) 1 hour run.

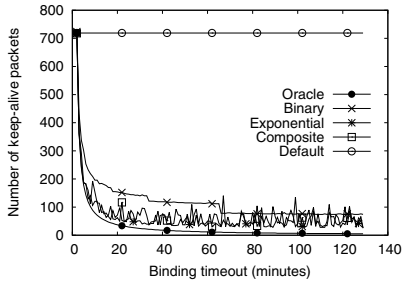


(c) 2 hour run.

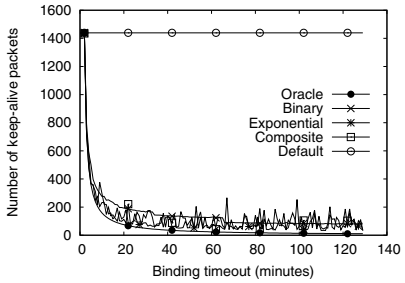
Fig. 10. Number of keep-alive packets sent during meeting or seminar scenario. ($p = 0.02$, $q = 0.10$).



(a) 8 hour run.



(b) 12 hour run.



(c) 24 hour run.

Fig. 11. Number of keep-alive packets sent during longer durations. ($p = 0.02$, $q = 0.10$).

[3] S. Guha, K. Biswas, B. Ford, S. Sivakumar, and P. Srisuresh, "NAT Behavioral requirements for TCP," October 2008, RFC 5382 (Best Current Practice).

[4] M. Stiemerling, E. Davies, C. Aoun, and H. Tschofenig, "NAT/Firewall NSIS Signaling Layer Protocol (NSLP)," October 2010, RFC 5973 (Experimental).

[5] M. Stiemerling, J. Quittek, and C. Cadar, "NEC's Simple Middlebox Configuration (SIMCO) Protocol Version 3.0," May 2006, RFC 4540 (Experimental).

[6] "Understanding Direct Push," http://technet.microsoft.com/en-us/library/aa997252(EXCHG.80).aspx, [Online; Last accessed on 22-Oct-2015].

[7] "Heartbeat Interval Adjustment," http://technet.microsoft.com/en-us/library/cc182270.aspx, [Online; Last accessed on 22-Oct-2015].

[8] "Apple Push Notification Service," https://developer.apple.com/library/ios/documentation/NetworkingInternet/Conceptual/RemoteNotificationsPG/Chapters/ApplePushService.html, [Online; Last accessed on 22-Oct-2015].

[9] "Google Cloud Messaging: Overview," https://developers.google.com/cloud-messaging/gcm, [Online; Last accessed on 22-Oct-2015].

[10] "Push Notifications (Windows Phone)," https://msdn.microsoft.com/en-us/library/hh221549.aspx, [Online; Last accessed on 22-Oct-2015].

[11] "Windows Push Notification Services (WNS) overview (Windows Runtime apps)," http://msdn.microsoft.com/en-us/library/windows/apps/hh913756.aspx, [Online; Last accessed on 22-Oct-2015].

[12] S. R. Gatta, K. Srinivasan, O. N. Ertugay, D. G. Thaler, D. A. Anipko, J. Vanturennout, M. S. Rahman, and P. R. Gaddehosur, "Keep alive management," November 2014, US Patent No. 8,892,710 B2.

[13] S. Herzog, R. Qureshi, J. Raastroem, X. Bao, R. Bansal, Q. Zhang, and S. M. Bragg, "Determining an efficient keep-alive interval for a network connection," February 2013, US Patent No. 8,375,134 B2.

[14] R. Seggelmann, M. Tuexen, and M. Williams, "Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS) Heartbeat Extension," February 2012, RFC 6520 (Standard).

[15] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) protocol version 1.2," August 2008, RFC 5246 (Standard).

[16] H. Haverinen, J. Siren, and P. Eronen, "Energy consumption of always-on applications in WCDMA networks," in *Proceedings of IEEE Vehicular Technology Conference*. IEEE, April 2007, pp. 964–968.

[17] N. Vallina-Rodriguez and J. Crowcroft, "Energy management techniques in modern mobile handsets," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 1, pp. 179–198, 2013.

[18] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009, pp. 280–293.

[19] D. J. Wetherall and A. S. Tanenbaum, "Computer Networks," 1996.

[20] M. Jain and C. Dovrolis, "End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput," in *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4. ACM, 2002, pp. 295–308.

[21] A. Akella, S. Seshan, and A. Shaikh, "An empirical evaluation of wide-area internet bottlenecks," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*. ACM, 2003, pp. 101–114.

[22] N. Hu and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 879–894, 2003.

[23] B. Melander, M. Bjorkman, and P. Gunningberg, "A new end-to-end probing and analysis method for estimating bandwidth bottlenecks," in *Global Telecommunications Conference, 2000. GLOBECOM'00*, vol. 1. IEEE, 2000, pp. 415–420.

[24] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathChirp: Efficient available bandwidth estimation for network paths," in *Proceedings of Passive and active measurement (PAM) workshop*, April 2003.

[25] M. Jain and C. Dovrolis, "End-to-end estimation of the available bandwidth variation range," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1. ACM, 2005, pp. 265–276.

[26] M. S. Rahman, "On Approaches to Detect Optimal Keep-alive Interval of TCP Connections." http://1drv.ms/1kktYFu, M.Sc. thesis, submitted to the Department of Computer Science and Engineering (CSE), Bangladesh University of Engineering and Technology (BUET). [Online; Last accessed on 22-Oct-2015].

[27] "Omnet++," http://www.omnetpp.org/, [Online; Last accessed on 22-Oct-2015].

SuperCrypt: A Technique for Quantum Cryptography through Simultaneously Improving Both Security Level and Data Rate

Kazi Sinthia Kabir, Tusher Chakraborty, and A.B.M. Alim Al Islam
Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
Email: { sinthia.096@gmail.com, tusherchakraborty@matholympiad.org.bd, alim_razi@cse.buet.ac.bd

***Abstract*—Secured data transmission has always been a matter of great interest and several encryption techniques for secured data transmission have been devised till now in this regard. However, with rapidly developing technology, the available encryption techniques are becoming more prone to attack every day. Focusing this issue, in this paper, we propose a new technique of encryption that will enhance the security level by a significant margin with the help of quantum computing. Additionally, our proposed technique also enhances data transmission rate through exploiting the notion of Superdense Coding. Such simultaneous improvement in both security level and data transmission rate, which we achieve through our proposed technique, is a rare trait for currently available encryption technique. We name our proposed technique SuperCrypt. We elaborate implementation issues pertinent for SuperCrypt such as recycling qubits and re-establishing entanglement. We also implement SuperCrypt in a discrete-event network simulator `ns-2` and evaluate the performance of SuperCrypt through analyzing the simulation results. The simulation results demonstrate significant performance improvement through SuperCrypt compared to available classical alternative. Further, we briefly present a more sophisticated and synchronized version of SuperCrypt that we plan to investigate in future.**

I. Introduction

Since the very early ages, secured transfer of information has been a matter of concern for people. With the rapidly developing technology, the need for secured data transfer is growing every day. For this reason, the study of cryptography is becoming more important day-by-day. At the ancient times, the main focus of cryptography was encryption. Though different methods such as encryption, authentication, non-repudiation, etc, have been applied in modern times, encryption still remains a very important part of cryptography.

There are several techniques of encryption in the literature such as one-time padding [1], RSA [2], AES [2], ElGamal [3], quantum cryptography [1], etc. However, recent studies show that none of these mechanisms is secure enough [1], [3]–[5]. In fact, RSA encryption, which was considered to be the most secure one among all the available alternatives, becomes vulnerable under the exploitation of quantum technology [1], [4]. Therefore, the search for efficient encryption mechanism is still going on.

To address this issue, in this paper, we propose a new technique for encryption that will provide enhanced security level. In addition to enhancing the security level, our proposed technique offers an increased data transmission rate, which is a kind of rare for conventional cryptography techniques. Here, we apply a two-level operation on the encryption key. Additionally, we exploit a unique feature of quantum computing, called Superdense Coding, in combination with the two-level operation that contributes in an increasing data transmission rate. We term our proposed technique SuperCrypt.

In SuperCrypt, first, the actual message is encrypted by exclusive-OR (XOR) operation with an encryption key. The encrypted message is then transmitted through the classical channel. Here, the bits of the key are first permuted and then encoded using Superdense Coding. The advantage of Superdense Coding is that, during the process, the information is transmitted to the receiving end within a very short delay due to having entangled photons. On the other hand, at the receiving end, the qubits are first measured as a part of decoding process. Afterwards, the measured qubits are re-permuted. This process reverts back the actual encryption key at the receiver end. Subsequently, the key can be used to decrypt the message. Thus, our proposed technique actually exploits the notion of symmetric key cryptography.

Based on our work, we make the following contributions in this paper:

- We propose a new technique of quantum encryption with increased data transmission rate. Here, we analyze the computational complexity of the proposed mechanism for an intruder.
- We investigate several design issues pertinent for our proposed technique such as re-establishment of entanglement
- We simulate the proposed technique in `ns-2`. here, we perform necessary modifications in the simulator to simulate our proposed technique. Our rigorous simulation study reveals different aspects of the performance of our proposed technique. To the best of our knowledge, we are the first to perform such discrete event simulation to evaluate the performance of a quantum networking based system.

II. Related Work

There are several techniques of encryption in the literature such as one-time padding [1], RSA [2], AES [2], ElGamal [3], quantum cryptography [1], etc. In one-time padding, though the level of security of transmitted data is very high, each one-time pad itself needs a secure way of transmission rendering the task of transmitting the pads a bottleneck in implementation of the technique [1]. Therefore, it is difficult (in many cases even impossible) to implement this technique in reality. On the

978-1-5090-0203-0/16/$31.00

other hand, the RSA technique has been believed to be a strong mechanism for encryption till now. However, recent studies show that this mechanism has become vulnerable to recent advancements of quantum technology [1], [4]. Consequently, other classical techniques such as AES, Elgamal, etc., have also started exhibiting vulnerability to the quantum technology.

Now, getting back to the one-time padding technique, quantum cryptography offers a secured technique for transmitting the one-time pads. One popular protocol of quantum cryptography is the BB84 protocol [1]. Unfortunately, several recent research studies demonstrate that this security is also breakable [1], [6]. Moreover, the mechanism of data transmission in quantum cryptography results in a 50% loss of qubits [1]. As a result, the use of BB84 protocol of quantum cryptography in one-time padding suffers from both limited security and limited performance.

Another protocol of quantum cryptography is called Ekert protocol [5] which might be used for better security. However, one limitation to quantum cryptography is that a single photon source actually emits two or more photons. This is called photon number splitting (PNS) [1]. The extra emitted photons can be captured by an intruder resulting in PNS attack [1]. The Ekert protocol states that a true single-photon source might be used to reduce the risk of PNS attack [1]. However, the protocol also states that the risk still remains if pair of photons are used. Moreover, the security of the protocol is higher for longer distances [5]. However, the protocol still depends on the choice of bases selection for measurement. Therefore, 50% loss of resources still exists. In addition to that, longer distance communications faces several losses in quantum channel [7].

Before the evolution of quantum cryptography, RSA technique was considered to be a highly-secured encryption mechanism. The main strength of this technique lies in the fact that factorization of the product of two large prime numbers is computationally very expensive. However, recent studies show that the Shor's algorithm of quantum computing can perform this factorization in polynomial time [8]. In fact, a practical experiment with recycled qubits demonstrated successful factorization of 21 using the quantum algorithm [4]. These studies reveal that, with the rapid development of quantum technology, the classical algorithm of RSA encryption is becoming vulnerable day-by-day.

AES or Advanced Encryption Standard algorithm [9] is based on a combination of both substitution and permutation of the message. Until May 2009, it was considered to be a secured mechanism. However, several studies showed that the security of AES encryption is breakable [9]–[11]. Additionally, Elgamal encryption system [3] is another well-investigated asymmetric key encryption system. The security of this algorithm depends upon the difficulty of computing discrete logarithms [3]. However, it has been proven that this system is also vulnerable under chosen ciphertext attack [12].

These studies show that none of the classical encryption mechanisms can be considered to be secure enough now-a-days. Therefore, we move forward to investigate the quantum computation and quantum cryptography mechanism in greater detail. before presenting the outcomes of our investigation, we present the basic of quantum computation and quantum cryptography in the following section.

III. Basics of Quantum Computing and Quantum Cryptography

Quantum computing is a computation theory that makes direct use of quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. The phenomena enables several specialized mechanisms such as Superdense Coding [13]. In our proposed technique, a very important step is Superdense Coding [14] of classical information with qubit. Therefore, we discuss a few necessary basics of qubits and Superdense Coding in this section. Additionally, as our focus in this paper is on cryptography, we briefly present basics of quantum cryptography in this section.

A. Quantum Computing

Quantum computing deals with quantum information, which is based on an analogous concept of bit called quantum bit or qubit. A classical bit has a state of either 0 or 1. A qubit, on the other hand, also has a quantum state that can be a superposition of both the classical states (0 and 1) at the same time. This quantum state is a linear combination of the classical states, which is often called superposition state.

The superposition state can be written as: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are probability amplitudes and can, in general, both be complex numbers. The states $|0\rangle$ and $|1\rangle$ are called computational basis states, which form an orthonormal basis for computation in a vector space [15]. With the help of the notion of superposition states, quantum computation allows a huge number of calculations that can be simultaneously carried out. A quantum computer with 400 basic units (qubits) could, for example, simultaneously process more bits of information than the number of atoms in the universe [16]. Such enormous processing power has driven towards devising new coding techniques that can deal with qubits. Superdense Coding is one of such techniques.

In quantum information theory, Superdense Coding refers to a technique used to send two bits of classical information using only one qubit [13]. The detail steps of Superdense Coding are discussed in later sections while elaborating our proposed method. In general, Superdense Coding requires entanglement between sender and receiver devices. Here, the notion of quantum entanglement [17] refers to a quantum mechanical phenomenon in which the quantum states of two or more objects have to be described with reference to each other even though the individual objects may be spatially separated.

Quantum entanglement occurs when particles such as photons, electrons, molecules, or even small diamonds interact physically in a certain way and then get separated. The interaction ensures that each resulting member of a pair is properly described by the same quantum mechanical description or state. The state can correspond to a number of factors [17] such as position, momentum, spin polarization, etc.

B. Quantum Cryptography

Quantum cryptography refers to the use of quantum mechanical properties to perform cryptographic tasks. The advantage of quantum cryptography lies in the fact that it allows the completion of various cryptographic tasks that are proven or conjectured to be impossible using only classical (i.e., non-quantum) computation [14]. The basic protocol of quantum

cryptography is usually explained as a method of securely communicating a private key from one party to another for using in one-time pad encryption [13].

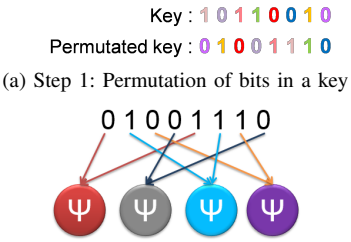
The traditional quantum cryptography relies on correct selection of bases for measurement of qubits. In this protocol, a sender encodes her one-time pads in strings of qubits through performing some quantum operations using particular bases. She then sends it over a public quantum channel. However, since only the sender knows the actual bases of her quantum operations, it is impossible for the receiver to distinguish all original states of the qubits. Here, as the receiver independently chooses own bases while receiving and decoding the received qubits back to the original pads, her probability of guessing the right bases is $\frac{1}{2}$. Therefore, to construct an n-bit one-time pad, $2n$ qubits are needed to be transmitted on an average. This is certainly a loss of qubits [13]. Such loss of qubits is a strong motivation behind our proposed technique. In the next section, we further discuss other motivations of our study.

IV. Motivation Behind Our Study

Among various encryption techniques available now-a-days, RSA is considered to be the most widely used one. It is also considered to be one of the most secure encryption technique for data transmission. However, recent studies demonstrate that the security of RSA technique can be easily broken by Shor's algorithm of quantum computing [1], [4]. Similar outcome is expected for other classical encryption algorithms. Consequently, quantum cryptography seems to be the only possible way-out for securing data transmission.

Security in traditional quantum cryptography is based on the selection of bases for measurement of photons, which carry the values of qubits. If an intruder fails to select the bases correctly, he cannot extract the data [1]. However, in recent times, quantum cryptography is also under attack. As stated earlier, a limitation to quantum cryptography is that a single photon source actually can emit two or more photons simultaneously. This is called Photon Number Splitting (PNS) [1]. The extra emitted photons can be captured by an intruder resulting in PNS attack [1]. Moreover, the security of traditional quantum cryptography is based on the fact that intruder fails to extract the information if he fails to select the proper basis of measurement [1]. However, in optically controlled quantum systems, there is no need to perform the measurement from any particular basis [6]. This weakens the security of traditional quantum cryptography. Therefore, we need a more secured encryption method pertinent for the quantum cryptography.

Besides, as we have already discussed in the previous section, doubled number of photons (i.e., $2n$ photons) are required to construct an n-bit pad or key in traditional quantum cryptography. This results in half data rate in transmission of the key [1]. As the key is used only once in continuous transmission using one-time padding technique, the resulting half rate can significantly threaten the performance of secured data transmission. Therefore, we attempt to propose a new technique that will simultaneously improve both the security level and the data transmission rate. In the next section, we present an overview of our proposed technique, which we name as SuperCrypt.



(a) Step 1: Permutation of bits in a key

Key bit positions		Photon position
1	5	1
4	8	2
2	7	3
3	6	4

(b) Step 2: Superdense Coding on the permuted bits

Fig. 1: Encoding process in the sender device

V. System-Level Overview of SuperCrypt

In our proposed technique, we attempt to devise a method for enhancing the security level of the key used in one-time padding encryption mechanism. The encryption in one-time padding starts through selecting a key at the sender, Alice. Alice performs an exclusive-OR (XOR) operation between the selected key and the message to be transmitted. She then transmits the encoded message. However, Alice needs to ensure the security of the key itself as the key needs to be transmitted over the same channel.

In SuperCrypt, Alice needs to perform two operations to enhance the security of the key to be transmitted. First, she performs a permutation operation on the bit sequence of the key. In the second step, she takes a pair of bits from the modified bit sequence and performs Superdense Coding on those. As a result, an n bit key is encoded in $\frac{n}{2}$ qubits in SuperCrypt. Fig. 1 presents the whole encoding process of SuperCrypt.

In Superdense Coding, if the sender wants to transmit a 2-bit message, e.g., 00, 01, 10, or 11 to the receiver, she first performs a single qubit operation on her qubits. The sender selects the operation according to the content of the message under transmission as follows:

- I gate operates on message 00
- X gate operates on message 01
- Z gate operates on message 10
- iY gate operates on message 11

Here, X, Y, and Z are the basic quantum gates [13]. Besides, applying I gate refers to doing nothing and applying iY gate refers to applying both X and Z gates together. Applying these gates transforms an EPR pair into the four Bell (EPR) states $|\Psi_{00}\rangle$, $|\Psi_{01}\rangle$, $|\Psi_{10}\rangle$, and $|\Psi_{11}\rangle$ respectively as [13]

- 00: $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle)\rightarrow\frac{1}{\sqrt{2}}(|00\rangle+|11\rangle)=|\Psi_{00}\rangle$
- 01: $\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle)\rightarrow\frac{1}{\sqrt{2}}(|10\rangle+|01\rangle)=|\Psi_{01}\rangle$

- 10: $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle) \rightarrow \frac{1}{\sqrt{2}}(|00\rangle-|11\rangle) = |\Psi_{10}\rangle$
- 11: $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle) \rightarrow \frac{1}{\sqrt{2}}(|01\rangle-|10\rangle) = |\Psi_{11}\rangle$

Here, we assume that the rules of permutation and Superdense Coding are decided or shared by both sender and receiver before establishing the system. As a result, the receiver already knows the operations needed to decrypt the key. It is worth mentioning that the rules need only a small piece of information to be decided or shared beforehand according to our assumption.

At the receiving end, the receiver, Bob receives the qubits transmitted by Alice. He performs the same operations performed by Alice in the reverse order. That means, he first decodes the Superdense Coded qubits, and then puts them in correct bit positions. We elaborate the steps of decoding process in more detail through the following cases:
Case 1: Sender's message is 00, which corresponds to $\Psi_{00}\rangle$ Bell state in transmission. The decoding process exhibits the following:

$$|\Psi_{00}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) = \frac{1}{\sqrt{2}}(|0\rangle|0\rangle + |1\rangle|1\rangle)$$

Step 1: Applying CNot gives $\frac{1}{\sqrt{2}}(|0\rangle|0\rangle + |1\rangle|0\rangle)$

Step 2: Applying H on the 1^{st} qubit gives

$$\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}((|0\rangle + |1\rangle)|0\rangle + (|0\rangle - |1\rangle)|0\rangle) =$$
$$\frac{1}{2}(|00\rangle + |10\rangle + |00\rangle - |10\rangle) = |00\rangle$$

Here, CNot and H are two quantum gates [13]. The receiver now measures both qubits to get the sender's message 00.
Case 2: Sender's message is 01, which corresponds to $\Psi_{01}\rangle$ Bell state in transmission. The decoding process exhibits the following:

$$|\Psi_{01}\rangle = \frac{1}{\sqrt{2}}(|10\rangle + |01\rangle) = \frac{1}{\sqrt{2}}(|1\rangle|0\rangle + |0\rangle|1\rangle)$$

Step 1: Applying CNot gives $\frac{1}{\sqrt{2}}(|1\rangle|1\rangle + |0\rangle|1\rangle)$

Step 2: Applying H on the 1^{st} Qubit gives

$$\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}((|0\rangle - |1\rangle)|1\rangle + (|0\rangle + |1\rangle)|1\rangle) =$$
$$\frac{1}{2}(|01\rangle - |11\rangle + |01\rangle + |11\rangle) = |01\rangle$$

The receiver now measures both qubits to get the sender's message 01.
Case 3: Sender's message is 10, which corresponds to $\Psi_{10}\rangle$ Bell state in transmission. The decoding process exhibits the following:

$$|\Psi_{10}\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle) = \frac{1}{\sqrt{2}}(|0\rangle|0\rangle - |1\rangle|1\rangle)$$

Step 1: Applying CNot gives $\frac{1}{\sqrt{2}}(|0\rangle|0\rangle - |1\rangle|0\rangle)$

Step 2: Applying H on the 1^{st} qubit gives

$$\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}((|0\rangle + |1\rangle)|0\rangle + (|1\rangle - |0\rangle)|0\rangle) =$$
$$\frac{1}{2}(|00\rangle + |10\rangle + |10\rangle - |00\rangle) = |10\rangle$$

The receiver now measures both qubits to get the sender's message 10.
Case 4: Sender's message is 11, which corresponds to $\Psi_{11}\rangle$ Bell state in transmission. The decoding process exhibits the following:

$$|\Psi_{11}\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle) = \frac{1}{\sqrt{2}}(|0\rangle|1\rangle - |1\rangle|0\rangle)$$

Step 1: Applying CNot gives $\frac{1}{\sqrt{2}}(|0\rangle|1\rangle - |1\rangle|1\rangle)$

Step 2: Applying H on the 1^{st} qubit gives

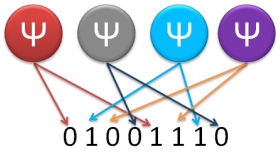
$$\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}((|0\rangle + |1\rangle)|1\rangle - (|0\rangle - |1\rangle)|1\rangle) =$$
$$\frac{1}{2}(|01\rangle + |11\rangle - |01\rangle + |11\rangle) = |11\rangle$$

The receiver now measures both qubits to get the sender's message 11.

This decoding process confirms providing Bob an n bit key from $\frac{n}{2}$ qubits. Subsequently, he performs the permutation as decided earlier and gets the original bit sequence. At the end of this phase, Bob has the exact key or one-time pad to decrypt the message. Fig. 2 presents the full decoding process.

We present all the operations at sender and receiver ends in Fig. 3. In this system, the two-step encryption of the key makes it extremely difficult for any intruder to decrypt and get the key. To demonstrate the level of difficulty, we analyze the complexity of decryption for an intruder in the next section.

Photon position	Key bit positions	
1	1	5
2	4	8
3	2	7
4	3	6



(a) Step 1: Superdense decoding on the permuted bits

Received key : 0 1 0 0 1 1 1 0

Reverse Permutated Key: 1 0 1 1 0 0 1 0

(b) Step 2: Re-permutation of bits in a key

Fig. 2: Decoding process in the receiver device

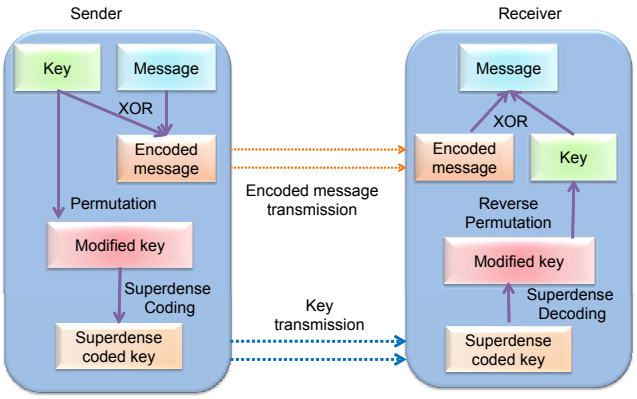


Fig. 3: Block diagram of our proposed encryption system SuperCrypt

VI. Strengths of Our Proposed Technique

Our proposed technique of encryption, SuperCrypt focuses on simultaneously improving both data transmission rate and security level. In this section, we elaborate the strengths of SuperCrypt technique in more detail.

A. Enhancement in Data Transmission Rate

SuperCrypt technique of encryption results in a higher data transmission rate than the existing mechanisms. This happens as one of the steps of the encoding phase involves Superdense Coding. Superdense Coding enables data transmission at double rate through transmitting $2n$-bit key using only n photons. Thus, this method supports an increased data transmission rate.

B. Higher Level of Security

As we have mentioned earlier, SuperCrypt simultaneously improves both security level and data transmission rate. The encoding mechanism in SuperCrypt involves permutation of the bits and selection of bits for Superdense Coding. This step results in a high computational complexity for an intruder to extract the transmitted information. To better quantify the computational complexity, the following section analyzes the complexity for an intruder to extract the information from an encoded message.

C. Computational Complexity for an Intruder

According to the decryption method of our proposed technique, the intruder, Trudy needs to perform two steps for successful decryption. First, he has to decode the Superdense Coded information and put the decoded bits in correct positions. Since we can select 2 bits from the n bits of of the key in ${}^nC_2$ possible ways, Trudy needs to try ${}^nC_2$ possible combinations. Then, he needs to perform permutation of the bit sequence to find out the correct bit pattern. For this step, he needs to try $n!$ combinations. Therefore, the intruder needs to try X number of bit patterns where,

$$X = n! \times^n C_2 = n! \times \frac{n^2 - n}{2}$$

This makes the computational complexity of the intruder to be $O(n^2 n!)$. Obviously, the complexity demonstrates that decoding a key by Trudy demands a significant amount of time in SuperCrypt ensuring a very high level of security. Nonetheless, we evaluate other operational performances of SuperCrypt through discrete event simulation. Before presenting the simulation results, we briefly illustrate some implementation issues of our proposed technique.

VII. Implementation Issues in SuperCrypt

While studying any entangled qubit-based system, one major consideration is that qubits are one time resources. If a qubit is measured, the corresponding particle will enter into an indeterminate state. As a result, the entanglement will fail to sustain. Therefore, to increase life-time of the system, we need to overcome this limitation. This happens as, in SuperCrypt, we exploit Superdense Coding, which is based on entanglement between sender and receiver.

Fortunately, recent studies have found that qubits can be recycled [8], which makes the qubits re-usable. Consequently, we need to consider re-establishment of entanglement only. However, as the sender and the receiver devices are spatially separated and no physical interaction is possible between the qubits of the two devices, we need to re-establish the entanglement without demanding any physical interaction between the two qubits.

There are a few methods of re-establishing entanglement between two remote qubits without any physical interaction between them. We briefly discuss such methods below:

A. Re-establishing Entanglement using Entanglement Swapping

Though creating entangled qubit pairs in an intuitive manner requires direct interaction of the particles, it is possible to create entanglement between photons that never coexisted in time. Entanglement swapping exploits this notion in accordance with utilizing an intermediate actor for restoring the sender's qubits. Entanglement swapping is performed in the following way [18]:

- The sender has a particle, which is entangled with a particle owned by the receiver.
- The receive teleports it to an intermediate actor.
- The sender's particle is entangled with the one of the intermediate actor.

Fig. 4 presents a pictorial representation of entanglement swapping. In the figure, the blue dotted lines represents that the connected particles are entangled. The purple dotted line represents that the particles of the sender and the intermediate actor are entangled though they have never co-existed. Here, the phenomena of permitting no co-existence enables re-use of the intermediate actor, and thus, to use only one intermediate actor for the complete system. Re-establishment of the states of sender's qubits using the only intermediate actor facilitates measuring qubits retaining the entanglement between sender and receiver.

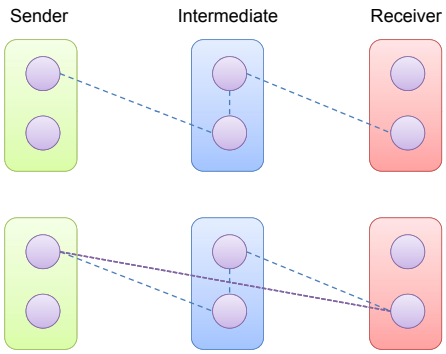


Fig. 4: Block diagram of re-establishing entanglement using entanglement swapping

B. Designing the Measurement in a Different Way

Another method of retaining entanglement is to use microwave radiation for measuring qubits and thereby entangle superconducting circuits separated spatially [19]. It might seem a bit confusing to produce entanglement through measurements since measurements are typically known to destroy quantum coherence. However, the method used in the research study presented in [19] focuses on designing the measurements in such a way that it cannot distinguish between all the states that the system might occupy, rather it distinguish only between different subsets of states.

The measurement process starts through preparing the system in a uniform coherent superposition of all possible states. Then, a measurement probabilistically yields a result corresponding to one of the subsets of the states. This action leaves the system in a fully coherent superposition of the states within that subset. This superposition can be further entangled. Note that, a similar process has been successfully conducted through an experiment preparing an entangled state with 100% probability [20]. Therefore, we can exploit such techniques to prepare entangle qubits that are spatially separated.

It is worth mentioning that, since entanglement is a very unique feature of quantum technology, various other studies have been performed to prepare entanglement between qubits without direct physical interaction. Several achievements in this field show that preparing entangled qubits is not a barrier anymore [21], [22].

C. Channel Errors in Quantum Channel

When we consider quantum networking using quantum teleportation, we consider that there will be no loss. However, in reality, we might face some losses in different phases of the transmission system such as at the sender device, over the transmission medium, and at the receiver device.

In case of loss over the transmission medium, the beam of photon generated at the sender side and transmitted to receiver side will become a spread beam due to diffraction [7]. This diffraction has adverse effect on the data represented by the photon [13]. Therefore, such diffraction is considered as vital loss in transmitted data. The effect of diffraction can be presented by the radius of photon. The radius p_d, of the purely diffraction-limited spot size of the beam incident upon a flat receiving plane at the receiver device is given by [7] the following

$$\rho_d = \sqrt{\frac{4L^2}{(kD_A)^2} + \left(\frac{D_A}{2}\right)^2}$$

where L is the path length over which the signal propagates, D_A is the diameter of the aperture of sender's transmitting instrument, and k is the wavenumber of the photon in the beam [7]. Here, wavenumber is the inverse of wavelength $i.e.$ $\frac{1}{wavelength}$.

In addition to diffraction loss, we may also experience other types of system losses. The most prominent one among all the system losses is the line attenuation of the quantum channel. The general expression for line attenuation, α_{fiber} is given as [7],

$$\alpha_{fiber} = 10^{-\frac{AL_{fiber}+b}{10}}$$

where L_{fiber} is the length of the channel connecting the sender and the receiver, A is the parameter that measures the intrinsic loss characteristic per unit length of the channel (provided by the manufacturer), and b is the "bulk loss" constant associated with the fiber. Here, the fiber is needed for re-establishing the entanglement of the qubits using the notion of entanglement swapping which we discussed earlier.

The above equations are used for simulation and analysis of the different losses that can occur in our system. The simulation observations are presented in the following section.

VIII. Experimental Evaluation of SuperCrypt

We evaluate the performance of our proposed technique through performing discrete-event simulation. Since specialized simulators for quantum networking are not available yet, we performed our simulation with the network simulator `ns-2` after performing necessary modifications from the quantum perspective. In this section, we discuss the modifications and outcomes of our simulation.

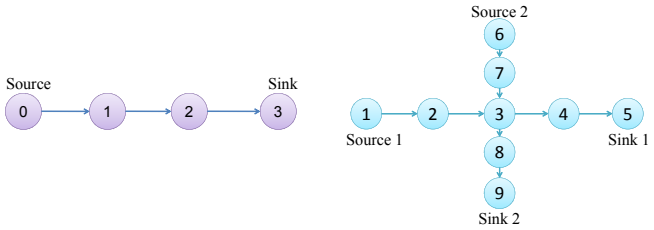
A. Simulator Modifications

To simulate our proposed quantum-based technique, we make some modifications in the classical network simulator `ns-2`. Since the basic difference between a quantum network and a classical network lies in the Physical Layer.

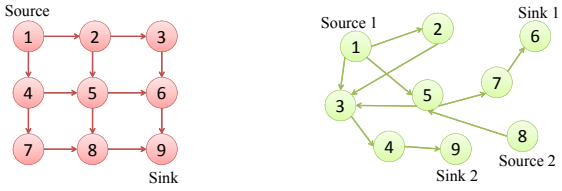
In our implementation, we consider the losses due to diffraction of the photon and the losses due to line attenuation, which we have already discussed in Section VII-C. Besides, we disregard any impact of interference or collision that is highly impactful in conventional networks such as wireless networks. When a packet arrives in its destination in our implementation, its success of reception depends on the combined loss due to diffraction and attenuation. If the loss is above a predetermined threshold, we discard the packet. Otherwise, we accept it and pass it to the next layer of the protocol hierarchy in `ns-2`.

B. Simulation Settings

We perform the simulation in two steps. For the first set of simulations, we consider classical networks. Here, we consider all the links to be simplex links. We set capacity of queue to



(a) Linear topology (b) Cross topology



(c) Mesh topology (d) Random topology

Fig. 5: Different network topologies used in our simulation

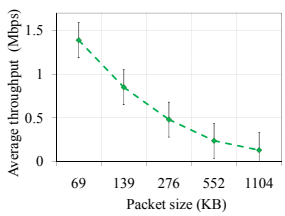
be 20 packets. We set the window size to 8000 packets. We also assume that the data rate of transmission interface card to be 0.3 Mb. The execution time is 124 seconds. With these settings, we perform simulation with varied channel delay and packet size.

For the second set of simulations, we consider the channel errors from the perspective of quantum networks. Here, we assume the packet size to be 256 kB. Besides, we adopt the channel delay to be 0.0125 ms considering very small delay in case of quantum networks. The other settings of `ns-2` remain same as the earlier one. In addition to that, for measuring the channel losses according the equations of Section VII-C, we assume that diameter of aperture of the sending device (D_A) is 3 cm, and wavelength is 1550 nm. We also assume that intrinsic loss constant A is 0.2 dB/km and bulk loss is 0.05 units.

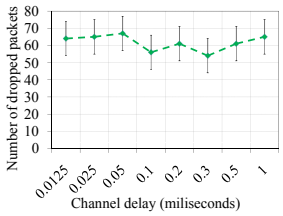
We simulate various network topologies for evaluating performance of SuperCrypt. Fig. 5 shows the various topologies. We first adopt a very simple network topology with a TCP source and a sink, (Fig. 5a). Here, node 0 is the TCP source and node 3 is the TCP sink. We adopt (DSDV) [2] as the routing protocol in our simulation. Besides, we use FTP over the TCP connection. However, for observing the network performance in response to different packet injection rate, we use CBR instead of FTP keeping all the other settings same as before.

Next, we adopt a few more complex network topologies for our simulation. Here, we consider a cross topology shown in Fig. 5b. In this topology, node 1 and node 6 are the source nodes and node 5 and node 9 are the sink nodes.

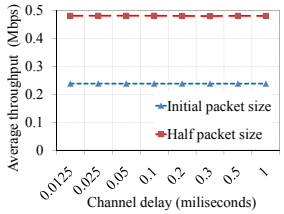
Besides, we adopt a mesh topology that is presented Fig. 5c. Here, node 1 is the source node and node 9 is the sink node. Additionally, we adopt a random topology presented in Fig. 5d. In this topology, node 1 and node 8 act as the source nodes and node 6 and node 9 act as the sink nodes. We keep other simulation settings same as the linear topology.



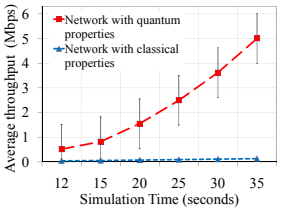
(a) Average throughput for different packet size



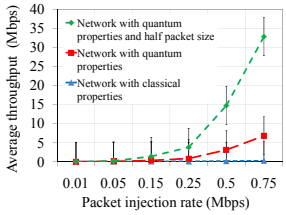
(b) Number of dropped packets for different channel delay



(c) Comparison of the average throughputs with different packet sizes



(d) Comparison of the average throughputs for quantum network and classical network



(e) Average throughput for different packet injection rate

Fig. 6: Simulation results for the linear topology

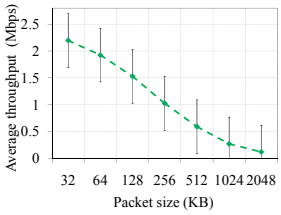
C. Simulation Results

While analyzing the simulation results, we first focus on the performance of the network for reduced packet sizes. Here, we attempt to investigate the impact of transferring $2n$ bits of information using only n bits. We plot the values of average throughput for different packet sizes in Fig. 6a. It can be easily seen from the graph that the network performance is much higher for shorter packet sizes.

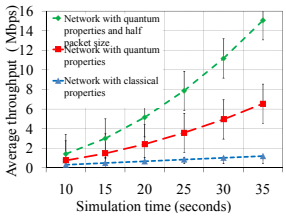
We also analyze the number of dropped packets in SuperCrypt. Fig. 6b demonstrates the impact on the number of dropped packets. Combining this graph with the earlier graph, we can conclude that delays of approximately 0.1 ms can result in higher network performance.

Next, we vary transmission delay for the channel and analyze the average throughput of the network. Since quantum entanglement transfers data with a very small delay, we focused on the shorter delays. Besides, we make our packet size half of that used in the previous simulation to see whether small-sized packets make any improvement in the performance or not. The comparison of the two simulation results is presented in Fig. 6c. This figure suggests that we get almost similar throughput irrespective of variation in the channel delay. Besides, the average throughput is significantly increased when the packet size is small.

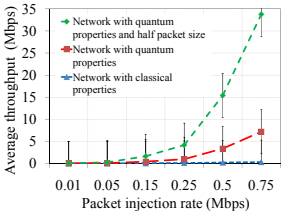
For the next simulation, we consider the effect of channel



(a) Average throughput for different packet size

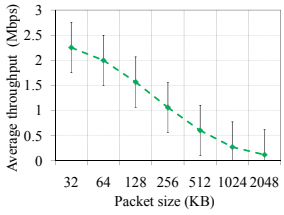


(b) Comparison of the average throughputs for quantum network and classical network

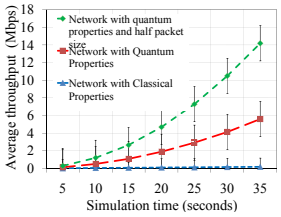


(c) Average throughput for different packet injection rate

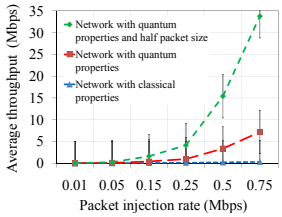
Fig. 7: Simulation results for the cross topology



(a) Average throughput for different packet size



(b) Comparison of the average throughputs for quantum network and classical network



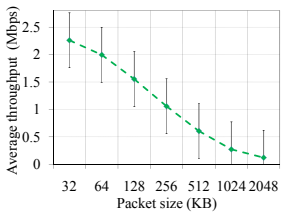
(c) Average throughput for different packet injection rate

Fig. 8: Simulation results for the mesh topology

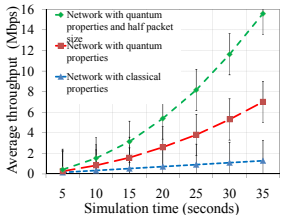
errors from the perspective of quantum networks and compare the outcomes with that of classical networks. We then plot average throughput for both quantum networks and classical networks for same simulation settings in Fig. 6d. This figure clearly shows that the performance of quantum network is much better than the classical one. We also observed the network performance for different packet generation rate for the network. The comparison of performance of classical network and quantum network is presented in Fig. 6e.

While the earlier simulation focuses on the performance of the simple linear network topology, our next set of simulation focuses on the performance of the complex topologies. First, we analyze the performance of a cross topology by measuring the average throughput for different packet sizes. Fig. 7a present outcomes of the analysis. Here, we find similar trend as we have already found for the linear topology. We also perform a comparative analysis of the design from quantum perspective and classical perspective. The comparative graph of Fig. 7b shows that the performance of quantum network is better than that of the classical one remaining in coherence of our earlier results. Next, we perform our simulation for observing the network performance with different packet generation rate. The comparison of performance of classical network and quantum network is presented in Fig. 7c exhibiting similar results as we have found for the linear topology.

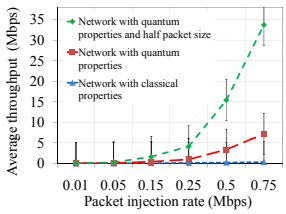
In our next simulation, we analyze the network performance for mesh topology. Similar to the pervious simulation results, we first observe the average throughput for different packet sizes. Fig. 8a presents simulation results in this regard. Besides, Fig. 8b and Fig. 8c shows the results of comparative study on quantum and classical perspectives. In both cases, we get similar results as we have found in the earlier cases. Nonetheless, the same set of analysis in the random topology provides similar results, which are presented in Fig. 9a, Fig. 9b, Fig. 9c.



(a) Average throughput for different packet size



(b) Comparison of the average throughputs for quantum network and classical network



(c) Average throughput for different packet injection rate

Fig. 9: Simulation results for the random topology

D. Simulation Findings

The simulation results presented in the previous section clearly demonstrate that the network performance gets significantly better when the channel delay and the packet sizes are reduced. Therefore, a channel with such properties can be used to design a higher performance network.

Moreover, when we replace the channel error properties of classical networks with that of quantum networks, the average throughput of the network gets significantly higher. Following this observation, we can state that the proposed quantum-based technique will be better in performance than the existing classical techniques.

Therefore, combining the computational complexity discussed in Section VI-C and the simulation results, we can summarize that the proposed SuperCrypt technique will simultaneously improve both data rate and security level.

IX. Future Work

The computational power of practical quantum computers is increasing day-by-day, which might cause threat even to our proposed technique in future. Therefore, we plan to incorporate a third level of encryption through further changing bases of the photons to make the system more secured. This incorporation will increase the level of security at the expense of introducing more complexity in the system.

Besides, in the study presented in this paper, we here mainly focused on encryption of the key while transmitting it through the quantum channel. An important aspect remains for SuperCrypt is that we need to maintain synchronization between transmission of classical information pertinent for the message and transmission of quantum information pertinent for the key, if we need to retain the classical method of message transmission. This may be needed to continue utilization of already-available classical channels. Now, to develop such a synchronization, we are planning to design a hybrid network that will use both classical and quantum channels.

Additionally, in recent times, tremendous progress has been made towards building real quantum devices. Therefore, we plan to perform real implementation of our proposed technique in near future.

X. Conclusion

Recent technological advancements in computing power leads towards breaking classical encryption techniques. Consequently, the state-of-the-art encryption techniques have started exhibiting their vulnerabilities. For example, RSA (the strongest encryption mechanism available to date) has become vulnerable after the emergence of extremely high computing power through quantum computing. Consequently, it becomes a necessity rather than an ambitious extension to come up with new encryption techniques that will offer more security sustaining system-level performance.

In this paper, we attempt to present a new technique of encryption, SuperCrypt to simultaneously improve both security level and data rate with the help of quantum computing and quantum networking. Here, we outline the theoretical aspects of the proposed system in accordance with portraying its implementation issues. We also evaluate the performance of the proposed technique through performing simulation in `ns-2`. We incorporate necessary modifications in the simulator that are needed to simulate our quantum based technique. Simulation results suggest that our proposed technique exhibits several advantages over the classical mechanisms. In future, we plan to implement our proposed technique in real systems.

XI. Acknowledgment

This work has been conducted at and partially supported by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. Besides, this work has been partially supported by the Ministry of Education, Government of the People's Republic of Bangladesh.

References

[1] A. V. Sergienko, *Quantum Communications and Cryptography*. Taylor and Francis, 2006.

[2] A. S. Tanenbaum, *Computer networks, 4th edition*. 2003.

[3] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *Advances in cryptology*, pp. 10–18, Springer, 1985.

[4] "University of bristol:quantum computing with recycled particles, sciencedaily. 23 october 2012." *www.sciencedaily.com/releases/2012/10/121023112515.htm* last accessed on 16 July, 2015.

[5] D. Naik, C. Peterson, A. White, A. Berglund, and P. Kwiat, "Entangled state quantum cryptography: Eavesdropping on the ekert protocol," *Physical Review Letters*, vol. 84, no. 20, p. 4733, 2000.

[6] C. H. Bennett and S. J. Wiesner, "Communication via one-and two-particle operators on einstein-podolsky-rosen states," *Physical review letters*, vol. 69, no. 20, p. 2881, 1992.

[7] G. Gilbert and M. Hamrick, "Practical quantum cryptography: A comprehensive analysis (part one)," *arXiv preprint quant-ph/0009027*, 2000.

[8] E. Martín-López, A. Laing, T. Lawson, R. Alvarez, X.-Q. Zhou, and J. L. O'Brien, "Experimental realization of shor's quantum factoring algorithm using qubit recycling," *Nature Photonics*, vol. 6, no. 11, pp. 773–776, 2012.

[9] A. Biryukov and D. Khovratovich, "Related-key cryptanalysis of the full aes-192 and aes-256," in *Advances in Cryptology–ASIACRYPT 2009*, pp. 1–18, Springer, 2009.

[10] H. Gilbert and T. Peyrin, "Super-sbox cryptanalysis: improved attacks for aes-like permutations," in *Fast Software Encryption*, pp. 365–383, Springer, 2010.

[11] A. Bogdanov, D. Khovratovich, and C. Rechberger, "Biclique cryptanalysis of the full aes," in *Advances in Cryptology–ASIACRYPT 2011*, pp. 344–371, Springer, 2011.

[12] V. Shoup, *Why chosen ciphertext security matters*. IBM TJ Watson Research Center, 1998.

[13] M. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. United States of America by Cambridge University Press, 10th anniversary ed., 2010.

[14] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, 2011.

[15] G. P. Berman, R. Mainieri, V. I. Tsifrinovich, and G. D. Doolen, *Introduction to quantum computers*. World Scientific, 1998.

[16] H.Bernien, B.Hensen, W.Pfaff, G.Koolstra, M.S.Blok, L.Robledo, and T.H.Taminiau, "Heralded entanglement between solid-state qubits separated by 3-meters," *NATURE*, 2012.

[17] A. Bokulich and G. Jaeger, "Philosophy of quantum information and entanglement," 2010.

[18] Megidish, E. A. Halevy, T. Shacham, T. Dvir, L. Dovrat, and H. S. Eisenberg, "Entanglement swapping between photons that have never coexisted," *Physical review letters*, 2013.

[19] N. Roch, M. E. Schwartz, F. Motzoi, C. Macklin, R. Vijay, A. W. Eddins, A. N. Korotkov, K. B. Whaley, M. Sarovar, and I. Siddiqi, "Observation of measurement-induced entanglement and quantum trajectories of remote superconducting qubits," *Physical review letters*, vol. 112, no. 17, p. 170501, 2014.

[20] D. Riste, M. Dukalski, C. Watson, G. de Lange, M. Tiggelman, Y. M. Blanter, K. Lehnert, R. Schouten, and L. DiCarlo, "Deterministic entanglement of superconducting qubits by parity measurement and feedback," *Nature*, vol. 502, no. 7471, pp. 350–354, 2013.

[21] D. E. Browne, M. B. Plenio, and S. F. Huelga, "Robust creation of entanglement between ions in spatially separate cavities," *Physical review letters*, vol. 91, no. 6, p. 067901, 2003.

[22] J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi, "Quantum state transfer and entanglement distribution among distant nodes in a quantum network," *Physical Review Letters*, vol. 78, no. 16, p. 3221, 1997.

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Full Papers

Security and Privacy in Communication

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Securing App Distribution Process of iOS Exploiting the Notion of Authentic Update

Sajeda Akter*, Farzana Rahman†, A. B. M. Alim Al Islam‡
*‡Department of CSE, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh
†Department of CS, James Madison University, VA, USA

Email: *sajeda24@yahoo.com, †rahma2fx@jmu.edu, ‡alim_razi@cse.buet.ac.bd

***Abstract*—iOS is, perhaps, considered as one of the most secured and reliable operating systems available now-a-days. However, its loopholes are coming into light in recent times causing a few security breaches such as the Masque attack. Even though prompt and pragmatic fixes for such breaches are of utmost significance, a fix for the Masque attack is yet to be proposed. To address this issue, in this paper, we propose a novel mechanism for guarding against the Masque attack. In our proposed mechanism, we exploit a synergy between authentication and non-repudiation to guard against different forms of Masque attack. Our exploitation leads towards a simple mechanism for preventing Replay attacks and for rejecting unauthenticated update, both of which generally paves the way of performing Masque attack. Consequently, the mechanism offers a pragmatic and easy-to-implement solution for the Masque attack.**

I. Introduction

Apple introduced a new mobile operating system in 2007, which is exclusively distributed for Apple hardware. Combining hardware, software, and advanced security features, Apple attempted to provide the best possible level of security. Till now, in mobile world, iOS is considered as one of the most secured operating systems [1].

iOS enforces secure boot chain, code signing, and run-time process security to ensure that only trusted code and apps can run on a device. Additional encryption and data protection protect personal and corporate information of the user.

Besides, mandatory code signing and app sandboxing of iOS prevent third-party apps from loading and executing unauthorized code. In addition, Apple also review the app to detect any suspicious behavior. Moreover, Digital Rights Management Technology limits the distribution of malicious apps by preventing users from sharing apps among arbitrary iOS devices [2]. All these measures taken by Apple provide customers a high level of confidence so that they feel fully secured.

From iOS 8, the system performs a code signature validation to protect itself and other apps from loading third-party code inside their address space [3]. After that, from late February 2014, iOS devices are facing different types of attacks. For example, after a few days of introducing WireLurker, FireEye[1] mobile security researchers discovered the Masque attack [4]. Exploiting this attack, a malicious app can replace a legitimate app without removing its local data, i.e, the malicious app can access cashed email and steal login credential through appearing as the original UI. This vulnerability exists both on jailbroken[2] and non-jailbroken devices [4]. As malwares do not go through the AppStore and generally use identical UI having same bundle identifier of the targeted app, built-in safeguards fail to detect the malwares. To address this issue, we propose a new mechanism that enables preventing this type of attack through extending some security measures.

In our proposed mechanism, we perform authentic update for all iOS apps. Here, we propose to include an update key in the property list for every app during its development. The update key should be signed using the developer's private key. When an app arrives with an identical bundle identifier of an existing app, i.e., an update of the app arrives, iOS would retrieve the update key using public key of the existing app. If the retrieved key is matched with the update key of the existing app, update will be allowed. The update will be rejected in case of any mismatch between the retrieved update key and the original update key.

Based on our work, we make the following set of contributions in this paper:

- We propose a new mechanism for authentic update in iOS app to prevent recent attacks such as the Masque attack.
- We present effectiveness of our approach in preventing the attacks. Here, we analyze in detail how different forms of attack can be prevented using our proposed mechanism. We also focus on the issue of Replay attack.
- Our proposed mechanism does not depend on the response from the user. It works even if the user response wrongly.

II. Background

iOS conventionally attempts to provide security from its core level. Secure boot chain, code signing, and run-time process security ensure that only Apple-signed code and app

[1]FireEye is a well known US network security company that provides automated threat forensics and dynamic malware protection against advanced cyber threats.

[2]Jailbroken devices refer to the devices that are free from the limitations and restrictions imposed on it by Apple.

978-1-5090-0203-0/16/$31.00 2016 IEEE

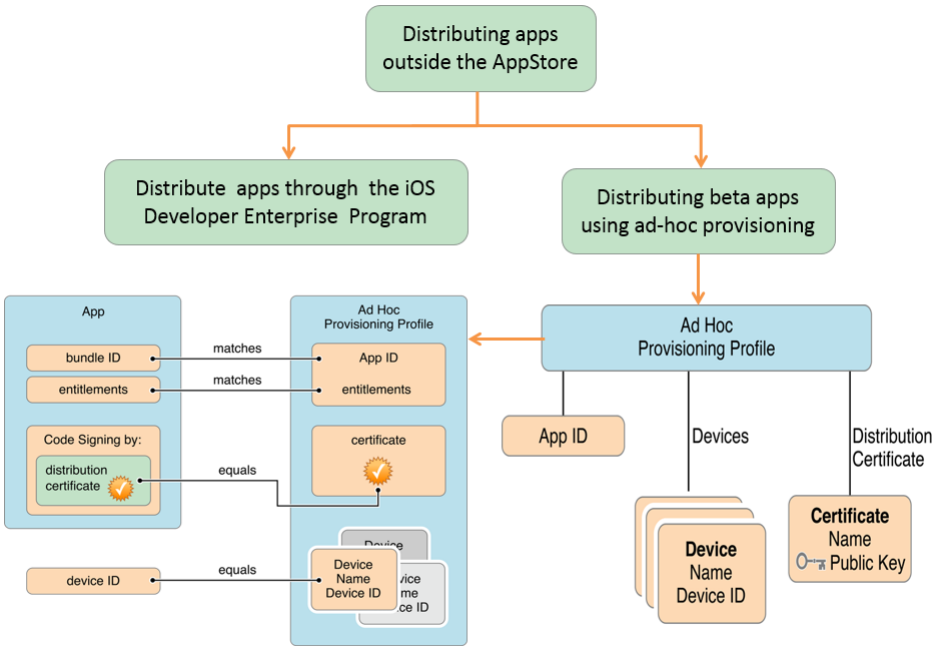


Fig. 1: Different approaches of app distribution outside the AppStore

can run on iOS devices. Additional encryption and data protection features perform as a safeguard to user data. To do so, Apple issues each certificate with a pair of keys (public and private) for every third-party app. Mandatory code signing, using the keys, prevents third-party app from loading malicious code. Here, iOS does not leave the third-party apps through only verifying certificates before giving access to the AppStore. Rather it also prevents dynamic attacks that attempt to inject malicious code at run-time. For this purpose, a single sandboxing profile is defined for every third-party app. Such profiling prohibits one app from accessing code and data of another app. Besides, several techniques such as trend analysis, dynamic lists, and content filtering are used by iCloud to automatically detect and block junk mail.

Inspite of adopting all these security measures, there still lie a few vulnerabilities with which intruders can attempt for different types of attacks. In recent time, attacks such as the Masque attack present an instance of such vulnerabilities.

III. Related Work

Once it was assumed that only jailbroken iOS devices are in threats, as they provide root privilege and permanently disable the code signing mechanism [5], [6]. On the other hand, non-jailbroken devices allow only Apple-signed app and all apps go through a strong vetting process. Inspite of providing such strong security, there still exist some vulnerabilities in iOS (both jailbroken and non-jailbroken) devices. Utilizing these vulnerabilities, adversary can create a major threat for end-users. Analyzing different threats and attacks, researchers have proposed different methods to enhance iOS security such that they can protect the end-users from the threats.

For example, an iOS device can be connected to a compromised device via USB or Wi-Fi for several reasons. Here, connected computer and communication channel both can become a new attack vector for the iOS device. The connected device can be instructed to install enterprise-signed or individual developer-signed malicious apps that are capable of stealing user's credentials from cookies or replacing the original apps [7]. As a remedy to this problem, Apple released a patch in iOS 7 to warn the users during making such connections for the first time. However, as users trust the compromised devices, this patch could not protect the iOS device [2].

On the other hand, several modern facilities or services, which are generally offered to make a smartphone much smarter, may be a medium of privacy leak and may cause several types of attacks [8]. According to the estimation of Mobile Thread Report 2011 [9], 33.9% of free iOS application had hidden capabilities to access user's location and 11.2% of them had capabilities to access personal contacts. Personal Hotspot (Tethering[3]) [8] and personal assistant over voice known as Siri [10], [8] are such types of services. Attackers can expose user's privacy through exploiting these services.

Through return-oriented programming[4], attackers can create a new control flow of a program at run-time by rearranging code gadgets[5]. Exploiting this approach, some apps (namely

[3]Tethering is a way of sharing internet connection of a smart device with other devices. Connection between two devices can be done through Bluetooth, Wi-Fi or USB cable.

[4]Return-oriented programming [11] is a computer security exploit technique. It allows attacker to execute malicious code even in the presence of security defense mechanisms.

[5]Code gadgets are small pieces of code, which possess the ability of performing a special task when they are chained together.

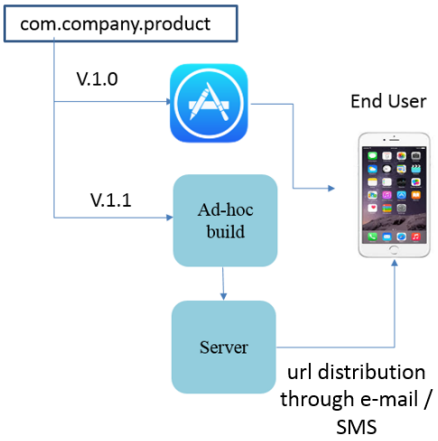


Fig. 2: Masque Attack

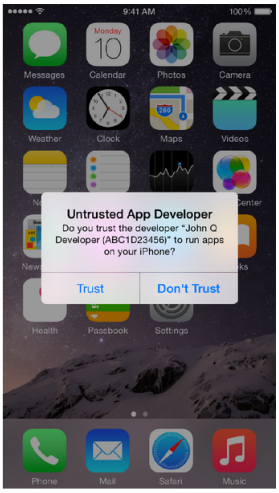


Fig. 3: iOS alert when opening a third-party app [4]

Jekyll[6] apps) can hide their malicious behavior and stay undetected during the review process and easily obtain Apple's approval. Once a Jekyll app has been approved and released into the App Store, it takes a devious turn. By rearranging signed code, Jekyll app can introduce malicious control flow that does not exist during the app review process. After installation, these apps share common characteristics having threats comprising of trojan and backdoor [5].

Now, as a developer submits a compiled app to Apple for review, Apple does not have the ability to view the actual source code. In hands-on review, which is one of the primary components of Apple's review process, Apple puts the app on a device to make sure that it meets the App Review Guidelines [12], [13] and does not violate any of the Apple's policies [12]. Another primary component of Apple's review process is static analysis, which looks for any indication of linking to private frameworks. If any link to private frameworks or a private API call exists, static analysis can detect it and the app gets rejected from the AppStore [13].

In Extended Application Sandboxing of iOS [14], a third-party app gets easy access into the private API inspite of existing strong vetting process. Using lazy binding[7] mechanism, the app could dynamically invoke private API at run-time. The solution proposed [14] to tackle this threat requires to implant an extended application sandboxing that contains a reference monitor to the original iOS application. The reference monitor mediates all access requests of an application to an external library. Here, a duplicate of the lazy and non-lazy symbol section, called "Shadow Table", is created. All calls to external functions are redirected to the reference monitor. The reference monitor decides on allowing or rejecting the invocation of the public API after performing necessary policy checks.

Additionally, there is another recent attack namely WireLurker [15], [16], which attacks both Mac OS and iOS. Here, while accessing iOS devices connected via USB with an infected computer, downloaded third-party malicious applications get installed onto the devices. The installation occurs in both jailbroken and non-jailbroken devices. Subsequently, the installed malicious apps can transfer data to and from the iOS devices over USB [15], [16].

One of the latest attacks for iOS is known as Masque attack [4]. As this attack is our main focus in this paper, we present a brief overview on it in the next section.

IV. Masque Attack

The Masque attack [4] exploits an in-house app, either distributed using ad-hoc provisioning or signed by developer's ID, to overwrite an existing original app installed from the AppStore. We present an analysis of this attack along with its vulnerabilities and existing counter measures.

A. Underlying Mechanism of the Masque Attack

Apple provides the opportunity of developing in-house apps to developers for their organizational use. Here, being outside the AppStore, app can be distributed following two different mechanisms (Fig. 1): (1) using ad-hoc provisioning, and (2) using iOS developer Enterprise Program.

In the case of using ad-hoc provisioning, an adversary needs to get the UDID of the target device. Here, the mechanism of getting UDID is nontrivial. Besides, this approach may limit the number of targeted devices. Consequently, attackers prefer the method of using enterprise-signed programs. On the other hand, enterprise-signed malicious apps are easier to distribute as they can be installed in any device.

For detailing the mechanism of Masque attack, let, a benign developer has developed an app with the bundle identifier "com.company.product" and has submitted it to the AppStore. An adversary targets the app, and gets the bundle identifier and version number of the target app. Then, he creates a malicious app with the same bundle identifier and the next version number. As he does not have the certificate, he does

[6]An app that possess two-sided characteristics, one side of which is good and the another is evil.

[7]In lazy binding, runtime address of a symbol is dynamically resolved at the frst time it is used

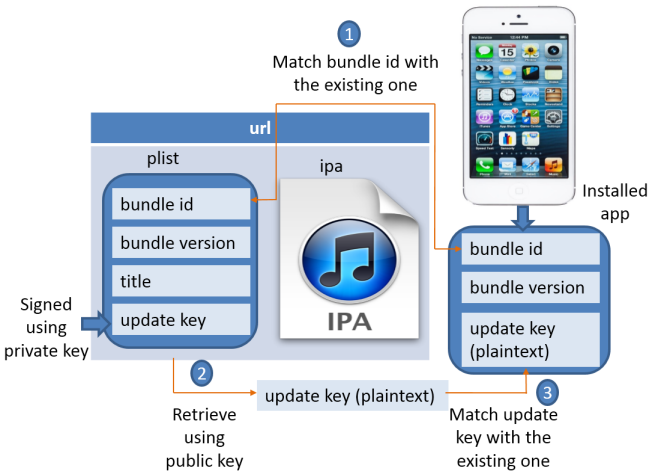


Fig. 4: Proposed mechanism for an authentic update

not go through the AppStore. He creates an ad-hoc build and upload it into their server. Then, he spreads out the url of the malicious app through e-mail or SMS with an attractive title. Although Apple has the ability to revoke enterprise at any time, no precaution is taken in case of installing this type of apps. Fig. 2 depicts the whole mechanism of Masque Attack.

B. Underlying Reasons of the Masque Attack

Before 2014, it was highly believed that Apple's vetting process is strong enough to protect users from any attack attempted by a malicious app. However, vulnerabilities of iOS have started coming to light one by one in recent times. These vulnerabilities may result in a great threat to the users. This happens due to several underlying reasons. Examples of such reasons are the following:

1) Enterprise-signed apps can be installed on any device without any restriction [17].
2) There is no MDM API[8] to get the certificate information for each app [4].
3) iOS does not enforce matching certificates for apps having the same bundle identifier [4].
4) No special security measure is taken for updating apps.

C. Existing Countermeasures for the Masque Attack

To date, no mitigation process has been adopted from the developer end to protect end-users from Masque attack. According to the corresponding authority, user's awareness is the only solution of this problem [17]. In this regard, end-users are encouraged to download apps only from known and trusted sources, and pay attention to warning messages appeared during downloading apps. When opening an app, if iOS shows an alert with "Untrusted App Developer" (as shown in Fig. 3), an end-user needs to click on "Don't Trust" and to remove the app immediately [18].

[8]Mobile Device Management architecture enables over-the-air provisioning, control and monitoring of remote iOS devices in real-time.

However, FireEye researchers later (in February 2015) has introduced Masque attack II that includes two different issues: (1) bypassing iOS prompt for trust and (2) iOS URL scheme hijacking. Here, the notion of bypassing iOS prompt for trust suggests that when a user clicks on a link, iOS either launches the enterprise-signed app without asking for "Trust" from the user or even ignores user's input of "Don't Trust" [19].

Besides, iOS allows apps from different developers to register the same URL scheme, that allows one app to communicate with other apps through a protocol. Currently there is no process for determining which app will be given more priority if more than one third-party app register to handle the same URL scheme [20]. Using this advantage, a malicious app can hijack ongoing communications among benign apps. Such hijacking can steal login credentials through performing mount phishing attack.

iOS 8.1.3 fixed the first issue (bypassing iOS prompt for trust), however, the second one (iOS URL scheme hijacking), is still unsolved [21]. Therefore, none of the current alternatives offers adequate protection against Masque attack.

V. Proposed Defense Mechanism

Mitigation of Masque attack is a challenging task as, in this attack a malicious app contains an identical bundle identifier of an existing authentic app and the bundle identifier is used in validating the app's signature. If we limit the number of devices allowed to install an enterprise-signed app to guard against Masque Attack, then the ultimate goal of the iOS Developer Enterprise Program will be hampered. On the other hand, rejecting an app for containing identical bundle identifier is impractical as it would also reject an authentic update. Moreover, as the malicious app needs not go through the AppStore, its certificate validation is difficult or near to impossible. Besides, as iOS is a closed-source, it does not allow any external monitoring.

As we can not prevent installing an app only for containing an identical bundle identifier, the only escape way remains is to ensure authentication when an update arrives. Consequently, we propose a novel mechanism comprising such authentic update of apps that would protect users from Masque attack.

A. System Model

The main goal of our solution is to ensure authentic update of an app that will effectively prohibit a malicious app from replacing an original app. Here, we devise our solution in such a way that the solution will not affect the conventional security system of iOS as well as will not limit the developer's opportunity to build in-house apps.

Fig. 4 depicts a high-level overview and flow of actions in our proposed authentication process. A *plist* manifest is associated with the *ipa* file. The *plist* contains bundle identifier and bundle version.

When an installation package arrives, a system gets information about the bundle identifier and version number of the corresponding app from the *plist*. Our proposed solution requires to have an additional update information in the *plist*

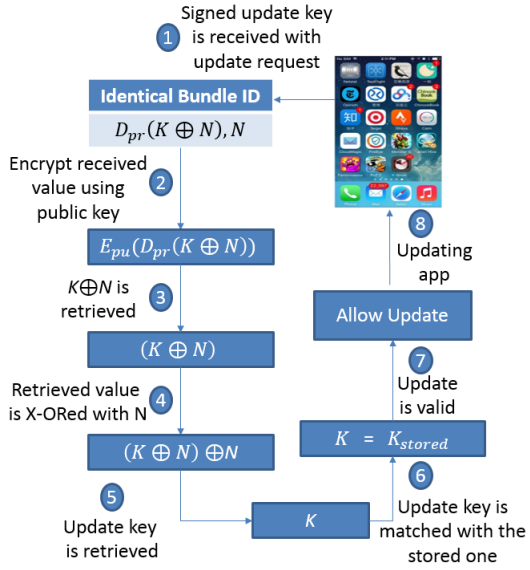


Fig. 5: Flow of actions in the proposed mechanism

similar to the previous version. Here, the update information is needed to be signed using the private key issued by Apple for the developer of the original app. System should already have the public key as the previous version of the app is already installed on the device. When an installation package arrives with the same bundle identifier of an installed app, the system fetches corresponding signed update key from the *plist*, retrieves it using the public key of the developer, and then matches it with the original one. If it gets matched, the system allows the update. Otherwise, the system rejects the update.

Fig. 5 depicts the flow of actions in our proposed mechanism in detail. Here, to protect Replay attack, we include a nonce[9] N with the update key K. When an authentic update request arrives, it should contain an update key K, being X-ORed with the N, which all together is signed using private key of the original developer. The system retrieves the X-ORed value using the corresponding public key. Now, the X-ORed value with N, i.e., K⊕N is again X-Ored with N to get the update key K. If the current key gets matched with the stored one, the system will allow the update action.

B. Design Considerations and Their Implications

We assume that the first version of an app is genuine and an attacker never attacks his own app. If the first version attempt for accessing the private API of another app, it would be an another kind of attack [14] that is different from the Masque attack. Solution of this kind of attack is already available [14].

Besides, in our proposed mechanism, we consider that the update key is signed using own private key. Here, no one except the original developer or developer team can have the private key. Consequently, our proposed mechanism successfully prohibits malicious apps from arriving as an update of another app and replacing the original app. Additionally, our proposed update mechanism incorporates the notion of non-repudiation. Consequently, in case of getting any update, originality of the source of the update can never be refuted later.

Note that, we do not adopt other authentication mechanisms available in the literature due to several reasons. For example, two-way authentication using challenge-response protocol [22] might be a candidate solution for our authentication, however, it exhibits own vulnerability under Reflation attack [23]. Besides, authentication using a shared key (for example, the mechanism of Diffie-Hellman [24]) could be another solution, however, it is prone to the Bucket Bridge [25] attack. Nonetheless, authentication using a key distribution center demands establishing such a center. Therefore, all such approaches lack in applicability in our case. Moreover, none of these approaches guarantee non-repudiation that we ensure in our mechanism. Next, we analyze effectiveness of the proposed mechanism.

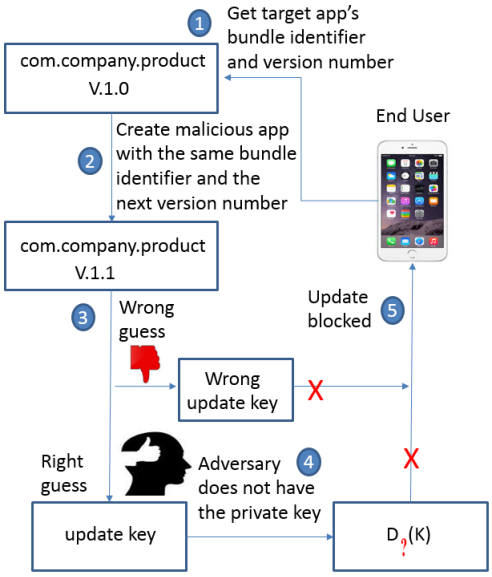
VI. Effectiveness of Our Mechanism

We analyze effectiveness of our proposed mechanism through focusing on a case where an adversary can get the bundle identifier and bundle version of a target app. After getting these information, the adversary can develop a malicious app with the same bundle identifier and the next version number. Nevertheless, the update key is not available to everyone. If the adversary could collect or guess the update key by any means, even then he should not be able to collect the private key issued by Apple for the original developer. Therefore, the adversary can never successfully finish the proposed authentication mechanism. This happens as, according to our proposed mechanism, iOS will not allow an update if it does not get an update key signed using the private key . Fig. 6(a) depicts such scenario of rejecting an unauthenticated app.

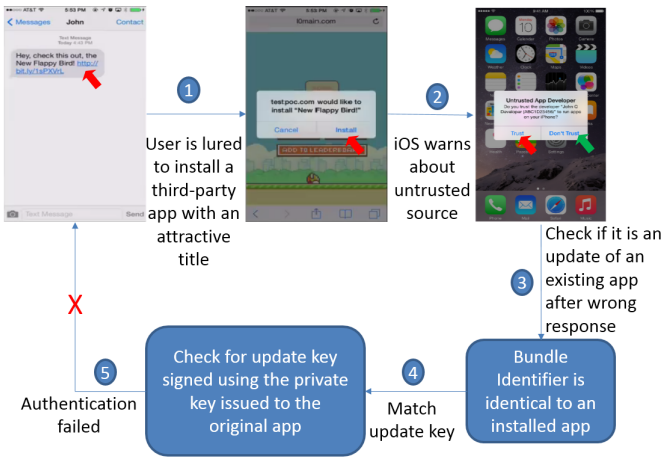
Note that, effectiveness of our proposed mechanism does not depend on the response from the user against the alert with “Untrusted App Developer”. Whatever the user chooses here (either “Trust” or “Don’t Trust”), in case of the newly arriving app exhibits an identical bundle identifier of an existing app, the system will check for an update key signed with private key of the original app. Fig. 6(b) depicts such a scenario. Therefore, the first part of Masque attack II i.e., Bypassing Prompt for Trust, also gets fixed in our proposed mechanism.

Besides, in our proposed mechanism, we also focus on the issue of Replay attack [26], as this is one of the most intuitive ways of bypassing authentication. Fig. 6(c) depicts the security environment offered in our proposed mechanism in case of Replay attack. Here, if only the update key K is signed with the private key, there lies a possibility of Replay attack. This happens as if an adversary can steal the signed update key through sniffing or by any other means, he can successfully perform an update of a targeted app impersonating the original

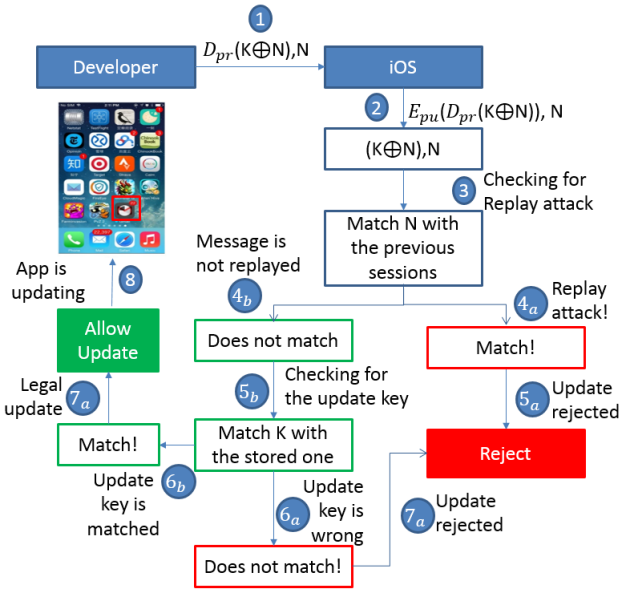
[9]A random value within a range, which must be unique for each communication. Therefore, by only checking if the current value has already been received earlier from this sender, replay attacks can easily be detected.



(a) Rejection due to not having the private key issued to the original developer ("?" denotes the unknown private key)



(b) Rejection even beyond wrong response from the user



(c) Preventing Replay attack in parallel to normal operation in the proposed mechanism

Fig. 6: Rejection of an unauthentic app in the proposed mechanism under different scenario

developer through replaying the signed update key. To protect such Replay attack, we adopt using the notion of nonce N in addition to the update key K. Here, K and N are X-ORed and signed using private key of the original developer. N is also sent as a plaintext so that the update key K can be retrieved letter. When an installation package arrives, N is matched with the nonces of previous sessions for ensuring that replay attack has not been occurred. Then, iOS retrieves K⊕N using the public key and X-ORs K⊕N with N to get the update key K. Now, the update key K is matched with the already-stored one to ensure authentication. If N has never been received from the developer of the targeted app, and update key is matched with the already stored one, update is allowed. The update is rejected in case of failing any of the condition. Fig. 6(c) presents a complete view of the authentication process in our proposed mechanism along with the prevention process of Replay attack in our proposed mechanism.

VII. Possible Limitations of Our Mechanism

Apple generally issues a signing identity for each app, which is used for code signing. The signing identity consists of a public-private key pair. The private key is stored only

in the keychain of the developer's machine and there is no way to retrieve the private key if it is lost. Though it is unlikely, if someone gets the private key of an app, he can distribute a malicious app impersonating the original developer. Therefore, the total responsibility of keeping the private key safe and secure resides on the developer of the app. Our proposed mechanism is unable to prevent this type of vulnerability, where the private key of the original developer gets compromised.

On the other hand, for URL Scheme Hijacking (second part of Masque attack II), it is necessary to install a malicious app on the target device. Now, in our proposed mechanism, our concern is to prevent illegal replacement of a benign app with a malicious app. Therefore, if a malicious app has already been installed in any other way (such as by own interest), our proposed mechanism will fail to prevent or fix it.

VIII. Conclusion

As certificate verification is a complex and lengthy procedure, iOS does not enforce certificate verification in case of updating an app. Rather the security is generally enforced by iOS when an app arrives through the AppStore. However, the security can get compromised while apps do not arrive from the AppStore, which might happen in case of enterprise-signed apps. Exploiting this vulnerability, adversary could attack a benign app. In this paper, we address this issue. Here, we propose a novel mechanism to prevent the attack (known as Masque attack) through including an update key, signed using the private key, for a specific bundle identifier. Our proposed mechanism is resilient in recognizing a malicious app even beyond user's wrong response. Besides, it can also prevent Replay attack. Therefore, we envision that, our proposed mechanism will be deployed in real systems in near future.

IX. Acknowledgment

This work has been conducted at and partially supported by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. Besides, this work has been partially supported by the Ministry of Education, Government of the People's Republic of Bangladesh.

References

[1] "iOS scores as most secure mobile OS in spyware report", http://www.cnet.com/news/ios-scores-as-most-secure-mobile-os-in-new-report/, Last accessed on May 24, 2015.

[2] Tielei Wang, Yeongjin Jang, Yizheng Chen, Simon Chung, Billy Lau, and Wenke Lee, "On the Feasibility of Large-Scale Infections of iOS Devices", 23rd USENIX Security Symposium, August 2014.

[3] "iOS Security Guide", https://www.apple.com/business/docs/iOS_Security_Guide.pdf, Last accessed on April 1, 2015.

[4] "Masque attack", https://www.fireeye.com/blog/threat-research/2014/11/masque-attack-all-your-ios-apps-belong-to-us.html, Last accessed on April 1, 2015.

[5] Tielei Wang, Kangjie Lu, Long Lu, Simon Chung, and Wenke Lee, "When Benign Apps Become Evil", 22nd USENIX Security Symposium, August 2013.

[6] Min Zheng, Hui Xue, Yulong Zhang, Tao Wei and John C.S. Lui, "Enpublic Apps: Security Threats Using iOS Enterprise and Developer Certificates", ASIA CCS15, 2015.

[7] B. Lau, Y. Jang, C. Song, T. Wang, P. H. Chung, and P. Royal. "Mactans: Injecting malware into ios devices via malicious chargers", In Black Hat USA, 2013.

[8] D. Damopoulos G. Kambourakis M. Anagnostopoulos S. Gritzalis J. H. Park, "User privacy and modern mobile services: are they on the same path?", Springer-Verlag London Limited 2012.

[9] Lookout Mobile Security, Mobile Threat Report (August 2011) https://www.lookout.com/_downloads/lookout-mobile-threat-report-2011.pdf. Last accessed 8 Sept 2015

[10] "About Siri", https://support.apple.com/en-us/HT204389, Last accessed on September 11, 2015.

[11] "Return-Oriented-Programming (ROP FTW)", https://www.exploit-db.com/docs/28479.pdf, Last accessed on May 4, 2015.

[12] "App Store Review Guidelines", https://developer.apple.com/app-store/review/guidelines/, Last accessed on May 24, 2015.

[13] Tao Wei, Min Zheng, Hui Xue &amp, Dawn SongFireEye, "APPLE WITHOUT A SHELL IOS UNDER TARGETED ATTACK", Virus Bulletin Conference September 2014

[14] Mihai Bucicoiu, Lucas Davi, Razvan Deaconescu and Ahmed-Reza Sadeghi, "XiOS: Extended Application Sandboxing on iOS", ASIA CCS'15, 2015.

[15] "WIRELURKER", https://www.paloaltonetworks.com/content/dam/paloaltonetworks-com/en_US/assets/pdf/reports/Unit_42/unit42-wirelurker.pdf, Last accessed on April 2, 2015.

[16] "iOS malware", http://researchcenter.paloaltonetworks.com/2014/11/wirelurker-new-era-os-x-ios-malware, Last accessed on April 1, 2015.

[17] "Masque attack", http://www.imore.com/masque-attack, Last accessed on April 1, 2015.

[18] "Guidelines for installing custom enterprise apps on iOS", https://support.apple.com/en-us/HT204460, Last accessed on May 19, 2015.

[19] "iOS Masque Attack Revived: Bypassing Prompt for Trust and App URL Scheme Hijacking", https://www.fireeye.com/blog/threat-research/2015/02/ios_masque_attackre.html, Last accessed on May 19, 2015.

[20] "Inter-App Communication", https://developer.apple.com/library/ios/documentation/iPhone/Conceptual/iPhoneOSProgrammingGuide/Inter-AppCommunication/Inter-AppCommunication.html, Last accessed on May 19, 2015.

[21] "About the security content of iOS 8.1.3", https://support.apple.com/en-us/HT204245, Last accessed on May 19, 2015.

[22] "Challengeresponse authentication", https://en.wikipedia.org/wiki/Challenge-response_authentication, Last accessed on 22 June, 2015.

[23] "Reflection attack", https://en.wikipedia.org/wiki/Reflection_attack, Last accessed on 22 June, 2015.

[24] " WHAT IS DIFFIE-HELLMAN?", http://www.emc.com/emc-plus/rsa-labs/standards-initiatives/what-is-diffie-hellman.htm, Last accessed on 22 June, 2015.

[25] "Bucket brigade", https://en.wikipedia.org/wiki/Bucket_brigade, Last accessed on 22 June, 2015.

[26] "Network Security", http://www.careerride.com/Networking-replay-attacks.aspx, Last accessed on May 24, 2015.

[27] "Alerts", https://www.us-cert.gov/ncas/alerts/TA14-317A, Last accessed on April 1, 2015.

[28] "Masque attack", http://www.imore.com/apple-comments-masque-attack, Last accessed on April 1, 2015.

[29] "iOS Security", https://www.apple.com/br/ipad/business/docs/iOS_Security_Oct12.pdf, Last accessed on April 1, 2015.

[30] "iOS Security", https://developer.apple.com/programs/ios/, Last accessed on April 1, 2015.

[31] "Sandboxing", https://developer.apple.com/app-sandboxing/, Last accessed on April 1, 2015.

[32] "iOS enterprise", https://developer.apple.com/programs/ios/enterprise/, Last accessed on April 1, 2015.

[33] "About App distribution", https://developer.apple.com/library/ios/documentation/IDEs/Conceptual/AppDistributionGuide/Introduction/Introduction.html, Last accessed on April 1, 2015.

[34] "Distributing Applications Outside the Mac App Store", https://developer.apple.com/library/mac/documentation/IDEs/Conceptual/AppDistributionGuide/DistributingApplicationsOutside/DistributingApplicationsOutside.html, Last accessed on April 1, 2015.

[35] “Beta Testing Your iOS App”, https://developer.apple.com/library/ios/documentation/IDEs/Conceptual/AppDistributionGuide/TestingYouriOSApp/TestingYouriOSApp.html, Last accessed on April 1, 2015.
[36] “Distributing iOS Developer Enterprise Program Applications”, https://developer.apple.com/library/ios/documentation/IDEs/Conceptual/AppDistributionGuide/DistributingEnterpriseProgramApps/DistributingEnterpriseProgramApps.html, Last accessed on April 1, 2015.

Optimal Allocation of 3G Budget for Smartphones Running Heterogeneous Applications

Saidur Rahman, Anika Anzum Prima and Md. Abdur Razzaque, *Senior Member, IEEE*
Green Networking Research (GNR) Group, Department of Computer Science and Engineering
Faculty of Engineering and Technology, University of Dhaka, Dhaka-1000, Bangladesh
Email: {sr.rifat, anikaprima13}@gmail.com, razzaque@du.ac.bd

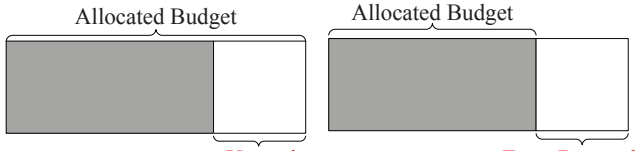
***Abstract*—Significant growth in number of smartphone users and applications running on them has been observed in recent decade. The applications require diverse amount of data bandwidths based on their interactivities. Data hungry applications demand huge plan whereas a background application is satisfied with a minimum amount. Thus, smartphone applications should be allocated their required budget in such a way that resource wastage is minimized and user experience is maximized. In this paper, we develop a prioritized and dynamic budget allocation policy for ensuring optimal amount of budget allocation to each application and improve system performance. In this regard, we formulate a linear programming optimization function that maximizes the utilization and minimizes resource wastage. We also develop runtime monitoring technique for estimating future bandwidth utilization. Experimental results confirm that system performance goes up using proposed algorithm and proves effectiveness of the algorithm.**

I. Introduction

The recent years have observed exponential growth in the usage of smartphones and applications running on them. The diversity of the applications is increasing day by day with the rapid development of smartphone technologies as well as various wireless access technologies. smartphones are powerful enough to run heterogeneous applications concurrently. The unique combination of features makes smartphones extremely usable and useful for different purposes. smartphone applications provide diverse kinds of services besides simple voice communication. smartphones are featured with music player, high megapixel cameras, better navigators, diverse sensors and so on. It is being accepted that in the future smartphones will take over all the other digital devices in next years such as laptops, desktop computers and notebooks. It is revealed that many people use 3/4/5G capable smartphones which allow the users to access the Internet from almost anywhere at anytime. The next generation telecommunication standards provide cost efficient, high quality, wireless multimedia applications and enhanced wireless communications. It offers greater security features and high data transmission rate at a low cost. Today's smartphone's indispensable part is Internet centric applications. These applications have heterogeneous sensitivities to delay-deadlines to environmental changes and different bandwidth utilization. [1], [2], [3], [4]

smartphone applications with internet accessibility are in the heart of user's digital life. The applications upload or download data via Wi-Fi or 3G network. When a user can access Wi-Fi communication, all the data packets in the applications buffer are uploaded to the destination server through Wi-Fi communication regardless of its priority. But, if Wi-Fi signal is absent or too low to upload the important data packets, then the system can autonomously switch to 3G communication. Every application needs to be allocated limited 3G budget according to their bandwidth utilization. A significant amount of data from some low priority applications may easily blow through expensive 3G data plan and cause exhaustive use of constrained bandwidth resources. These applications can do a large amount of downloading or updating, e.g., weather updates, application updates, social networking application updates, etc.. These low priority and background applications can quickly chew up large volume of bandwidth from fixed data plan [5]. Consequently, all important or more sensitive applications will be deprived of uploading data or performing their tasks properly. Another problem is that a significant amount of data may remain unused at the end of the data plan for lack of proper distribution of budget among the applications. Therefore, it is essential to set a proper budget plan for smartphone applications that can apply smart policies to reduce the wastage of bandwidth resources as well as increase the user experiences. Resource provisioning is another key consideration. If an application is allocated less budget than it requires then that results in *under-provisioning*. On the other hand, if an application is allocated so high amount of budget that a significant amount of budget remains unused at the end of data plan period, that results in *over-provisioning*, as shown in Fig. 1. Our aim is to ensure efficient and systematic 3G budget utilization for each application for avoiding penalties incurred due to both over- and under-provisioning.

Efficient and effective utilization of bandwidth is a challenging task. Most of the state-of-the-art works [6], [7] did not consider dynamic budget allocation for smartphone ap-



(a) Over-provisioning (b) Under-provisioning

Fig. 1. Budget allocation penalties

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

plications. Bandwidth requirement is satisfied for constant bit rate and variable bit rate connection and connection blocking probability is kept low as well [7]. In [6], a heuristic solution to the problem of allocating budgets to sensitive and non-sensitive applications has been developed. However, it did not handle over- and under-provisioning problems while allocating budgets. Thus, they failed to make difference among applications with much diverse budget requirements, causing degraded performance in achieving better bandwidth utilization as well as many applications may be deprived of required bandwidth due to poor allocation policies.

In this work, we explore dynamic budget allocation policy which analyzes the budget usage behavior of each application and decides how to allocate resources for each application such as maximization of bandwidth will be ensured. Our proposed scheme of resource allocation ensures judicious amount of budget allocation for each application so that more important applications are not hampered performing their tasks. It also increases the overall bandwidth utilization for the applications.

The key contributions of this work can be summarized as follows:

- In this paper, we develop a prioritized and dynamic 3G budget allocation technique for smartphone applications.
- A bandwidth allocation function is formulated using linear programming optimization that maximizes budget utilization of all applications while minimizing allocation error as much as possible.
- A runtime monitoring and measurement scheme for estimating budget utilization has been developed using Weighted Average Usage Prediction (WAUP) method.
- We also recommend the amount of budget for future data plan using ARIMA (Autoregressive Integrated Moving Average) model. We carry out numerical evaluations to study the effectiveness of the proposed optimal budget allocation policy.
- The results show that the proposed model provides better performances than a number of state-of-the-art models.

The rest of the paper is organized as follows. Section II describes some of the works related to our topics of interest. The Section III presents system model. In section IV, we formulate optimization problem and propose dynamic budget allocation scheme. The Section V presents the result of performance evaluation and conclusions are drawn in Section VI.

II. Related Works

Allocating bandwidth among the competing users or devices is a challenging problem and it has been studied in the literature for many networks. In [7], the authors presented a utility based bandwidth allocation algorithm for multiple services in the heterogeneous wireless access networks consisting of WMAN, 3G cellular network and WLAN. Bandwidth is allocated to a new arrival connection in heterogeneous wireless environment depending on utility fairness. The researchers of [8] have proposed a smart bandwidth allocation algorithm based on smartphone users' personality traits and channel condition. Based on one user's data usage, the service provider could estimate this user's probability of each personality trait using diagnostic inference, and then based on predictive inference to calculate this user's usage of bandwidth in the future. The researchers of [9] have dealt with bandwidth disposition problem for heterogeneous networks. Their proposed method determines the amount of disposed bandwidth and upgraded or downgraded sequence of bandwidth is quantified by using Upgrade Rank or Downgrade Rank function.

The authors of [6] have introduced online 3G budget algorithm that decides which sensory data should be uploaded via 3G communication while others will be uploaded or downloaded later when Wi-Fi access point is encountered. Their optimization scheme ensures efficient 3G budget utilization but the algorithm causes large amount of computational overheads. Therefore, the approach is both computational resource and energy hungry. Also they have proposed a heuristic algorithm and the main focus of their proposed algorithm is to split overall 3G budget in each time cycle into two pieces: reserved budget and flexible budget. Sensitive applications use reserved budget and non-sensitive applications use flexible budget. If reserved budget runs out then sensitive applications take help from flexible budget. But this two-state classification (sensitive and non-sensitive) of the applications decreases the dynamicity and flexibility of bandwidth allocation. In addition to that the budget allocation strategies for heterogenous applications following their urgency have not been explicitly discussed and analyzed.

III. System Model

In this section, we present the system model for 3G budget utilization. We consider that a smartphone is connected with the Internet either using Wi-Fi access point or by 3/4/5G mobile Internet connection. The smartphone uses 3G bandwidth budget for urgent application usage whenever no Wi-Fi access point is available at nearby. The smartphone applications are allowed to buffer the data packets at local device till it is connected with any AP. In the case, the buffer space of the mobile phone is exhausted, it stops data collection process. When a user is in the range of a Wi-Fi access point, all the backlogged data packets in the buffer are uploaded to the destination server through Wi-Fi communication regardless of its priority. However, if Wi-Fi signal is absent or too low to upload the important data packets, then the system switches to 3G communication.

We assume that a user has fixed budget for 3/4/5G Internet connection (e.g., 3GB monthly, 1GB weekly package). The amount of the budget data plan for each of the applications is proportional to how much important the application is. That is, a real-time and interactive application needs more bandwidth and it may not tolerate significant delay; on the other hand, some low priority applications may be delayed and reduced amount of data budget can be allocated. In this work, we dynamically prioritize all the applications running in the mobile device by estimating bandwidth usage behavior of the applications. The more bandwidth an application uses,

the higher it's priority is. We exploit autogregressive integrated moving average (ARIMA) formulae for estimating the runtime usage of resources by different applications and recommend a user the most appropriate amount of monthly data plan (to be discussed in detail in section IV-D).

We also assume that the budget allocation algorithm periodically runs every after t time. It tries to avoid overprovisioning as well as underprovisioning so as to maximize the resource utilization and application performance. Each application falls in one of the n application types with different priorities $p_1, p_2, p_3, \ldots, p_n$. In this case, we use higher values for higher priority applications.

IV. Proposed Model

In this section, we present the proposed dynamic budget allocation strategy for heterogeneous applications running in a smartphone. The proposed budget allocation policy dynamically expands or shrinks the amount of bandwidth allocated to different applications over time based on the usage behavior of the data plan. Our budget allocation optimizes the bandwidth resource utilization as well as reduces the penalties incurred due to over- and under-provisioning. We exploit Weighted Average Usage Prediction (WAUP) method to more accurately infer the bandwidth usage in future time intervals. We use Autogressive Integrated Moving Average (ARIMA) model for recommending appropriate amount of monthly data plan for a user.

A. Optimization Problem Formulation

The problem of optimal allocation of bandwidth to the mobile applications is translated as maximizing the utilization of resources while minimizing the penalties incurred due to over- and under-provisioning. And, this policy needs to be maintained for all applications in all allocation intervals. Therefore, the optimization function is a linear programming (LP) problem, defined as follows:

Maximize:

$$Z = \sum_{i=1}^{n} \sum_{t=1}^{T} (U_{i,t} - C_{i,t}) \tag{1}$$

subject to:

$$x_{i,t} \leq r_{i,t}, \tag{2}$$

$$y_{i,t} \leq f_{i,t}, \tag{3}$$

$$x_{i,t} + y_{i,t} \leq r_{i,t} + f_{i,t} \tag{4}$$

Here, $U_{i,t}$ is the resource utilization of application i at time interval t and $C_{i,t}$ is the corresponding over-provisioning penalty, if there is any. Given that the $r_{i,t}, x_{i,t}$ and $y_{i,t}$ are the amount of reserved budget, used reserved budget and used flexible budget for application i at time interval t, the utilization and penalties are defined as follows:

$$U_{i,t} = \begin{cases} 0 & \text{if } \frac{x_{i,t}+y_{i,t}}{r_{i,t}} > 1 \\ 1 & \text{if } \frac{x_{i,t}+y_{i,t}}{r_{i,t}} \leq 1 \end{cases} \tag{5}$$

$$C_{i,t} = \begin{cases} 0 & \text{if } \frac{x_{i,t}+y_{i,t}}{r_{i,t}} > 1 \\ \frac{(r_{i,t}-x_{i,t})}{r_{i,t}} & \text{if } \frac{x_{i,t}+y_{i,t}}{r_{i,t}} \leq 1 \end{cases} \tag{6}$$

We observe from Eq. (5) that the system performance decreases with the increasing usage of flexible budget $y_{i,t}$ and the Eq. (6) states that penalty increases with the gap between the amount of reserved budget $r_{i,t}$ and used reserved budget $x_{i,t}$. In summary, the more approriate amount of budget that we can allocate which just meets the requirement of an application, the more the system performance is increased and vice-versa. The constraints (2) and (3) are corresponding to bandwidth usage constraints for reserved and flexible budgets, i.e., usage must be bounded by the proportionately allocated amount for an application i. The constraint (4) states that the constraints (2) and (3) follow additive rule.

B. Budget Allocation Policy

The overall 3G budget B is split into two parts in each time cycle: *reserved* budget (B_1) and *flexible* budget (B_2). Initially ($t = 0$), their values are determined as follows:

$$B_1 = \alpha \times B, \tag{7}$$

$$B_2 = B - B_1, \tag{8}$$

where, α is a control parameter that determines how much of the total budget is to be kept in reserved portion. When the data plan period starts every application is allocated a certain amount of reserved budget based on their priority assuming all the applications have equal bandwidth usage for k time cycles. If n applications are running then,

$$r_{i,t} = \frac{p_i \times \bar{b}_{i,t}}{\sum_{i=1}^{n}(p_i \times \bar{b}_{i,t})} \times B_1, \tag{9}$$

where, p_i, $r_{i,t}$ and $\bar{b}_{i,t}$ are the priority, reserved budget and the estimated bandwidth usage of i'th application, respectively, within the time cycle. The detail estimation process of $\bar{b}_{i,t}$ is presented in Section IV-C. We assume that in the first time cycle t_1, an application i has used $x_{i,t}$ amount of data from the reserved budget. So, remaining reserved budget is $B_1 - \sum_{i=1}^{n} x_{i,t}$.

In the case, an application is run out of it's reserved budget within the current time cycle, then flexible budget is allocated to it from B_2. If $f_{i,t}$ denotes the extra budget requirement for i'th application in t time cycle then,

$$f_{i,t} = \frac{p_i \times T^{'}}{T \times \sum_{i=1}^{n} p_i} \times B_2 \tag{10}$$

If each application uploads or downloads $y_{i,t}$ amount of data using flexible budget then remaining flexible budget is $B_2 - \sum_{i=1}^{n} y_{i,t}$. So remaining total budget after t_1 time cycle,

$$B = (B_1 - \sum_{i=1}^{n} x_{i,t}) + (B_2 - \sum_{i=1}^{n} y_{i,t}) \tag{11}$$

This is the budget for next time cycle that means the assignment is additive. We now calculate total budget by adding remaining flexible and reserved budgets for the current time cycle. From the second time cycle, the reserved budget is calculated according to following equation:

$$B_1 = \frac{T^{'} + \alpha \times (T - T^{'})}{T} \times B \tag{12}$$

where, T is the budget validation time and $T^{'}$ is present time. The Eq. 12 helps us to dynamically update the reserved budget amount B_1 following the historical usages. It also minimizes the wastage of bandwidth later at the end of the data plan period. The control parameter α plays an important role to start with minimum reserved amount from the first day of data plan and to increase gradually. Therefore, it minimizes both the over- and under-provisioning penalties. The value of α depends on execution frequency of the budget allocation algorithm compared to the total data plan period. For performance evalution, we have set $\alpha = \frac{t}{T}$, where, t is the time interval of executing allocation algorithm.

C. Estimation of Budget Usage

We calculate budget usage ratio $b_{i,t}$ after each usage interval, t, as follows,

$$b_{i,t} = \frac{x_{i,t} + y_{i,t}}{r_{i,t} + f_{i,t}} \tag{13}$$

Thus, the Eq. 13 refers to how much of the allocated budget is used by an application i. We need to predict the allocated bandwidth budget usage of each application so as to infer the judicious amount of budget to be allocated in the upcoming time cycle. The possible amount of usage of the budget by an application in the next time cycle typically depends on its historical usage patterns. And the most recent usage behavior puts more impact on the future usage estimations. In this work, we exploit Weighted Average Usage Prediction (WAUP) method similar to WALI model [10], [11] that works as follows. The WAUP measures the average bandwidth usage of i'th application in the current time cycle as a weighted average of last m time cycles as follows:

$$\bar{b}_{i,t} = \frac{\sum_{j=1}^{m}(w_j \times b_{i,j})}{\sum_{j=1}^{m} w_j} \tag{14}$$

For weights w_j:

$$w_j = \begin{cases} 1 & \text{if } 1 \leq j \leq \frac{m}{2} \\ 1 - \frac{j - \frac{m}{2}}{\frac{m}{2}+1} & \text{if } \frac{m}{2} < j \leq m \end{cases} \tag{15}$$

For $m = 8$, this gives weights of 1, 1, 1, 1, 0.8, 0.6, 0.4, 0.2 for w_1 through w_8, respectively where the most recent four samples are equally weighted.

D. Budget Recommendation using ARIMA

ARIMA(Autoregrassive Integrated Moving Average) is a common and effective method as one kind of time series prediction method. ARIMA($p,\ d,\ q$) models are first introduced by Box and Jenkins in 1970 [12] for purposes of modeling time series data. The model is the combination of autoregression and a moving average models. The full form of ARIMA can be written as [13], [14]

$$B_l^{'} = c + \phi_1 B_{l-1}^{'} + \cdots + \phi_p B_{l-p}^{'} + \theta_1 e_{l-1} + \cdots + \theta_q e_{l-q} + e_l \tag{16}$$

ARIMA($p,\ d,\ q$) can also be written as

$$B_l^{'} = c + \sum_{i=1}^{p} \phi_i B_{l-i}^{'} + \sum_{i=1}^{q} \theta_i e_{l-i} \tag{17}$$

or by using lag polynomial operator:

$$\bigtriangledown^d B_l^{'} \phi(L) = \theta(L) e_l \tag{18}$$

where,
$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \cdots - \phi_p L^p$
$\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \theta_3 L^3 - \cdots - \theta_p L^q$

$B_l^{'}$ is correlated normally distributed random variable, e_l is an uncorrelated Gaussian noise, θ_l is moving average coefficient and L is the lag operator.

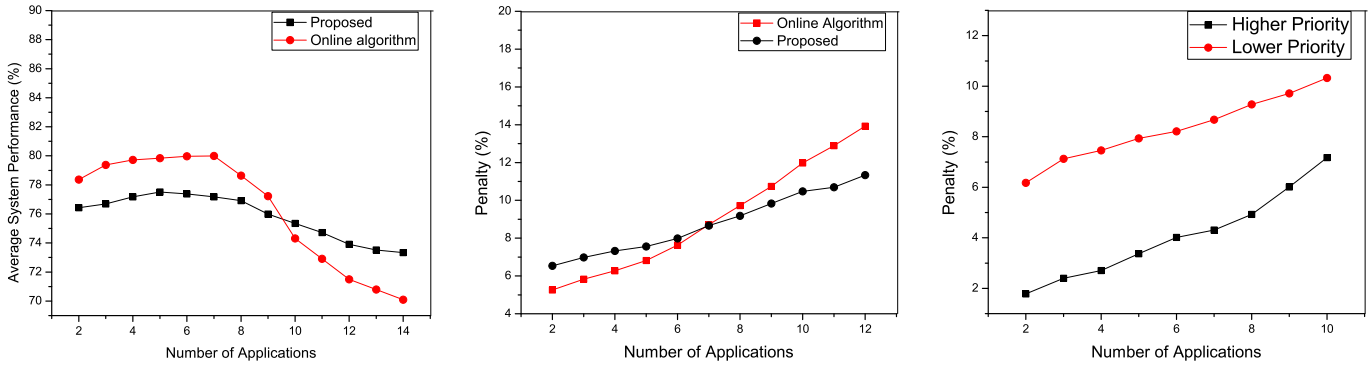
For predicting data plan period, ARIMA model is the better option for forcasting the data. Here, $B_l^{'}$ is the estimated data budget for lth data plan period. If the data is not stationary, then $B_l^{'} = B_l - B_{l-1}$.

V. Numerical Evaluation

In this section, we study the effectiveness of the proposed dynamic budget allocation policy compared to an online 3G budget algorithm [6] through numerical evaluations. We have implemented both the budget allocation algorithms using C++ programming language. We assume many applications are running in a mobile device and they have diverse bandwidth requirements. The number of applications that are active on a mobile device and it's monthly data plan are randomly chosen from a wide range of $2 \sim 12$ and $3 \sim 10$GB, respectively, with uniform distribution. The arrival and departures of applications are exponentially distributed. As a result, the duration for which an application keeps it active varies greatly from others. We also emulate that the mobile user does not use 3G data plan continuously for Internet accessibilities, rather sometimes it uses Wi-Fi access points for data transfer. The total data plan period is assumed to be 720 hours (i.e., 30 days) and the budget allocation algorithm execution time interval t is chosen 4-6 hours. For each of the graph data points, we run the program 20 times for different random inputs and take the average of the results.

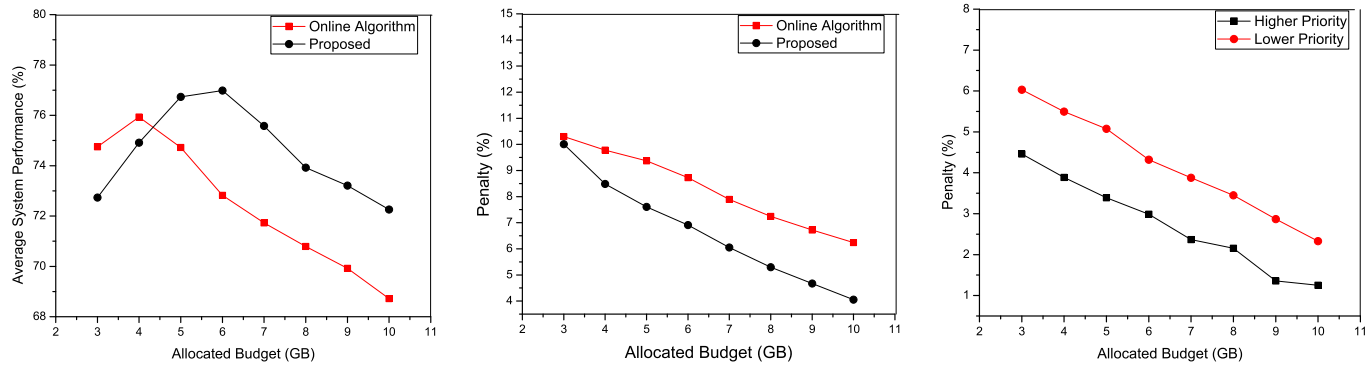
We have studied the following two metrics: *average system performance* and *penalty* for varying number of applications running on the smartphone. Eq. (1) defines system performance denoting the difference between resource utilization and over-provisioning penalty of each application in all time cycles. Our aim is to upgrade system performance by maximizing utilization and minimizing penalty. The *average system performance* is measured using Eq. (1) for all applications and then the average is taken for graph data points. The penalty is measured as the percentage of applications that could not be run due to shortage of bandwidth during the experiments. The average is taken for all time periods and all applications.

As shown in Fig. 2(a), the average system performance linearly increases with the number of applications in both the studied budget allocation algorithms. However, the performance of online 3G budget allocation algorithm starts



(a) System performance (b) Execution penalty (c) Execution penalty

Fig. 2. Performance studies for increasing number of mobile applications.



(a) System performance (b) Execution penalty (c) Execution penalty

Fig. 3. Performance studies for increasing amount of 3G bandwidth budgets.

decreasing when the number of applications is 8 and above. On the other hand, the proposed optimal budget allocation policy offers as high as almost 80% performance for higher number of applications. This happens because of its higher capability of accommodating diverse applications with different priorities and dynamically adjusting the bandwidth allocation to the applications following their historical usage pattern.

The graphs of Fig. 2(b) depict that the percentage of penalty increases with increasing number of applications but the penalty offered by proposed algorithm remains relatively low comparing to online 3G budget algorithm.

The graphs of Fig. 2(c) depict that the percentage of applications that are deprived of required amount of bandwidth allocation increases exponentially for lower priority applications. In this case, the results are caused by excessive underprovisioning penalty. However, for the higher priority applications, the percentage is significantly low, which is expected theoretically as well.

Fig. 3(a) shows that the average system performance offered by proposed algorithm grows with larger amount of budget but starts decreasing due to over-provisioning penalty when the allocated budget is 7 and above. However, the average performance offered by online algorithm remains low because of their incapability of proper distribution of budget among applications and handling over- and under-provisioning. Fig. 3(b) depicts that underprovisioning penalty decreases with larger amount of allocated budget as the applications are allocated judicious amount of budget. Fig. 3(c) depicts that underprovisioning penalty decreases with larger amount of allocated budget for both higher and lower priority applications.

VI. Conclusion

In this paper, we proposed dynamic bandwidth allocation algorithm for heterogeneous smartphone applications. The proposed data plan usage policy maximizes the utilization of all applications while minimizing over- and under-provisioning. We exploit the application's behavior and recommend future data plan after long term analysis. Experimental results confirmed that our proposed scheme gives optimal solution and potentially brings benefits to users. Our model also gives the better performance to distribute the budget among the applications where penalty will be minimized.

ACKNOWLEDGEMENTS

This work is supported by a grant for the "Innovative Project (2013-2014)" - "ICT Assisted Safe Driving for Mitigating Road Accidents in Bangladesh", funded by the Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh. Dr. Md. Abdur Razzaque is the corresponding author of this paper.

REFERENCES

[1] Aaron Smith. U.s. smartphone use in 2015. In *The Smartphone Difference*. Pew Research Center, April 2015.

[2] Trent D. Buskirk and Charles Andrus. Smart surveys for smart phones: Exploring various approaches for conducting online mobile surveys via smartphones. In *Journal*. Survey Practice, 2012.

[3] Ahmad Rahmati and Lin Zhong. Studying smartphone usage: Lessons from a four-month field study. *IEEE Transactions on Mobile Computing*, 2012.

[4] Oliver Amft and Paul Lukowicz. From backpacks to smartphones: Past, present, and future of wearable computers. *IEEE Pervasive Computing*, 8(3):8–13, 2009.

[5] Statistics and market data on mobile internet & apps. http://www.statista.com/markets/424/topic/538/mobile-internet-apps/, (accessed July 16, 2015).

[6] Hengchang Liu, Shaohan Hu, Wei Zheng, Zhiheng Xie, Shiguang Wang, Pan Hui, and Tarek F. Abdelzaher. Efficient 3g budget utilization in mobile participatory sensing applications. In *INFOCOM*, pages 1411–1419. IEEE, 2013.

[7] Changqing Luo, Hong Ji, and Yi Li. Utility-based multi-service bandwidth allocation in the 4g heterogeneous wireless access networks. In *2009 IEEE Wireless Communications and Networking Conference, WCNC 2009, Proceedings, Budapest, Hungary, 5-8 April 2009*, pages 1915–1919, 2009.

[8] Junjie Chen, Qilian Liang, and Jie Wang. Bandwidth allocation based on personality traits on smartphone usage and channel condition. In *The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems Part III*, pages 273–282, 2015.

[9] Hui-Min Huang and Ying-Hong Wang. Bandwidth management method for heterogeneous wireless network. *WTOC*, 7(4):267–276, April 2008.

[10] Sally Floyd, Mark Handley, Jitendra Padhye, and Jörg Widmer. Equation-based congestion control for unicast applications. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '00, pages 43–56, New York, NY, USA, 2000. ACM.

[11] Md. Abdur RAZZAQUE, Choong Seon HONG, and Sungwon LEE. Autonomous traffic engineering for boosting application fidelity in wireless sensor networks. In *IEICE Trans. Commun, Vol. E93-B*, pages 2990–3003. IEICE, November 2010.

[12] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco, CA, 1970.

[13] Jie Wu Xin Jin, Yao Dong and Jujie Wang. An improved combined forecasting method for electric power load based on autoregressive integrated moving average model. In *2010 International Conference of Information Science and Management Engineering*, pages 476–480. IEEE, 2010.

[14] Mohamed Maaroufi Noreddine Citroen, Mohammed Ouassaid. Long term electricity demand forecasting using autoregressive integrated moving average model: Case study of morocco. In *1st International Conference on Electrical and Information Technologies (ICEIT)*, pages 59–64. IEEE, 2015.

A New cost-effective approach for Battlefield Surveillance in Wireless Sensor Networks

Ensuring maximum destruction and efficient monitoring

Fariha Tasmin Jaigirdar
Dept. of Computer Science and Engineering, Daffodil International University
Dhaka, Bangladesh
fariha.cse@diu.edu.bd

Mohammad Mahfuzul Islam
Dept. of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh
mahfuz.cse.buet@gmail.com

***Abstract*—Assuring security (in the form of attacking mode as well as in safeguard mode) and at the same time keeping strong eye on the opposition's status (position, quantity, availability) is the key responsibility of a commander in the battlefield. Battlefield surveillance is one of the strong applications of Wireless Sensor Networks (WSNs). A commander is not only liable to his above responsibilities, but also to manage his duties in an efficient way. For this reason, ensuring maximum destruction with minimum resources is a major concern of a commander in the battlefield. This paper focuses on the maximum destruction problem in military affairs. In [1] the authors proposed two novel algorithms (Maximum degree analysis and Maximum clique analysis) that ensure the efficiency and cost-effectiveness of the above problem. A comparative study explaining the number of resources required for commencing required level of destruction made to the opponents has been provided in the paper. In this paper the authors have come forward with another algorithm for the same problem. With the simulation studies and comparative analysis of the same example set the authors in this paper demonstrate the effectiveness (in both the quality and quantity) of the new method to be best among the three.**

Keywords—wireless sensor network; military application; intersection point algorithm; maximum destruction; minimum resource; battlefield monitoring

I. INTRODUCTION

Recent advancement of technology everyday reveals new door towards the next step of research in different area. Wireless Sensor Network (WSN) though not a new area in the era of research lovers, but of course adding new dimensions day by day by it is different branches of applications. WSN consists of numerous physically distributed autonomous devices, known as *sensors* used for sensing and monitoring the physical and/or environmental conditions. A WSN uses a gateway that provides wireless connectivity both to the wired world and distributed networks [2]. WSN proves its excellency in different areas of applications including ocean and wildlife monitoring, industrial process monitoring, home automation, traffic control, healthcare applications, building safety and earthquake monitoring, and battlefield surveillance [1][2][3].

Different scenarios of military applications include battlefield (well-defined enemy), monitoring oppositions area, operations in urban environments, other than war, i.e., peacekeeping, disaster relief etc. In every scenario a commander has to perform his duties efficiently with minimum budget. Thus, a commander's responsibility can be summed up with the following points

- Maximum destruction of the opponents with minimum resources (in battlefield),
- Placing minimum watch tower to monitor the maximum area of his opponent (in battlefield),
- Maximum usage of environment using minimum soldiers (urban environment),
- Minimum usage of resources (food truck for example) for meeting the need for maximum number of people (disaster relief).

Sensors can be deployed in friendly as well as in unfriendly environments. For meeting military applications of WSNs, the later environment is preferred as sensors can be deployed from aircraft or under the sea. In [1] the authors placed the problem as deploying minimum number of resources (i.e., tank, bomb, watch tower etc.) for destroying maximum amount (in terms of opponent's soldiers and/or opponent's resources) of opponent's area/resources. Two cost-effective and efficient algorithms were proposed in [1] to solve the problem.

With the advancement of technology, intruders are planning new ideas to harm the network setup of a battlefield and make the network quality down for having control over the network. Efficient and cost-effective deployment of active sensor nodes as well as finding the optimum location is one of the key problems in battlefield [1]. A cost-effective deployment of the sensor nodes (tank, mine, bomb etc.) can guarantee the minimum or nearly minimum number of these resources needed to destruct the opponent's area. This paper

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

demonstrate a new deployment strategy named *intersection point algorithm* that guarantees the desired result (minimum number of resources and their locations) to be best among the three.

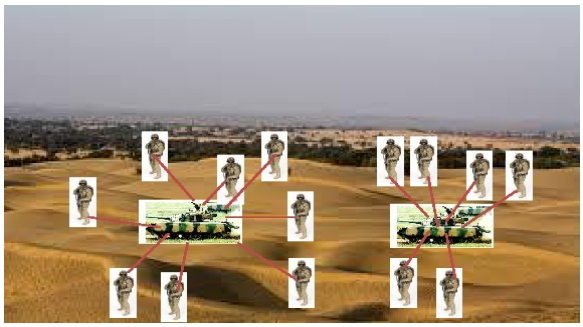


Fig. 1. Destroying fourteen soldiers with only two tanks

Using this technique a commander can achieve his goal of maximum destruction with minimum resources/placing minimum number of watch tower to cover maximum area of his opponents in the battlefield. Fig. 1 portrays the scenario where by deploying only two tanks maximum destruction is possible as the deployment is in the most dense region.

Finding the minimum location of military resources is an NP-hard problem [1]. Therefore authors in this paper are focusing on finding out the most prominent solution of the problem. They have emphasized the problem by exploiting and merging the techniques of finding the intersection point [4], minimum set cover problem [5] and the greedy approach [6].

In [1] F. T. Jaigirdar and M. M. Islam proposed two strategies named maximum clique analysis and maximum degree analysis to solve the problem. The approximation technique that the authors are going to propose in this paper is intersection point algorithm. This is easier to implement then those of two. A comparative analysis is also presented here to show the effectiveness of the new approach.

The rest of the paper is organized as follows: Section II discusses related works on the title; Section III discusses the methodology with detail description of every portion, whereas the simulation and comparative analysis has been given in Section IV. Some concluding remarks are given in Section V.

II. Literature Review

The idea discussed in this paper can be slightly related to coverage problem of WSN, but in reality it is not. The coverage problem schemes are mainly of two types: area coverage and target or point coverage. The area coverage problem explores the solution to cover the entire area of a WSN, while point coverage problem, a special case of area coverage problem, focuses on determining the exact position of sensor nodes to provide efficient coverage application for a limited number of targets [1].

Another problem that has drawn the attention of the authors while discussing the proposed approximation algorithm is Minimum Enclosing Circle (MEC) [7]. In MEC, the aim is to find out the smallest area enclosing circle in a given region to cover all the nodes of the network. By definition, the proposed problem seems very close to MEC, but they differ in the basic structure. Here, in MEC problem, the radiuses of the circles are arbitrary, whereas, the deployed sensor's transmission ranges in research have fixed range R. intersection points.

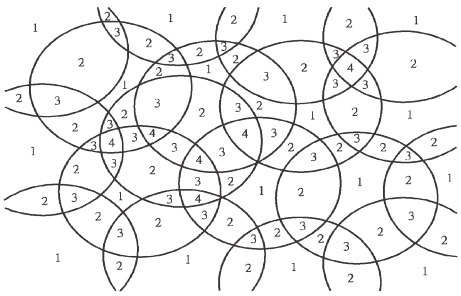
In this paper the authors are concerned with another algorithm, center selection problem, where the goal is to find out the center from some given sites.

Center selection problem finds a set of k centers C so that maximum distance r(C) from a site to the nearest center is minimum [8]. The site can be any source within a region. This problem deals with finding k centers by applying the algorithm k times, where k is a given constant. On the other hand, the author in this paper focus on finding out the number of such centers where by placing malicious node all other soldiers, i.e., site are covered.

III. Methodology

From the view of a WSN, the first consideration factor is the sensor. It can be easily understood that the authors need to emphasize on the deployment sensors (minimum resources) and the sensors of target (opponent's area resources). The goal is to cover maximum targeted sensors with minimum or near minimum number of resources. In [1] the authors assume that they know the network topology and all the sensors in the network have the same transmission range. To fulfill the aim, the main consideration factor is the strategy that directs to the destination more accurately and efficiently. F. T. Jaigirdar and M. M. Islam in their previous paper proposed two approximation techniques along with the complexity. To limit the searching area and reach the goal promptly and efficiently in this paper they illustrate a new strategy named *intersection point algorithm*. The authors prove the Excellency of the new approximation algorithm with its ease of implementation and reduced complexity as well as with more accurate result.

To reach the desired goal, the prime task is to find out the most crowded region in order to get the optimum locations of the deployed sensors. The strategy searches for the intersection points of the disks and place the sensor node (minimum resource) only in that intersection points for maximum destruction. By the proposed deployment strategy, it would be easier to find out the best locations as well as the minimum or near minimum amount of resources needed to destroy all the sensor nodes(opponent's area) in the network.



978-1-5090-0203-0/16/$31.00 ©2016 IEEE

Fig. 2. Intersection regions of the sensor nodes

An intersection region of two circles are the place where the distance between two circles is less than or equal to 2*R*, where *R* is the radius of the circle. In Fig. 2, different intersection regions of the intersecting circles are shown for understanding the intersection point's impact on the sensor nodes covering range.

A. Finding the intersection points

At first, from every sensor node, the entire sensor node's distance is counted to determine which sensor nodes are neighbors' to it, i.e., the intersection points of two sensors are calculated respectively. Here, the strategy finds out all nodes u, v such that dist (u, v) ≤ 2R, where R is the transmission range of the malicious node. In the Fig. 3, the sensor nodes are shown in small circle, and their transmission ranges are shown in larger circle. The intersection points (i1 to i9) are arrow marked in the figure. It should be noted that, the intersection point that has been taken once, should not be added if it comes with the next iteration.

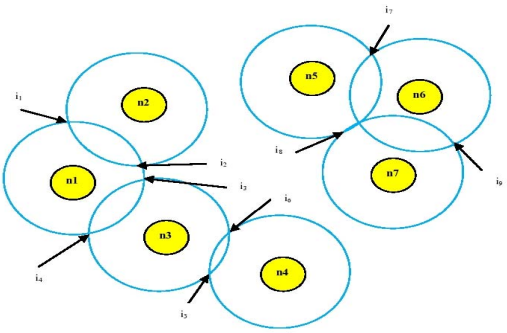


Fig. 3. Intersection points of the deployed sensor nodes

As a result, when n1 is searching for its neighbors, it finds i1 and i2 as intersection point with sensor node n2 and so on. But in turn for n2, it checks distance with n1 and find that these intersection points have already been taken and don't need to consider as new intersection points.

B. Placing the jamming nodes

After finding out the intersection points, the next task is to place the malicious or jamming node in the intersection points. It should be noted here that, the number of these jamming or malicious node is actually the minimum number of resources a commander needed in the battlefield for maximum destruction in the opponent's area (the minimum number of watch tower needed to monitor the opposition's area). To decide the exact deployment position, the strategy needs to search for the highest number of covering sensors by that intersection point. Greedily, the algorithm searches for that intersection point that covers the maximum number of sensor nodes in it's transmission range. Upon placing the malicious node in that position the intersection point strategy removes the covering nodes by that intersection point and reapply the method in order to find out the required goal. In Table 1, the intersection point, the node covered by intersection points and total number of covered sensors have been shown and this table is updated periodically to reach the final destination.

From the table, it can be noticed that, intersection point i2, i3 and i8 have the highest number of covering sensor nodes. So, the jamming node can be placed in any of these three points. Suppose the jamming node is placed in the position of i2. So, n1, n2 and n3 can be jammed or destroyed. Then these three nodes is removed from the table and the table is updated accordingly. In this way, the jamming node is placed in the intersection points and the strategy maintains the greedy approach and finally end with minimum set cover problem to find out the minimum or near minimum number of jamming or malicious node needed to jam or destroy all the sensors in the network.

TABLE I. INTERSECTION POINT'S COVERING SENSOR NODES AND THEIR TOTAL COUNTING

Intersection Point	Node covered	Total
i_1	n1,n2	2
i_2	n1,n2,n3	3
i_3	n1,n2,n3	2
i_4	n1,n3	3
i_5	n3,n4	2
i_6	n3,n4	2
i_7	n5,n6	2
i_8	n5,n6,n7	3
i_9	n6,n7	2

C. Algorithm

This section portrays the algorithm of intersection point scheme. In the algorithm, there are two parts. At first it finds out the intersection points of the sensor nodes and then it places the malicious or jamming node in the maximum covered point. Finally it finds out the minimum number of jamming node needed to completely jam or destroy the entire network. Here Algorithm determine_intersection_points() performs the first part, i.e., finding out the intersection points and the later part of the strategy is performed by the Algorithm Intersection_Point_Scheme().

Algorithm Intersection_Point_Scheme (NN, R)

```
//num_unaffected=number of nodes unaffected,
//high_freq=highest number of nodes covered in its
//transmission range(R), high_freq_x,high_freq_y = represents
//the coordinates of highest frequency point, R=transmission
//range
1.  determine_intersection_points (NN, R)
2.  num_unaffected=NN
3.  While num_unaffected > 0 do
4.      high_freq:= 0
5.        for i := 0 to ip_count
6.            count:= 0
7.                for k := 0 to num_unaffected
```

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

```
8.                    if distance( ip [i] [0], ip [i] [1], nodes
    [unaffected_nodes [k]] [0],
9.                  nodes [unaffected_nodes [k]] [1]) <= R
10.                 count++
11.                   endif
12.                 endfor
13.                if count > high_freq
14.                   high_freq:= count
15.                   high_freq_x:= ip [i] [0]
16.                   high_freq_y:= ip [i] [1]
17.                endif
18.          endfor
19.   for j := 0 to num_unaffected
20.   if distance(high_freq_x, high_freq_y, nodes
    [unaffected_nodes[j]] [0],  nodes[unaffected_nodes[j]
    [i])>R
21.          temp [i++]:= unaffected_nodes [j]
22.         endif
23.   endfor
24.   for j:= 0 to i
25.         unaffected_nodes [j]:= temp [j]
26.   endfor
27. endwhile
```

Algorithm determine_intersection_points (NN, R)

```
//NN=number of nodes, R=transmission range,
//ip_count=intersection point count
1.   for i := 0 to NN
2.      for j := 0 to NN
3.         determine x1,y1,x2,y2 as node's co-ordinates
4.          d:= distance (x1,y1,x2,y2)
5.              if d <= 2R
6.                 find xx1, yy1,xx2,yy2 as intersection
    points
7.                    for each intersection point do
//check whether the new intersection point already exists
//in the array.
8.                  yes:= 0
9.                  for i:= 0 to ip_count
10.                    if xx = ip [i] [0] and yy = ip [i ][1]
11.                          yes:= 1
12.                      else
13.                        ip [ip_count] [0]:= xx
14.                         ip [ip_coubt] [1]:= yy
15.                     ip_count++
16.                    endif
17.                 endfor
18.             endif
19.        endfor
20.  endfor
```

D. Complexity

The authors have assumed that two sensors are not in the same place. Considering, Number of sensors=N, R=transmission range, CD =Cost for calculating distance, CI= Cost for calculating intersection points and D=dimension of the network.

Step 1: At first all the nodes are deployed randomly and from each sensor node all other sensor node's distance are calculated. So, the first task here is the distance calculation. As the number of sensor nodes are N, the distance calculation at this step is $N \times N = N^2$.

Step 2: The reason behind calculating distance between two nodes is that if distance, $d \leq 2R$, then those two nodes intersect with each other. So, the number of intersection points, NIP is

$$\frac{\Pi (2R)^2}{D^2} \times (N-1) \times 2 \times (1/2) \times N$$

$$NIP = (4\prod N(N-1)(R/D)^2 \quad (1)$$

Here, NIP is calculated dividing the intersecting area by the dimension of the network. There are two intersection points by connecting two nodes, so the above statement at first multiplied by two and again divide by two, as there are also two points which will count twice with same node. For calculating the total number of intersection points, the whole statement is multiplied by N.

Step 3: After getting the intersection points, the strategy searches for that intersection point, where by placing a jamming node maximum interruption is possible. In this stage distance from every intersection point to every sensor node needs to be calculated. As a result, the distance calculation for this step is

$$4\Pi N (N-1) (R/D)^2 \times N$$

Here, other comprising costs for further processing have neglected and that is why is not added in calculating the total complexity of the strategy.

Finally, Total Complexity $= N^2CD + 4\Pi N (N-1) (R/D)^2 CI + 4\Pi (N-1) (NR/D)^2 CD$

$$[N^2 + 4\prod (N-1)(NR/D)^2]CD + 4\prod N(N-1)(R/D)^2 CI \quad (2)$$

IV. Simulation Results and Comperative Analysis

This section illustrates the simulation arrangement to evaluate the performance of the intersection point approximation technique. Different considering factors are added here for clear understanding the network scenario as well as the different network parameters ranges.

A. Simulation Setup

The authors have developed a simulation software in Java environment using the discrete event simulation toolkit, SimJava. The simulations were carried out in an Intel Xeon processor with 2GB RAM. The environment, number of sensors, step size for placing grids and other simulation parameters were chosen carefully to ensure that the real environment is to be reflected through simulation. To set up the simulation environment, a 2D space having size of D×D is

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

used and *N* sensor nodes deployed in that space. The number of nodes, *N* is changed to several values for examining the effects of the algorithm. The dimensional length, D also varied into different ranges for analyzing the effects of the algorithm on different space size, i.e., from small to large network.

B. Simulation Parameters

The different parameters used for simulation environment are the transmission range *R*, i.e., a malicious node's covering range of receiving and transmitting signals, number of sensor nodes *N* and total networking area or dimension, *D*. To increase the accuracy of simulation results, the outcome of the simulation has been obtained at least five times for the same snapshots of the experimental environments and parameters and then the average is taken. Different values of the changing parameters used in simulation are listed in Table 2.

TABLE II. VALUES OF THE PARAMETERS

Parameters	Values
Transmission Range, R	5, 8, 10, 25, 40,70
Number of Nodes, N	70,100,150,200,300
Dimension, D	100, 150, 200, 300

C. Results and comperative analysis

The authors have started the experimental analysis by keeping the number of nodes fixed, 300. Other two parameters, i.e., transmission range, *R* and network topology or dimension; *D* are changed in this scenario that is illustrated in Fig. 4. From the figure it can be noticed that as the transmission range increases, the number of resources needed to jam the network decreases accordingly.

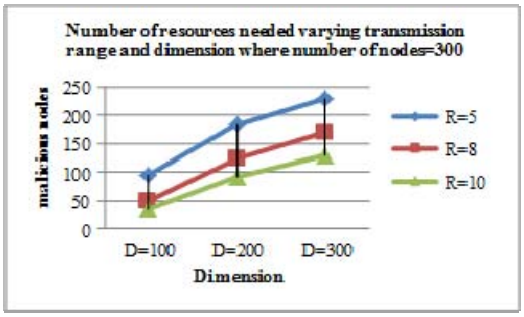


Fig. 4. Results by changing transmission range and dimension

The reason for such a result is with the increasing value of transmission range, more nodes can be in the range of the intersection points (i.e., by placing a malicious node at that point, all the sensors that are within the transmission range of that point can be destroyed easily) and number of resources is lessening accordingly.

The second scenario varies the dimension and number of nodes, while transmission range remains fixed, 10. Another important parameter of the network is number of sensing nodes, *N*. For establishing different applications of WSN, in most of the cases, a sensor network with maximum number of nodes deployed is needed. From the Fig. 5 and Fig. 6, it can be noticed that, in most of the cases, as number of nodes increase the number of required resources increase accordingly. This is because, with higher number of nodes, more jamming nodes are necessary to embrace them meeting the corresponding criteria. Finally, the result is verified by different values of transmission range and number of nodes, keeping the dimension or network topology fixed to 150×150. A last concerning criterion of the network is its dimension or network topology, D. In a large area, where nodes are placed randomly it is normally happens that they are placed in a scattered manner and that's why more resources are required to cover all the nodes in the network. Fig. 4 and Fig. 5 shows that as the dimension increase, therefore, in most of the cases, the number of resources needed increase

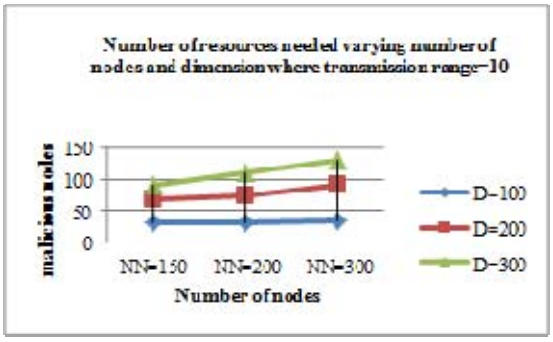


Fig. 5. Results by changing dimension and number of nodes

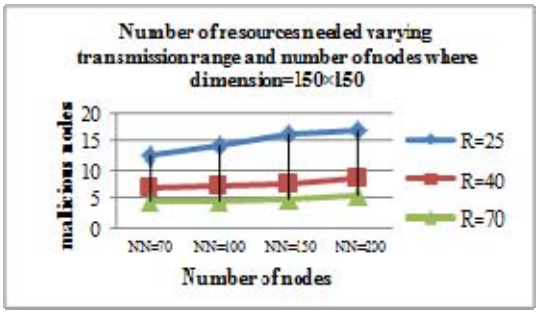


Fig. 6. Results by changing number of nodes and transmission range

In [1], the authors have described these scenarios for maximum clique analysis and maximum degree analysis algorithm. The algorithms were compared in respect of performance analysis. In this paper, the authors compare the three approaches that surely help to determine the acceptability of the new Intersection Point Algorithm. Fig. 7 depicts the comparison scenario by changing the values of number of nodes from 100 to 300.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

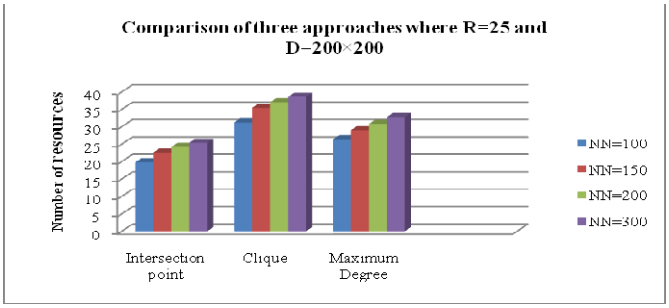


Fig. 7. Different approximation results by changing the values of number of nodes

It can be easily determined from the Fig.7 that in every value for number of nodes (N=100,150,200,300), number of resources required is minimum for intersection point algorithm in comparison to maximum degree and maximum clique algorithm. For example, while number of sensors, N=200, the minimum number of resources needed for intersection point algorithm is 24, whereas in maximum degree it is 31, and 37 for maximum clique approach.

The authors have portrayed another comparative scenario in Fig. 8 by changing the values of transmission range. The number of nodes needed for this is fixed to 200 and the dimension is 200×200. Here, for R=40, the minimum number of resources needed for intersection point is 13, 15 for maximum degree analysis and 18 for maximum clique approach. As anyone has a short glance to the figure, he/she can easily understand that among three approaches, intersection point algorithm is the best.

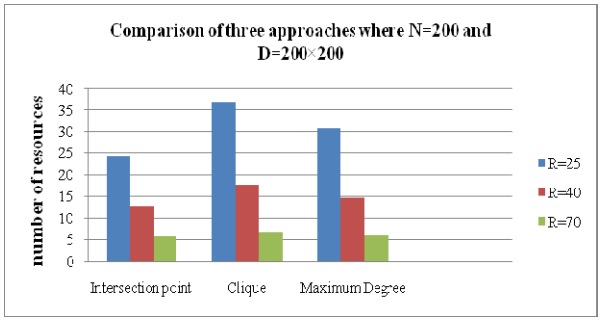


Fig. 8. Different approximation results by changing the values of transmission range

V. CONCLUSIONS

A commander in the battlefield is responsible for not only securing his troops and resources, but also for assulting his opponents in a planned way. He also has to be very aware of the resources he has in the battlefield as measuring the cost-effectiveness while planning is also a major concern. So, maximum destruction/ careful monitoring with minimum/near minimum number of resources in the battlefield is a challenging task for a commander. In [1] the authors have solved the problem by two techniques, one is maximum clique analysis and other is maximum degree analysis. This paper is focused on finding a new approach, intersection point algorithm, that is proved to be best among the three techniques. With simulation results and comparative analysis the authors have successfully proved its excellency in this regard. The algorithm proposed in this paper are suitable to be applied in battlefield cost-effectively and efficiently.

References

[1] F. T. Jaigirdar and M. M. Islam, "Assurance of the maximum destruction in battlefield using cost-effective approximation techniques", *Journal of Networks*, vol. 7, no. 12, pp. 1967-1977, December 2012.

[2] Y. Wang, Y. Zhang, J. Liu and R. Bhandari, "Coverage, connectivity, and deployment in wireless sensor networks" in *Recent Development in Wireless Sensor and Ad-hoc Networks (Signals and Communication Technology)*, S. Patnaik, X. Li and Y-M. Yang, Springer, pp. 25-44, 2014.

[3] C.–Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities and challenges," *Proceedings of IEEE*, vol. 91, pp. 1247-1256, 2003.

[4] Finding intersection points, http://mathworld.wolfram.com/Circle-CircleIntersection.html, last visited 9th February, 2015.

[5] R. Hassin, and A. Levin , "A better-than-greedy approximation algorithm for the minimum set cover problem," *SIAM Journal on Computing*, pp. 189-200, 2005.

[6] Greedy algorithm, *http:// en.wikipedia.org/wiki/ Greedy_algorithm*, last visited 6th June, 2014.

[7] A. Karmakar, S. Roy and S.Das, "Fast computation of smallest enclosing circle with center on a query line segment," Conference paper in *Information Processing Letters*, January, 2007.

[8] B. Gopalakrishnan, T. Yoshii and S. M. Dappili, "Decision Support System for mstching center selection", *Journal of manufacturing Technology Management*, Vol. 15, Issue 2, 2004.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Full Papers

Secured and Efficient Mobile Computing

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Enhancing the Embedding Payload by Handling the Affair of Association and Mapping of Block Pixels through Prediction Errors Histogram

A H M Kamal and Mohammad Mahfuzul Islam
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
kamal@jkkniu.edu.bd and mahfuz@cse.buet.ac.bd

A H M Kamal
Department of Computer Sciene and Engineering
Jatiya Kabi Kazi Nazrul Islam University
Mymensingh, Bangladesh
kamal@jkkniu.edu.bd

***Abstract*—Images are intentionally distorted in many reversible data hiding schemes. This distortion is performed either by encrypting before data embedment or by applying histogram association mapping (HAM) during data concealment. The later one minimizes the processing complexity and demolishes the requirement of sharing secret keys. In the HAM, each block's pixels belong to a segment in the gray scale are associated and mapped by the message bits to another gray part. The quantity of embedded bits are determined by the range of block pixels, i.e. limit within which pixels spread out. Though, smaller valued ranges contribute much to embedding payload, the frequency of such ranges decreases in archetype images. However, a predictor, if applied on the block pixels, huddles its prediction errors together. Consequently, the range of absolute values of the prediction errors become smaller. The proposed scheme, therefore, applies HAM to prediction errors and shifts the block pixels accordingly. Experiments were conducted using three different standard image datasets to examine the enhancement in the embedding payload and to investigate the performance on different block-sizes. The investigation reveals that the average payloads in the proposed scheme is 1.5 to 2 times of the competing one in all the cases.**

Keywords— Histogram association mapping, intentional image distortions, prediction errors, embedding payload, reversible data hiding scheme and block range.

I. INTRODUCTION

In the image steganography, images are used as a carrier of a secret message. Secret data are implanted inside into an image by a set of embedding rules to deceive the sensing capability of unauthorized viewers [1]. The image before and after data embedment is termed as 'cove' and 'stego' image respectively. This image steganography are classified into two broad groups - reversible [1-11] and irreversible [12-15] - depending on whether the secret data extractor can reconstruct the original cover image from the stego one or not. The reversible data hiding schemes are applied when the secret message and the cover image are equally important to the post processing stages in the receiver end.

Though most of the reversible data hiding schemes are devoted to preserve the image quality to their best level [5-11, 16-18], a good number of endeavors are also observed in the literature to destroy the image quality intentionally [1-4]. In the purposely image quality degradation schemes, the information of the cover image are annihilated so that no illicit person or device can retrieve the original contents. Such schemes play very vital roles in medical and forensic applications; in transmission of legal documents, evidence or report by law-enforcing agencies and in communicating for copyrights and certificates. The purposely quality degradation based schemes destroy the cover information either by the embedding rules during the data embedment [1] or by encrypting before the data embedment [2-4]. The encryption is not a part of the steganographic process and thus, it just increase the processing complexity. Besides, modified cover contents by a smaller encryption key can be decrypted by repeated tries. Therefore, destroying the cover information by the embedding rules are more rational.

This image distortion can be achieved in three ways. (1) The embedding is performed into the pixel values or into the prediction errors, generated by a predictor during its estimation of the cover pixels, by shifting these up to a larger embedding layers respectively in the image histogram or prediction errors histogram [10]. However, it does not ensure the shifting of pixel values by an equal amount. There, many pixels either remain unchanged or move to near values. Consequently, cover information is partially accessible or fully guessable. Thus, these schemes do not serve the objectives of intentionally image degradation based data hiding policy. (2) In the second case, introduced by Ong *et al.* in 2014 [1], a block of pixels are equally shifted by an amount in the gray scale by the embedding rules. The number of bits that are embedded into a block depend on range of block (maximum valued pixel - minimum valued pixel + 1). The

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

smaller range results in yielding higher embedding capacity. That scheme destroys the cover information during the data embedment. This process is known as histogram association and mapping (HAM). This is a very fruitful mechanism. This scheme both destroys the image quality noticeably and enhances the embedding capacity. (3) Many embedding process annihilate the cover image before data embedment. First, each the pixel in an image is encrypted by an encrypted key, e.g. Liao *et al.* in 2014 [4]. This encryption process fully raze the cover information. The secret information are then concealed into that encrypted image.

It is noticeable in [1] that more bits will be planted into the block by the scheme if the range becomes smaller. That range depends on the values of the pixels of the block. Therefore, it cannot be minimized by the data hider. Consequently, the embedding payload cannot be enhanced, if applications demand too. That dilemma is solved, in this paper, by introducing a predictor and measuring the range values from the prediction errors where prediction errors are measured from the differences of the cover values and predicted values. The authors have observed that the range of the prediction errors in their absolute values is much smaller than the range of the pixel values of the working block. Hence, the proposed novel scheme has boosted up the embedding payload by employing the range of absolutes of the prediction errors rather than the range of pixel values. That proposed scheme dominates its competing one [1] by the embedding payload and the image quality.

The remaining article is organized into more five sections. Section II is to describe the related works with which the proposed work is compared. Section III is devoted to illustrate the proposed scheme. The results of the proposed scheme from various perspectives are demonstrated and discussed in section IV. There are many steganalyzers which can detect the perturbation in the stego image. The detection rate by the steganalyzers is increased with the growth of distortions in the stego image. The target of the article is to distort the stego image as much as it is possible. A steganalyzer, which applies generalized Benford Law, is experimented in section V to test the rate of stego detection, indeed the rate of distortions. Finally, the article is concluded in section VI.

II. Related Works

In this section, *HAM* scheme of Ong *et al.* [1] and Liao *et al.'s* scheme of embedding into encrypted image [4] are shortly explained. The first one destroys the image quality during the data embedment by the embedding rules while the next one razes the cover information before starting the data concealment process by encrypting the image.

A. HAM Scheme of Ong et al.

In *HAM,* first, the range of the block, (maximum valued pixel-minimum valued pixel+1), are calculated. Say, the range is *R*. The gray scale, i.e. 0 to 255, is divided into $P = 2^{8-\lceil \log_2^R \rceil}$ parts where $\lceil . \rceil$ stands for mathematical ceiling. In this block, $b = 8 - log_2^R$ bits are concealed by the embedding rules. The pixels of the block associated to a gray part are shifted to one of the *P* parts depending on the pattern of the bits in the secret message chunk. For example, if $32 \le R < 64$ and pixel values ranges within 64 to 127, as is shown in Fig. 1, then gray scale is partitioned into 4 parts. In this scenario, 2 bits are embedded into the block, i.e. *b*=2. Depending on the patterns of 2 bits message chunk, i.e. {00, 01, 10,11}, the block pixels are associated and mapped to partitioned range {64, 127}, {128, 191}, {192, 255} and {0, 63} respectively.

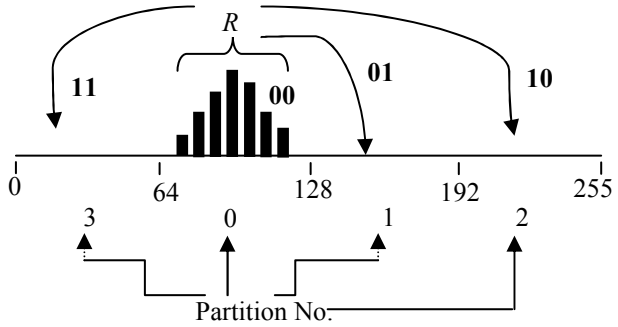


Fig. 1: Data embedment in HAM.

From the aforesaid discussions, it is investigated that the number of bits *b* that can be embedded into a block depends on the value of range and *b* varies within 1 bit to 8 bits, e.g. *b*=1, 2, 3, ... , 8 for *R*={65 to 128}, {33 to 64}, {17 to 32}, {9 to 16}, {5 to 8}, {3 to 4}, {2} and {1} respectively. For *R*>128, a special case known as non-reflective block (NRB), pixels are shifted by a fixed value both to the left and right side in the gray scale and one bit is embedded then.

B. Liao et al.'s Scheme of Embedding into Encrypted Image

Each of the pixels of the cover image are encrypted by an encryption key. All the pixels are encrypted by an XOR operation between each pixel and the encryption key. This encryption process fully destroys the cover information. This encrypted image is partitioned into equal-sized blocks. The block pixels are then separated into two equal-sized sets S_0 and S_1. The 3 LSBs in S_0 are flipped to embed '0' bit while the same number of LSBs are flipped in S_1 to embed '1'. In each image block a single bit is concealed. Thus, the scheme lessens the embedding capacity.

III. Prediction Errors Based HAM Scheme

The reason of applying the range of absolute values of prediction errors, say R_e, rather than the range of pixel values *R* has already been explained in the above section. In this section the pre-processing steps, the embedding procedure and the data extraction as well as image recovery is described.

First, the image *I* of size $h \times w$ is partitioned into blocks of size $m \times n$. Data is embedded into and extracted from each block independently and the procedure is same for all the

blocks. Hence, all the processes are discussion for $m \times n$ pixels only, i.e. for a single block only.

A. Measuring the Predicted Values and the Prediction Errors

4-connected neighborhood is applied to predict the values of *k*-th block $B^k, 1 \le k \le \left(\frac{h \times w}{m \times n} \right)$. The four corner pixels are estimated from the average of two neighbor pixels. The other border pixels are estimated from the average of three neighbor pixels. The remaining, inner, pixels are predicted from the average of four neighbor pixels. Mathematical floor function is applied to make integers from the fractional values. Let the predicted values are $pB_{i,j}, 1 \le i \le m, 1 \le j \le n$. Then the prediction errors is computed by eq. (1)

$$pE_{i,j} = B_{i,j} - pB_{i,j} \qquad (1)$$

The absolute value of prediction errors is $pE^A_{i,j} = ABS\left(pE_{i,j} \right)$.

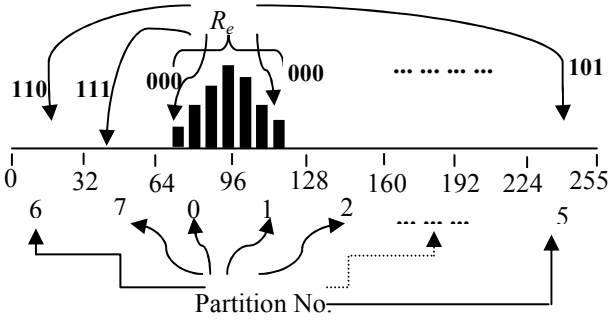


Fig. 2: Prediction error based HAM scheme

B. Computing Ranges

The range of block's $pE^A_{i,j}$ is $R_e = \max\left(pE^A_{i,j} \right) - \min\left(pE^A_{i,j} \right) + 1$ where *max* and *min* function returns the maximum and minimum values respectively among the absolute errors in $pE^A_{i,j}$. The gray scale partition range R_P is calculated by eq. (2) and number of gray scale partitions by eq.(3) as is shown in Fig. 2 for R_P=32.

$$R_P = 2^{\lceil \log_2^R \rceil} \qquad (2)$$

$$P = 256 / 2^{\lceil \log_2^R \rceil} = 2^{8 - \lceil \log_2^R \rceil} \qquad (3)$$

C. Data Embedment Process

First the gray scale is partitioned into *P* parts. Say, the minimum value of the block is Min_c. A partition whose lower gray value is smaller than Min_c and upper gray value is greater than Min_c is numbered as '0'. A partition labeler forwards to the right direction numbering each partition as 1, 2, 3, ..., until it reaches the last partition. Thereafter, it moves to the leftmost partition and again move to right direction until it labels '(1-*p*)'. In the Fig. 2 the labels are 6, 7, 0, 1, 2, 3, 4 and 5. The number of message bits *b* that are embedded into that block B^k is $b = 8 - log_2^P$. Let this message chunk of *b*-bits is *m*. The binary message chunk *m* is converted to decimal *d*. The modified block S^k is formed by eq. (4). Many of the pixels in S^k may exceeds the upper gray extreme 255, e.g. if B^k moves to last partition. Hence, the final stego block $\tilde{S}^k$ is formed by eq. (5).

$$S^k = B^k + d * R_P \qquad (4)$$

$$\tilde{S}^k = \mathrm{mod}\left(S^k, 256 \right) \qquad (5).$$

Overflow of pixel values in S^k is tracked by the eq. (6).

$$O^k_{overflow} = \begin{cases} 1 & if \quad S^k_{i,j} > 255 \quad for\ any\ i, j \\ 0 & otherwise \end{cases} \qquad (6)$$

All the stego blocks, $\tilde{S}^k$ are concatenated to form the stego image $\tilde{I}$. The embedding process is demonstrated with an example in the appendix.

D. Side-Information

Most of the reversible applications manage side-information. Side-information usually assists the extractor to extract the secrets as well as to reconstruct the cover image from the stego image by providing sufficient information to the extractor. These side-information are also communicated either by embedding these to a part of the stego image or by sending these separately through alternate mechanism. The management of side-information, like selection of parameters and allocation of bits for each one, compression techniques, sending these to the destination, etc., are described in many of the literatures including [1, 11]. Therefore, we are avoiding that management of side-information.

In the competing scheme [1], side-information of minimum and maximum value of each block, the number of shifted pixels in left and right in *NRB* are managed by 32 bits of side information for each block. In our case, it is only of 9 bits. The proposed scheme manages 8 bits side information to track Min_c , and 1 bit for $O^k_{overflow}$.

E. Secret Data Extractions and Cover Reconstructions

It can be observed that during the data embedment the eq. (5) shifted the overflowed pixel to the lower values of the gray scale. The following eq. (7) is applied before message extraction to return to the state produced by eq. (4). Let the maximum and the minimum value of the stego block is Max_s and Min_s and stego block range $R_s = \max_s - Min_s + 1$. The equation (7) is executed for $O^k_{overflow}$ ='1'.

$$S_{i,j}^{k}=\begin{cases}\tilde{S}_{i,j}^{k}+256 & if \quad R_s>128 \quad and \quad \tilde{S}_{i,j}^{k}<\max_s/2\\ \tilde{S}_{i,j}^{k}+256 & if \quad R_s\le 128\\ \tilde{S}_{i,j}^{k} & otherwise\end{cases} \quad (7)$$

Now, the cover block minimum, Min_c is collected from the first 8-bits of side-information. Again, applying the stated policies in the above subsections *A*, *B* and *C* on $S_{i,j}^{k}$ few other parameters $pE_{(i,j)}^{A}, R_P, P$ and *b* are computed. The present stego displacement *dS*, regarding to cover block, in $S_{i,j}^{k}$ is measured by the eq. (8). Then, the cover block is reconstructed in eq. (9) by again subtracting the amount of stego displacement from the $S_{i,j}^{k}$. The decimal value of the secret is found by eq. (10).

$$dS=\min\left(S_{i,j}^{k}\right)-\min_c \quad (8)$$

$$B_{i,j}^{k}=S_{i,j}^{k}-dS \quad (9)$$

$$d=2^{\left\lfloor \log_2^{dS/R_P} \right\rfloor} \quad (10)$$

Finally, *d* is converted to binary *m* and sufficient '0's are appended to the left of *m* to make it *b*-bits length. The whole process is demonstrated with an example in the appendix.

The process is repeated for the other blocks in the stego image.

IV. Experimental Results and Discussions

The experiments were done using 5000 CalTech101 images, 500 BOSS image datasets and 200 images collected by authors. The images were resized to 240x240. Then, the embedding payload and the peak signal to noise ratio (PSNR) were analyzed in our experiment and these were investigated on various block sizes. The results are demonstrated in the following. The demonstrated results justify the claim of boosting up the embedding payload and improving the distortions in the image quality by the proposed scheme.

A. Effect of Range Measured from the Prediction Errors

Why the authors have computed range from the prediction errors is analytically observed in Fig. 3. Most of the prediction errors are condensed to '0' or close to '0' whereas the pixel values are distributed over the gray range in about a flat rate. Which implies that the pixel values in the block are distributed in a wider range. Hence, the *R* becomes larger. On the contrary, prediction errors are distributed into a smaller range. As a result R_P becomes smaller.

A primary test is performed to check the behavior of ranges computed from both the prediction error and the pixel values. The higher valued ranges decrease the embedding payload. The scheme embeds a single bit into a block when the range is greater than 63. If the scheme produces more ranges greater than 63, the probability of holding bigger values by the other ranges are also higher. Therefore, the more the scheme generates *R* or R_P greater than 63, the more it reduces the embedding payload. Hence, total blocks whose ranges are greater than 63 are counted and tabulated in the Table 1 separately for [1] and the proposed scheme. It is observed that for all the image block sizes and in all the image datasets, the quantity of these blocks in [1] is multiple of *x* of the proposed scheme where *x* ranges from 3 to 6.36. Thus, it infers that the other ranges are also smaller in the proposed method. Another noticeable point is that the BOSS image dataset provides less number of higher ranged blocks. The reason is that most of the BOSS images are captured for a single objects where the variations in the color values are smaller. Consequently, the quantity of higher ranged blocks are smaller.

Table 1: Average number of blocks in each image dataset whose block range is greater than 63 at various block sizes

Image Database	Scheme	Maximum NBR in Block Size			
		3x3	5x5	8x8	12x12
BOSS	Ong *et al.*	435	312	189	115
	Proposed	110	69	41	27
CalTech 101	Ong *et al.*	1266	776	416	240
	Proposed	388	236	137	80
Authors collection	Ong *et al.*	1136	801	455	249
	Proposed	198	126	77	49

B. Analysis on Embedding Payload

The experimental results demonstrated in Fig. 4 states that the proposed scheme noticeably dominates the others by the achieved payload. The payload of first 100 images of CalTech101 image dataset are depicted along y-axis. Among the compared schemes Liao *et al.*'s one presents low embedding capacity because in this method only a single bit of information is planted in each image block. Ong *et al.*'s scheme presents a better payload due to their attempts of embedding multiple bits in each image block. The proposed method further improves its embedding payload by a factor of about two regarding to Ong *et al.*'s one. This is achieved by applying the range of the absolute values of the prediction errors during the selection of number of embedded bits in a block. As only Ong *et al.*'s scheme competes with the proposed scheme regarding to the concern of payload, these two are further compared in Fig. 5 for different image block sizes. The embedding payload in the proposed scheme is much higher than the obtained payload in Ong *et al.*'s scheme. The average payload in each image datasets are drawn along y-axis and the results against different block sizes are shown along x-axis. The results of the proposed scheme and the Ong *et al.*'s scheme is grouped in the figure for each block size. The results for the block size 2x2 is not depicted in the figure just to let the results for block size of 8x8 and 12x12 visible. The payload in 2x2 is about 2.5 times of the payload in 3x3. The

payload for block size of 3x3, in the Fig. 5, is more than the payloads in the other block sizes because in block size 3x3, the number of processed blocks are more in quantity than the others and data are embedded into each block separately. For the same reason, the payload is highest in block size 2x2 among the experimented block sizes. With comparison to Ong *et al.*'s scheme, the average payload increases in the proposed scheme in {BOSS image dataset, CalTech101 image dataset and authors collected images} by a factor of {1.31, 1.46, 1.47}, {1.31, 1.51, 1.58}, {1.33, 1.62, 1.77}, {1.35, 1.73, 1.93} and {1.38, 1.83, 2.04} respectively for image block of size 2x2, 3x3, 5x5, 8x8 and 12x12. Thus, the improvement in the embedding payload by our scheme ranges from 1.31 to 2.04 depending on the category of images and size of image block.

In Fig. 6, results of 5000 images of CalTech101 image dataset are demonstrated to check whether the payload of any individual image in the proposed scheme falls bellow the corresponding payload in Ong *et al.*'s scheme. The proposed scheme successfully passes the test. The minimum payload that are obtained in each dataset are also compared for different block sizes separately in Fig. 7. There, it is noticed that our scheme dominates the competing one. Thus, it can be concluded that the proposed scheme certainly enhance the embedding payload and the matter of that improvement is free from size of blocks and sources of images.

C. Analysis on PSNR

The target of the scheme is to destroy the image quality intentionally and notably. Fig. 8 delineates that the cover image, here an apple, is destroyed in both the Ong *et al.*'s scheme and in the proposed scheme in a scale such that nothing is recognizable from the stego images. Thus, visually the target is achieved. To test statistically, the *PSNR* of the resulted stego images are measured by the following eq. (11).

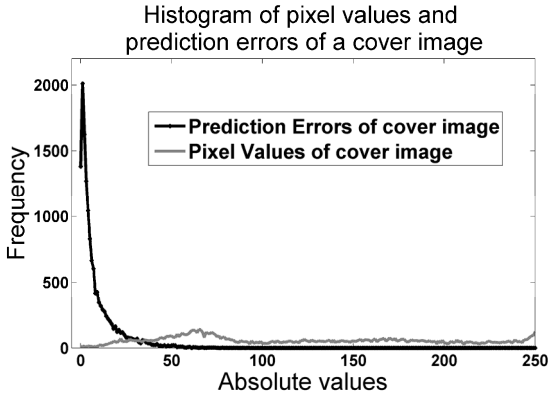


Fig. 3: Histogram drawn from the prediction errors and pixel values of a cover image of CalTech101 image database

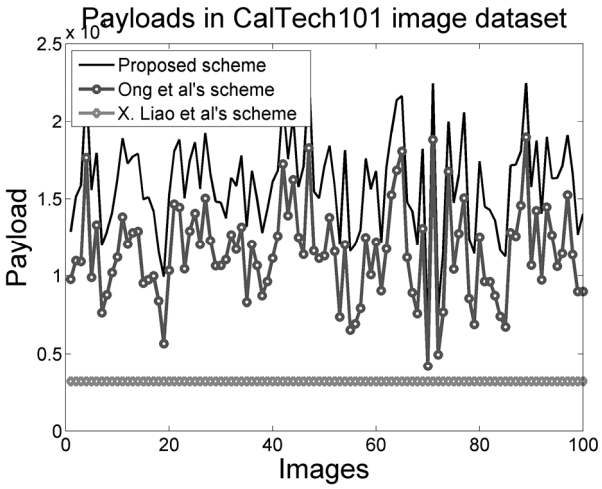


Fig. 4: Comparison of payload between the proposed scheme and the scheme of Ong *et al* and Liao *et al*. The results are presented from the first 100 images of CalTech101 image dataset.

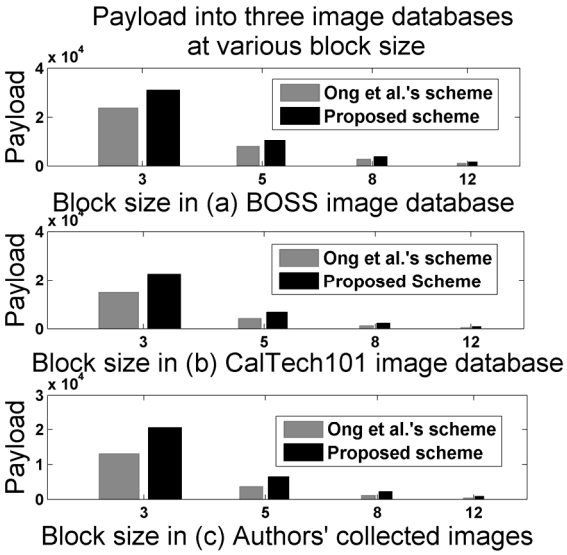


Fig. 5: Average payload obtained into three image databases at different block sizes.

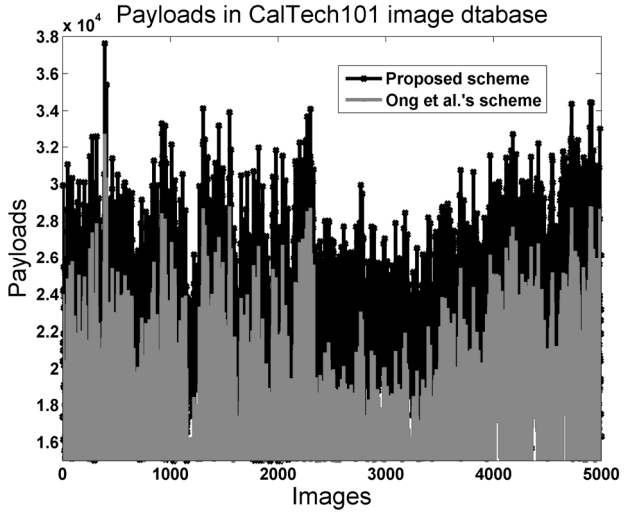


Fig. 6: Achieved payloads in 5000 images of CalTech101 image dataset

$$PSNR = 10\log\left(\frac{255^2}{MSE}\right) \qquad (11)$$

Where *MSE* stands for mean-square-errors and

$$MSE = \frac{\sum_{i=1}^{row\_size}\sum_{j=1}^{column\_size}\left(I_{i,j} - \tilde{I}_{i,j}\right)^2}{row\_size * column\_size}$$

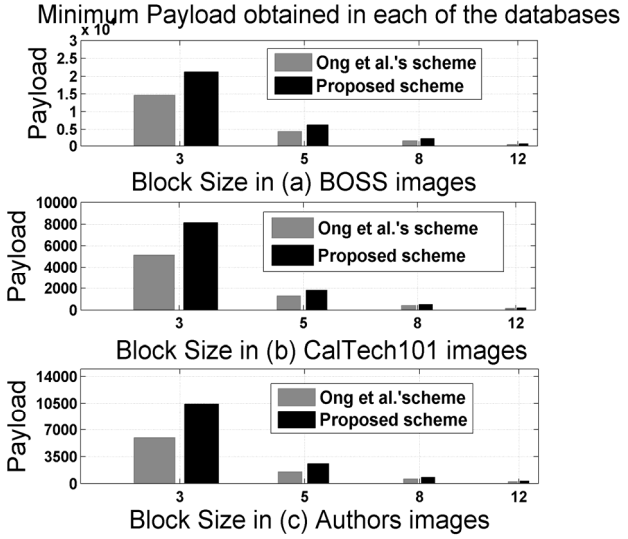
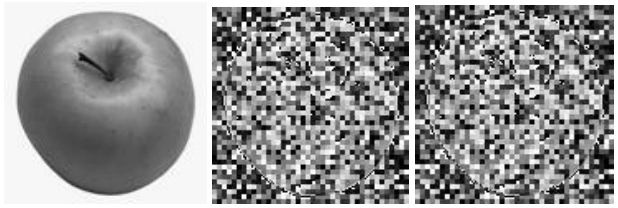


Fig. 7: Minimum payload obtained in each of the databases



(a) Cover (b) Stego by Ong *et al.* (c) Stego by Proposed

Fig. 8: Distorted stego images in (b) and (c) which are obtained from cover image (a) by the Ong *et al.*'s and proposed scheme respectively

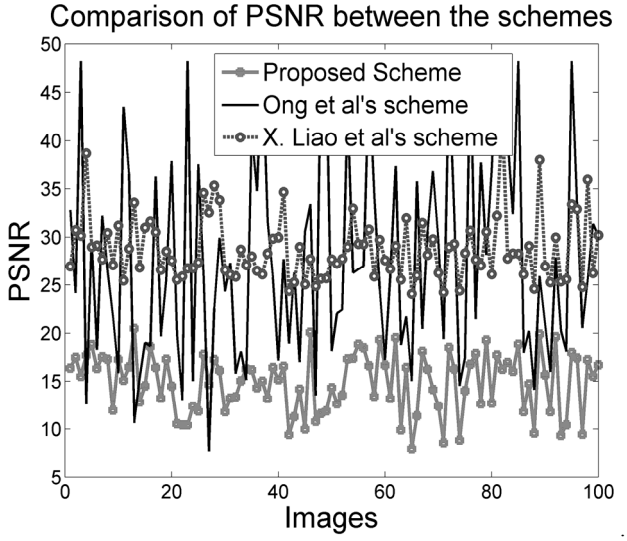


Fig 9: Comparison of PSNRs achieved in first 100 images of CalTech101 image dataset by different schemes.

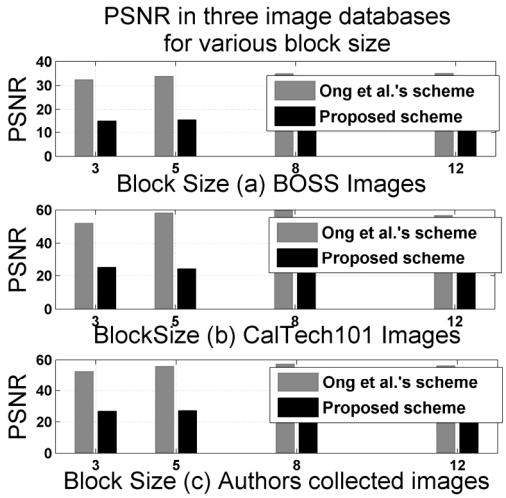


Fig. 10: Database and block size wise PSNR

The results are presented in Fig. 9 to Fig. 10. In Fig. 9, it is noticeable that PSNR in Liao et al.'s scheme does not vary in a wider range because only 3 LSBs are flipped of the 50% of the pixels, i.e. MSE is roughly $(8^2 \times Image\_size)/2$. On the other hand, PSNR in Ong et al.' scheme depends on the pattern of implanted bits and the range pixel values and prediction errors. This is why the PSNR varies in a wider range from image to image. Again, both the smaller range and the eq. (5) in the proposed scheme allow the modified pixels to be destroyed more. As a result worsening PSNR is presented by the proposed scheme. In Fig. 10, it is examined that the PSNR in all block sizes for the same category of images are very closed and increased a little bit for bigger blocks. In bigger blocks, the probability of appearing wider ranges is higher. Then, number of gray partitions, *P*, become smaller. Consequently, average shifting amount decreases. Besides, shifting a block to its neighbor partition ensure less distortion for a large bundle of pixels. For these reasons, *PSNR* is a bit improved in larger blocks. Among the categories, BOSS image dataset shows worsening PSNR because in BOSS images most of the ranges are smaller valued (the reason is explained earlier). Then, lots of gray scale partitions are generated and thus the scope of shifting blocks to different gray range, depending on bits pattern, is increased. Hence, in a frequent manner, depending on the bits pattern, the blocks are moved to a far gray scale level. Consequently, the *PSNR* is decreased. Again, *PSNR*s in the proposed scheme is much smaller than the same in Ong *et al.*'s scheme hopefully for only two reasons. Firstly, the ranges in the proposed scheme are smaller and thus the blocks, in a frequent manner, are shifted to distance gray parts. Secondly, the proposed scheme by eq. (5) has shifted lots of larger values, greater than 255, to very smaller values, like complementation. As a result, PSNR is sharply affected by these pixels. However, such events are not happened in Ong *et al.*'s scheme in a frequent manner

V. Checking Stego Distortions Applying Steganalyzer

Steganalyzers are applied both by the intruders or devices to check the existence of hidden contents in an image and by

the researchers to check the resistance of a stego image against statistical attacks. In both the cases, the steganalyzer presumes that the stego generator manages the image quality to its best level by lessening distortions. Hence, the detection threshold applied in the analyzer becomes small. Therefore, stego images generated by our scheme should be detected very successfully as the proposed scheme destroys the image quality intentionally. In this section, this is checked by a very effective steganalyzer which employs generalized Benfords law (gBL) in its detection algorithm.

A. What is gBL?

In a large set of natural numbers, it is observed that the probability of appearing 1 as a first significant digit is more than the probability of appearing 2 and so on, i.e. *P(1)>P(2)> ... ... >P(9),* where *P(n)* is the appearing probability of *n as* a significant digit and *n=1, 2, ..., 9*. Expected values of each *n* are measured by following eq. (12), known as *gBL* [19].

$$p(n) = N.\log_{10}(1+\frac{1}{s+n^q}), n = 1,2,\ldots,9 \ldots\ldots\ldots (12)$$

where *N*, *q* and *s* defines the accuracy of the relation at different compression quality factors. A goodness-of-fit of these three parameters *N*, *q* and *s* are measured using Matlab curve fitting toolbox in [19] for each of the different quality factors 50, 60, 70, 75, 80, 90 and 100. The goodness-of-fit values for *N*, *q* and *s* are used in (12) to estimate the expected value by it. For example, for the compression quality factor of 75 the values are fitted to *N*=1.396, *q*=1.731 and *s*= -0.3549.

B. Stego Detection by gBL

In the experiment, the expected values of all the digits, 1 to 9, are computed by eq. (12) in Matlab. The average appearing rate of each digit as a first significant one in the cover and stego image are also computed separately. Say, this is μ_i^C and μ_i^S . The percentage of differences between expected values and average appearing values for all the digits *i* are measured both for cover and stego image separately by computing $d_i^C = \frac{(p_i - \mu_i^C)}{\mu_i^C} 100\%$ and $d_i^S = \frac{(p_i - \mu_i^S)}{\mu_i^S} 100\%$. Then, the difference d_i between d_i^C and d_i^S, i.e. $d_i = d_i^C - d_i^S$, should be nominal if the changes in the stego image due to data embedment are not happen in a large scale. A threshold T_i is used to classify an image as a cover or stego one by the following eq. (13).

$$ImClass = \begin{cases} cover \ if & d_i < T_i, \quad i = 1,2,\ldots\ldots,9. \\ stego & otherwise \end{cases} (13)$$

For the different quality factors, the minimum value of T_2, i.e. for the digit 2, is listed in [19]. For the compression factor of 75%, it is defined to 3 for T_2.

Table 2: Rate of distortion detection by *gBL* in Ong *et al*.'s scheme and proposed scheme.

Image Database	Scheme	Detection rate (%) by *gBL*			
		Block Size			
		3x3	5x5	8x8	12x12
BOSS	Ong *et al.*	72	67	72	76
	Proposed	71	71	68	69
CalTech 101	Ong *et al*.	86	77	94	97
	Proposed	86	81	95	98
Authors collection	Ong *et al*.	88	81	99	97
	Proposed	89	77	96	97

C. Analysis on the Results Detected by gBL

The detection rate by *gBL* for block size of 3x3, 5x5, 8x8 and 12x12 are tabulated in the Table 2. The detection rate between the schemes varies by (1.4%, 0%, 1.1%) for block size 3x3, (6%, 5.2%, 4.9%) for block size 5x5, (5.6%, 1.1%, 3%) for block size 8x8 and (9.2%, 1%, 0%) for block size 12x12 in BOSS, CalTech101 and authors collected images respectively. The variations are nominal and none is greater than 10%. The detection rates are ranged from 67% to 99%. This implies that images are distorted enough. The detection rate increases in CalTech101 images w.r.t. BOSS dataset and in authors collected images w.r.t. CalTech101 dataset because last two dataset contains more complex scene and randomness of pixel values is more w.r.t. BOSS images. Hence, shifted pixels in the stego image has changed the frequencies of digits as a first significant digit in a large scale which then guided *gBL* to be detected more successfully. Finally, it is concluded here that the stego images are distorted adequately.

VI. Conclusion

The proposed scheme enhances the embedding payload up to 2 times and it also distorts the stego image notably, thus stego detection rate is higher. Hence, in the field of electronic healthcare, forensic, similar applications and agencies like military and law-enforcing where both the hided message and the cover image are equally secrets while the volume of secret message is large, that proposed scheme will satisfy all the demands successfully. Therefore, in the field of distortion based reversible data hiding scheme, we believe that it will be marked as a notable contribution in the research arena. In our next work, a new scheme applying multi-cycles in the prediction policies, i.e. predicting again on the prediction errors, will be employed so that the values of ranges can be forced to be valued within a limit. If, it is possible, the capacity and the distortion rate will be increased more

Appendix

A. Data Embedment Procedure

Here $Min_c = 206$. Let *m*=101. Then, *d*=5. The process of hiding *m* into cover clock B^k of (a) is shown in steps (b)-(f).

225	222	220
217	215	211
212	210	206

(a) Cover block B^k

225	222	220
217	215	211
212	210	206

(b) Predicted block pB^k

6	2	4
0	0	-2
-1	-1	-4

(c) *pE* by eq. (1)

6	2	4
0	0	2
1	1	4

(d) Stego pE^4 where R_P=8

265	262	260
257	255	251
252	250	246

(e) from eq. (4)

9	6	4
1	255	251
252	250	246

(f) from eq. (5)

9	6	4
1	255	251
252	250	246

(g) Stego block

265	262	260
257	255	251
252	250	246

(h) by eq. (7)

259	260	256
257	255	253
253	251	250

(i) Predicted stego values pS^k

6	2	4
0	0	2
1	1	4

(j) Stego prediction errors pE^4 where R_P=8

225	222	220
217	215	211
212	210	206

(k) Cover block by eq.

B. Data Extraction and Cover Reconstruction Procedure

Data extraction and recover of cover block from stego block in (g) is shown in (h)-(k). Here, $R_s = 255$ where R_s>128 , Max_s=255 and $O^k_{overflow} = 1$. From (j) it is also computed that R_P=8, P=32, b=3. In this scenario $dS = min(S^k_{(i,j)}) - Min_c$ = 246-206=40. By eq.(10) $d = 2^{\left\lfloor \log_2^{dS/R_P} \right\rfloor}$. Thus, m=101.

ACKNOWLEDGMENT

The research work is done by the support of ICT division of the ministry of Post, Telecommunication and Information Technology of Government of Bangladesh through their Information and Communication Technology Fellowship program. Hence, the authors are happy to acknowledge that remarkable support.

REFERENCES

[1]. Ong, SimYing, Koksheik Wong, and Kiyoshi Tanaka., (2014), A Scalable Reversible Data Embedding Method with progressive quality degradation functionality., *Signal Processing: Image Communication,* 29 (1), 135-149.

[2]. X. Liao, C. Shu, Reversible, (2015), Data Hiding in Encrypted Images Based on Absolute Mean Difference of Multiple Neighboring Pixels, J. Vis. Commun. Image R. doi: http://dx.doi.org/10.1016/j.jvcir. 2014.12.007

[3]. Zhang, Xinpeng, et al., "Efficient reversible data hiding in encrypted images.", *Journal of Visual Communication and Image Representation,* vol. 25, no. 2, pp. 322-328, 2014.

[4]. Liao, Xin, and Changwen Shu., (2015), "Reversible Data Hiding in Encrypted Images Based on Absolute Mean Difference of Multiple Neighboring Pixels.", *Journal of Visual Communication and Image Representation,* vol. 28, pp. 21-27, 2015.

[5]. Hong, Wien, (2012), "Adaptive reversible data hiding method based on error energy control and histogram shifting.", *Optics Communications,* 285.2, 101-108.

[6]. Leung, Hon Yin, et al., (2013), "Adaptive reversible data hiding based on block median preservation and modification of prediction errors.", *Journal of Systems and Software,* 86.8, 2204-2219.

[7]. Fu, De-Sheng, et al., (2014), "Reversible data hiding based on prediction-error histogram shifting and EMD mechanism.", *AEU-International Journal of Electronics and Communications,* 68.10, 933-943.

[8]. Ou, Bo, et al., (2013), "Reversible data hiding based on PDE predictor.", *Journal of Systems and Software,* 86.10, 2700-2709

[9]. Yang, Wei-Jen, et al, (2013), "Efficient reversible data hiding algorithm based on gradient-based edge direction prediction.", *Journal of Systems and Software,* 86.2, 567-580

[10]. Zhenfei Zhao, Hao Luo, Zhe-Ming Lu, Jeng-Shyang Pan, (2011), Reversible data hiding based on multilevel histogram modification and sequential recovery., Int. J. Electron. Commun. (AEÜ), 65, 814– 826

[11]. Hong, Wien, and Tung-Shou Chen, (2010), "A local variance-controlled reversible data hiding method using prediction and histogram-shifting.", *Journal of Systems and Software,* vol. 83, no. 12, pp. 2653-2663.

[12]. Kamal, A. H. M., and M. Mahfuzul Islam, (2014), "Facilitating and securing offline e-medicine service through image steganography.", *Healthcare Technology Letters,* 1.2, 74-79

[13]. Lee, Chin-Feng, and Hsing-Ling Chen, (2010), "A novel data hiding scheme based on modulus function.", *Journal of Systems and Software,* vol. 83, no. 5, pp. 832-843.

[14]. Hong, Wien, Tung-Shou Chen, and Chih-Wei Luo, (2012), "Data embedding using pixel value differencing and diamond encoding with multiple-base notational system.", *Journal of Systems and Software,* vol. 85, no. 5, pp. 1166-1175

[15]. Hong, Wien, and Tung-Shou Chen, (2012), "A novel data embedding method using adaptive pixel pair matching.", *Information Forensics and Security, IEEE Transactions on,* vol. 7, no. 1, pp. 176-184.

[16]. Seyyedi, Seyyed Amin, and Nick Ivanov. "High Payload and Secure Steganography method Based on Block Partitioning and Integer Wavelet Transform." *International Journal of Security and Its Applications, under review* (2014).

[17]. Ong, SimYing, KokSheik Wong, and Kiyoshi Tanaka. "Scrambling–embedding for JPEG compressed image." *Signal Processing* 109 (2015): 38-53.

[18]. Karim, Mustafa S. Abdul, and KokSheik Wong. "Data embedding in random domain." *Signal Processing* 108 (2015): 56-68.

[19]. Andriotis, Panagiotis, George Oikonomou, and Theo Tryfonas., (2013), "JPEG steganography detection with Benford's Law.", *Digital Investigation,* 9(3), 246-257.

Enhancing Security in Specialized Use of Mobile IP

Iftakhar Ahmad*, Kazi Sinthia Kabir*, Tanzila Choudhury†, and A.B.M. Alim Al Islam*
*Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.
†Department of CS, Prairie View A&M University, Texas, USA
Email: { ifty58, sinthia.096, choudhury.trina}@gmail.com, alim_razi@cse.buet.ac.bd

***Abstract*—Several security threats are present in the satellite communication system now-a-days. Specially, after incorporating Route Optimization between end hosts by informing correspondent node and home agent about mobile node's current location through binding updates in Mobile IPv6, security threats have become alarming. Vulnerability of the binding updates under several attacks is the main reason of these security threats. The attacks include sending unauthorized and unnecessary data to mobile node, redirection of traffic, denial-of-service attack, man-in-the middle attack, bombing attack, etc. The effect of these attacks is severe on various sophisticated mobile nodes such as satellite, air-crafts, naval ships, drones, etc. Several approaches have been used to prevent these attacks. However, any suitable solution to all these attacks is yet to be found. For these reason, we propose a new solution to all these attacks through incorporating a notion of run-time authentication between the sophisticated mobile nodes. Our proposed solution ensures that no unauthorized agent can manipulate and send modified data packets to the mobile nodes. Consequently, our system of run-time authentication can be used in any highly-secured network system to ensure authentic data communication.**

I. Introduction

Satellite communication is one of the most important ways of communication at present. It has vital applications in telephony, weather forecasting, satellite television, in-flight Internet, navigation (GPS), military communications, etc. Such applications involve communication among various mobile nodes such as aircrafts, naval ships, drones, etc. Even a satellite may act as a mobile node as it continuously moves in an orbit. Fig. 1 shows a generalised diagram of communication among such mobile nodes (MN). Here, a correspondent node (CN) sends data packets to the mobile nodes through satellite.

As an attempt to modernize satellite communication, mobility protocols have been incorporated with it [1]. Mobile IP is an example of such a mobility protocol, which uses home agent for mobility management. The protocol requires signaling among the mobility agents, mobile node, and the correspondent node for its operation. Previously, Mobile IP had no Route Optimization between the end hosts and all the traffic were passed through the mobility agents [1]. However, recent mobility protocols, such as Mobile IPv6 has incorporated the notion of Route Optimization [2] between the end hosts through informing the correspondent node and the home agent about the mobile node's current location using the binding updates. After introducing the Route Optimization, satellite communication has become more vulnerable because the binding updates can be open to the unauthorized people. It is needless to say that satellite communication systems need to be very secured and authentic. Therefore, to enhance the security of satellite communication, vulnerability of the binding updates should be removed.

To address this issue, we propose a new solution that limits the phenomena of modification, manipulation, or sending data packets by unauthorized users to mobile nodes. In our solution, we exploit the notion of quantum computing. Additionally, we propose a run-time authentication mechanism in the solution. The synergy between quantum computing and run-time authentication enables inhibiting active intruders to a great extent, which we validate through experimental evaluation.

Based on our work, we make the following contributions:

- To make the satellite communication systems that use Mobile IP more secured, we propose a method of run-time authentication. Our proposed system limits unauthorized access in an ongoing communication among mobile nodes to a great extent. Besides, the proposed system facilitates preventing any active intruder from intervening the communication between mobile nodes and sending modified data packets to them.
- We experimentally evaluate the performance of the proposed run-time authentication system. Here, we perform discrete-event simulation in `ns-2` to analyze various aspects of the performance.
- Finally, we point out advantages and limitations of our proposed system. Besides, we present real-life implementation issues and our future work.

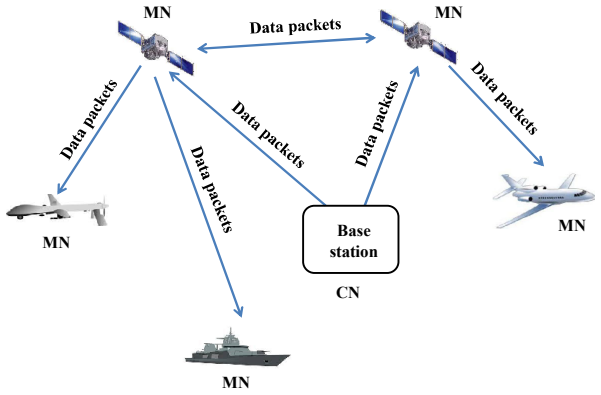


Fig. 1: General communication system among mobile nodes

978-1-5090-0203-0/16/$31.00 © 2016 IEEE

II. Background

Satellite communication is indispensable in today's advanced era of communication. From the inception of satellite communication, new and modern technologies have been being incorporated with it. In conjunction with this, lately, latest version of Mobile IP, i.e. Mobile IPv6 has been incorporated with satellite communication. The Mobile IPv6 includes Route Optimization [2] between the end hosts. This Route Optimization is performed using the binding updates. The correspondent node and the home agent are informed about the mobile node's current location through the binding updates. In earlier mobility protocol, whenever the location of a mobile node changes i.e. the mobile node comes under the coverage of a foreign network from the home network or from another foreign network, it obtains a care-of address from the foreign agent of the current foreign network. This care-of address along with the address of the home agent, form the binding update. This binding update is kept stored in the server of the home agent.

Later on, whenever a data packet is sent to the mobile node from any correspondent node, initially the packet is sent to the home agent of the mobile node. The home agent encapsulates the received data packet with the home agent's address and the care-of address of the mobile node and then sends the encapsulated packet to the foreign agent. The foreign agent then delivers the packet to the mobile node after decapsulation. If the mobile node wants to send data packets to the correspondent node, it can directly send the data packets to the correspondent node through the foreign agent. Here, all the data packets from the correspondent node follow an un-optimized route (through the home agent) to the mobile node. This leads to longer routing path as well as degraded performance.

To overcome this performance penalty, Mobile IPv6 allows a mode of operation in which the mobile node and the correspondent node can exchange data packets directly, bypassing the HA completely after the initial setup phase. This is called Route Optimization. In this process, mobile node sends the binding update (BU) to the correspondent node informing the newly acquired care-of address along with its home address. The correspondent node caches the binding of the mobile node's home address with the care-of address. Later on it can send any packets destined for the mobile node directly to it at this care-of address [2]. The process of obtaining care-of address and performing Route Optimization by exchanging binding updates is shown in the sequence diagram of the Fig. 2.

After introduction of Route Optimization, satellite communication has become vulnerable to several attacks. The reason behind this vulnerability is that, the binding updates can be easily accessed by unauthorized people. When an unauthorized user obtains the binding update of a mobile node (that can be either satellite or any mobile host), he/she can easily misuse it. For example, if the binding updates are not authenticated, then the attacker can use spoofed binding update (BU), thereby, misinforming correspondent node about the mobile hosts current location. The process of acquiring fake or spoofed binding update from mobile node or corresponding node by an attacker and then sending unauthorized or unnecessary or modified data packets to the them is shown by a sequence diagram in Fig. 3. This process

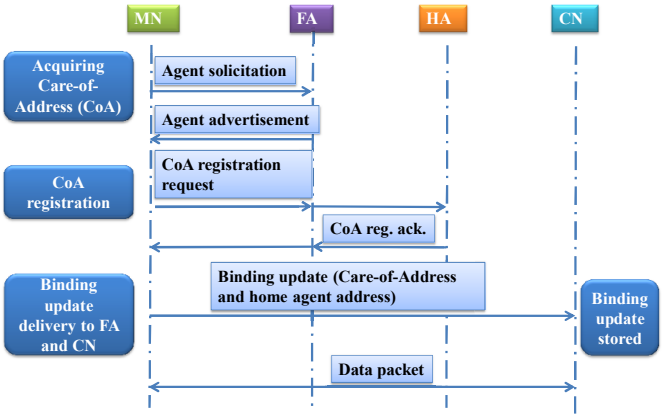


Fig. 2: Sequence diagram of the existing system

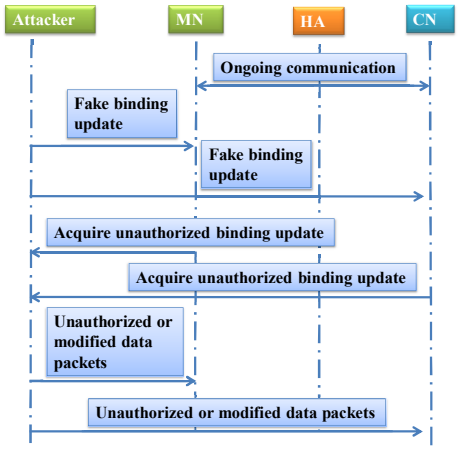


Fig. 3: Sequence diagram of how an attacker can send modified data to mobile nodes in the existing system

of acquiring false binding update may lead to Traffic Redirect attack, Man-in-the-middle attacks, Bombing attack, home agent poisoning, etc. [3] and thus compromising the secrecy and integrity of data packets. For all the above mentioned attacks, unauthenticated use of binding updates is diagnosed as the main reason.

In the above mentioned Traffic Redirect attack, the attacker may send a fake binding update message to the correspondent node claiming that a mobile node has changed its care-of address due to its movement to a new location. Afterwards, the correspondent node will start sending packets to the new care-of address and the victim mobile node will not get any more traffic [3]. Similarly, in case of Man-in-the-middle attacks the attacker sends spoofed binding update message to the correspondent node and the correspondent node updates the cache entry of the victim mobile node. Therefore, the correspondent node will start sending the packets to the attacker instead of the actual mobile node and the attacker

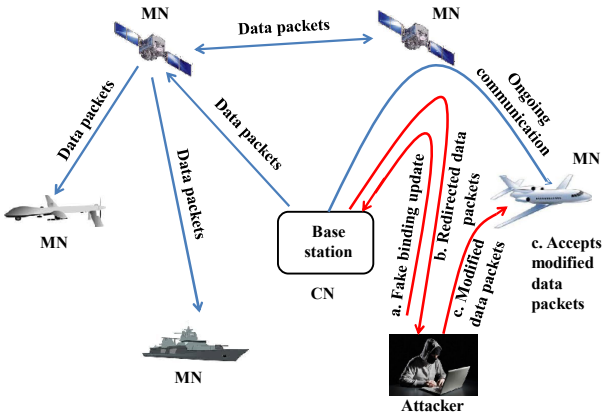


Fig. 4: Process of sending modified data by an attacker in the existing system

may learn the confidential information of the message. Also the attacker may modify the packet before forwarding it to the victim mobile node. Thus, the attacker might act as a man-in-the-middle, getting all the important private data destined to the victim mobile node, without the knowledge of the concerned parties. Moreover, the attacker can send modified control and command messages to the satellite and alter its operation sequence [3]. This types attacks may have catastrophic impacts on various delicate communication systems involving the satellite. Fig. 4 demonstrates the process of sending modified data packets to sophisticated mobile nodes by an attacker in the Man-in-the-middle attack.

In the Bombing attack, mentioned earlier, huge amount of unsolicited data traffic are sent to the victim mobile node. This results in wastage of the bandwidth as well as degradation of performance [3]. For example, the attacker may establish a connection with a streaming server and starts to download a stream of data. After getting the sequence number, the attacker may claim that its location has changed and send a fake binding update. In this binding update the attacker will use the IP address of the victim mobile node. Consequently, subsequent data packets from the streaming server will be directed to that victim node. Thus, the victim mobile node will start receiving unwanted and unnecessary data packets. However, in such attacks, the victim node will not accept those unsolicited packets and will not send the acknowledgement. Consequently, the communication will stop. Unfortunately, the attacker can spoof acknowledgement packets towards the server as it knows the initial sequence number, thereby making a continuous flow of data streams sent to the victim and making the scenario worse. Therefore, this type of attacks may exhibit devastating effects on the resources of the mobile node resulting in denial-of-service.

In another earlier mentioned attack, namely home agent poisoning, attacker spoofs the binding update stored in the server of the home agent. If this binding is accepted by the home agent, all the subsequent communication with the mobile node will be interrupted. Furthermore, there will not remain any scope to communicate with the victim mobile node as any query or data packet to it will be sent to the attacker [3].

III. Related Work

In the existing literature, several approaches have been attempted to prevent the attacks on mobile node or home agent. Goal of the existing protection mechanisms is to mitigate or prevent the possible attacks. Such mechanisms should be computationally less expensive so that they can be implemented in mobile nodes with low processing power. Moreover, the mechanisms need to have lower latency so that the existing communication is not hampered.

One of the existing approach to reduce the number of attackers is Return Routability protocol [3]. This protocol is used before each binding update is sent to correspondent node and they are exchanged among the mobile nodes, home agent, and correspondent node. Here, the home agent receives a message from the mobile node and forwards it to the correspondent node. It also receives a message from correspondent node and sends it back to the mobile node. Similar message is exchanged between mobile node and correspondent node. This protocol requires less CPU processing power and it can reduce the number of attacks.

However, some security vulnerabilities exist in this protocol on the path between the home agent and the correspondent node. The reason for such vulnerabilities is that the correspondent node can be any node in the Internet and no prior relationship or security association exists between these nodes. The attackers may access the data packets transferred between the two nodes and can learn the secret that is necessary for spoofing the binding update. Therefore, this protocol is a relatively weak routing-based authentication method and does not protect against all possible attacks.

Another existing approach to reduce the vulnerabilities of the communication between satellite and mobile node is to use IPSec protocols such as Authentication Header (AH) protocol [4] and Encapsulating Security Payload (ESP) [5] protocol. The AH protocol ensures that the binding update is originated from the actual mobile node, not from any unauthorized agent. In this protocol, a pre-configured IPSec security association is done between the mobile node and the home agent or between the mobile node and the correspondent node. This procedure helps to authenticate the exchange of binding update. The security associations can be established through Internet Key Exchange (IKE) [6] with certificate authentication. However, the use of AH protocol does not ensure the data integrity or privacy of the contents. Therefore, ESP protocol [5] can be used since it provides confidentiality, data origin authentication, connectionless integrity, anti-replay service, etc., through data encryption and its own authentication scheme. Also, it can be used in conjunction with the AH protocol.

However, the use of these IPSec protocols cannot solve the location verification problem [3]. As a result, using only the IPSec protocol to secure binding updates between a mobile node and its correspondent node is not enough to secure the data traffic among mobile nodes [3].

There are some other approaches to secure the data traffic between the mobile nodes such as IKE based schemes [3], use of cryptographically generated address [3], stateless approach [3], certificate based approach, [3] etc. Although some of

these approaches are promising and can provide security to the data traffic, however, these have several limitations. Some of these protocols demand very complex and power-consuming operations that may not be suitable for mobile nodes. Additionally, in case of some of the protocols, CPU and memory requirement is pretty high. As a result, it becomes difficult to implement such protocols for mobile nodes.

We will use the notion of quantum entanglement [7] for run-time authentication. There has been some approaches in the literature for authentication in network systems using quantum entanglement. One approach is quantum authentication protocol using entangled states [8]. In this protocol the sender and the receiver share a sequence of EPR(Einstein-Podolsky-Rosen) pairs [9] as the authentication key. To authenticate each other, one creates auxiliary qubits and make them interact with the authentication key. After measurement in the selected basis, he or she affirm the others identity. Here, no one without the authentication key can pass the authentication process successfully [8]. However, this is protocol does not provide run-time authentication and an intruder can eavesdrop the connection and destroy the existing connection.

Another approach in existing literature is Quantum identification system [10] where the BB84 QKD is used to share an identification sequence (IS) as common secret information. After sender and receiver share these secret codes, they use a classical channel. First, the sender sends the first IS to the receiver and the receiver verifies it. Second, the receiver sends the second IS to the sender for verification. Finally, the sender repeats the first step and the receiver verifies that the sender is authentic [10]. However, this protocol requires an additional authentication because the BB84 needs an authentication before the communication starts.

In the protocol proposed by Kuhn [11] a trusted server shares a secret key with the sender and the receiver. Authentication between each party and the server is made by a classical authentication protocol. Therefore, a trusted third-party is required here.

There are some other authentication protocols that use quatum entanglement [7]. However, these protocols do not provide run-time authentication. Moreover, many authentication protocols of existing literature require a trusted third-party to exchange authentication key between the sender and the receiver. Therefore, we are going to propose a run-time authentication system that do not require a trusted third-party and provides a mechanism to authenticate each data packet of a classical network.

IV. Basics of Quantum Computing

Quantum computing is a computation system that makes direct use of quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. A Quantum computer uses qubits, rather than representing information as 0s or 1s as classical computers do. A Qubit or quantum bit is an analogous concept of bit. A classical bit has a state of either 0 or 1. However, a Qubit can be a 1 or a 0 or both at the same time. This is called quantum superposition of states which can be written as

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

where, α and β are probability amplitudes and can, in general, both be complex numbers. The states $|0\rangle$ and $|1\rangle$ are called computational basis states, which form an orthonormal basis for computation in a vector space [12].

Quantum teleportation is a significant feature of quantum computation. Quantum teleportation relocates the state of a material rather than relocating the material itself. Nevertheless, given the fact that all elementary particles are identical, quantum teleportation of one electron to another one has the same effect as relocating the electron itself [13].

Quantum entanglement is another important feature of quantum computing. It is a quantum mechanical phenomenon that occurs when pairs or groups of particles are generated or interact in ways such that the quantum state of each particle cannot be described independently. Rather, their states have to be described with reference to each other, even though the individual objects may be spatially separated, at different places. The interaction ensures that the particles are properly described by the same quantum mechanical description or state. The state corresponds to a number of important factors such as position, momentum, spin polarization, etc. When a change occurs in the state of one of the entangled particles, similar change is observed in the other particles [7].

With the help of the notion of quantum entanglement, quantum networks are built. Quantum physical effects such as entanglement, can be used to detect eavesdropping, to improve the shared sensitivity of separated astronomical instruments, or to create distributed states that will enable numerical quantum computation over a distance using teleportation. Quantum networks, like classical networks, involves nodes and links, and a layered communication architecture with the individual protocols communicating vertically up and down a protocol stack and horizontally with peers [14].

V. Working Methodology of Our Proposed System

The communication among various sophisticated mobile nodes such as aeroplanes, naval ships, drones, etc., needs to be highly-secured. However, the use of binding update exchanges between such mobile nodes makes data traffic between them vulnerable to attackers. Moreover, state-of-the-art approaches do not offer any highly-secured protection mechanism from the attackers. Therefore, we propose a new approach in this section with a view to ensure highly-secured data communication between sophisticated mobile nodes. We elaborate our proposed approach below.

In our proposed system, there will be a run-time authentication between the correspondent node and mobile node. Here, similar to the existing system, the correspondent node acquires the binding information of the mobile node through home agent of the mobile node. After completion of the initial setup, if the location of the mobile node changes, the correspondent node and mobile node exchanges a binding update. During the initial setup, a special kind of highly-secured connection sets up between the correspondent node and the mobile node. run-time authentication takes place using this special connection. Whenever the correspondent node sends some data packets to the mobile node, run-time authentication will take place to check whether the data is coming from the legitimate source i.e., the intended corresponding node. As a result of this run-time authentication, no unauthentic data can be sent to the mobile node by an intruder. For example, the correspondent node, in a highly-secured network system, wants to send data to a mobile node. Now, if there is no protection mechanism between them then an intruder may

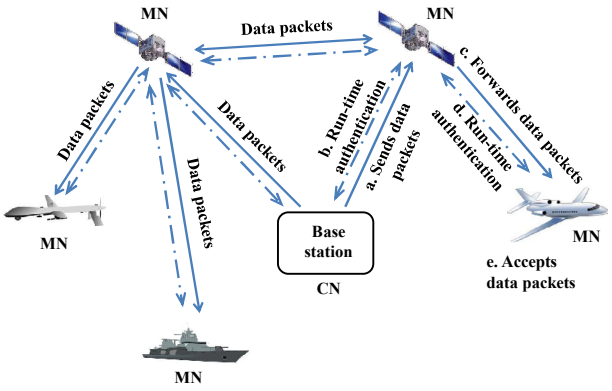


Fig. 5: Operations of run-time authentication in network scenario

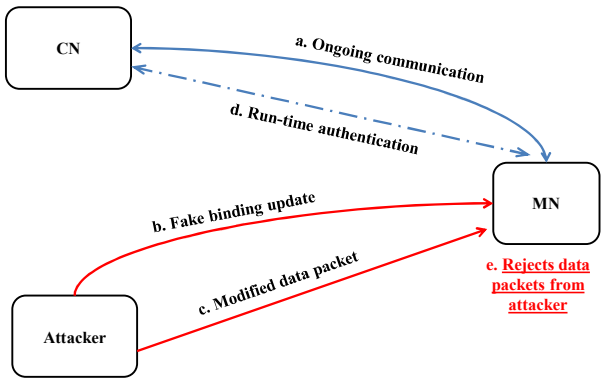


Fig. 6: Failed attempts by an active intruder in our proposed system

interven their communication, read the data packets that are being transmitted through the network, alter the contents of these data packets, and finally send them to the mobile node. Thus, the mobile node will recieve the modified data packets, without realising that these data packets came from an intruder, not from the actual sender. However, according to our proposed system, the two communicating nodes, first perform run-time authentication then start transmitting data. If the authentication is successful, only then the incoming data packets will be accepted by the receiving mobile node. It is very evident that, the core of our proposed system is the special connection between the mobile node and the correspondent node by which run-time authentication is performed. Therefore, it is needless to mention that this special connection must be very secured and it should have the capability to check if an intruder has intervened the connection. Fig. 5 shows a generalized diagram of communication among various sophisticated mobile nodes with run-time authentication.

Now, after incorporating the run-time authentication with the existing system, if an attacker sends spoofed or fake binding update to a mobile node and then attempts to send modified data packets to it, with a view to intervene the ongoing communication between mobile node and correspondent node, the mobile node first attempts to perform run-time authentication with the attacker in disguise of a correspondent node. The run-time authentication eventually fails since the modified data packets came from the attacker, not from the correspondent node. Therefore, the mobile node eventually rejects the incoming data packets. Consequently, the attacker, which is an active intruder, fails to send modified data packets to a sophisticated mobile node. Fig. 6 demonstrates the whole scenario, where the attacker's attempt for active participation is rejected.

VI. Applicability of Our Proposed System

The necessity for highly-secured and authentic communication among various sophisticated mobile nodes is of utmost importance. To reduce or eliminate the vulnerability of the binding update, exchanges between such mobile nodes, we propose to establish a run-time authentication mechanism between the correspondent node and the mobile node using the notion of quantum entanglement. Because with the help of quantum entanglement [7] secured communication is possible and any sort of eavesdropping can be detected easily. Therefore, in our proposed system we use the quantum entanglement to exchange authentication keys between the sender and the receiver. Moreover, our proposed system of run-time authentication can be used in any highly-secured network systems to prevent the active intruders. We elaborate the mechanism of the proposed run-time authentication system below.

Our proposed run-time authentication system augments the classical network with the notion of quantum entanglement [14]. In this system, each of the correspondent node and the mobile node shares some pairs of entangled particles. These pairs of particles enable the run-time authentication. Whenever a correspondent node wants to start transmission of data packets to the mobile node, it first sends an intent through a classical data packet to the receiving mobile node. After receiving the intent the mobile node produces a random authentication code and sends it along with a packet ID to the sender or the correspondent node through the pairs of entangled qubits [14]. Here, the receiver uses Superdense Coding scheme [15] to encode the authentication code and packet ID pair to the qubit. Recently, the Superdense Coding scheme has been experimentally demonstrated [16] which gives experimental evidence of the Superdense Coding scheme. The mobile node also stores the authentication code and packet ID as a pair in a queue. The sender optically reads [17] the code and packet ID from the entangled qubit pairs. After getting the authentication code and packet ID, the sender or the correspondent node combines this code and packet ID with the classical data packet and finally transmits the data packet through the classical channel to the receiving mobile node. When the mobile node receives the data packet, it compares the authentication code and the packet ID from the queue. If the packet ID and the authentication code match

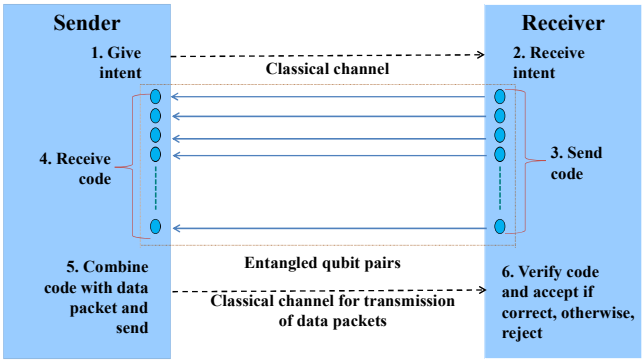


Fig. 7: Steps of operations of the proposed run-time authentication system

with the stored packet ID and authentication code pair, it accepts the data packet. Otherwise, it rejects the data packet. Fig. 7 depicts these steps of operation of our proposed system.

This process goes on until all the data packets are transmitted. Multiple sets of qubit pairs will be used to send many authentication code and packet ID pairs concurrently. When we measure the qubits optically, entanglement will be destroyed. However, as we are measuring the entanglement optically [17], we can re-establish the entanglement in the qubit pairs optically. The protocol that will be used can prepare entanglement between two remote particles [18]. The qubit pairs will be entangled again immediately after measurement at the senders side and will be ready for exchanging authentication code and packet ID pair for subsequent data packets. Thus, by exchanging and comparing the authentication code and packet ID pairs run-time authentication will be implemented. As the intruder cannot get the authentication code, data packets coming from it can never be accepted by the receiving mobile node. The intruder may manage to access the data packets through traffic redirection attack [3] and copies the authentication code and packet ID and try to send modified data packets to the receiver. To prevent the intruder from doing so, the receiver will not accept the data packets after a certain time interval, even if the authentication code and packet ID pair matches with a pair stored in the queue. Therefore, implementing our proposed run-time authentication system, it can be ensured that no active intruder can send modified or malicious data packets to the receiver. The simulation results and findings of our proposed system is presented in the following section.

VII. Experimental Evaluation

We have performed discrete-event simulation of our proposed run-time authentication system. Simulator for quantum networking is not available yet. Therefore, we have performed our simulation with the widely-adopted network simulator `ns-2`. We have performed simulations for both medium-speed mobile nodes and high-speed mobile nodes. In this section, we discuss the outcomes of our simulation.

A. Simulation Settings

We have adopted a network with two nodes both simultaneously acting as TCP sources and sinks. We have also deployed a malicious node or intruder which is a TCP source and attempts to send data packets to former two nodes. All these nodes are mobile nodes. The network topology is shown in Fig. 8. We have set the window size to 8000 packets. Besides we have adopted wireless channel and Two-Ray Ground Propagation Model [19] in our simulation.Additionally, we have used the physical layer and MAC layer 802.11. Nonetheless, we have adopted DSDV as the routing protocol. We have set the initial packet size to 452 kB and simulation time to 100 seconds. For the medium-speed mobile nodes, we have set the speed of the mobile nodes to 110 kmph and for the high-speed mobile nodes we have set the speed of the mobile nodes to 11000 kmph.

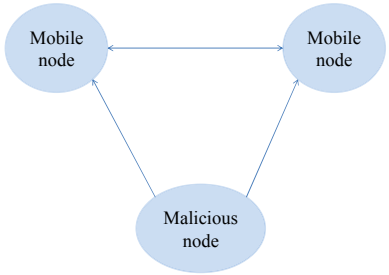


Fig. 8: Network topology used in simulation

B. Simulator Modifications

In a quantum network using quantum teleportation, we consider that there will be no loss in the channel. However, in reality, some losses may occur in different phases of the transmission system such as at the sender device, transmission medium, or the receiver device. In the transmission medium, diffraction may occur and the beam of photon generated at the sender's side and transmitted to receiver side becomes a spread beam [20]. This diffraction has adverse effect on the data represented by the photon [21]. Therefore, to incorporate this loss we have modified the physical layer of `ns-2`. Besides, a quantum network experiences line attenuation in the quantum channel [20]. We have also incorporated this loss in our simulation.

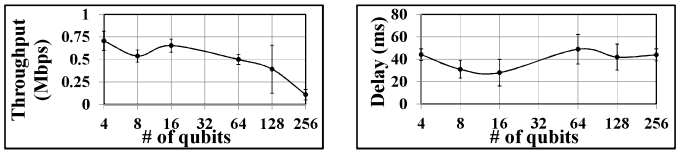
C. Simulation Results for Medium-Speed Mobile Nodes

We have performed the simulation by varying several factors such as number of qubits, size of queue used for storing authentication code and packet ID pair, threshold on

delay, etc., for the medium-speed mobile nodes. For each parameter settings, we calculated the results by averaging the values found in five iterations. The standard deviation is also shown in each graph using error bars.

Fig. 9 shows simulation results for various number of qubits. Firstly, we have calculated the per-node throughput for various number of qubits keeping other factors constant. The number of qubits was varied from 4 to 256. Here, the size of authentication code was double of the number of qubits. This happens as Superdense Coding [15] has been adopted to encode the classical bits into the qubits in the proposed system. Fig. 9a shows the change in the per-node throughput in response to a variation in the number of qubits. Here, we find that the per-node throughput exhibits significant values up to qubits to be sixteen. After that, for an increase in the number of qubits, per-node throughput decreases as overhead for the authentication code increases.

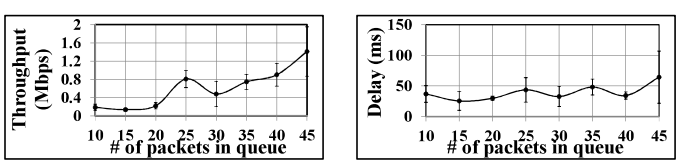
Next we have calculated, the average ene-to-end delay for various number of qubits. Fig. 9b shows the end-to-end delay. This figure suggests that initially the ene-to-end delay is high for smaller number of qubits due to transmission of lots of packets in `ns-2`. However, minimal delay occurs when number of qubits is between eight and sixteen. Then, again end-to-end delay rises for the overhead of the authentication code.



(a) Per-node throughput (b) End-to-end delay

Fig. 9: Network performance for various number of qubits used in authentication code for medium-speed mobile nodes

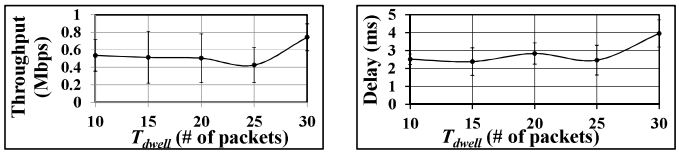
We have also varied size of queue used for storing authentication code and packet ID pair, keeping other factors constant. Fig. 10 shows the network perormance. Here, Fig. 10a shows change in per-node throughput in response to a variation in the size of queue. Here, we find that per-node throughput exhibits an increasing trend in response to an increase in the size of queue. If the size of queue is too small, number of packets accepted by the receiver becomes small and eventually, per-node throughput decreases. Besides, Fig. 10b shows change in the end-to-end delay for various queue sizes. This figure portrays minor variation in delay in response to changing queue size.



(a) Per-node throughput (b) End-to-end delay

Fig. 10: Network performance for various number of packets in queue for medium-speed mobile nodes

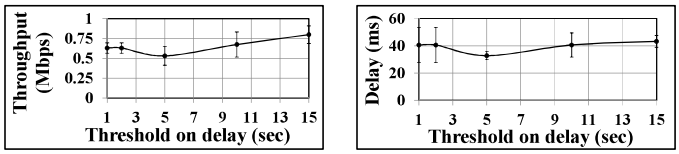
We have also varied the threshold of time interval after which a packet with valid authentication code is discarded. The time interval was measured by a counter which is relative to the number of packet generation in `ns-2` . Fig. 11 shows the simulation results for this case. Fig. 11a and Fig. 11b



(a) Per-node throughput (b) End-to-end delay

Fig. 11: Network performance for various threshold on counter for medium-speed mobile nodes

show change in per-node throughput and end-to-end delay for various threshold on delay which is denoted by T_{dwell}, after which a packet is discarded even though it's authentication code might be correct. Fig. 11a per-node throughput remains almost constant up to the counter value twenty and shows an increasing trend when the counter value is more than twenty five. Additionally, Fig. 11b shows that end-to-end delay shows an increasing trend when the counter value is more than twenty five.



(a) Per-node throughput (b) End-to-end delay

Fig. 12: Network performance for various threshold on delay for medium-speed mobile nodes

Moreover, we varied the threshold on delay, after which a data packet with a valid authentication code is discarded. Thus the proposed system prevents an active intruder from accessing a data packet, modify it and send it to the receiver with the packet's authentication code as these procedure will require some minimal time. We have varied the threshold of dealy from one to fifteen seconds in `ns-2`. Fig. 12 shows the simulation results. Fig. 12a shows that throughput increases when threshold on delay is equal or greater than ten. Because when the threshold is less than ten, many valid data packets, arriving after reasonable delay, are discarded and eventually per-node throughput decreases.

Fig. 12b shows change in end-to-end delay for various values of threshold of the delay. This figure depicts that end-to-end delay also increases when threshold on delay is equal or greater than 10.

In the next subsection we show network performance for high-speed mobile nodes.

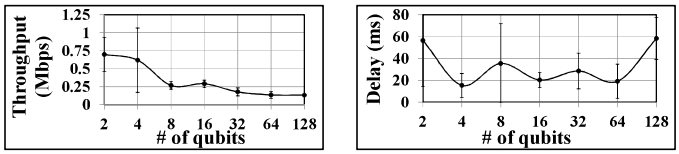
D. Simulation Results for High-Speed Mobile Nodes

We have also performed the discrete-event simulation for high-speed moblie nodes such as satellite. In this case, we

have varied several factors such as number of qubits, size of queue used for storing authentication code and packet ID pair, threshold on delay of incoming packets, etc. Similar to the case of medium-speed mobile nodes, here, for each parameter settings, we have calculated the results by averaging the values we have found in five iterations and the standard deviation is shown in each graph using error bars.

Fig. 13 shows network performance for various number of qubits. Firstly, we calculated the per-node throughput for various number of qubits, keeping other factors constant. Here, the number of qubits is varied from 2 to 128. Fig. 13a shows the change in the per-node throughput in response to a variation in the number of qubits. Here, per-node throughput shows a decreasing trend when the number of qubits used for authentication code is more than four.

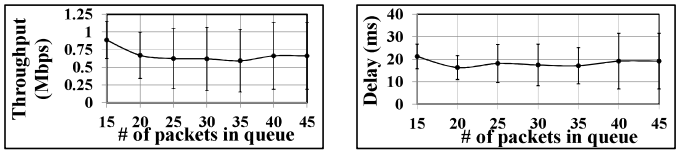
Then, the average ene-to-end delay has been calculated for various number of qubits. Fig. 13b suggests initially the end-to-end delay is high for smaller number of qubits due to transmission of lots of packets in `ns-2`. However, minimal delay occurs when the number of qubits is between four and sixty four.



(a) Per-node throughput (b) End-to-end delay

Fig. 13: Network performance for various number of qubits used in authentication code for high-speed mobile nodes

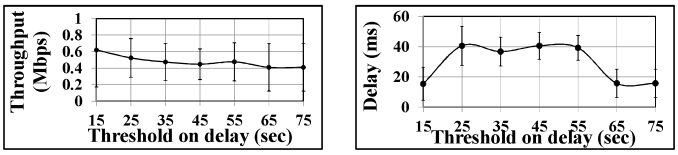
We have also varied size of queue used for storing authentication code and packet ID pair, keeping other factors constant. Fig. 14 shows the network performance. Fig. 14a shows change in per-node throughput with in response to a variation in the number of packets stored in queue. Fig. 14b shows change in the end-to-end delay for the various number of packets stored in queue. Here, both per-node throughput and end-to-end delay do not fluctuate much with the change in the number of packets stored in queue.



(a) Per-node throughput (b) End-to-end delay

Fig. 14: Network performance for the various number of packets stored in queue for high-speed mobile nodes

Moreover, we have varied the threshold on delay, after which a data packet with a valid authentication code and packet ID pair is discarded. Here, the data packets have been discarded to prevent an active intruder from accessing a data packet, modify it and send it to the receiver with the packet's authentication code as, these procedure will require some minimal time. We have varied the threshold of dealy from one to fifteen seconds



(a) Per-node throughput (b) End-to-end delay

Fig. 15: Network performance for various threshold on delay for high-speed mobile nodes

in `ns-2`. Fig. 15 shows the simulation results. Fig. 15a shows the change in per-node throughput and Fig. 15b shows the change in end-to-end delay for various values of threshold on delay.

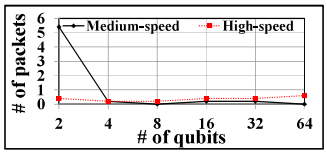


Fig. 16: Number of data packets received from the intruder for both medium-speed and high-speed mobile nodes

Later on, we have calculated the number of data packets received from the intruder for both medium-speed and high-speed mobile nodes. Here, the number of qubits used for the authentication code have been varied from 2 to 64. Fig. 16 shows that, if the number of qubits used for the authentication code is two, a few data packets are received from the intruder for medium-speed mobile nodes. However, when the number of qubits used for the authentication code increases, the amount of data packets received from the intruder becomes negligible. Therefore, our proposed run-time authentication system prevents the intruder from sending unauthentic data packets successfully to mobile nodes and thus it is justified that our propsed system limits unauthorized access in an ongoing communication among mobile nodes to a great extent.

VIII. Advantages and Limitations of Our Proposed System

We have proposed a run-time authentication system for the medium-speed and the high-speed mobile nodes. Advantage of our system is that no active intruder can send data packets to the receiver successfully. Because, the correspondent node accepts data packets with the valid authentication code and packet ID pair. An intruder do not have quantum entanglement with the correspondent node. Therefore, he or she cannot get the authentication code from the receiver or the correspondent node and eventually, the receiver discards the data packets from the intruder. Besides, our prosposed system of run-time authentication can be used in any highly-secured network systems. Moreover, we do not need any trusted third-party [22] to exchange authentication key between the sending mobile node and the receiving mobile node. Our proposed system also do not require to exchange any authentication key before starting the communication.

Limitaions of our proposed system is that, if the number of concurrent intruders becomes very high, it may become

vulnerable. Moreover, quantum network [14] is not widely-adopted at present. However, the equipments to implement the quantum network is expected to be widely availabe in near future. Therefore, quantum network can be widely used in near future as, quantum computing and networking is a fast growing technology field in modern era.

IX. Real Life Implementation and Future work

In future, we will provide the analytical model of our proposed system. Moreover, we will try to implement our proposed system of run-time authentication using real mobile nodes. Furthermore, our proposed system of run-time authentication can be used in any other communication system where highly-secured and authentic communication is required. Our proposed system makes any type of existing communication system more secured along with alleviating the vulnerabilities of binding updates of the current Mobile IP. In future, we will try to present all the real life implementation issues of our proposed system for both mobile nodes and generalized highly-secured network systems more elaborately.

X. Acknowledgment

This work has been conducted at and partially supported by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. Besides, this work has been partially supported by the Ministry of Education, Government of the People's Republic of Bangladesh.

XI. Conclusion

Day-by-day necessity of authentic communication among various sophisticated mobile nodes is increasing. Mobile nodes such as drones, aircrafts, naval ships etc., require highly-authentic communication system. If such sophisticated mobile nodes are attacked and fooled to accept malicious and modified data packets, the consequence may be catastrophic in some cases. Therefore, these sort of mobile nodes should always use immensely-secured and authentic communication systems.
To address this issue we present a new system in this paper. Here, we exploit two key notions, namely quantum computing and run-time authentication, to enhance the security. Our proposed solution of incorporating run-time authentication with the existing classical network will ensure highly-secured communication among the above-mentioned sophisticated mobile nodes. We have evaluated performance of our proposed technique through discrete-event simulation in `ns-2`. Simulation results reveal different aspects of performance of our proposed system. Finally, we have pointed out advantages and limitations of our proposed system along with our future work.

References

[1] H. Tsunoda, K. Ohta, N. Kato, and Y. Nemoto, "Supporting ip/leo satellite networks by handover-independent ip mobility management," *Selected Areas in Communications, IEEE Journal on*, vol. 22, no. 2, pp. 300–307, 2004.

[2] D. Johnson, C. Perkins, J. Arkko, *et al.*, "Mobility support in ipv6," 2004.

[3] M. Atiquzzaman and M. S. Hossain, "Security issues in space networks," 2011. esto.nasa.gov.

[4] S. Kent, "Ip authentication header," 2005.

[5] S. Kent, "Ip encapsulating security payload (esp)," 2005.

[6] C. Kaufman, P. Hoffman, Y. Nir, and P. Eronen, "Internet key exchange protocol version 2 (ikev2)," tech. rep., RFC 5996, September, 2010.

[7] A. Bokulich and G. Jaeger, "Philosophy of quantum information and entanglement," 2010.

[8] X. Li and D. Zhang, "A quantum authentication protocol using entangled states as authentication key.," *WSEAS Transactions on Computers*, vol. 5, no. 5, pp. 830–835, 2006.

[9] C. H. Bennett and S. J. Wiesner, "Communication via one-and two-particle operators on einstein-podolsky-rosen states," *Physical review letters*, vol. 69, no. 20, p. 2881, 1992.

[10] M. Dušek, O. Haderka, M. Hendrych, and R. Myška, "Quantum identification system," *Physical Review A*, vol. 60, no. 1, p. 149, 1999.

[11] D. R. Kuhn, "A hybrid authentication protocol using quantum entanglement and symmetric cryptography," *arXiv preprint quant-ph/0301150*, 2003.

[12] G. P. Berman, *Introduction to quantum computers*. World Scientific, 1998.

[13] H. J. Kimble, "The quantum internet," *Nature*, vol. 453, no. 7198, pp. 1023–1030, 2008.

[14] R. Van Meter, "Quantum networking and internetworking," *Network, IEEE*, vol. 26, pp. 59–64, July 2012.

[15] C. H. Bennett and S. J. Wiesner, "Communication via one-and two-particle operators on einstein-podolsky-rosen states," *Physical review letters*, vol. 69, no. 20, p. 2881, 1992.

[16] B.-H. Liu, X.-M. Hu, Y.-F. Huang, C.-F. Li, G.-C. Guo, A. Karlsson, E.-M. Laine, S. Maniscalco, C. Macchiavello, and J. Piilo, "Experimental demonstration of efficient superdense coding in the presence of non-markovian noise," *arXiv preprint arXiv:1504.07572*, 2015.

[17] C. G. Yale, B. B. Buckley, D. J. Christle, G. Burkard, F. J. Heremans, L. C. Bassett, and D. D. Awschalom, "All-optical control of a solid-state spin using coherent dark states," *Proceedings of the National Academy of Sciences*, vol. 110, no. 19, pp. 7595–7600, 2013.

[18] M. Abdi, P. Tombesi, and D. Vitali, "Entangling two distant non-interacting microwave modes," *Annalen der Physik*, vol. 527, no. 1-2, pp. 139–146, 2015.

[19] J. B. Andersen, T. S. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *Communications Magazine, IEEE*, vol. 33, no. 1, pp. 42–49, 1995.

[20] G. Gilbert and M. Hamrick, "Practical quantum cryptography: A comprehensive analysis (part one)," *arXiv preprint quant-ph/0009027*, 2000.

[21] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press, 2010.

[22] J. M. Kizza, *Guide to computer network security*. Springer, 2009.

Tradeoffs Between Sensing Quality and Energy Efficiency for Context Monitoring Applications

Sujan Sarker, *Student Member, IEEE*, Amit Kumar Nath, *Student Member, IEEE* and
Md. Abdur Razzaque, *Senior Member, IEEE*
Green Networking Research (GNR) Group, Department of Computer Science and Engineering
Faculty of Engineering and Technology, University of Dhaka, Dhaka-1000, Bangladesh.
Email: {sujancse16, amitnath1614}@gmail.com, razzaque@du.ac.bd

***Abstract*—Mobile devices have gone far beyond their basic usage like calling, texting, etc. New generation mobile devices like smartphones contain a rich set of embedded sensors which enable them to perform sensing operations in different domains. To get user contextual information, both human centric & participatory sensing can be helpful. In order to accurately capture, recognize and classify various user states, sensors need to operate more in active mode. However, continuous sensing results in huge energy consumption, decreasing the battery lifetime. Hence, a tradeoff in between the sensing accuracy and energy-efficiency is required. In this paper, we have developed novel strategies to make this tradeoff. User contexts are monitored and multiplicative increase and multiplicative decrease (MIMD) approach is exploited to dynamically adjust the sensing frequency following the sensing quality requirements of the applications. We have developed a real-time android application to evaluate the effectiveness of the proposed method in terms of sensing accuracy and energy-efficiency. The results show a good level of improvement compared to a state-of-the-art work.**

I. Introduction

Mobile devices, especially smart mobile phones have emerged as the main electronic device not only for voice communication but also for running heterogeneous real-time applications. Today's mobile devices have a large number of embedded sensors like GPS, Bluetooth, camera, accelerometer, ambient light sensor, magnetic compass, Wi-Fi, etc. Data readings from these sensors may put tremendous impact in developing many applications, e.g., social networking, health care, transportation safety, smart office, smart home, etc.

Mobile sensing applications are of two categories: *participatory sensing* in which the parameters (how, when, what, where to sense) are determined actively by the users; and, *opportunistic sensing* in which sensing operations are performed without involving the user. In this paper, we have made our observations based on opportunistic sensing, although participatory sensing can also be enabled for the proposed method. Human beings are generally involved in a large variety of diverse context events. As mobile phones have become a constant part of our daily lives, inferring a particular context can be performed using the built-in sensors of the mobile phones. For instance, Jigsaw [1] has been developed for performing real-time monitoring of events that are driven by location, activity and sound. A noise mapping system named Ear-Phone [2] has been developed to create different area based noise maps. Another fine example is www.sensorly.com [3] which is a website offering access to coverage maps for various wireless networks while the information itself is entirely community powered.

978-1-5090-0203-0/16/$31.00

However, there is a dilemma in the extensive use of mobile devices for sensing purposes. Optimal calibration of sampling frequency is a critical issue in context monitoring applications due to over sensing and under sensing. In case of over sensing, redundant samples are taken increasing the energy consumption and calculation overheads. On the other hand, under sensing might result in generating inaccurate results. In cases where user states remain unchanged for a long time, frequent sampling would result in consuming excessive energy. However, frequent sampling would be required if the user is in a dynamic state.

The situation becomes worse when multiple sensors are powered on simultaneously. In such situations battery life would be significantly reduced. Various researches suggest that modern mobile phones aren't capable of supporting all on-board sensors at the same time. For example, Nokia N95, a mobile phone that supports about 10 hours for phone conversation, loses all battery charge within only 6 hours if GPS is on continuously [4]. Therefore, a tradeoff in between the energy consumption and the sensing accuracy is of utmost requirement for the successful penetration of context monitoring applications. Another example can be found from the built-in accelerometer sensor in HTC Touch Pro which is activated with a fixed sampling frequency. While the sensor itself should consume less than 1 mW when data samples are received by the phone; total power consumption of the device is increased by 370 mW in such situations [5]. Most of the recent studies on participatory and opportunistic sensing [3] emphasized on the design and implementation of a particular sensing application. Also, operation specific adjustable duty cycle assignment and using adaptively changing sampling periods have been proposed as a suitable approach [6]. Other approaches like CoMon [7] shares phone sensors among adjacent users thus improving sensing capacity and reducing overall energy consumption. Another approach, ACE [8] introduced techniques for reducing energy consumption by leveraging context patterns occurred in real life.

In this paper, we introduce a context monitoring scheme and multiplicative increase and multiplicative decrease (MIMD) based adaptive sensing frequency scaling so as to achieve higher accuracy even as expending reduced amount of energy from the resource-constrained sensor devices. The main contributions of this paper are summarized below.

- We have developed a framework for dynamic sensing and monitoring applications that makes a tradeoff in between sensing accuracy and the energy-efficiency.
- An intelligent context monitoring scheme has been developed using k-means clustering algorithm that can differentiate between redundant and useful data. Thus, it can decrease processing overhead as well as the energy consumption.
- MIMD based dynamic controlling of the sensing frequency makes the system more adaptive with the application environment and thus it enhances the system performance.
- The results of our performance evaluation study, carried out through implementing android application, depict that the proposed sensing strategy outperforms a state-of-the-art method in terms of sensing accuracy and energy-efficiency.

The remaining of the paper is organized as follows. Section II provides a study on the related works in this area of research. The proposed model and mechanism have been presented in Section III and Section IV gives an insight of our algorithms for efficient sensor management. Section V presents the performance evaluation of the algorithm. Finally, the paper is concluded in Section VI.

II. Related Works

For performing user activity recognition tasks, different embedded sensors on mobile devices like GPS, Bluetooth, accelerometer etc have been explored in details. Several research works have been carried out for recognizing user states accurately while reducing the energy consumption. Most of the existing works provide partial solutions for the energy accuracy tradeoff.

Wang et al. [4] proposed Energy Efficient Mobile Sensing System (EEMSS), which is a sensor management system that improves the battery life of the mobile devices by powering a minimum set of sensors applied with duty cycles. However, this approach uses a highly directional sensor which cannot cover the entire area. Also sensors have fixed duty cycles when they are active and aren't adjustable to different user behavior.

Optimization of sensor duty cycles to minimize user state estimation errors has been studied in [9]. Maintaining an energy consumption budget is also taken into consideration in this study. "SeeMon" is introduced in [10] which provides a hierarchical sensor management system. Energy efficiency and reduced computational complexity is achieved by the system through performing continuous context detection only when changes appear during the context monitoring.

In order to provide a solution for the accuracy vs. energy consumption tradeoff, a dynamic sensor selection scheme is demonstrated in [11]. Another approach to handle this tradeoff is provided in [12], [13], [14] which uses different sampling period schemes. These schemes are used to query sensor data in continuous sensing mode in mobile devices. They used a function based approach which can be dynamically adopted. They changed the sampling interval using some pre-defined advance and back-off functions (linear, quadratic, exponential etc). Depending on the stability of the context, the parameters switched among different functions of sampling intervals. Sensors cannot support sensing at any random sampling frequency, thus this type of methods are not applicable. In [15], a system named "SenseLess" that saves energy by sensing localization applications is described. Constandache et al. [16] suggested that humans can be profiled based on mobility patterns and using this information location can be predicted. "EnLoc", the system that was proposed, achieves localization accuracy with realistic energy budget. Approaches suggesting context monitoring mechanism along with adaptive sampling and duty cycling was presented in [5], [6] that exploited additive increase additive decrease (AIAD) approach.

In this paper, we provide a system model and necessary algorithms to solve the accuracy vs. energy consumption tradeoff. Our concept has some similarities with [5]. While their approach was novel, there were some shortcomings. For example, they introduced a parameter called t_{suf}. This parameter was used to adjust duty cycle and sampling frequency by measuring the stability of the context inference up to the time period denoted by t_{suf}. Due to comparison with min and max values for duty cycles and sampling frequency, convergence becomes slow. In the proposed algorithms, authors adjusted the parameters with reference to t_{suf} while not quantifying the amount of change in user context. As a result exact measurement cannot be ensured. Our approach is distinguishable from other proposed methods by the following key elements: *First*, we provide a system model to illustrate the overall process. *Second*, we introduce an algorithm for controlling data acceptance and rejection in the buffers which are used to store and compare sensed values. Decision making for currently sensed data is carried out according to this algorithm. *Third*, we introduce another algorithm for proper scheduling of sensors. Dynamic frequency calibration and duty cycle assignment operations are executed. *Finally*, we set the required parameters and calculate the efficiency and accuracy of the system.

III. System Model and Assumptions

We consider a smartphone with N number of sensors whose set is denoted by S_N. Let A_n is the set of n sensors required by a context monitoring application A to extract contextual information correctly where $A_n \in S_N$ and $n \leq N$.

The overall system model and it's functional components are illustrated in Fig. 1. n sensors produce discrete raw data which enter into a processing pipeline. Output of this pipeline indicates possible change in user context or not.

The first component of the processing block is a pre-processing structure which filters out necessary information

TABLE I
NOTATION TABLE

Symbol	Description
S_N	Set of all sensors
A_n	Set of sensors for a particular application
F	Set of available frequencies
D	Set of available duty cycles
B_i	Set of buffered values of sensor i
X_j	j^{th} data tuple sensed by the sensor
T_{in}	Sensor initialization time
T_{adj}	Sampling frequency and duty cycle adjustment time
T_c	Active cycle time
T_{ter}	Sensor termination time
T_{run}	Active running time
T_s	Sensing interval
t_s	Sampling period
n_i	Required samples between i^{th} and $(i-1)^{th}$ samplings
ψX	Max. allowable diff. betn. two consecutive samples
n_{req_i}	Supported number of samples in between i^{th} and $(i-1)^{th}$ samplings
μ_k	Mean of the k^{th} cluster
σ_k	SD of k^{th} cluster
N_{max}	Max. samples taken in one cycle
N_c	Total number of active cycles
$\mathcal{B}$	Base
$\mathcal{N}_{cur}$	Generated Number with base $\mathcal{B}$ and N_{req_i} as digits
$\mathcal{N}_{max}$	Generated Number with base $\mathcal{B}$ and N_{max_i} as digits
f_{min}	Minimum sampling frequency
f_{max}	Maximum sampling frequency
d_{min}	Minimum duty cycle
d_{max}	Maximum duty cycle
f_{next}	Calibrated frequency for next cycle
d_{next}	Assigned duty cycle for next cycle
$\varphi(f_{next})$	Frequency mapping for f_{next}
$\varphi(d_{next})$	Duty cycle mapping for d_{next}

from the raw sensor data. There is a dedicated buffer for each sensor. Each Buffer contains contextual information and stores N previous samples. Let B_i denotes the set of buffered values for sensor A_i. Here, $\mid B_i \mid \gg N$ where $N_{min} \leq N \leq N_{max}$. Buffers also consist of comparison variables which are obtained by statistical analysis on previously stored data such as standard deviation, variance etc. In the middle of the processing pipeline, there is a decision making component. Intelligent decision maker classifications are applied here among current and previous contextual information to determine whether further processing on the currently sensed data will be carried out or not. The feature extraction block extracts new and different contextual information to determine a possible user state transition.

The user state detection block recognizes user state transitions. This block also notifies the application about a new user state transition and also selects the required set of sensors for the recognition of current user state and possible user state transitions. The dynamic frequency and duty cycle adjuster block keeps track of the operational frequency and the length of the duty cycle of the sensor. To maintain tradeoff between energy efficiency and sensing accuracy, duty cycle and sampling period of the sensor can be adjusted dynamically. The notations used in this paper are summarized in Table I.

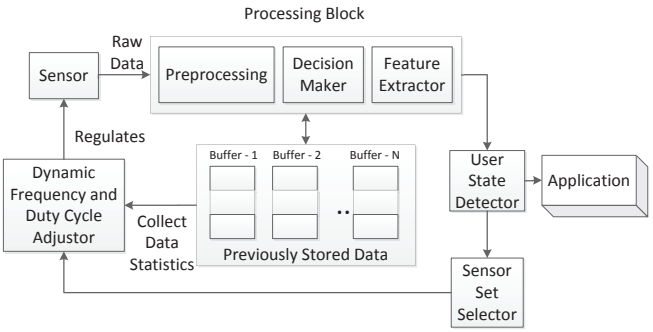


Fig. 1. The System Model

IV. PROPOSED STRATEGIES FOR ENERGY EFFICIENCY

Based on operation methods, smartphone sensors can be divided into 2 categories. The first category consists of sensors that support changing the sampling periods adaptively. Accelerometer and microphone belong to this category. Second category sensors like GPS, Wi-Fi, Bluetooth etc don't support setting different sampling periods [6]. The proposed techniques cannot be adapted to the second category sensors since their sampling frequencies are not allowed to be adjusted dynamically.

Despite the presence of several sensors in the mobile phones, we cannot use all of them simultaneously, the main reason being draining the battery life extensively. However, the energy requirements for different sensors vary significantly. According to studies [15], accelerometer is a sensor that works faster than other sensors. Accelerometer is followed by Bluetooth, microphone, GPS, Wi-Fi and video camera. As a result it would be wise to select accelerometer as the default sensor. Our proposed model provides an efficient way to improve the user context recognition vs. power consumption tradeoff.

Our basic model consists of 2 parts: a context monitoring mechanism and a dynamic frequency calibration system.

A. Context Monitoring Mechanism

In order to get user contextual information properly, the context must be monitored continuously. Such tasks cause heavy workloads which result in hampering the analysis of user context and reducing the battery life. Also redundant repetition of same information must be avoided.

In this paper, we have proposed an intelligent computational method to recognize contextual information of the user. Conventional methods process all raw sensor data, carrying them out through the whole processing pipeline. To reduce this overhead for required operation sequences, our proposed mechanism notifies the application about the user's state only when a state transition occurs. When a raw sensor data arrives, it is compared with previously buffered data to decide whether it indicates a possible change of user's current state or not. Decision on the raw data is taken using the following mechanism.

Let, for each sensor $A_i \in A_n$ there is a buffer B_i of size N which stores N previously sensed data where $B_i = \{X_1, X_2...X_N\}$ and X_j denotes the j^{th} data tuple sensed by sensor A_i. We applied K means clustering algorithm to partition B_i into K clusters, $C = \{C_1, C_2..., C_K\}$ using Eq. 1.

$$\underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{1}$$

where μ_k denotes the mean of k^{th} cluster. Here k is the number of clusters generated by the algorithm thus the value of k is an important consideration. A small value of k will affect the accuracy of the context monitoring mechanism. On the other hand, a large value of k will lead to high computational overhead. We adopt a common approach that is choosing the value of k such that adding another cluster does not achieve much better gain in terms of minimizing within-cluster sum of squares (WCSS). Though our proposed mechanism does not restrict the type of algorithm used for clustering, in practice we found k means algorithm to be simple and sufficient to serve our purpose.

Let X_j is the current sample and X_{j-1} is the previous sample stored in the buffer. Now we calculate mean (μ_k) and standard deviation (σ_k) of k^{th} cluster, C_k such that $X_{j-1} \in C_k$ using Eq. 2 and Eq. 3.

$$\mu_k = \frac{\sum_{i=1}^{|C_k|} x_i}{\mid C_k \mid} \tag{2}$$

$$\sigma_k = \sqrt{\frac{1}{\mid C_k \mid} \sum_{i=1}^{|C_k|} (x_i - \mu_k)^2} \tag{3}$$

Now when sensor A_i senses a raw data X_j, whether this data is redundant is decided using Algorithm 1.

Algorithm 1 Decision making for currently sensed data

1. C_k = FindClusterWithPreviousValue(C,X_{j-1})
2. Calculate μ of C_k using Eq. 2
3. Calculate σ of C_k using Eq. 3
4. Calculate deviation of X_j from μ_k as, $D_{cur} = \mid X_j - \mu_k \mid$
5. **if** $D_{cur} \geq \sigma_k \times \xi$ **then**
6. $B_i \leftarrow B_i \cup \{X_j\}$
7. Accept X_j for further processing
8. **else**
9. Reject X_j
10. **end if**

B. Sensor Operation Structure

In Fig. 2, sensor operation structure is illustrated. It starts with sensor initialization time which is denoted by T_{in}, the required time for waking a sensor up and getting acknowledged response that the sensor is ready to operate. For other sensors a shorter period is sufficient to power them up and to set their initial system requirements before sampling operations begin. T_{adj} is the time required for adjusting sensor's current operational frequency f and current duty cycle d (portion of time of a cycle spend on sampling). If T_c denotes the time required for an operational cycle, then the number of samples N, taken in this cycle can easily be calculated as, $N = f \times d \times T_c$.

After initialization, sensors start capturing contextual information and continue this operation repeatedly until the starting of termination time T_{ter}. T_{ter} is the time required to terminate the operation of a sensor. After T_{ter}, sensor shuts down until a new duty is assigned. The total running time denoted by T_{total} is the time required from waking a sensor up until shutting it down. Now, sensor's active running time for taking samples t_{run} can be calculated as, $T_{run} = T_{total} - (T_{init} + T_{ter})$. The number of active cycles run by a sensor between T_{init} and T_{ter} can be calculated as, $N_{cycle} = \frac{T_{run}}{T_c}$. T_s denotes sampling interval within which time a sample is taken and can be calculated as, $T_s = \frac{T_c}{N}$. Here, $t_s (\leq T_s)$ is the sampling period and is given by, $t_s = \frac{1}{f}$.

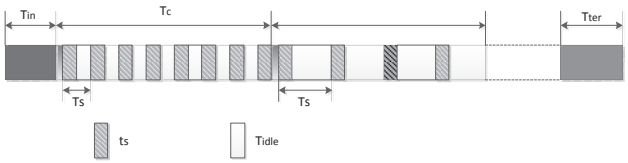


Fig. 2. Sensor Operation Structure

C. Dynamic Frequency Calibration

Let, X_i and X_{i-1} denote the i^{th} and $(i-1)^{th}$ data tuples respectively. Now, due to low sampling frequency some necessary data might be missed. On the other hand, in case of high sampling frequency some redundant data could be captured. So, the actual number of required readings between these two consecutive samples can be calculated as, $n_i = \frac{|X_i - X_{i-1}|}{\psi X}$. Here, ψX is the maximum allowable difference between two consecutive samples to extract contextual information accurately. But as the sampling frequency is limited to $[f_{min}, f_{max}]$ the maximum and minimum numbers of samples that can be taken during T_s are $N_{max} = (T_s + t_s) \times f_{max}$ and $N_{min} = (T_s + t_s) \times f_{min}$ respectively. So, the actual number of required samples is given by,

$$N_{req_i} = \begin{cases} min(N_{max}, n) & \text{if } n > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$N_{req_i} = \lfloor (N_{req_i} + 0.5) \rfloor \tag{5}$$

In our proposed mechanism, the duty cycle length and sampling frequency are adjusted dynamically for each sensor. To reduce energy consumption, a pair of duty cycle and sampling frequency is assigned to each sensor. In this situation a tradeoff between energy consumption vs. accuracy arises, which has to be balanced efficiently. Too long sampling intervals cause insufficient sampling to represent real conditions which may lead to incorrect user state recognition. On the other hand, a short sampling interval enables a larger number of samplings thus consuming more energy. In case of duty cycle, a short

duty cycle length will save energy, increasing sleeping period of a sensor. However, this causes higher detection latency and leads to false detection of user states. A long duty cycle length will increase data accuracy while wasting more energy. Our proposed mechanism dynamically adjusts sampling frequency and duty cycle to balance this tradeoff.

We define base $\mathcal{B}$ as,

$$\mathcal{B} = N_{max} + 1 \tag{6}$$

Now, n_{req_i}, which denote the required number of samples between i^{th} and $(i-1)^{th}$ samples and can be calculated using Eq. 6. We consider a number system of base $\mathcal{B}$ whose digits can be any values between 0 and $\mathcal{B}-1$. Now, for $N-1$ consecutive samples we generate two numbers $\mathcal{N}_{cur}$ and $\mathcal{N}_{max}$ taking N_{req_i} and N_{max_i} as digits respectively, where $0 \leq N_{req_i}, N_{max} \leq (\mathcal{B}-1)$.

$$\mathcal{N}_{cur} = \sum_{i=1}^{N-1} N_{req_i} \times \mathcal{B}^{i-1} \tag{7}$$

$$\mathcal{N}_{max} = \sum_{i=1}^{N-1} N_{max} \times \mathcal{B}^{i-1} \tag{8}$$

Now, using $\mathcal{N}_{cur}$ and $\mathcal{N}_{max}$ we can determine sampling frequency for the next cycle using the following equation.

$$f_{next} = \lfloor \frac{\mathcal{N}_{cur}}{\mathcal{N}_{max}} \times (f_{max} - f_{min}) + f_{min} \rfloor \tag{9}$$

Now, according to f_{next}, the duty cycle for the next cycle d_{next} can be adjusted as,

$$d_{next} = \lfloor \frac{f_{next} - f_{min}}{f_{max} - f_{min}} \times (d_{max} - d_{min}) + d_{min} \rfloor \tag{10}$$

Since sensor's frequency can't be calibrated to any continuous value, we must map the f_{next} and d_{next} pair to the available discrete values. Let F and D denote the sets of all available frequencies and duty cycles. So, the required mapping function can be defined as,

$$\varphi(f_{next}) = \begin{cases} f_i & |f_i - f_{next}| < |f_{i+1} - f_{next}| \\ f_{i+1} & |f_i - f_{next}| > |f_{i+1} - f_{next}| \end{cases} \tag{11}$$

where $f_i, f_{i+1} \in F$, $f_i \leq f_{next} \leq f_{i+1}$ and $i = 1, 2, ... \mid F \mid -1$. If $\mid f_i - f_{next} \mid = \mid f_{i+1} - f_{next} \mid$ then the mapping function can be modified as,

$$\varphi(f_{next}) = \begin{cases} f_i & \text{for power sensitive app.} \\ f_{i+1} & \text{for accuracy sensitive app.} \end{cases} \tag{12}$$

Using similar mapping function, we can map d_{next} to $\varphi(d_{next})$. Now the sensors sampling frequency is calibrated to $\varphi(f_{next})$ and $\varphi(d_{next})$ is assigned as the active duty cycle in the next cycle, as summarized in Algorithm 2. Overall sensor operation scheme is presented in Algorithm 3.

Algorithm 2 Dynamic Frequency Calibration

1. **INPUT:** B, f_{max}, f_{min} and ΔX
2. **OUTPUT:** $\varphi(f_{next})$ and $\varphi(d_{next})$
3. **for all** $X_i \in B$, i = 1, 2, 3, ..., $\mid B \mid$-1 **do**
4. calculate n_i and N_{req_i} using Eq. 4 to Eq. 5
5. **end for**
6. Calculate $\mathcal{N}_{cur}$, $\mathcal{N}_{max}$ and $\mathcal{B}$ using Eq. 5 to Eq. 8
7. Calculate f_{next} and d_{next} using Eq. 9 to Eq. 10
8. Find $\varphi(f_{next})$ and $\varphi(d_{next})$ using Eq. 11 and Eq. 12

Algorithm 3 Sensor Operation Scheme

1. **repeat**
2. Run Algorithm 2
3. Set sampling frequency and active duty cycle to $\varphi(f_{next})$ and $\varphi(d_{next})$ respectively
4. **repeat**
5. Take Sample for sampling interval T_S (sample is taken in t_s time and sensor remains idle for rest of the T_S before taking a new sample)
6. **until** active cycle time reaches to $T_c - T_{sch}$
7. **until** no of active cycle reaches N_{cycle}

V. Performance Evaluation

In this section, we study the performance of our proposed system with comparison to the methods described in [5]. For our convenience, we denote the methods described in [5] as ASDC (Adaptive Sampling and Duty Cycling). Performance measurement is done in terms of accuracy and power efficiency.

A. Simulation Environment

A smartphone application is implemented in order to evaluate the performance of the proposed system. The application collects contextual data using accelerometer sensor. Samsung Galaxy S-DUOS smartphone is used as the target device. Android Studio is used as software development tool. The target device supports a 3-axis accelerometer which is used to collect sensor data following the specifications provided in [5]. Parametric values for the simulation environment are summarized in Table II.

B. Performance Metrices

1) Accuracy: Accuracy between i^{th} and $(i-1)^{th}$ samples can be calculated as

$$\alpha_i = \begin{cases} \frac{1}{n_i} & \text{if } n_i > 0 \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

TABLE II
Simulation Parameters

Parameter	Value
F	{10Hz, 15Hz, 20 Hz}
D	{1, 0.75, 0.5}
T_c	2 sec
T_{total}	1 hour
ψX	0.1

Then, the average accuracy for N samples taken in T_{run} time is calculated as follows,

$$\alpha = \frac{1}{N} \times \sum_{i=1}^{N-1} \alpha_i \tag{14}$$

2) Power Efficiency: Power Efficiency is defined as the ratio of remaining energy and initial energy budget. For calculation of power efficiency ζ we adopt power consumption analysis of [5].

C. Simulation Result

Simulation results demonstrate satisfactory performance. In Fig. 3, we observe that the accuracy increases until it reaches its saturation level in both ASDC and our proposed method. However, the rate of increase in our proposed method is greater than that of ASDC. This is caused by the fact that, in ASDC, after every t_{suf} time a sensor adjusts its duty cycle and sampling frequency to the next possible index. Thus, it requires more time to converge to the appropriate duty cycle and sampling frequency. However, in our proposed method, according to the contextual information, sampling frequency and duty cycle are dynamically scaled for next time. So, convergence becomes faster and thus increasing the accuracy.

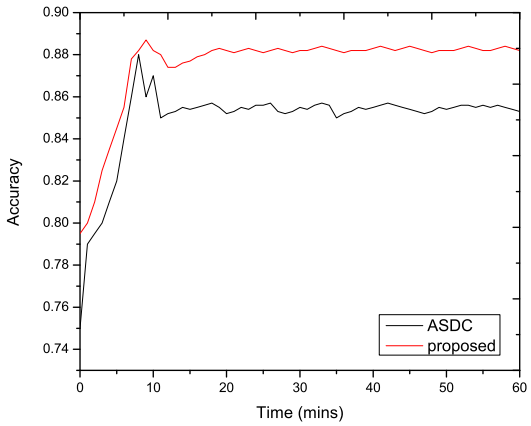


Fig. 3. Time vs. Accuracy

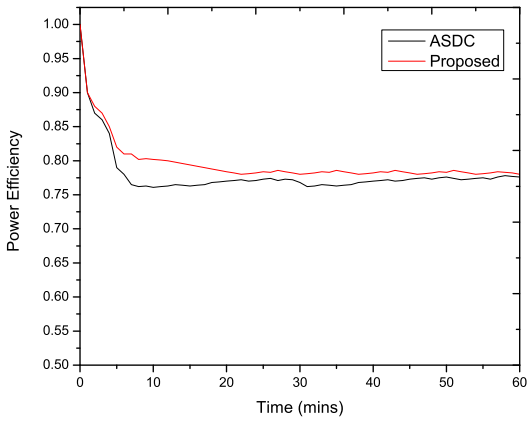


Fig. 4. Time vs. Power Efficiency

The Figure 4 shows that power efficiency decreases in time in both ASDC and our proposed method. But the rate of decreasing is slower in our proposed method. Since the convergence is faster in our method, it reduces unnecessary sampling thus reducing power consumption.

VI. Conclusion

We have presented novel approaches in this paper to address the energy accuracy tradeoff for smartphone sensing. Several algorithms have been developed to improve the energy efficiency of context monitoring applications while capturing accurate contextual information. The use of MIMD in adaptively scaling the sensing frequency has been explored and the results prove that it is more effective than AIAD (additive increase additive decrease). Implementation results showed satisfactory performance improvement both in accuracy and energy domains. Now, we are exploring the formulation of an optimization function to further improve tradeoff levels whenever the application environment parameters greatly vary over time. In addition to that the system stability analysis would be an indispensable part of this work.

Acknowledgements

This work is supported by a grant for the Research Fellowship (2014-2015) funded by the Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh. Dr. Md. Abdur Razzaque is the corresponding author of this paper.

References

[1] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '10. New York, NY, USA: ACM, 2010, pp. 71–84. [Online]. Available: http://doi.acm.org/10.1145/1869983.1869992

[2] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: An end-to-end participatory urban noise mapping system," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, ser. IPSN '10. New York, NY, USA: ACM, 2010, pp. 105–116.

[3] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *Comm. Mag.*, vol. 48, no. 9, pp. 140–150, Sep 2010.

[4] Y. Wang, J. Lin, M. Annavaram, Q. A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh, "A framework of energy efficient mobile sensing for automatic user state recognition," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '09. New York, NY, USA: ACM, 2009, pp. 179–192.

[5] O. Yurur, C. Liu, X. Liu, and W. Moreno, "Adaptive sampling and duty cycling for smartphone accelerometer," in *Mobile Ad-Hoc and Sensor Systems (MASS), 2013 IEEE 10th International Conference on*, Oct 2013, pp. 511–518.

[6] Ö. Yürür and W. Moreno, "Energy efficient sensor management strategies in mobile sensing," *ISTEC General Assembly 2011*, 2010.

[7] Y. Lee, Y. Ju, C. Min, S. Kang, I. Hwang, and J. Song, "Comon: Cooperative ambience monitoring platform with continuity and benefit awareness," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 43–56. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307641

[8] S. Nath, "Ace: Exploiting correlation for energy-efficient and continuous context sensing," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 29–42. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307640

[9] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram, "Markov-optimal sensing policy for user state estimation in mobile devices," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*. ACM, 2010, pp. 268–278.

[10] S. Kang, J. Lee, H. Jang, H. Lee, Y. Lee, S. Park, T. Park, and J. Song, "Seemon: scalable and energy-efficient context monitoring framework for sensor-rich mobile environments," in *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, 2008, pp. 267–280.

[11] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, "Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection," in *Wireless Sensor Networks*. Springer, 2008, pp. 17–33.

[12] K. K. Rachuri and C. Musolesi, Mirco an d Mascolo, "Energy-accuracy trade-offs in querying sensor data for continuous sensing mobile systems," in *Proc. of Mobile Context-Awareness Workshop*, vol. 10. Citeseer, 2010.

[13] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "SociableSense: Exploring the Trade-offs of Adaptive Sampling and Computation Offloading for Social Sensing," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom'11)*, Las Vegas, USA, September 2011.

[14] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: A mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ser. UbiComp '10. New York, NY, USA: ACM, 2010, pp. 281–290.

[15] F. Ben Abdesslem, A. Phillips, and T. Henderson, "Less is more: energy-efficient mobile sensing with senseless," in *Proceedings of the 1st ACM workshop on Networking, systems, and applications for mobile handhelds*. ACM, 2009, pp. 61–62.

[16] I. Constandache, S. Gaonkar, M. Sayler, R. R. Choudhury, and L. Cox, "Enloc: Energy-efficient localization for mobile phones," in *INFOCOM 2009, IEEE*. IEEE, 2009, pp. 2716–2720.

Conflicting Goal Constrained Architecture of a Heterogeneous Mobile Sensor Network

Afroza Sultana
Dept of CS, University of Texas at Arlington
Texas, USA
Email: afroza.sultana@uta.edu

Mahmuda Naznin
Dept. of CSE, BUET
Dhaka, Bangladesh
Email: mahmudanaznin@cse.buet.ac.bd

Rifat Shahriyar
Dept. of CSE, BUET
Dhaka, Bangladesh
Email: rifat@cse.buet.ac.bd

***Abstract*—Maintaining coverage and connectivity is a conventional research problem in a sensor network. However, it is still a challenging research field due to the power constraint and connection vulnerability of the battery powered sensing devices. However now a days a network comprises of different mobile devices of different ranges. Still now, there is no comprehensive optimal solution for providing coverage with desired connectivity in a mobile wireless sensor network. In our paper, we present an efficient coverage model for heterogeneous mobile sensing devices maintaining the required connectivity. We compute the optimal positioning of the mobile devices based on Linear Programming, maintaining the required connectivity. We validate our claim by experimental results.**

I. Introduction

A smart sensing system is composed of low powered, low cost different types of sensing devices [1]. If the sensing nodes do not have mobility, after power failure, the coverage hole or dis-connectivity is created. Not only that, the coverage hole or gap can be formed during the deployment also. Adding mobile sensing nodes, we gain more flexibility in sensor network system as they have self-healing and self-optimizing capability [2]. Mobile sensing devices can be used also to physically collect or transport data or to repair and recharge the static sensing nodes in the network. In *homogeneous sensing system*, we assume that all of the sensors have the similar property and in *heterogeneous sensing system*, sensors have different properties; they may have different sensing and transmission ranges, they may have different sensing properties, different energy consumption rate etc. Since now a days, Internet of Things is the suitable example of heterogeneous smart sensing system [1], [2]. However, in a heterogeneous mobile network to provide the optimal coverage with a required degree of connectivity is challenging since if we keep more nodes nearby it provides connectivity but may lack enough coverage, on the contrary, if we deploy them a bit far away, they may not provide the required degree of connectivity. In Section II, we discuss the related research, in Section III, we provide the problem formulation and our proposed solution, Section IV provides experimental results and finally, we give the conclusion in Section V.

978-1-5090-0203-0/16$31.00

II. Related Works

In [6], the authors first illustrate coverage computation when a single sensor nodes is randomly deployed. They consider a homogeneous sensor network where all sensors have the same sensing area. Then, they compute the fraction of area non-covered as a function of number of N sensors. Then, they compute the *probability distribution function* of the fraction of area covered by exactly and at least given K sensors. However, in this paper, it is shown that for heterogeneous network scenario, the computational cost increases exponentially. Therefore, an approximation method is used to reduce the cost where coverage is achieved but connectivity is not guaranteed.

In [7], the authors consider *constrained coverage* for a network whose constituent nodes are all autonomous mobile robots with sensors. The *Constrained Coverage* is the problem of finding a deployment configuration which maximizes the collective sensor coverage while satisfying one or more constraints. They design their network with homogeneous mobile sensing devices which are capable of omni-directional motion. In their deployment algorithm, the authors construct virtual forces between nodes so that each node can attract to provide connectivity or can repel its neighbour sensor nodes to cover more area. By using a combination of these opposite forces each node maximizes its coverage while maintaining a degree of at least K connectivity. However, this algorithm is applicable for homogeneous mobile sensor network.

In [9], the authors present a solution that utilizes a small number of mobile sensors added to static nodes to improve coverage performance in sensor networks. They use Integer Linear Programming (ILP) method, Computational Geometry and Spiral-In algorithm. Though the ILP based method achieves optimal result, it costs high computation and their work do not consider connectivity issue.

In [3], the authors discuss connectivity issues in a mobile sensor network. They design two types mobile sensor nodes. One type is controllable and referred to as *robots*. Another type is uncontrollable and referred to as *mobile sensor nodes*. They use *K-redundancy* to determine the communication characteristics of a dynamic wireless sensing network system. A node N is $K - redundant$ when it is removed from the network, at least K more nodes needs to be removed from the network to remove any communication route between

any two neighbours of N. $K - redundancy$ for nodes with only one neighbour is defined their connectivity as zero for practical purposes. In order to detect the redundancy of nodes, they use the information in hello messages. In this approach, each node stores two hop neighbourhood information. They present two dynamic approaches to fill up the required level of connectivity. They define (i)*reactive recovery*, where idle mobile sensors are directed towards the locations to reconnect mobile nodes to their neighbours after disconnection happens, (ii) *proactive recovery*, where idle mobile robots are directed towards the low redundancy regions before a disconnection event to increase the response time. They find that a robot near a lower redundancy node would be more likely to provide faster time than a robot near a higher redundancy node. For this reason, in proactive approach, they locate robots near the low redundancy nodes. This is achieved by propagating the direction of the lowest redundancy nodes towards idle robots. After redundancies of nodes are determined, each node includes this information in its hello messages. As these messages propagate in the network, each node obtains information about the directions of the regions which have low connectivity. Thus, when there is an idle robot in the communication range of a node, it navigates the robot in the direction of low redundancy regions. While combining these redundancy values from different sources, they always choose the minimum value. In case of active repairing, a mobile node directs a robot to a location to re-establish the communication with its neighbour. But in their method, of propagating message towards the robot in case of reactive repairing, the path may not be the shortest path. As robot information is propagated one hop at each hello period, the node which is N hops away has the location of the robot which was $N*(helloperiod)$ ago. Therefore, the robot location information is not updated. In their method, it is not possible to tell whether or not a node is out of network due to power failure or its intentional departure.

In [4] and [5], the authors propose algorithm for minimizing the number of additional nodes required and the distance they need to move for construction of a topology with desired levels of fault-tolerance or connectivity. They model the communication topology as a graph on the nodes, assuming a fixed range for each device. Each device is considered to be a vertex in the graph, and an edge exists between two vertices if they are within each other transmission range. They call the nodes they want to construct a topology on as $terminal$ nodes. They design a scenario where they can control the movement of only a subset of nodes, and those nodes are used just to provide the desired connectivity among the terminal nodes. They consider the number of relays required for constructing a topology that has $K - edge(vertex)$ disjoint paths every pair of terminals and assume that the terminals move at a slow time-scale. They use the approximation algorithm. However, their method only provides approximations. Derivation of an approximation ratio for $K > 2$ found from their proposed algorithms is a big challenge. However, their algorithms do not work in a heterogeneous sensor network.

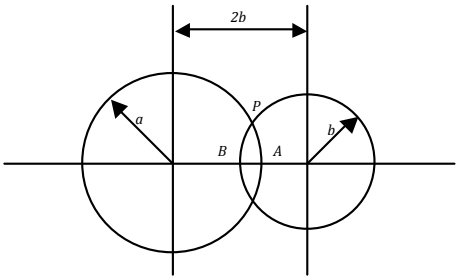


Fig. 1. Two sensors with sensing range a, b respectively

Therefore, it is challenging to have coverage architecture with the desired level of connectivity in a heterogeneous mobile sensor system.

III. Problem Formulation

In this section, we describe the problem domain and preliminary definitions in details. Let us assume that, sensors are deployed randomly on ad-hoc basis. Usually, to have a grid and predefined position based deployment is relatively expensive [2]. Therefore, if the nodes are adjust their position in such a way that they can maximize their coverage area, it would definitely save cost [3], [4]. There might be some applications, where a specific level of connectivity is also required. In that case, the deployment is more challenging since having the best coverage with the required degree of connectivity since these cause conflicting goals. In this paper, we give an algorithm for given N mobile sensor nodes with different radial sensing ranges R_ss and different radio communication range R_cs, so that they deploy themselves maximizing the coverage of the network with the constraint that each node has at least K neighbours.

A. Preliminaries

In this section we provide some preliminaries. Two heterogeneous sensor nodes S_i and S_j are called $neighbour$ of each other if the Euclidean distance between them is less than or equal to min(R_{c_i}, R_{c_j}). They are called *optimal neighbour* of each other if the Euclidean distance between them is exactly min(R_{c_i}, R_{c_j}).

$Theorem$: For any three heterogeneous sensor nodes S_a, S_b and S_c with communication ranges R_{c_a}, R_{c_b} and R_{c_c} respectively, the coverage will be the maximum if the network is designed such a way that both S_a and S_c are the optimal neighbours of S_b as there is no overlapped portion in their sensing areas.

Let the sensing ranges of S_a and S_b be a and b. For any connected graph, communication range R_c at least twice of its sensing range R_s, $R_c >= 2R_s$ [8]. The equations of their coverage areas can be written as follows.

$$x^2 + y^2 = a^2 \tag{1}$$

$$(x - 2b)^2 + y^2 = b^2 \tag{2}$$

At their intersection point,

$$x = \frac{(a^2 + 3b^2)}{4b}$$

The area overlapped by Equations (1) and (2) = 2 x Area of BPA

The area of BPA is,

$$= \int_b^{\frac{a^2+b^2}{4b}} \sqrt{b^2 - (x-2b)^2}\, dx + \int_{\frac{a^2+b^2}{4b}}^{a} \sqrt{a^2 - x^2}\, dx$$

$$= \frac{\pi(a^2+b^2)}{4} - \sqrt{\frac{10a^2b^2 - a^4 - 9b^4}{4}}$$

$$-\frac{a^2}{2} sin^{-1}\frac{a^2+3b^2}{4ab} + \frac{b^2}{2} sin^{-1}\frac{a^2-5b^2}{4b^2}$$

$$= f(b) \tag{3}$$

Let Equation (3) be a function of b; differentiating it we get $f^{'}(b)$

$$= \frac{\pi b}{4} - \frac{20a^2b - 36b^3}{8\sqrt{10a^2b^2 - a^4 - 9b^4}} - \frac{\frac{24ab^2 - (a^2+3b^2)4a}{16b^2}}{2\sqrt{1 - (\frac{a^2+3b^2}{4ab})^2}}$$

$$+bsin^{-1}\frac{a^2-5b^2}{4b^2} + \frac{\frac{4a^2(-10b) - (a^2-5b^2)8b}{16b^2}}{2\sqrt{1 - (\frac{a^2-5b^2}{4b^2})^2}}$$

$$= \frac{\pi b}{4} - \frac{5a^2b - 9b^3}{2\sqrt{10a^2b^2 - a^4 - 9b^4}} - \frac{a(24ab^2 - 4a^3 - 12ab^2)}{8b\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$+bsin^{-1}\frac{a^2-5b^2}{4b^2} + \frac{-40b^3 - 8a^2 + 40b^3}{8\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$= \frac{\pi b}{4} - \frac{5a^2b^2 - 9b^4 - a^4 + 3a^2b^2 + 2a^2b^2}{2b\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$+bsin^{-1}\frac{a^2-5b^2}{4b^2}$$

$$= \frac{\pi b}{4} - \frac{\sqrt{10a^2b^2 - a^4 - 9b^4}}{2b} + bsin^{-1}\frac{a^2-5b^2}{4b^2}$$

For maximization or minimization gradient, $f^{'}(b) = 0$. If b = a, we get $f^{''}(b)$

$$= \frac{\pi}{2} + sin^{-1}\frac{a^2-5b^2}{4b^2} + \frac{\frac{4a^2(-10ab) - (a^2-5b^2)8b}{16b^3}}{\sqrt{1 - (\frac{a^2-5b^2}{4b^2})^2}}$$

$$-\frac{20a^2b^2 - 36b^4 - 20a^2b^2 + 2a^4 + 18b^4}{4b^2\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$= \frac{\pi}{2} + sin^{-1}\frac{a^2-5b^2}{4b^2} - \frac{2a^2}{\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$-\frac{2a^4 - 18b^4}{4b^2\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

$$= \frac{\pi}{2} + sin^{-1}\frac{a^2-5b^2}{4b^2} - \frac{2a^4 - 18b^4 + 8a^2b^2}{4b^2\sqrt{10a^2b^2 - a^4 - 9b^4}}$$

As, $f^{''}(a) > 0$, at $b = a$, $f(b)$ is the minimum and the value is zero. Thus, since $(a-b)$ increases $f(b)$ increases, that means the overlapped area of Equations (1) and (2) increases, and the coverage decreases.

As, $R_{c_a} > R_{c_b} > R_{c_c}$, the coverage will be the maximum if the network is designed such that both S_a and S_c are the optimal neighbours of S_b but there is no overlapped portion in their sensing areas. [Proved]

For example, we take three heterogeneous sensor nodes A, B and C of sensing ranges 3, 2 and 1.5 respectively (see Fig.2). In Figure 2, we show that the overlapped area (coverage) will be the minimum (the maximum) if the network is designed in such that both of A and C are the optimal neighbours of B but there is no overlapped sensing area between A and C.

B. Optimal Neighbour Selection

We now provide the optimal coverage by formulating the problem as LP problem. Given N heterogeneous mobile sensor nodes, the objective is to deploy them so that the resulting configuration maximizes the coverage of the network with the constraint that each node has at least K neighbours.
Let x_{ij} be a 0 or 1 binary variable defined as:

$$x_{ij} = \begin{cases} 1 & \text{if Sensor } S_i \text{ and Sensor } S_j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

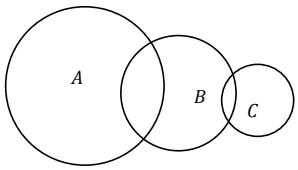
The objective of the LP is to maximize the coverage minimizing the neighbours to K :
Maximize

$$\sum_{i,j \in N} A_{ij} * x_{ij}. \tag{4}$$

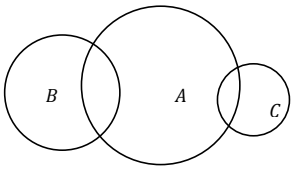
Subject to:

$$\sum_{i,j \in N} x_{ij} \geq K \tag{5}$$

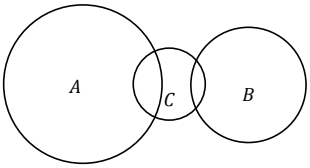
$$A_{i,j}, \text{ area by neighbours } S_i \text{ and } S_j, (\pi {R_{s_i}}^2 \cup \pi {R_{s_j}}^2) \neq 0 \tag{6}$$



(a) Case 1: A and C are optimal neighbours of B.



(b) Case 2: B and C are optimal neighbours of A and no overlapped sensing area between B and C



(c) Case 3: A and B are optimal neighbours of C and and no overlapped sensing area between A and B

Fig. 2. Finding optimal neighbours for different cases.

C. The Deployment Method

For deployment algorithm, we use two forces F_{cover} and F_{degree} defined in [6]. Mathematically, we modify the forces as follows so that it can handle heterogeneous sensors.

We consider a network of N nodes $1, 2, 3, \ldots N$ at position $x_1, x_2, \ldots, x_n$ respectively. Their communication ranges are $R_{c_1}, R_{c_2}, \ldots, R_{c_n}$ and sensing ranges are $R_{s_1}, R_{s_2}, \ldots, R_{s_n}$ respectively. Let Δx_{ij} represents the Euclidean distance between nodes i and j i.e. $\Delta x_{ij} = ||x_i - x_j||$

Modified F_{cover} and F_{degree} as described as follows.

$$F_{cover}(i,j) = (\frac{-K_{cover}}{\Delta x_{ij}^2})(\frac{x_i - x_j}{\Delta x_{ij}})$$

$$F_{degree}(i,j) = \begin{cases} (\frac{K_{degree}}{(\Delta x_{ij} - R_{c_i})(\Delta x_{ij} - -R_{c_j})})(\frac{x_i - x_j}{\Delta x_{ij}}) \\ \text{if critical conditions} \\ 0 \\ \text{otherwise} \end{cases}$$

where K_{cover} and K_{degree} are the force constraints. The resultant force between the nodes i and j is

$$F(i,j) = F_{cover}(i,j) + F_{degree}(i,j)$$

and node i will experience a net force of

$$F_i = \sum_{allneighbours j} F(i,j)$$

TABLE I
SIMULATION PARAMETER

Parameters	Value
Network Size	500m x 500m
No. of Sensor Nodes	10-400
Sensing Range	1-4 m
Degree of Connectivity	1-4

The equation of v motion for node i is formulated as:

$$x_i(t) = \frac{F_i - v x_{i-1}}{m}$$

where v is a damping factor and m is the virtual mass of the node which is assumed to be 1.

Having described the equation of motion for the node, we discuss our choices of the four parameters K_{cover}, K_{degree}, v and η.

1) K_{cover}

Let us consider two nodes i and j be communicating and repelling each other. As the distance d between them increases, the combined coverage of the node increases, reaches a maximum of $\pi(R_{s_i}^2 + R_{s_j}^2)$ at $d = R_{s_i} + R_{s_j}$. This implies, for $d > R_{s_i} + R_{s_j}$ repelling does not improve coverage. We therefore pick a value for K_{cover} such that $d = R_{s_i} + R_{s_j}, ||K_{cover}|| \approx 0$

2) K_{degree}

At $d = \eta(R_{c_i} + R_{c_j})$ we want $||F_{cover} + F_{degree}|| = 0$, i.e.

$$\frac{-K_{cover}}{d} + \frac{K_{degree}}{(d - R_{c_i})(d - R_{c_j})} = 0$$

$$K_{degree} = \frac{(d - R_{c_i})(d - R_{c_j})}{d^2} K_{cover}$$

IV. EXPERIMENTAL RESULTS

In this section, we provide the experimental results to validate our method. We implement the algorithm in OMNeT++ [10]. The coverage area of a sensor node is shaped as a circle, and the sensing range is varied randomly for different sensors. Network area is considered as a square area where nodes with different sensing and communication ranges are randomly deployed. Each of the nodes is capable of having 2D motion and sensing [11]. We run the simulations for different degree of connectivity. We measure the elapsed time in millisecond. The reported results are the average of 10 simulation runs. Simulation parameters are listed in Table I.

We assume that all nodes have capability with desired connectivity. Therefore, we start with a network with initial configurations, where each node has a degree greater than the required degree K so that the algorithm ensures 100% of the nodes have a degree of at least K. Network nodes are randomly deployed and we use synthetic data sets which is randomly generated. From our experiments, we see that each node selects its neighbours in such a way that the

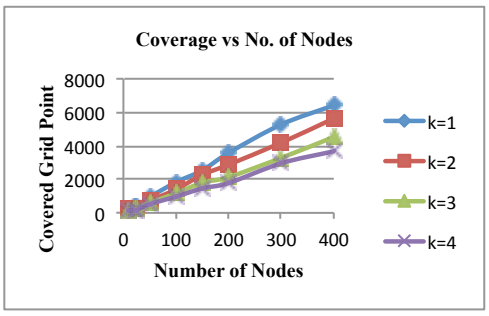


Fig. 3. Coverage for the different node density.

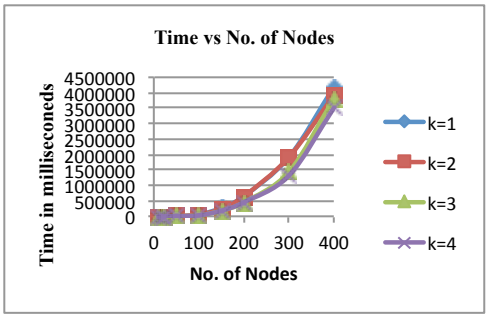


Fig. 4. Time for the different node density.

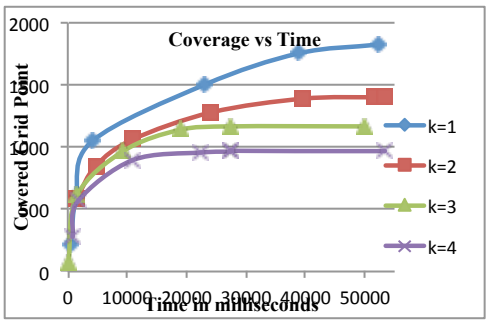


Fig. 5. Coverage vs Time for different connectivity.

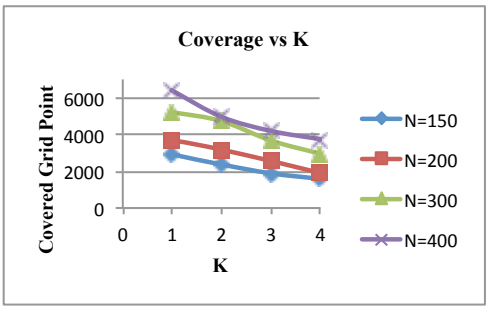


Fig. 6. Variation of coverage with the degree of connectivity for different no. of nodes.

overlapped area between them is minimized. Therefore, the mechanism provides the maximized coverage with satisfying the constraints after applying the theorem stated in Section III. In our experiment, we vary no. of sensor nodes, no. of connectivity requirement and simulation time. For performance measurement we use different metrics called *Network Lifetime*, *Covered Grid Point*, and *No. of Neighbours* of the node. *Network lifetime* is the simulation time how long network can provide coverage with K connectivity, *no. of neighbours* denotes the connectivity, a *covered grid point* denotes the point of the network location under any sensor node's sensing range. Network is divided into unit square grid.

The Impact of the no. of Sensors

We discuss the impact of the number of the network nodes on the performance metric. In Fig. 3, we find that if the number of deployed sensor nodes increases coverage improves which is natural.

In Fig. 4 the impact on the network life time is shown. Figures 3 and 4 show the variation of the coverage and network life time with the number of nodes for different values of K. The coverage and lifetime increases almost linearly as the number of nodes increases.

The Impact of the Connectivity

We now discuss the impact of the connectivity K on the performance metrics. We find that if the network connectivity K is increased, coverage is better.

In Fig. 5 the trade-off between coverage and network life time is shown. Fig. 5 shows the variation of the coverage with time for different values of K. The coverage increases rapidly in first 9-10 milliseconds than saturates to a stable value within 38-53 milliseconds. This is because, initially all nodes have more than K neighbours and so they spread out uninhibitedly to improve the coverage until the degree constraints active and restrict their motion. As the time increases, coverage decreases.

In Fig. 6 coverage for varying connectivity of the network is shown. Here we find, coverage decreases as connectivity increases. It is usual since more nodes are to stay close for getting connectivity and therefore, they can not move to provide adequate coverage.

Fig. 7 shows the impact on connectivity for varying K in course of time. Since time increases, no. of connectivity decreases. We find, the average connectivity decreases as time elapses. This is natural as sensor nodes are depleted after sometime. Therefore, the values show a decreasing curve.

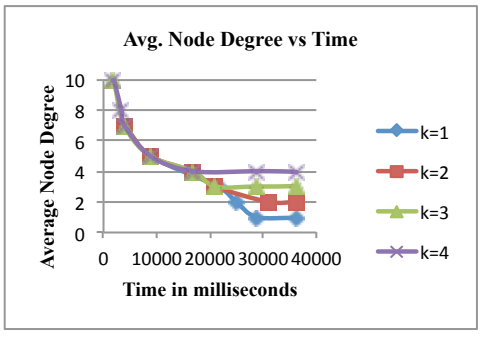


Fig. 7. Average Node Degree vs. Time, for different connectivity.

V. Conclusion

In our paper, we present a self adjusting deployment in a heterogeneous mobile sensing system with the required connectivity degree as well as the maximizing the collective coverage. We provide the optimal positioning of the mobile sensor nodes based on Linear Programming with maintaining the required connectivity. We validate our model with experimental results. We find that our model can provide the coverage and required connectivity successfully. This protocol is fully autonomous and distributed and can be applied any heterogeneous sensing networking system.

References

[1] I. F. Akyldiz, W. Su, Y. Sankarsubramaniam, and E. Cayirci, Wireless sensor networks: A survey, *Computer Networks*, Elsevier, vol. 38, pp. 393-422, 2002.

[2] F. Koushanfar, S. Meguerdichian, M. Potkonjak, and M. Srivastava, Coverage problems in wireless ad-hoc sensor networks., *In Proc. of IEEE INFOCOM*, pp.1380-1387, 2001.

[3] N. Atay and B. Bayazit, Mobile wireless sensor network connectivity repair with K-redundancy, *Tech. Rep., WUCSE-2007-48*, Dept. of Computer Science and Engineering, Washington University in St. Louis, USA, 2007.

[4] A. Kashyap and M. Shayman, Relay placement and movement control for realization of fault-tolerant ad-hoc networks, *In Proc. of International Conference on Information Sciences and Systems*, 2007.

[5] A. Kashyap, S. Khuller, and M. Shayman, Relay placement for higher order connectivity in wireless sensor networks, *Ad-Hoc Networks*, Elsevier, 2000.

[6] L. Lazos and R. Poovendran, Coverage in heterogeneous sensor network, *In Proc. of IFIP/IEEE International Symposium on Modelling and Optimization in Mobile, Ad-Hoc and Wireless Networks (WiOpt)*, 2006.

[7] S. Poduri and G. S. Sukhatme, Constrained coverage for mobile sensor networks, *In Proc. of IEEE International Conference on Robotics and Automation*, LA, USA, pp. 165-172, 2004.

[8] G. Xing, X. Wang, Y. Zhang, C. Lu, R. Pless, and C. Gill, Integrated coverage and connectivity configuration for energy conservation in sensor networks, *ACM Transaction on Sensor Network*, vol. 1, no. 1, 2008.

[9] M. Zhang, X. Du, and K. E. Nygard, Improving Coverage Performance in Sensor Networks by Using Mobile Sensors, *In Proc. of IEEE International Conference on Military Communication (MILCOM)*, USA, 2005.

[10] Discrete Event Simulator, OMNeT++, https://omnetpp.org/

[11] Moog Crossbow, Milpitas California, http://www.moog-crossbow.com/

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Full Papers

Optimized Communication Protocols

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Game Theoretic Downlink Resource Scheduling for Self-coexisting Cognitive Radio Networks

Sayef Azad Sakin, Md. Abdur Razzaque, *Senior Member, IEEE*
Green Networking Research (GNR) Group, Deptartment of Computer Science and Engineering
Facutly of Engineering and Technology, University of Dhaka, Dhaka - 1000, Bangladesh
Email: sayefsakin@gmail.com, razzaque@du.ac.bd

***Abstract*—Cognitive Radio Networks (CRNs) provide a convenient way to access under-utilized TV bands for wireless radio users. Different co-located networks must have to coexist by accessing different parts of the incumbent free available spectrum in an opportunistic manner. In this paper, we propose a fully distributed non-cooperative game theoretic scheme for resource scheduling in self-coexisting cognitive radio networks. We formulate the downlink resource scheduling problem in self-coexistent CRNs as a non-linear convex optimization problem, which is NP-hard in real world network scenario. An alternate greedy algorithm is developed that produces sub-optimal solution to the resource scheduling problem by dividing it into downlink channel allocation and power assignment sub-problems. We propose a novel utility function for the game to achieve high throughput and an algorithm to guide through the Nash Equilibrium state. Theoretical proof for the Nash Equilibrium has been presented and performance improvements compared to the state-of-the-art works have been depicted through simulation studies.**

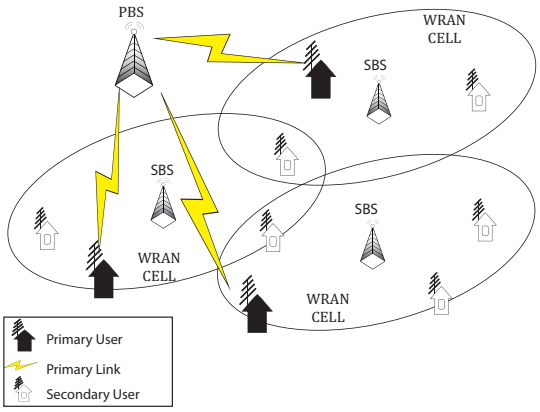


Fig. 1. IEEE 802.22 network architecture

I. Introduction

Cognitive Radio (CR) [1] is a cutting edge technology to increase efficient spectrum utilization between many radio users. It is widely considered as a promising technology to deal with the spectrum shortage problem caused by the current inflexible spectrum allocation policy. Cognitive Radio enabled devices sense the environment, learn from history, make intelligent decisions based on gathered environmental knowledge and historic results. This enables the CR devices to use the unused frequency spectrum efficiently without disturbing the licensed users. The licensed users are considered as primary users (PU) and unlicensed users i.e. cognitive radio enabled devices are considered as secondary users (SU).

CR-OFDMA is a multiple access technique refers to Orthogonal Frequency Division Multiple Access for CR networks. It enables the CR devices to maintain connection in vacant sub-carriers and turn off the sub-carriers where PU is active. Thus OFDMA technology provides a wider range and controlled access over spectrum bands occupied by the primary incumbents. For its dynamic and adaptive nature, it has been adopted by the popular next generation networks like LTE, WiMAX etc [5]. The IEEE 802.22 wireless regional area network (WRAN) is considered as the first wireless standard based on CR-OFDMA technology [15] that operates over the licensed TV bands (54MHz-862MHz). It is an infrastructure based WRAN, where a base station (BS) controls stationary wireless subscribers called customer premise equipments (CPE)(Figure-1). Each BS forms a single wireless network (CR cell) and it manages the medium access in its own WRAN.

Significant research studies on primary-secondary spectrum etiquettes [1] efficiently characterize primary incumbents activities in different regulatory domains. IEEE 802.22 WRAN entities use Geo-location databases with spectrum sensing techniques for white space tracking. These research studies mostly ignored or partially covered the secondary-secondary spectrum etiquettes like maintaining quality of service (QoS) or fairness between secondary users within co-located CR network cells. Self-coexistence refers to the ability of ensuring interference free transmission among neighboring homogeneous CR network cells. The IEEE 802.22 standard proposes time division multiple access (TDMA) based self-coexistence mechanisms like - *Dynamic Resource Renting and Offering (DRRO)*, *Adaptive On Demand Channel Contention (AODCC)* [15] etc. DRRO suffers with recursive contention, where AODCC removes channel contention by dividing time frames between different networks. Areas with densely populated primary incumbents, open channels have higher demands for different CR networks, which leads to multiple similar CR networks to acquire same channel at the same time. Therefore, efficient dynamic channel access among co-located CR

978-1-5090-0203-0/16/$31.00

network cells is necessary for reducing interference.

Graph coloring based resource allocations like Sengupta et. al. [13] requires centralized spectrum manager. In real life, co-located CR cells can be operated by different service providers where centralized spectrum manager doesn't possible. In [14] [6], they switch between channels based on expected utility cost and model it with *Modified Minority Game (MMG)* and *Potential Game* respectively. Though they improve channel utilization, maintaining power between CR cells and ensuring intra-cell fairness are ignored. Kaushik et. al. [2] propose joint power and bandwidth allocation in IEEE 802.22 base cognitive LTE networks, again their work also limited by a centralized spectrum manager. Therefore, an efficient distributed approach for resource allocation (channel, power etc.) ensuring QoS requirements and intra-cell fairness for the IEEE 802.22 WRAN is needed for next generation networking.

In this paper, we address the power control problem in co-located uncoordinated cognitive radio networks (Figure-1). The network cells can be partially or fully overlapped that compete for the spectrum resources and try to find out efficient power allocation to spectrum bands ensuring minimized interference from other coexisting network cells. Here, throughout the network a narrow-band interference free common control channel (CCC) is used for transferring control messages between BSs and CPEs. The objective is to gain intra-cell fairness by maintaining quality of service in CPEs. Minimized global power consumption and increased throughput in CPEs are also considered. Our game theoretic power allocation scheme aims to achieve better channel utilization ratio while minimizing power loss of the base stations.

The key contributions of our work is that, we formulate the self-coexistence problem as a non-linear convex optimization problem that suggests neighboring base stations provide network resources to secondary users in non-interfering manner. Then, we provide a greedy channel allocation strategy and develop a game theoretic model for power allocation with utility function for each network cell, based on network throughput and power consumption. We also propose best response function for each step of the game considering best QoS, protection of primary incumbents and achieving minimum SINR requirement of each SUs. We provide algorithm based on the game model with best response function, which can guide each base stations toward the Nash Equilibrium point. We also provide illustrative figures of our simulation results and show that our greedy approach can achieve expected throughput as well as minimum overall power consumption in each network cell.

The rest of the contents of this paper are as follows. In section II existing works on this literature is reviewed. The network model is analyzed and problem is formulated In section III. In section IV the proposed solution methodology and the game model are described. In section V performance of proposed scheme is evaluated by simulation results, followed by conclusion in section VI.

II. Related Works

The first cognitive radio based wireless regional area network standard IEEE 802.22 [15] proposes self-coexistence protocols and algorithms with its initial standard definitions. It proposes two methods for network cells to discover each other - Network entry - initialization and Normal Operation. The self-coexistence mechanism in IEEE 802.22 standard uses Self-Coexistence Window (SCW) and the Co-existence Beacon Protocol (CBP). These are included in a communication protocol which implements beacon transmissions among network cells. The IEEE 802.22 inter-BS coexistence mechanism [15] consists of four stages: spectrum etiquette, interference-free scheduling, dynamic resource renting and offering (DRRO), and adaptive on-demand channel contention (AODCC).

There have been a lot of works to improve the self-coexistence mechanism of IEEE 802.22 standard. In [13] [7] they address the self-coexistence problem in graph coloring perspective. They turn network cells into nodes, interference among network cells into vertex, form the graph and solve the graph as a graph coloring problem where channels are considered as colors. Authors of [12] treat channel allocation as optimization problem and use optimization algorithms such as integer linear programming to solve objective function.

Sengupta et al. [14] propose distributed game theoretic solution. Here network cells are game player who play a non-cooperative game where each cell try to reach into nash equilibrium point by channel switching. In [6], they use a physical interference model to identify overlapping network cells and model a potential game for channel assignment.

For downlink channel allocation in cognitive environment, authors in [4] consider general cognitive networks where neighboring cells use partial frequency reuse (PFR) scheme to access channel. In [8] they propose channel/power allocation with global knowledge of active CPEs by *Dynamic Interference Graph Allocation (DIGA)* and with local knowledge of active CPEs by *Two-Phase Resource Allocation (TPRA)*. Although they provide embedded primary incumbent parameters, their methodology lacks fairness of assignments and also needs primary users cooperation.

III. Problem Definition

A. System Model

We consider a multicell CR-OFDMA network as shown in Fig-1, where N number of cognitive network cells coexist, represented by $\mathcal{N} \equiv \{1, 2, \ldots, N\}$. Total frequency spectrum band (6MHz TV Band) is divided into K separate orthogonal sub-channels with equal length bandwidth B. In each network cell, there is a Base Station (BS) serving stationary Customer User Premises (CPEs) opportunistically to use one of the K sub-channels. There are total C CPEs in the total area where each CPE $c \in \mathcal{C}$ can initiate one or more sessions. Here, a session is defined as an unique connection demand of CPE which will be fulfilled by the BSs. Each CPE $c \in \mathcal{C}$ in cell $n \in \mathcal{N}$ can maintain multiple sessions at a time and each session

$s \in \mathcal{S}_c^n \equiv \{1, \ldots, S_c^n\}$ can operate on any of the channels. Each channel $k \in \mathcal{K} \equiv \{1, \ldots, K\}$ will not be assigned to more than one sessions in the same network cell at a time. It is considered that a standard mechanism [9] is used for ensuring common control channels (CCC) are being contention free. Also a spectrum sensing scheme [16] is used at the physical layer and the results obtained from it are assumed to be correct.

In this paper, we only consider the downlink channel (from BS to CPE). A BS searches for incumbent free channels and assigns downlink sub-channels to CPEs. CPEs with more need of frequency bandwidth may initiate multiple sessions and thus gain access of more bandwidth. A list of major mathematical symbols is given in Table I.

TABLE I
MAJOR NOTATIONS

Symbol	Definitions
$\mathcal{S}_c^n$	Set of sessions initiated by CPE $c \in \mathcal{C}$ under network $n \in \mathcal{N}$
$\mathcal{S}_c$	Set of session initiated by CPE $c \in \mathcal{C}$
$\mathcal{S}^n$	Set of session under BS $n \in \mathcal{N}$
$\mathbf{S}$	Set of all sessions in total vicinity area. $\bigcup_{n \in \mathcal{N}} \mathcal{S}^n$
D	$\{d_k^s\}_{S^n \times K}$ where d_k^s is a binary indicator having '1' if downlink channel $k \in \mathcal{K}$ is assigned to session $s \in \mathbf{S}$ and '0' otherwise.
$h_{s,k}^n$	Channel gain from BS n to session s in channel k. $n \in \mathcal{N}, s \in \mathbf{S}, k \in \mathcal{K}$
p_k^n	Transmission power of BS n in channel $k \in \mathcal{K}$.
P^n	$\{p_k^n\}_{K \times 1}$ Transmission power matrix for BS $n \in \mathcal{N}$.
p_{max}^n	Maximum power level of BS $n \in \mathcal{N}$.
$\mathbf{P}$	Power allocation vector $\{P^1, P^2, \ldots, P^N\}$.
N_0	Average Gaussian noise power.
γ^s	Minimum SINR required to achieve a certain bit error rate (BER) performance for each session.
σ_k^s	Calculated SINR value of a session s in channel k. $s \in \mathbf{S}, k \in \mathcal{K}$
θ^s	Quality of Service requirement (in terms of throughput per bandwidth) for session $s \in \mathbf{S}$.
q_k^s	Throughput of session $s \in \mathbf{S}$ in channel $k \in \mathcal{K}$.
I_k^s	Interference inflicted in session $s \in \mathbf{S}$ in channel $k \in \mathcal{K}$.

B. Downlink Channel Allocation and Power Control

Each CPE periodically senses the available channels and shares the sensing results to its BS. For any power allocation vector $\mathbf{P}$ interference measured at any CPE (session) s in channel k is,

$$I_k^s = \sum_{a=1, a \neq n}^{N} h_{s,k}^a \; p_k^a, \qquad \forall s \in \mathcal{S}^n, \forall k \in \mathcal{K}, \tag{1}$$

where, n is the BS to which the session s is registered. Received signals from other BSs occur interference if BS n chooses to transmit in channel k. The Eq.1 gives us total interference in channel k sensed by session s. Now, we can calculate the achieved Signal-To-Interference-plus-Noise-Ratio (SINR) as follows,

$$\sigma_k^s = \frac{h_{s,k}^n \; p_k^n}{I_k^s + N_0}. \tag{2}$$

Each CPE sends this sensing information (σ_k^s) to its associated BS, which then calculates expected throughput by Shannon Capacity theorem,

$$q_k^s = Blog_2(1 + \sigma_k^s). \tag{3}$$

A BS also calculates the expected maximum achievable throughput for each session $s \in \mathcal{S}^n$ if it uses the maximum transmission power p_{max}^n in channel k,

$$Q_k^s = Blog_2(1 + \frac{h_{s,k}^n \; p_{max}^n}{I_k^s + N_0}). \tag{4}$$

Now, we generate the relative throughput as,

$$R_k^s = \alpha \frac{q_k^s}{Q_k^s} - \beta \frac{p_k^n}{p_{max}^n}. \tag{5}$$

The relative throughput represents the relationship between power and throughput in each session. After normalizing throughput and power with coefficients α and β, respectively, it gets the difference between them. Since our objective is to maximize the throughput with minimum transmission power, we take the difference from the normalized value of throughput to normalized value of power. Here, increment of α with constant β value increases the throughput ratio which emphasizes on throughput by compromising power. Conversely, reduction of α with constant β value compromises throughput and it tends to go negative with minimum power increment. On the other hand, higher β value with constant α emphasizes on power by compromising throughput. Lower β value with constant α eliminates power increment effect on the relative throughput. Therefore, α, β act as a control parameters for each CPE by which we can control throughput and power effects on the relative throughput equation.

Then the problem of downlink channel allocation and power control can jointly be formulated as follows:

$$\max_{p_k^n} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}^n} \sum_{k \in \mathcal{K}} R_k^s \tag{6}$$

subject to:

$$0 \leq p_k^n \leq p_{max}^n, \qquad \forall k \in \mathcal{K}, \; \forall n \in \mathcal{N} \tag{7}$$

$$\sigma_k^s \geq \gamma^s, \qquad \forall s \in \mathbf{S}, \; \forall k \in \mathcal{K} \tag{8}$$

$$log_2(1 + \sigma_k^s) \geq \theta^s, \qquad \forall s \in \mathbf{S}, \; \forall k \in \mathcal{K} \tag{9}$$

$$d_k^s \in \{0, 1\}, \qquad \forall s \in \mathbf{S}, \; \forall k \in \mathcal{K} \tag{10}$$

$$\sum_{k \in \mathcal{K}} d_k^s \leq 1, \qquad \forall s \in \mathcal{S}^n, \; \forall n \in \mathcal{N} \tag{11}$$

$$\sum_{s \in \mathcal{S}^n} d_k^s \leq 1, \qquad \forall k \in \mathcal{K}, \; \forall n \in \mathcal{N} \tag{12}$$

$$(\sum_{s \in \mathcal{S}^n} d_k^s) \; p_k^n \leq p_{max}^n, \qquad \forall k \in \mathcal{K}, \; \forall n \in \mathcal{N} \tag{13}$$

Here, BSs from all network cells will gather relative throughput values R_k^s from each session s for each channel k. The objective function in (6) is to maximize the total sum of relative throughputs for all sessions and all channels. That is, it always tries to choose the session which achieves maximum throughput with minimum power increment.

Transmission power of a cell is bounded by the maximum transmission power of BS in constraint (7). We take separate

variables for each BS without loss of generality, although maximum transmission power can be the same for all BSs. Constraint (8) and (9) ensure the quality of service (QoS) in each session. Constraint (10) is a domain constraint for channel allocation. Constraint (11) depicts that at most one channel can be assigned to a session. Constraint (12) limits channel assignment to at most one session under a BS. These constraints ensure intra-network interference and eliminates over assignment problem hence ensuring fairness. Constraint (13) limits the power constraint only to assigned sessions.

Now, it is easy to visualize that the above problem belongs to the class of non-linear convex optimization problem. For single channel and fewer number of CPEs, this can be solvable in linear time by converting it to a sequence of approximating linear programs [3]. For highly co-located CPEs and higher number of channels, this problem turns into NP-hard. Again, we need a global coordinator for accommodating and running the optimization solver. Therefore, distributed approach is an urgent need to solve this convex optimization problem.

C. Game Theoretic Definitions

In wireless networking game theory is a common mathematical tool for resource scheduling and analysis of interactive decision processes [10]. In game theory, there are 3 basic components:

- Player set $\mathcal{N}$.
- Strategy space $\mathbf{A} = \Pi_{i\in\mathcal{N}}\mathcal{A}_i$, where $\mathcal{A}_i$ is a set of strategies available to player $i \in \mathcal{N}$.
- Set of utility functions, $\mathcal{U} = \{u_i(\mathbf{a})\}$, where $u_i(\mathbf{a})$ depicts player i's utility under a strategy profile $\mathbf{a} \in \mathbf{A}$.

A strategy profile is depicted as $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ where $a_i \in \mathcal{A}_i$ and $a_{-i} = \mathbf{a} - a_i$ is the strategy profile of all players except player i and also $\mathbf{a} = (a_i, a_{-i})$. Each player chooses the best strategy profile from their available strategy space under the belief that other players also choose their best. The collection of all players' best strategies form a steady state at which no player has any incentive to deviate from the strategy that is different from their best available ones. This steady state is called Nash Equilibrium [11].

Definition III.1. A strategy profile $\mathbf{a}^* = (a_i^*, a_{-i}^*)$ is a Nash Equilibrium if

$$u_i(\mathbf{a}^*) \geq u_i(a_i, a_{-i}^*), \quad \forall a_i \neq a_i^*, \forall i \in \mathcal{N}.$$

From Definition III.1, no player can improve its utility by deviating from a Nash Equilibrium. To converge into Nash Equilibrium point each player finds its best strategy from its available strategy space with the help of Best Response function.

Definition III.2. $BR_i(a_{-i})$ is the Best Response Function of player i if

$$BR_i(a_{-i}) = \{a_i : u_i(a_i, a_{-i}) \geq u_i(a_i', a_{-i})\} \quad \forall a_i' \neq a_i.$$

The best response function gives the best (in terms of highest utility) strategy given all possible strategies s_i from other players.

Definition III.3. A strategy profile $\mathbf{a}^* = (a_i^*, a_{-i}^*)$ is a Nash Equilibrium if

$$a_i^* = BR_i(a_{-i}^*) \quad \forall i \in \mathcal{N}.$$

This gives us an approach for finding Nash Equilibrium (NE) by only calculating the best response for each player, then it finds a strategy by definition III.3.

IV. Proposed Solution Architecture

We propose a distributed approach to solve the convex optimization problem stated in Eq. 6. Our algorithm aims to achieve the following goals: (i) the transmission power should be allocated in such a way that both the inter-cell and PU interferences are minimized; (ii) efficient channel allocation to maximize system throughput; (iii) fairness should be achieved by all CPEs in a cell; and, (iv) QoS requirement is achieved in each session. We develop a *Distributed Self-coexistence and non-cooperative Power allocation Game* (DSPG) that divides the NP-hard problem of Eq. 6 into two sub-problems: greedy channel assignment and power allocation with global knowledge of active CPEs.

For channel assignment, we have used a greedy approach which assigns channel to a session having maximum relative throughput.

$$\{d_k^s\}_{|\mathcal{S}^n|\times K} = \begin{cases} 1, & \text{if } s = \arg\max_s R_k^s \\ 0, & \text{otherwise} \end{cases} \quad s \in \mathcal{S}^n, k \in \mathcal{K} \tag{14}$$

We formulate the power allocation problem as a non-cooperative game, where, network BSs are the players of the game. Each network tries to find out the optimal power allocation for the selected CPEs.

The game running in each BS $n \in \mathcal{N}$ is defined as $\mathcal{G}(\mathcal{N}, \mathbf{P}, \{U_n\})$. Here, $\mathbf{P} = \{P^1, P^2, \ldots, P^N\}$ is the power allocation vector, $P^n = \{p_k^n\}_{1\times K} = \{1\}_{1\times|\mathcal{S}^n|}.\{p_k^s\}_{|\mathcal{S}^n|\times K}$, where, p_k^s depicts the transmission power level of BS $n \in \mathcal{N}$ for session $s \in \mathcal{S}^n$ in channel $k \in \mathcal{K}$. A BS runs the game for a channel $k \in \mathcal{K}$, if it does not sense any PU signal on the channel and $\sigma_s^k \geq \gamma_s,\ s \in \mathcal{S}^n$. This constraint is added to remove hidden incumbent problem. During data transmission, if the condition $\sigma_s^k \geq \gamma_s$ is failed or the BS senses PU signal on channel k, it restrains all CPEs transmissions and re-initiated the game and channel allocation procedure. After achieving NE for any channel, the corresponding channel is assigned to the session with maximized utility value $U_n(\mathbf{P})$.

A. The Utility Function

As stated above, each BS $n \in \mathcal{N}$ performs as a player of the game and tries to maximize its own objectives independently defined in Eq. 6. Therefore, the utility function corresponding to objective in Eq. 6 is defined as follows,

$$U_n(\mathbf{P}) = \sum_{s\in\mathcal{S}^n}\sum_{k\in\mathcal{K}} R_k^s. \tag{15}$$

Now, it is easy to visualize that, if any session experiences higher interferences, then the BS power for the session should

be increased so as to allow the session to achieve the minimum throughput. On the other hand, interference-minimized sessions would gain higher throughput for minimum power. In each iteration of the game, a BS tries to maximize its utility by allocating optimal power for its sessions. Given the above utility function in Eq. 15, the BSs eventually converge to a Nash Equilibrium after several iterations (under certain conditions).

The strategy profile for player $n \in \mathcal{N}$, denoted as $\mathbf{P} = (P^n, P^{-n})$, where, $P^{-n} = (P^1, P^2, \ldots, P^{n-1}, P^{n+1}, \ldots, P^N)$. The payoff to player n for choosing strategy profile $\mathbf{P}$ is $U_n(P^n, P^{-n})$. At NE point, each player $n \in \mathcal{N}$ should satisfy the following condition,

$$U_n(P^{n,*}, P^{-n,*}) \geq U_n(P^n, P^{-n,*}). \tag{16}$$

B. Game Formulation

Given the power based utility function in Eq. 15, each BS $n \in \mathcal{N}$ selects its power vector P^n to maximize its $U_n(P^n, P^{-n})$. The non-cooperative game is formulated as:

$$\max_{P^n} U_n(P^n, P^{-n}), \quad \forall n \in \{1, 2, \ldots, N\}, \tag{17}$$

subject to the constraints in Eq. 7 to Eq. 13.

Theorem 1. The unique solution of the game defined as $G(\mathcal{N}, \mathbf{P}, \{U_n\})$ establishes a Nash Equilibrium.

Proof: According to fixed point theorem in game theory, two conditions must be satisfied for the existence of Nash Equilibrium in a game [11]:

- The strategy space $\mathbf{P}$ should be non-empty, compact and convex subset of certain Euclidean Space.
- The payoff function $U_n,\ n \in \mathcal{N}$ should be continuous in $\mathbf{P}$ and quasi-concave in P^n.

Here, strategy space is a set of power values. In utility function, we consider the assigned transmission power in channels as the control variable, which is bounded by 0 and p^n_{max}. So, it is definitely non-empty, closed and bounded, i.e., a compact and convex set.

According to Eq. 15, U_n is a linear combination of p^n_k in P^n. Therefore, U_n is continuous in $\mathbf{P}$. Furthermore, it is trivial to show that U_n is also concave on its strategy set. ■

C. Best Response Function Formulation

For a given channel assignment matrix D and power allocation matrix $\mathbf{P}$, we can get the objective function for each BS $n \in \mathcal{N}$ as follows,

$$M = \max_{p^n_k} \sum_{s \in \mathcal{S}^n} \sum_{k \in \mathcal{K}} R^s_k, \tag{18}$$

subject to the constraints in Eq. 7 to Eq. 13 and we can rewrite the SINR constraint in Eq. 8 as,

$$p^n_k \geq \frac{\gamma^s (I^s_k + N_0)}{h^n_{s,k}} = c^s_k, \tag{19}$$

and the QoS constraint in Eq. 9 as

$$p^n_k \geq \frac{(I + N_0)(2^{\theta^s} - 1)}{h^n_{s,k}} = d^s_k. \tag{20}$$

Algorithm 1 Iterative Power Allocation

```
Initialize P = 30, D = 0
l ← 0
while l ≤ MaxIter do
    l ← l + 1
    Set the backoff timer for a random time interval
    while The backoff timer expires do
        The countdown of backoff timer.
    end while
    Send Beacon Signal in available channels.
    Get sensing results ∀k ∈ K, ∀s ∈ S^n
    Estimate D using Eq. 14
    Calculate Interference using Eq. 1
    Calculate P^n using best response function in Eq. 23
    if ||P^n_l − P^n_{l−1}|| / P^n_{l−1} < ω then
        Break.
    end if
end while
```

Therefore, the Lagrangian of Eq. 18 can be written as,

$$\begin{aligned} L(\mathbf{P}, \mu) = & M - \sum_{k \in \mathcal{K}} \lambda_k (c^s_k - p^n_k) \\ & - \sum_{k \in \mathcal{K}} \mu_k (p^n_k - p^n_{max}) \\ & - \sum_{k \in \mathcal{K}} \rho_k (d^s_k - p^n_k), \end{aligned} \tag{21}$$

where, λ_k, μ_k, ρ_k are the Lagrangian multipliers (non-negative real numbers). The Karush-Kuhn-Tucker (KKT) conditions for user i are given by,

$$\begin{aligned} \nabla_{p^n_k} M + \lambda_k - \mu_k + \rho_k &= 0, \\ \lambda_k (p^n_k - c^s_k) &= 0, \\ \mu_k (p^n_{max} - p^n_k) &= 0, \\ \rho_k (p^n_k - d^s_k) &= 0, \\ \lambda_k, \mu_k, \rho_k &\geq 0. \end{aligned} \tag{22}$$

From these conditions, we can easily derive the best response function for a base station $n \in \mathcal{N}$ as follows,

$$BR_n(P^{-n}) = P^n = \{p^n_k\}_{1 \times K}$$

$$p^n_k = \left[\frac{\alpha p^n_{max}}{\beta ln(2) log_2\left(1 + \frac{h^n_{s,k} p^n_{max}}{I^s_k + N_0}\right)} - \frac{I^s_k + N_0}{h^n_{s,k}} \right]^+ \tag{23}$$

Each Base Station runs Algorithm. 1 to achieve the Nash Equilibrium point. Here a backoff timer has been used to synchronize BSs in the game. When backoff timer expires, each BS receives sensing results from its CPEs, estimates D using Eq. 14, assigns channels and calculates the best response function. ω is used as a control parameter for terminating the iteration.

V. Performance Evaluation

We present representative simulation results for the proposed scheme, compare its performance with related state of the art works. We use Network Simulator 3 (NS-3) as the simulation tool and design corresponding IEEE 802.22 modules (channels, ports, users) for the cognitive simulation environment. We consider an area of total $110 \times 110km^2$ in which the IEEE 802.22 networks are deployed. The service area is further divided into $N = 7$ overlapping network cells each of which $30km$ in radius. A BS is deployed at the center of the network and deploy random number of CPEs ranging from 5 to 70 across the entire region using uniform distribution. Also there is placed random number of PUs across the area. The transmission range of each network is chosen randomly between $25km$ to $30km$.

We measure PUs effect on convergence cost and show the result in Figure-2. Here we take 7 network cells with 40 CPEs across the network. We present a comparison between our proposed approach DPSG with *Modified Minority Game (MMG)* [14] and *Queue Based Control (QBC)* [2]. In this case we can see that QBC converges slowly than MMG. Queue Based Contention mainly focuses on reducing delay, loss and total transmission power by compromising iteration time. For increased number of PUs, its branch-and-bound (B&B) algorithm creates more branches to fathom, which leads to increased iteration number. In modified minority game, it switches channel based on the expected cost incurred in each iteration. It needs several iterations only for collecting sensing information in different channels. Although for lower number of PUs it requires more iterations, performance improves for higher number of PUs. While in DSPG BS collects sensing information for all channels from each CPE, which helps to decide the perfect channel for next iteration based on our greedy approach. Therefore it shows average performance irrespective of the number of PUs.

We measure the average throughput per CPE in Figure-3. Here we can see that with 15 PUs, increasing the number of BS can increase average throughput since, more CPEs will be covered by BS and intra-cell collision reduces. Throughput decreases in terms of increased number of CPEs for same number of networks, because more CPEs will produce more network demands and more chance for inter-cell interference. Also we notice that some of the CPEs get fewer chance to access the bandwidth and thus average throughput reduces.

TABLE II
Simulation Parameters

Parameter	Value
Nodes Mobility	fixed
Area	$110km^2$
Operational Channel	470MHz - 476MHz
Bandwidth	6 MHz
Duplexing Method	TDD
Modulation Type	16-QAM
Coding Rate	1/2
Transmission Range	25km to 30km
Sensing Range	50km to 75km
Maximum Transmission Power	40 dBm
Propagation Loss Model	Friis
Noise Model	AWGN
Path Loss Exponent	2
Maximum Number of BS	7
Maximum Number of CPE	70
Maximum Number of PU	25

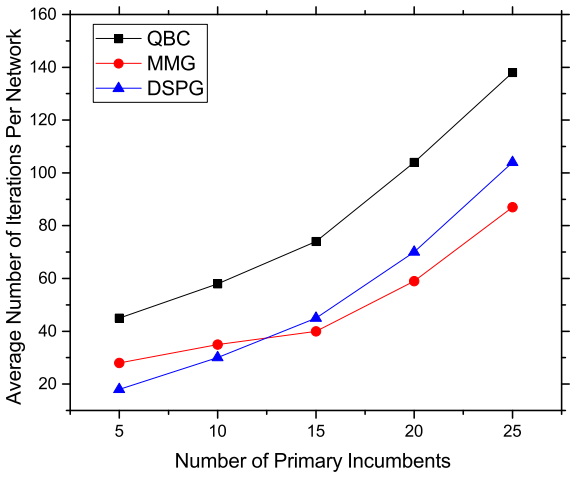


Fig. 2. Iterations needed for varying PUs

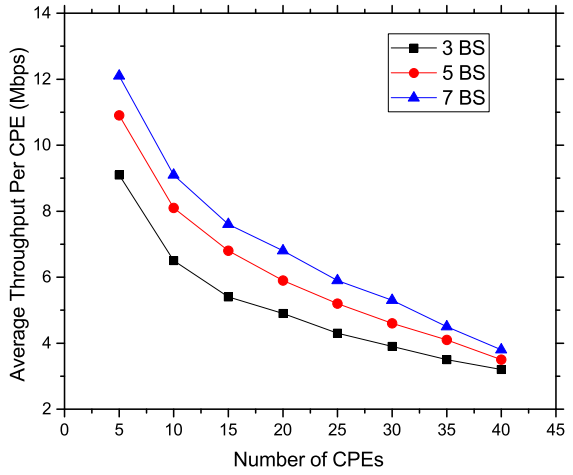


Fig. 3. Average Throughput in CPEs

We also show transmission power values in Figure-4 after achieving Nash Equilibrium for all channels in each BS. Here we choose 3 BS with 7 channels and 7 CPEs where maximum CPEs are resided in overlapping regions. We can see that, BS puts higher transmission power value for CPEs residing in lower interference regions. BS puts lower transmission power for CPEs residing in severe interference regions, since providing higher transmission values will be wastage of power.

VI. Conclusion

We have formulated the resource scheduling problem in self-coexistent cognitive networks as a non-linear optimization problem and further subdivided the problem as greedy user selection and game theory based near-optimal power assignment

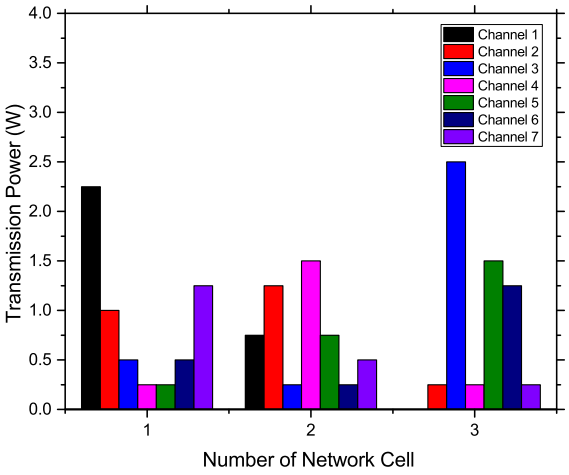


Fig. 4. Transmission Power in Different Base Stations

problem. We have used a greedy user selection equation based on their relative throughput. After choosing sessions for CPEs, we have used iterative water-filling algorithm to find out near-optimal power allocation through competitive non-cooperative game. Our solution method doesn't need primary users co-operation and need no message passing between neighbor networks. We also provide theoretical proof of our game framework. Our best response function ensures live updating of assigned power based on network environment. We also provide extensive simulation using NS-3 and hence provide practical improvements of our game framework compared to the state-of-the-art works.

ACKNOWLEDGEMENTS

This work is supported by a grant for the Research Fellowship (2013-2014) funded by the Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh. Dr. Md. Abdur Razzaque is the corresponding author of this paper.

REFERENCES

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[2] A. Asheralieva and K. Mahata, "Joint power and bandwidth allocation in IEEE 802.22 based cognitive LTE network," *Computer Networks*, vol. 71, pp. 117–129, 2014.

[3] M. Avriel, *Nonlinear programming: analysis and methods*, ser. Prentice-Hall series in automatic computation. Prentice-Hall, 1976. [Online]. Available: http://books.google.com.bd/books?id=815RAAAAMAAJ

[4] K. W. Choi, E. Hossain, and D. I. Kim, "Downlink subchannel and power allocation in multi-cell OFDMA cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 7, pp. 2259–2271, 2011.

[5] M. Ergen, *Mobile Broadband - Including WiMAX and LTE*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[6] V. Gardellin, S. K. Das, and L. Lenzini, "A fully distributed game theoretic approach to guarantee self-coexistence among wrans," in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*. IEEE, 2010, pp. 1–6.

[7] Y. Ge, J. Sun, S. Shao, L. Yang, and H. Zhu, "An improved spectrum allocation algorithm based on proportional fairness in cognitive radio networks," in *Communication Technology (ICCT), 2010 12th IEEE International Conference on*. IEEE, 2010, pp. 742–745.

[8] A. T. Hoang and Y.-C. Liang, "Downlink channel assignment and power control for cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 8, pp. 3106–3117, 2008.

[9] B. F. Lo, "A survey of common control channel design in cognitive radio networks," *Physical Communication*, vol. 4, no. 1, pp. 26 – 39, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874490710000406

[10] A. MacKenzie, L. DaSilva, and L. DaSilva, *Game Theory for Wireless Engineers*, ser. Synthesis lectures on communications. Morgan & Claypool Publishers, 2006. [Online]. Available: https://books.google.com.bd/books?id=iMCkbGnbVWsC

[11] M. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994. [Online]. Available: http://books.google.com.bd/books?id=5ntdaYX4LPkC

[12] T. W. Rondeau and C. W. Bostian, *Artificial intelligence in wireless communications*. Artech House, 2009.

[13] S. Sengupta, S. Brahma, M. Chatterjee, and N. Sai Shankar, "Self-coexistence among interference-aware IEEE 802.22 networks with enhanced air-interface," *Pervasive and Mobile Computing*, vol. 9, no. 4, pp. 454–471, 2013.

[14] S. Sengupta, R. Chandramouli, S. Brahma, and M. Chatterjee, "A game theoretic framework for distributed self-coexistence among IEEE 802.22 networks," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*. IEEE, 2008, pp. 1–6.

[15] C. R. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. J. Shellhammer, and W. Caldwell, "IEEE 802.22: The first cognitive radio wireless regional area network standard," *Comm. Mag.*, vol. 47, no. 1, pp. 130–138, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1109/MCOM.2009.4752688

[16] E. Z. Tragos, S. Zeadally, A. G. Fragkiadakis, and V. A. Siris, "Spectrum assignment in cognitive radio networks: A comprehensive survey." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1108–1135, 2013.

A Genetic Algorithm for Virtual Machine Migration in Heterogeneous Mobile Cloud Computing

Md. Mofijul Islam, *Student Member, IEEE* and
Md. Abdur Razzaque, *Senior Member, IEEE*
Green Networking Research (GNR) Group
Department of Computer Science and Engineering
University of Dhaka, Dhaka-1000, Bangladesh.
Email: akash.cse.du@gmail.com, razzaque@du.ac.bd

Md. Jahidul Islam
Department of Computer Science and Engineering
United International University, Bangladesh
Email: jahid@cse.uiu.ac.bd

***Abstract*—Mobile Cloud Computing (MCC) improves the performance of a mobile application by executing it at a resourceful cloud server that can minimize execution time compared to a resource-constrained mobile device. Virtual Machine (VM) migration in MCC brings cloud resources closer to a user so as to further minimize the response time of an offloaded application. Such resource migration is very effective for interactive and real-time applications. However, the key challenge is to find an optimal cloud server for migration that offers the maximum reduction in computation time. In this paper, we propose a Genetic Algorithm (GA) based VM migration model, namely GAVMM, for heterogeneous MCC system. In GAVMM, we take user mobility and load of the cloud servers into consideration to optimize the effectiveness of VM migration. The goal of GAVMM is to select the optimal cloud server for a mobile VM and to minimize the total number of VM migrations, resulting in a reduced task execution time. Additionally, we present a thorough numerical evaluation to investigate the effectiveness of our proposed model compared to the state-of-the-art VM migration policies.**

I. Introduction

Cloud Computing is an emerging technology, where users can offload their computing-intensive tasks from local resource-constrained computing platforms to a more resourceful computing platform. Due to the availability of Mobile Devices (MD) and unprecedented advancement of wireless technologies, Mobile Cloud Computing (MCC) [6], [9], [12] has emerged as the most effective way to utilize cloud resources for executing resource-hungry mobile applications.

As MDs are resource constrained, the use of MCC virtually increases their capacity to accommodate applications that demand rich computational resources. Using MCC, MDs can offload these applications to a resourceful cloud server for faster execution [2], [3], [7]. However, long distance between cloud servers and the MDs may increase the response time for interactive applications resulting in an increase in the total execution time. To alleviate this problem, *cloudlet* is proposed in [13]. Cloudlet is a resourceful local cloud that brings remote cloud resources closer to the mobile user. By offloading tasks in the nearest cloudlet, the user can decrease total task execution time.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

Virtualization technology introduces a middle layer between the hardware and software layer in a cloudlet. This allows us to share the hardware resources by means of VM. Resources (CPU, memory, network bandwidth, etc.) in a cloudlet are provisioned to these VMs. Resource provisioning in cloud computing is a well-studied area [1], [4], [5], [10], [15], [16]. However, mobility in MCC introduces several challenges to maintain an acceptable Quality of Service (QoS) level when provisioning cloud resources. A mobile user can move from one Access Point (AP) to another AP and therefore can increase the distance between its current location and the cloudlet, where the task is provisioned. As a result, it increases the task execution time. To address this issue, we propose a VM migration technique for heterogeneous MCC system following the user mobility pattern. That is, when a user moves from one cloudlet to another cloudlet, the resource or VM must be migrated to the cloudlet that is nearest to the user.

Let us consider the following scenario: a blind user is executing an application that takes an image from his surroundings. Then, the application processes the image in the cloudlet and gives a response to the user's local client. That is, it is the type of application where a user continuously uploads some data and the cloud server processes this data to provide response back to the user. Now, if the blind user moves away from the current cloudlet, then he will experience a delayed response from the mobile application executing in the cloudlet. Therefore, the overall performance of the application will also degrade. To avoid this performance degradation, it is necessary for the system to adopt a VM migration method to choose a cloudlet for migrating the VM that is currently closer to the user.

User mobility is not the only reason that forces a VM to migrate. Increased load at the cloud server can also initiate a VM migration. VM migration due to the cloud serve load occurs when a cloud server provisions more VM than their actual physical resources. When a VM is under-provisioned, it increases the task execution time, as it gets lower resource than its capacity. VM migration is an effective approach for maintaining task execution deadline for mobile users. It is

more critical when a user executing an interactive application in cloud, where interaction time between the user and the remote cloud resource are low. There are few research work conducted on VM migration in MCC [8], [11], [14], [17]. However, these methods are more suited for static task execution. A VM migration model that effectively considers user mobility and the load of the cloudlet is yet to be proposed in the literature.

In this work, we propose a VM migration (VMM) model based on not only the user mobility but also the load of the cloudlet. The objective is to select the optimal cloud server for a mobile VM in addition to minimizing the total number of VM migrations, resulting in a reduced task execution time. We use Genetic Algorithm (GA) to identify the optimal target cloudlet. The main contributions of our work are given bellow:

- A Genetic Algorithm (GA) based VMM model is proposed which allows us to identify the candidate cloud servers for the provisioned VM. This feature of our model helps to maximize the total utility of the MCC system.
- User mobility is considered for selecting the optimal cloudlet for a VM so that the VM can be provisioned to a cloudlet closer to user's current location.
- A VM will be migrated to a cloudlet which is not overloaded. This feature enables a cloud server to maintain a proper distributions of cloud resources among the provisioned VM.

The remaining paper is organized as follows: in Section II, we discuss on the state-of-the-art work on VM or task migration in MCC. Next, in Section III, we present the network model and assumptions. Next to that, in Section IV, we present the problem formulation of our VM migration model. Subsequently, in Section V, we present a numerical evaluation of our proposed model to investigate the effectiveness of our proposed model compared to the state-of-the-art models. Finally, we draw the conclusions and mention our future work in Section VI.

II. Related Work

In MCC, user can offload (part of) a task in cloudlet where cloud allocates resources to VM for executing that task. There are several methodologies used in the literature to detect when an user should offload task in a cloudlet [1]–[3], [7], [10], [15], [16]. However only a few works have been conducted in the field of VM migration for MCC systems. A Traffic aware cross-site VM migration model is proposed in [11]. In this model, when multiple VMs are needed to be migrated, an arbitrary sequence of VM migration congests the inter-site links bandwidth and thus reduces the number of successful VM migration. Then, the VM migration problem is formulated as a Mixed Integer Linear Programming (MILP) problem and a heuristic algorithm is used to get approximate optimal result.

Besides, a mobility induced service migration for MCC is proposed in [17]. In this work, a threshold based optimal service migration policy is developed where the VM migration problem is modeled as a Markov decision process (MDP). This proposed system considered that the mobile users follow a one-dimensional asymmetric random walk mobility model. A service is migrated from one micro-cloud to another when a user is in states bounded by a set of predefined thresholds.

In addition, three lightweight task migration models are developed in [8]:

- *Cloud-wide task migration* where the task migration decision is made by central cloud which maximize the cloud provider objectives
- *Server-centric task migration* where all the migration decision is made by the server where the task is currently executing
- *Task based migration* where the migration is initiated by the task itself

In this approach, the migration decision is made after each decision epoch. After each epoch, the migration controller makes decision on possible cloudlets, based on the user mobility and remaining task execution time. This proposed method considers the *increasing data volume transfer time* during the task migration from one cloud to another cloud.

In summary, most of the VM migration methodologies do not effectively consider the user mobility with the overloaded condition of a particular cloudlet server in heterogeneous MCC. Therefore, it increases the service downtime, especially for those applications where the user interaction interval between the user mobile application and the remote cloud is very low. Besides, when we migrate a VM, we have to consider the load of the target cloudlet as well. If we do not consider the load of cloudlet, then the total number of VM migration will be increased. Also, we have to consider the user mobility because if the distance between the user current location and resource increases, then it also increases the service response time. To the best of our knowledge, we are the first to utilize efficient searching methodology of Genetic Algorithm (GA) to optimize VM migration model considering user mobility and load of the cloudlets for a heterogeneous MCC system.

III. Network Model

We assume a Mobile Cloud Computing (MCC) environment where a set of M APs make the backbone network. Each AP is connected to any of the L cloudlets, denoted as $S = \{S_1, S_2, S_3.....,S_L\}$. We present a sample network scenario in Fig. 1. We assume that multiple access points may have common cloudlet. These cloudlets are connected using the backhaul network where bandwidth between the cloudlet i and j is denoted as B_i, j.

All the cloudlets are monitored by the master remote cloud. A cloudlet server i has fixed processing power C_i and memory M_i. Each cloudlet provisions P number of VMs, denoted as $V = \{V_1, V_2, V_3.....,V_P\}$. User can offload their task to a dedicated VM. We further assume that each user is mobile and execute a task in different VMs and there are no inter-dependency among those VMs. In the remaining of this paper, we use task and VM interchangeably.

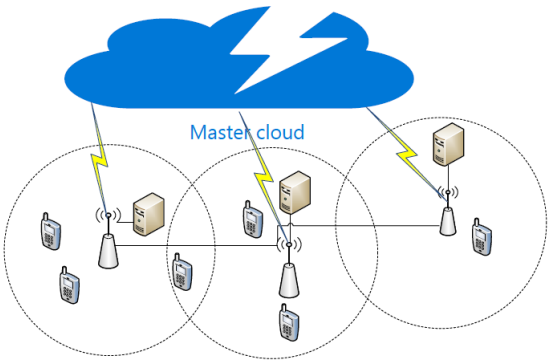


Fig. 1. Network Model

IV. GAVMM: Genetic Algorithm based Virtual Machine Migration

In VM migration of a cloudlet, we have to remap the VMs to new cloudlets so that the task execution and interaction time is reduced. This problem is NP-hard problem [18] and thus, there are no algorithm that can provide a guaranteed optimal solution in polynomial time. For this reason, we use a methodology that can ensure near-optimal solutions in a bounded time. Evolutionary algorithms such as GAs are known to be very successful in such cases. However, to ensure good performance of GA, we need to choose suitable genetic operators and tune its parameters perfectly.

In Fig. 2, we present our algorithm, i.e., GA based VM Migration (GAVMM). In the following sub-sections, we elaborate different modules of GAVMM.

A. Chromosome Representation

To use GA for VMM problem, we have to first code the problem into a appropriate form. In classical GA, each population contains a set of individuals which is represented by a set of chromosomes. Each chromosome contains a fixed number of bits. In our problem, each solution of VMM is represented by a chromosome $C = \{C_1, C_2, C_3 ....., C_N\}$ of length N. Each bit in a chromosome represents a VM and the value of this bit is used to identify the target cloudlet where the VM can be migrated. The length of chromosome N is equal to the total number of task in the cloudlet.

B. Population Initialization

We generate initial population G_p which evolves in each generation of GA. First, we generate the accessible set A_k^i for each VM k in cloudlet i. A_k^i represents the set of all access points from which user U_k is getting wireless signal. Then, the cloudlet calculates the accessible probability $P_{k,j}^A$ for each VM k in AP j, that represents *how likely user can attach to another access point*. We calculate $P_{k,j}^A$ using the following equation:

$$P_{k,j}^A = \frac{A_{k,j}^i}{\max_{\forall j \in A_k^i} A_{k,j}^i} \quad (1)$$

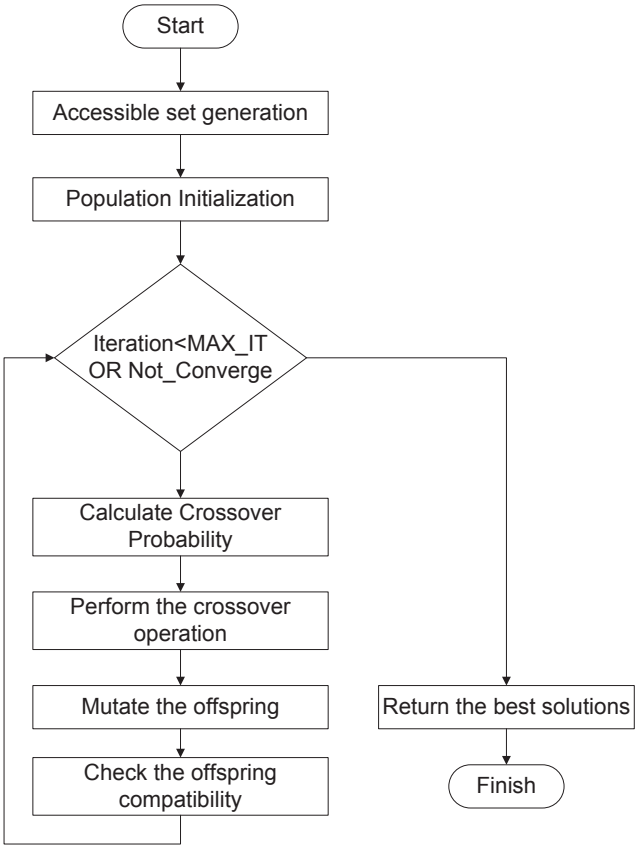


Fig. 2. GAVMM Workflow

We use Algorithm 1 to generate the initial population where we use the previous best solution as a seed of the current solution to increase the converge speed.

Algorithm 1 Population generation, at cloudlet S_i

INPUT: Accessible set for each VM and the previous best solution.
OUTPUT: Initial number of population.
Calulates the $P_{k,j}^A$ for each VM k in AP j
$\beta_b \leftarrow$ previous best solution.
if (previous best solution is available) **then**
    Use the β_b as a seed with the access probability $P_{k,j}^A$.
    If any user $U_k \in \beta_b$ is not present in current then we can not use seed for that user.
    AP used in β_b will get higher access probability.
    Using the rotation strategy select a AP for each VM in the chromosome gene.
    In this rotation strategy, AP with higher access probability has higher chance for getting selected.
else
    Generate the population using just rotation strategy.
end if

C. Fitness Function

We have to evaluate all the individuals in a population using a Fitness Function (FF). We use the following fitness function in our model:

$$F_n = \sum_{\forall_{k \in C}} U^i_{n,k} \tag{2}$$

Here, F_n is the fitness value for n^{th} individual and U^i_k is the utility of VM k, which is defined by the following equation:

$$U_{n,k} = J^i_{n,k} + T^i_{n,k} + R^i_{n,l} \tag{3}$$

Here,

$J^i_{n,k}$ is the minimum normalized task completion time of VM k for the n^{th} individual when we migrate the VM from cloudlet i to another cloudlet j. When we do not migrate a VM i, it will be equal to j. We calculate $J^i_{n,k}$ using the following equation:

$$J^i_{n,k} = \min_{\forall_{j \in A_{k,i}}} \left\{ \frac{J^{i,j}_{n,k}}{\max_{\forall_{j \in A_{k,i}}, t \in G_p} J^{i,j}_{t,k}} \right\} \tag{4}$$

Besides, $T^{i,j}_{n,k}$ VM transfer time of VM k for the n^{th} individual from the cloudlet i to cloudlet j. When cloudlet i is the same as cloudlet j, the VM transfer time will be equal to zero, that is $T^{i,i}_k = 0$. We calculate minimum normalized VM transfer time $T^i_{n,k}$ using the following equation:

$$T^i_{n,k} = \min_{\forall_{j \in A_{k,i}}} \left\{ \frac{T^{i,j}_{n,k}}{\max_{\forall_{j \in A_{k,i}}, t \in G_p} T^{i,j}_{t,k}} \right\} \tag{5}$$

In an interactive application the data footprint increase with the task execution time, so it enables the cloudlet to determine the VM migration. $R^{i,j}_{n,l}$ is the VM load correlation degree of l type VM between the cloudlet i and cloudlet j, which helps the system to predict the load of the cloudlet. We calculate minimum normalized VM load correlation $R^i_{n,l}$ using the following equation:

$$R^i_{n,l} = \min_{\forall_{j \in A_{k,i}}} \left\{ \frac{R^{i,j}_{n,l}}{\max_{\forall_{j \in A_{k,i}}, t \in G_p} R^{i,j}_{t,l}} \right\} \tag{6}$$

To calculate normalized $J^i_{n,k}, T^i_{n,k}$, and $R^i_{n,l}$ we use min-max normalization technique.

D. Genetic Operators

In this section we discuss about the genetic operators. After the calculation of fitness value for each individual we have to select two individuals for the crossover operation. At beginning of the selection, the selection probability P^S_n for each individual n is calculated using the following equation:

$$P^S_n = \frac{F_n}{\sum_{\forall_{t \in G_P}} F_t} \tag{7}$$

We use a rotation strategy to select two individuals for crossover operation. In this selection, an individual with higher P^S_n has higher chance for getting selected. After selecting two individuals, we perform mutation operation on the new offspring with the mutation probability P^m. After that, we alleviate offspring which are not compatible with VM's accessible set. Then, we replace an individual from G_P with lowest fitness by the offspring with the highest fitness value. To maintain the stochastic property of VMM problem, we replace an individual from G_P with the offsprings such that the offspring fitness values are greater than the individual fitness values. The whole procedure is listed in Algorithm 2.

Algorithm 2 Genetic Algorithm based VM Migration

INPUT: Fitness Function.
OUTPUT: Remapped cloudlet for each VM.
$G_P \leftarrow Population Initialization()$
while (iteration $<$ MAX_IT and Not_Converge) **do**
  Select two individual based on P^S_n
  Perform the crossover operation.
  Alleviate the incompatible offspring.
  $\alpha_b \leftarrow$ Best offspring.
  $\alpha_r \leftarrow$ Random offspring.
  $\beta_w \leftarrow$ Worst individual in G_P
  $\beta_r \leftarrow$ Random individual in G_P which fitness is less than α_r.
  Replace β_w and β_r in G_P with α_b and α_r respectively.
  Repeat with iteration $\leftarrow$ iteration +1
end while
Return the best individual.

Using GAVMM, we get the best solution β_s which remapped the all the VM or task to other cloudlet. If cloudlet for VM k in β_s is not equal to the previous cloudlet for this VM, then we initiate a VM migration event. However, if the cloudlet for VM k in β_s is equal to the previous cloudlet for this VM but the AP is different then we just initiate a wireless AP handover event. Otherwise, we run the VM k in the current cloudlet.

V. Numerical Evaluation

In this section, we assess the performance of our proposed GAVMM methodology for VM migration. To evaluate the performance we compare our GAVMM method with three VM migration methods, No Migration, Load-centric migration and Gready Migration.

- In *No VM Migration* method, the system does not migrate any VM in any situation. Although the cloudlet can be overloaded or user can move to another AP, the system does not initiate any VM migration event. In this approach, cloudlet which provisions the VM, executes the task and forwards the result through the intermediate APs. Thus, the user can experience a longer response time or interaction time and reduces the user experience.
- On the other hand, *Gready VM Migration* method migrates a VM based on the wifi signal strength received by the user. In this approach, system initiates a VM migration for VM to a cloudlets from which an user is getting highest signal strength. However, it does not

consider the load of the cloudlet to initiate VM migration. Therefore, it increases the total number of migrations for overloaded cloudlets.

- Load centric VM migration is a server centric migration methodology where each server selects a task and migrates to a target server to improve the task execution time. During this migration process a target server is chosen based on the load of the target server and it uses single VM migration approach.

A. Simulation Environment

To conduct simulation for evaluating the performances of GAVMM, we used ten cloudlets, each having [1.5-3.0] GHz processor, [8-16] GB memory and [250-350GB] SATA hard disks. All the cloudlets are attached with one or multiple access points and the cloudlets are interconnected with [2-20] Mbps Ethernet communication link. Users access the cloudlet resources using mobile devices. Each user device is connected with cloudlet AP through WiFi IEEE 802.11g interface. The parameters used in Genetic Algorithm are listed in Table V-A.

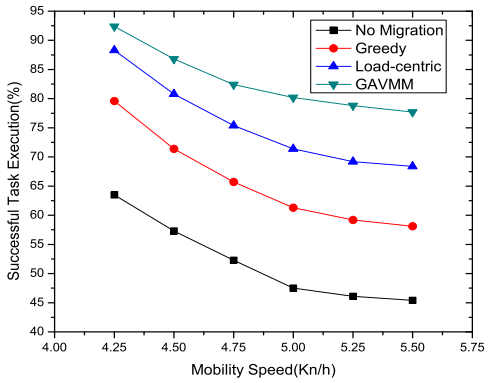
TABLE I
GENETIC ALGORITHM PARAMETERS

Parameters	Value
Population size	80
Maximum iteration MAX_IT	1200
Mutation probability	0.01
Simulation Time	1200 Seconds

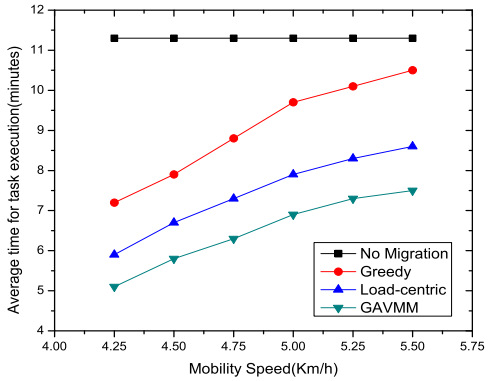
B. Simulation Result

We study the performance of our proposed GAVMM method by varying the user mobility. In simulation we only consider pedestrian walking model.

1) Impacts of user mobility: We measure the successful task execution and the average task execution time by varying the user mobility ranging from 4.25Km/h to 5.5Km/h in pedestrian walking model. The graphs in Fig. 3(a) show that the successful task execution decreases with the user mobility speed. The successful task execution decrease rapidly in no-migration and greedy approach as it do not consider the user mobility and the target server load. On the other hand load-centric VM migration shows better result compared to no-migration and greedy approach, as it takes account the load of the server. The GAVMM approach shows better performance compared to other approach because it jointly migrated a set of VMs from the current cloudlet to the target cloudlets. All the state-of-the-art VM migration approach uses single VM migration methodology, but the GAVMM approach uses a joint-multiple VMs migration approach. As a result the overall performances of all VMs are improved and the total number of successful task execution is also increased. GAVMM approach is also considering user mobility during the selection of VMs to be migrated and thus effectively reduce the task execution time.



(a) Successful task execution



(b) Average task execution Lifetime

Fig. 3. Impacts of user mobility

2) Impacts of average time for task execution: We study the performance of GAVMM on total number of VM migration by varying the average time for task execution ranging from 3 to 18 minutes. The graphs in Fig. 4 state that the total number of VM migration rapidly increase with average task execution time in greedy VM migration approach, because greedy approach does not consider either load of the server or the user mobility to select the appropriate cloudlet for each VM. Load-centric VM migration gradually increases the total number of VM migration as it only consider the load of the server and it uses the single VM migration approach. Compared to these VM migration approach total number of VM migration in GAVMM gradually goes to steady state.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have developed a mobility and load aware virtual machine migration algorithm for heterogeneous mobile cloud computing systems. The performance improvement that can be achieved if a VM migration policy explores both the user mobility and load of the cloudlet in lieu of exploiting any one of them has been thoroughly studied. We develop a genetic algorithm based VM migration mechanism, called GAVMM, which effectively brings the cloud resources close to the user as well offers load-balancing among the cloudlet computing resources. The proposed GAVMM serves as a theoretical basis for the investigation of other VM migration methods in heterogeneous MCC.

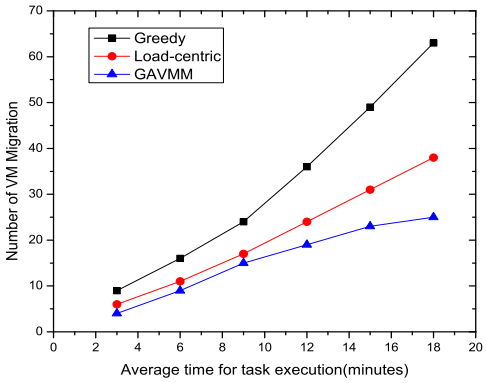


Fig. 4. Impacts of average time for task execution on total number of VM migration

Currently, we are working on how to further minimize the task execution time through a light weight optimization framework that facilitates joint VM migration, i.e., to allow consolidation of many VMs that are highly non-correlated on a single cloudlet computing resources.

ACKNOWLEDGEMENTS

This work is supported by a grant for the Research Fellowship (2013-2014) funded by the Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh. Dr. Md. Abdur Razzaque is the corresponding author of this paper.

REFERENCES

[1] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in *Integrated Network Management, 2007. IM '07. 10th IFIP/IEEE International Symposium on*, May 2007, pp. 119–128.

[2] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: Elastic execution between mobile device and cloud," in *Proceedings of the Sixth Conference on Computer Systems*, ser. EuroSys '11. New York, NY, USA: ACM, 2011, pp. 301–314.

[3] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: Making smartphones last longer with code offload," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010, pp. 49–62.

[4] A. Das, T. Adhikary, M. Razzaque, and C. S. Hong, "An intelligent approach for virtual machine and qos provisioning in cloud computing," in *Information Networking (ICOIN), 2013 International Conference on*, Jan 2013, pp. 462–467.

[5] A. K. Das, T. Adhikary, M. A. Razzaque, E. J. Cho, and C. S. Hong, "A qos and profit aware cloud confederation model for iaas service providers," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC '14. New York, NY, USA: ACM, 2014, pp. 42:1–42:7.

[6] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[7] L. Gkatzikis and I. Koutsopoulos, "Migrate or not? exploiting dynamic task migration in mobile cloud computing systems," *Wireless Communications, IEEE*, vol. 20, no. 3, pp. 24–32, June 2013.

[8] ——, "Mobiles on cloud nine: Efficient task migration policies for cloud computing systems," in *Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on*, Oct 2014, pp. 204–210.

[9] A. Khan, M. Othman, S. Madani, and S. Khan, "A survey of mobile cloud computing application models," *Communications Surveys Tutorials, IEEE*, vol. 16, no. 1, pp. 393–413, First 2014.

[10] J. Li, K. Bu, X. Liu, and B. Xiao, "Enda: Embracing network inconsistency for dynamic application offloading in mobile cloud computing," in *Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing*, ser. MCC '13. New York, NY, USA: ACM, 2013, pp. 39–44.

[11] J. Liu, Y. Li, D. Jin, L. Su, and L. Zeng, "Traffic aware cross-site virtual machine migration in future mobile cloud computing," *Mobile Networks and Applications*, vol. 20, no. 1, pp. 62–71, 2015.

[12] M. Rahimi, J. Ren, C. Liu, A. Vasilakos, and N. Venkatasubramanian, "Mobile cloud computing: A survey, state of art and future directions," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 133–143, 2014.

[13] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, Oct 2009.

[14] T. Taleb and A. Ksentini, "An analytical model for follow me cloud," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 1291–1296.

[15] H. N. Van, F. Tran, and J.-M. Menaud, "Sla-aware virtual resource management for cloud infrastructures," in *Computer and Information Technology, 2009. CIT '09. Ninth IEEE International Conference on*, vol. 1, Oct 2009, pp. 357–362.

[16] L. Wang, F. Zhang, A. V. Vasilakos, C. Hou, and Z. Liu, "Joint virtual machine assignment and traffic engineering for green data center networks," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 3, pp. 107–112, Jan. 2014.

[17] S. Wang, R. Urgaonkar, T. He, M. Zafer, K. Chan, and K. Leung, "Mobility-induced service migration in mobile micro-clouds," in *Military Communications Conference (MILCOM), 2014 IEEE*, Oct 2014, pp. 835–840.

[18] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Full Papers

Communication over Ad-hoc Networks

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Fault Tolerant Optimized Broadcast for Wireless Ad-hoc Networks

Mamtaj Akter
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
Email: mamtaj@ari.buet.ac.bd

Alimul Islam
Department of Computer Science and Engineering
University of Information Technology & Sciences
Dhaka-1212, Bangladesh
Email: faysal.alimul@gmail.com

Ashikur Rahman
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
E-mail: ashikur@cse.buet.ac.bd

***Abstract*— Many broadcasting protocols for ad-hoc wireless networks perform poorly in situations when network becomes unreliable or untrusted. In real environment, any node may suddenly become unavailable due to power failure or mobility, or be unable to receive or forward broadcast packets due to noise/interference. To improve the reliability of broadcasting in an environment with limited trust, a protocol dubbed as *multicover dominant pruning* (MDP) [2] has been proposed where all nodes cover their 2-hop neighbors multiple times. However, MDP incurs too much redundancy; the same broadcasting task can be achieved with reduced number of packet forwarding. In this paper, we propose three better fault-tolerant heuristics dubbed as multicover total dominant pruning (MTDP), multicover partial dominant pruning (MPDP) and multicover improved dominant pruning (MIDP). All these heuristics not only utilize 2-hop neighborhood information more effectively to reduce redundant transmissions but also significantly improve the reachability of nodes in an unreliable ad-hoc wireless network. Extensive simulation experiments have been conducted to evaluate the efficiency of the proposed heuristics. Performance analysis shows that the proposed heuristics outperforms the multicover dominant pruning broadcasting algorithm.**

Keywords—wireless ad-hoc network, broadcast, fault-tolerance, broadcasting protocol, reachability, redundancy.

I. INTRODUCTION

Wireless ad hoc networks are formed by wireless hosts without requiring any pre-existing infrastructure where hosts may be mobile in nature. In such networks, each nodes act not only as senders or receivers but also as routers or forwarders. Routes between any two nodes may potentially contain multiple hops. Ad hoc networking is rightly called an "art of networking without network". Due to its two key advantages namely ease and speed of deployment and decreased dependence on infrastructure, ad hoc networks are becoming popular now-a-days. The applications of ad hoc networks range from military use in battlefields, medical sectors, emergency operations, civilian environments, to personal area networking.

In a mobile ad hoc network, the task of broadcasting is to send a packet from a source node to all other nodes in the network. Such broadcasting task is a very fundamental operation. It is a one-to-all operation which creates a multi-hop scenario, where packets originated from the source host are received and further relayed by several intermediate hosts before reaching all the nodes in the network since all mobile hosts may not be within the transmission range of the sender host due to the limited radio power. Broadcasting has many applications in graph-related problems and distributed computing problems. It is also widely used to solve many network layer problems. Due to host mobility in wireless ad-hoc network, hosts require to perform broadcasting more frequently, e.g., for paging a particular host, sending an alarm signal and for service discovery as well. Broadcasting is also performed for finding a route to a particular host in several routing protocols and many other unicast routing protocols. Some protocols even use broadcasting for actual data transmissions.

A straightforward broadcasting method is blind flooding where each node forwards the broadcast packet exactly once whenever it receives the packet for the first time. Clearly, this costs n number of transmissions in a network with n hosts. In wired networks, the shortest path tree and *Steiner tree* [24] method are used to construct broadcast tree where each node broadcasts a packet to its neighbors whenever it receives the packet along the shortest path from the source node. The broadcast tree constructed by this blind flooding method becomes the shortest path tree in wired networks. Therefore, blind flooding is simple and relatively efficient if it is used in wired networks. Besides, in wired network, one transmission refers to one reception whereas in wireless networks, whenever a node transmits a packet, all of its neighbors residing in its transmission area receive the packet. Thus one transmission may cause multiple receptions as there may be several hosts within the transmission range of a host and it becomes worse when the omnidirectional radio propagation is used. So, although theoretically, such naive blind flooding ensures a complete coverage, in wireless networks it becomes very much resource-consuming and also causes serious redundancy, contention and collision. Collectively, these three problems are known as the *broadcast storm problem* [6].

Many broadcast algorithms have been proposed [1], [3-10] to overcome the shortcomings of blind flooding. These

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

algorithms utilize neighborhood and/or history information to reduce redundant transmissions. The *dominant pruning* (DP) algorithm [5] is one of the promising approaches that utilizes 2-hop neighborhood information to reduce redundant (re)transmissions. To reduce more redundancy, two improved algorithms named *total dominant pruning* (TDP) and *partial dominant pruning* (PDP) have also been proposed in [1]. Both of these algorithms utilize neighborhood information more effectively. However, all of these optimized algorithms have been proposed considering only the *ideal* environment. In real life scenario, a wireless ad-hoc network may operate (rather) in an unreliable environment where a node may fail to carry out its forwarding duties for various reasons, e.g., due to power failure, noise in the environment etc. Sometimes a node may intentionally un-cooperate other nodes or behave selfishly. The optimized algorithms such as PDP or TDP are not applicable in such unreliable environment as they cannot ensure reachability of a broadcast packet to every node in the network due to their "single-cover" characteristics. To address the problems of reachability, unreliable environment demands fault tolerant design of broadcasting protocols. To make broadcasting algorithms fault tolerant, we need to increase redundancy. So, the fault-tolerant broadcasting algorithm needs a tradeoff between increased fault tolerance and reduced redundancy. Many fault-tolerant protocols have been proposed in [14-22], but these are aimed at protecting unicasting routing protocols. For broadcasting, multicover dominant pruning (MDP) which is the first fault-tolerant broadcasting algorithm has been proposed in [2]. MDP improves the reachability of nodes in an unreliable ad-hoc wireless network by facilitating multiple receptions of data packets by the same host. Although MDP ensures fault tolerance, the number of packet forwarding is too large; sometimes it can be as large as the blind flooding.

In this paper, we propose three novel efficient fault-tolerant broadcasting algorithms which not only provide fault tolerance feature but also cut down the number of packet transmissions as well. To this end, we apply the concept of PDP and TDP to reduce redundant retransmissions and modify them to provide fault-tolerance. In particular, we propose three fault tolerant algorithms dubbed as *Multicover Total Dominant Pruning* (MTDP), *Multicover Partial Dominant Pruning* (MPDP) and *Multicover Improved Dominant Pruning* (MIDP). Here, we incorporate some tunable parameters in the design so that they can continue to operate in both ideal and real (unreliable) environments without any modifications. Simulation results of applying these three algorithms show performance improvements in terms of redundancy and reachability compared with the original multicover dominant pruning and single-covered broadcasting protocols respectively.

The rest of the paper is organized as follows: Section II discusses some related works on fault-tolerance and reducing broadcast redundancy. Three proposed fault-tolerant optimized broadcasting algorithms are described in Section III. Simulation results are presented in Section IV. Finally, Section V concludes the paper and outlines our future work.

II. Related Work

In this section, we will review some works related to the broadcasting problem. A significant number of works have been done to alleviate the redundancy problem, assuming the environment is ideal. These algorithms utilize neighborhood and/or history information to reduce number of forwarding. In [5], Lim and Kim prove that creating a minimum flooding tree is an NP-complete problem. The problem of constructing broadcast tree that minimizes the number of packet forwarding is very similar to the *minimum connected dominating set* (MCDS) problem [23]. Since this is an NP-complete problem, an approximation algorithm AMCDS proposed in [12]. But this approximation is not suitable for ad hoc wireless networks, as it is applicable only when each host knows the global network topology. Obtaining overall topology information in ad hoc wireless network is very costly and power hungry. Two approximation algorithms: *self pruning* and dominant pruning (DP) which can work using small topology information are proposed by Lim and Kim in [5]. Although self-pruning is winner in terms of overhead and complexity, DP algorithm is more promising heuristic due to its better optimization capability. Notations used in this paper to describe the heuristics are listed in Table I.

The dominant pruning (DP) algorithm which is one of the promising proactive approaches utilizes 2-hop neighborhood information to reduce redundant (re)transmissions. Each host can obtain neighborhood information through sending periodic "Hello" packets, each of which contains the sender node's id and the list of its neighbors. When the source node u generates a broadcast packet, it selects some of its neighbors for forwarding from $B(\emptyset,u)=N(u)$ to make sure that all of its exactly two-hop away nodes in $U(\emptyset,u)=N(N(u))-N(u)$ receives the packet. While creating forward node list, a node tries to engage as minimum number of neighbors as possible to cover all of its 2-hop neighbors and hence, it maps the *B-set* and *U-set* into the so-called *set cover problem*. For the selection of forwarding nodes, the *greedy set cover* algorithm [13] is used. The selected set of neighbors create the forward list *F*-set and the IDs of forwarding nodes are piggybacked in the packet header. After receiving the broadcast packet, the forwarding node that is requested to relay the packet determines a new forwarding list. When v receives the packet, it searches in the packet header whether it's id is in the $F(\emptyset,u)$ or not. If yes, at first it calculates *B-set* and *U-set* according to the DP algorithm. Among the nodes in $N(N(v))$, v does not need to cover them all since u is the source node, nodes in $N(u)$ have already received the packet, and nodes in $N(v)$ will automatically receive the packet after v rebroadcasts the

TABLE I. SUMMARY OF NOTATIONS

Notation	Description
u	Originator node (Sender)
v	Intermediate node (Receiver)
$N(u)$	Set of all one-hop neighbors of node u
$N(v)$	Set of all one-hop neighbors of node v
$N(N(v))$	Set of all one-hop and two-hop neighbors of v
$F(u,v)$	Set of forwarding nodes of v
$B(u,v)$	Set of nodes from which v will create the forwarding node set F(u,v)
$U(u,v)$	Set of uncovered neighbors that are exactly two-hop away from v
$S(u)$	Set of nodes that is covered by node u only once

packet. As a result, the uncovered two-hop neighbor set of v becomes $U(u,v) = N(N(v)) - N(u) - N(v)$. After that, v starts constructing its forwarding node list $F(u,v)$ from $B(u,v) = N(v) - N(u)$ to cover all the nodes in $U(u,v)$ so that each node of U-set can receive the data packet at least once. Initially, v sets $F(u,v)=\emptyset$ and $B(u,v) = N(v) - N(u)$. Then, in each iteration, v selects a node w from $B(u,v)$, such that $w \notin F(u,v)$ and w has the maximum number of neighbors in $U(u,v)$. Next, v inserts w in $F(u,v)$ and sets $U(u,v)= U(u,v)- N(w)$. When $U(u,v)$ becomes empty, the iteration stops.

For more optimization, two improved algorithms named total dominant pruning (TDP) and partial dominant pruning (PDP) have been proposed by Lou and Wu in [1]. Both of these algorithms utilize neighborhood information more effectively. To reduce redundancy, both of these algorithms focus to cut down the size of U-set which in turn minimizes the size of F-set. In Total Dominant Pruning (TDP) algorithm, each node piggybacks their 2-hop neighbor list in the header of the broadcast packet. When v gets $N(N(u))$ with the packet, v can easily deduct this set from its own 2-hop neighbor list $N(N(v))$ as u has already covered all the nodes in $N(N(u))$. Thus, $U(u,v)$ is reduced to $U(u,v)=N(N(v))-N(N(u))$. Although TDP algorithm shows a remarkable improvement compared to DP algorithm in terms of number of forwarding, it consumes more overhead as each node needs to piggyback their 2-hop neighborhood information with the data packet and therefore, the size of the broadcast packet increases. In terms of Partial Dominant Pruning (PDP) algorithm, unlike the TDP algorithm, no neighborhood information of the sender is piggybacked with the broadcast packet. However, this algorithm deducts all the nodes that are deducted in DP algorithm and besides this, some more nodes are excluded from the two hop neighbor set of v. Since the neighbors of the common neighbors of u and v exist in the two-hop neighbor set of u (i.e., $N(N(u)\cap N(v)) \subseteq N(N(u))$), this set can be excluded from $N(N(v))$. Therefore, the two-hop neighbor set U in PDP algorithm is $U(u,v)=N(N(v))-N(u)-N(v)-P$, where $P=N(N(u) \cap N(v))$. Among these three optimized broadcasting protocols DP is the best for the complexity whereas TDP is the most optimized heuristic, but in terms of overhead, PDP is the real winner although it has an additional computational cost of calculation of the P-set. Therefore, the PDP algorithm is considered as one of the promising optimized broadcast protocols for wireless ad-hoc networks and due to its less overhead requirement, it is applied in ad hoc wireless mesh networks for multimedia streaming applications [11].

The first fault tolerant broadcasting protocol, multicover dominant pruning (MDP) has been proposed in [2] by Rahman et al. MDP improves the reachability of nodes in an unreliable ad-hoc wireless network by ensuring multiple receptions of data packets. This algorithm works as like as the DP algorithm with one exception. Here, when v receives a broadcast packet from u, v at first calculates U-set and B-set like the DP algorithm. Here, like DP, $U(u,v)=N(N(v))-N(v)-N(u)$ and $B(u,v)=N(v)-N(u)$. Then it maps the B-set and U-set into the set multicover problem unlike the DP where set single cover mapping is used. Thus, node v starts constructing its own $F(u,v)$ such that each node in $U(u,v)$ is covered by at least m number of nodes in $F(u,v)$. Since $m = 2$ is the best choice according to [2], throughout the paper we will consider the matter of multicover to the case $m = 2$. When $m =2$, the MDP algorithm works like DP with double-cover which means nodes of U-set will receive data packet from at least two forwarding nodes. The ultimate idea behind this double-cover is that in an unreliable network, if a forwarding node fails to forward the data packet to the destination node, at least another forwarding node can do the job. For example, in a connected network, a source node w generates a broadcast data packet where $\{x, y\}$ and $\{z\}$ are w's exactly one hope and two-hop neighbors respectively. In double-covered broadcasting protocol like MDP, w will select both x and y as forwarding nodes to convey the data packet to z. Therefore, due to the unreliable network, if any node between x and y becomes uncooperative, at least there will be another forwarding node which can forward the packet to z. On the other hand, in the single-covered broadcasting protocols like DP, TDP and PDP, w will select only one node x or y to forward the packet to node z. If that only forwarding node becomes inactive for any reason, z will not get the packet. Thus, MDP improves the reachability of broadcast packets by ensuring multiple receptions of the same data packets. Although MDP ensures fault tolerance, it incurs too much redundancy.

III. Enhanced Multicover Dominant Pruning Algorithms

In this section, we propose three enhanced dominant pruning algorithms: the multicover total dominant pruning (MTDP) algorithm, the multicover partial dominant pruning (MPDP) algorithm and the multicover improved dominant Pruning (MIDP).

Our proposed algorithms exploit the multicover characteristics of MDP to ensure fault-tolerance but use the concept of the optimized broadcast protocols PDP and TDP to address the excessive redundancy problem of MDP. A generic algorithm of the three enhanced multicover dominant pruning broadcasting protocols is given in Algorithm 1. Here this algorithm presents an extension of the greedy algorithm introduced in [5] for finding multicover ($m=2$). The algorithm starts by creating the set $U(u,v)$ which will be varied for the individual heuristics. For example, in MDP, the set $U(u,v) = N(N(v))-N(u)-N(v)$. The forwarding node set $F(u,v)$, the set $S(v)$ (the set of nodes that are covered exactly once by the node v) and a temporary set Z are initialized with empty. Then, for each node w of the set $U(u,v)$, the algorithm will check that whether it exists in the set $S(u)$. If it exists, that node w will be marked as *1*, else it will be marked as *0*. If we sum up from step 7 to 12, in every iteration, it finds the node $w_k \in B(u, v)$ with the largest number of neighbors that exist in $U(u,v)$. Then in step 13, w_k is added to the forwarding set $F(u,v)$. After that, each node of the set C_k which is the set of neighbors of w_k is remarked: *0* to *1*, *1* to *2*. Mark *0* means that the node has not been covered yet, *1* indicates a single cover, and *2* stands for double cover. If it is marked as *2*, it is added to the set Z and deleted from all the sets C_i of K. Eventually when Z becomes equal to the set $U(u,v)$, the algorithm will find out the nodes of the set Z which are marked *1* in order to add them to the set $S(v)$ and then the process stops. If Z is still not equal to $U(u,v)$, it goes to step 10 again. Here, the two obtained sets $S(v)$ and $F(u,v)$ are piggybacked by node v in the broadcast packet before rebroadcasting.

A. Multicover Total Dominant Pruning (MTDP) Algorithm

Likewise MDP, the node v makes its F-set so that it can cover all the nodes in $U(u,v)= N(N(v))-N(u)-N(v)$ twice. In MTDP, it applies the minimization technique of TDP to cut down the number of forwarding. In particular, node v will exclude all 2-hop neighbors of its previous broadcasting node u from its U-set. Thus, node v will reduce the size of its U-set to $U(u,v)= N(N(v))-N(v)-N(N(u))$. However, if any node in $N(N(u))$ is covered by its previous broadcasting node u only once, the node v must try to cover that node one more time, even though it is in $N(N(u))$. If $S(u)$ is the set of nodes that is covered by u only once, u will piggyback this set in the broadcast packet. Then, the 2-hop neighbor set that needs to be covered by v's forward node list $F(u,v)$ is increased to $U(u,v)=N(N(v))-N(v)-(N(N(u))-S(u))$. Like MDP, the set $B(u,v)$ will remain same, $B(u,v)=N(v)-N(u)$. In Fig. 1, v and w both nodes are the neighbors of node u. When u broadcasts the packet, it piggybacks its single covered node set $S(u)$ in the header of the packet ($S(u)$ is the area with blue and purple color in the circle of v and w). After receiving the packet, v will construct its U-set from its two-hop neighbor set and also add the $S(u)$ nodes with the U-set. Here, v deducts its $B(u,v)$ from the U-set as they will receive the packet during the rebroadcasting. After all the calculations, the final U-set of node v will be as like as the blue colored area in the Fig. 1. The algorithm of MTDP remains the same as the generic algorithm for enhanced multicover dominant pruning (Algorithm 1) where in step 1, the set $U(u,v)=N(N(v))-N(v)-(N(N(u))-S(u))$.

Algorithm 1: Generic Enhanced Multicover Dominant Pruning

1. Create the set $U(u,v)$
2. $F(u,v)= \emptyset$, $S(v)=\emptyset$, $Z = \emptyset$.
3. For each node $w \in U(u,v)$do
4. If $w \in S(u)$
5. $Mark(w) \leftarrow 1$.
6. Else $Mark(w) \leftarrow 0$.
7. For each node $w_i \in B(u,v)$ do
8. Create the set C_i such that $C_i= N(w_i) \cap U(u,v)$.
9. Let $K = \{C_1,C_2,...,C_n\}$.
10. Suppose, C_k is the set such that,
11. $|C_k|=\max_{Ci \in K}\{|C_i|\}$
12. If $|C_k|=\emptyset$ then exit.
13. $F(u,v)=F(u,v)\cup\{w_k\}$
14. For each node $x \in C_k$ do
15. If $Mark(x) = 0$ then
16. $Mark(x) \leftarrow 1$
17. Else if $Mark(x) = 1$ then
18. $Mark(x) \leftarrow 2$
19. $Z = Z \cup\{x\}$
20. For each $C_i \in K$ do
21. $C_i= C_i-\{x\}$
22. $K = K -\{C_k\}$
23. If $Z = U(u,v)$
24. For each node $y \in Z$ do
25. If $Mark(y)=1$ then
26. $S(v)=S(v)\cup\{y\}$
27. Exit.
28. Otherwise go to step 10.

The extra cost of the MTDP algorithm is that 2-hop neighborhood information and the set of single-covered nodes of each sender are piggybacked in the broadcast packet. Therefore, it consumes more bandwidth than any other broadcast algorithms.

B. Multicover Partial Dominant Pruning (MPDP) Algorithm

In the multicover partial dominant pruning algorithm, likewise MDP, when a node v makes the list F of forwarding nodes, it will try to cover all the nodes in $U(u,v)= N(N(v))-N(u)-N(v)$ twice so that each node of U-set can receive the same data packet from two different forwarding nodes in F-set. However, like the concept of PDP broadcasting algorithm, the node v will exclude some more nodes from its U-set to reduce the redundancy of MDP. These nodes are the neighbors of each node in $X=N(u)\cap N(v)$. Thus, the node v will shrink the size of its U-set to $U(u,v)= N(N(v))-N(u)-N(v)-N(X)$. Nevertheless, if any node of $P=N(X)$-set exists in $S(u)$, the node v must try to cover that node one more time to ensure double cover. Then the 2-hop neighbor set that needs to be covered by v's forward node list F is increased to $U(u,v)=N(N(v))-N(v)-N(u)-(P-S(u))$. Like all other algorithms, the set $B(u,v)$ will remain same, $B(u,v)=N(v)-N(u)$. In Fig. 2, v and w both nodes are the neighbors of node u. $S(u)$ is the area with blue and purple color in the circle of v and w . The U-set of node v is the blue colored area in the Fig. 2. The algorithm of MPDP remains the same as the generic algorithm for enhanced multicover dominant pruning (Algorithm 1) where in step 1, the set $U(u,v)=N(N(v))-N(v)-N(u)-(P-S(u))$.

The MPDP algorithm does not increase the size of the broadcast packet (only single-covered node set of u is piggybacked with the packet header), compared with the other algorithms, it eliminates more redundant transmissions. The only additional computational cost for the MPDP algorithm is that each forward node v needs to calculate set P.

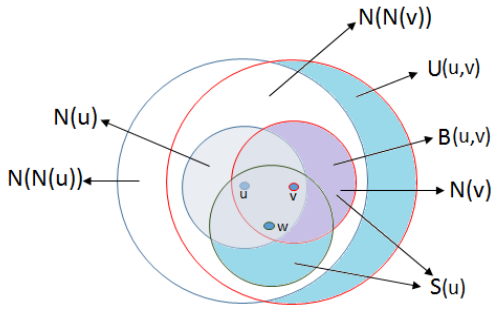


Fig. 1. Multicover Total Dominant Pruning Algorithm.

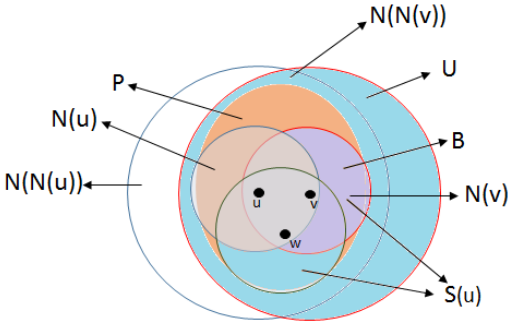


Fig. 2. Multicover Partial Dominant Pruning Algorithm.

C. *Multicover Improved Dominant Pruning (MIDP) Algorithm*

Before discussing MIDP, we need to clear out the concept of *Improved Dominant Pruning (IDP)*. At first without loss of generality let us assume that node *v*'s id is greater than node *w*'s id and both nodes have received the broadcast packet from *u* and both of these nodes are the forward nodes of *u*. Like the PDP algorithm, when *v* constructs its *U*-set, it will exclude *N(u), N(v), N(N(u)∩N(v))* from its two hop neighbor set. Moreover, it will exclude some more nodes. Node *v* will see that its id is greater than the other forwarding node *w*'s id, so it can deduct the nodes which are the neighbors of the common neighbors of both *w* and *v* as *w* must cover them. Therefore, the two-hop neighbor set *U* in this IDP algorithm is $U(u,v)=N(N(u))-N(u)-N(v)-P-Q$, where $P=N(N(u) \cap N(v))$ and $Q=N(N(v) \cap N(f_i))$, for each node $f_i \in F(\emptyset,u)$, where $F(\emptyset,u)$ is the forwarding list of *u*. *P* and *Q*, are the area with orange and yellow color respectively and blue colored area is the *U-set*, are shown in Fig. 3.

Fig. 4 illustrates our proposed third algorithm MIDP. Here, we suppose that node *w, z and v are the neighbors of node u and* node *v*'s id is the greatest among the other two and all of these three neighbors exist in the forwarding list of node u. Like MPDP, node *v* will at first exclude the neighbors of each node in $X=N(u)\cap N(v)$ from its *U*-set. Then MIDP will exclude some more nodes to minimize the number of forwarding further. Let, the three forwarding nodes *w, z* and *v* have id *1, 2* and *3* respectively. As *v* has the third highest id among all, *v* will assume that the two co-forwarding nodes *w* and *z* belonging lower id than *v* must cover the yellow colored area *Q* which represents the neighbors of the common neighbor of these three forwarding nodes and here $Q=N(N(w)\cap N(v)) \cap (N(z)\cap N(v))$. Thus, the size of the *U*-set of *v* reduced to $U(u,v)= N(N(v))-N(u)-N(v)-P-Q$ *where* $P=N(X)$. Here, both *w* and *z* which have the lowest ids must have to cover *Q* to ensure double cover. This is the special case of MIDP that

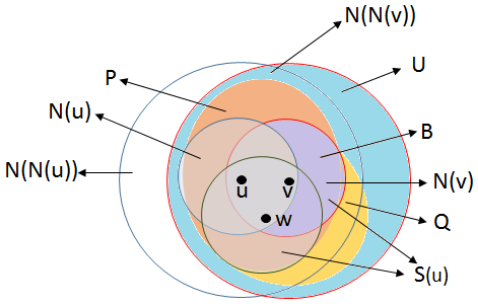


Fig. 3. Improved Dominant Pruning Algorithm.

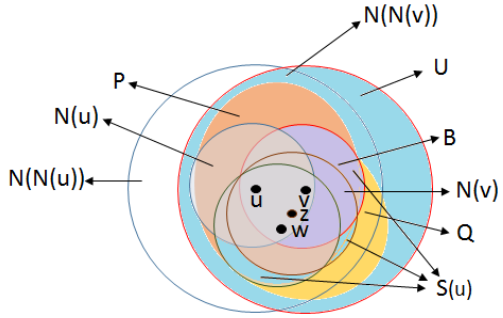


Fig. 4. Multicover Improved Dominant Pruning Algorithm.

after receiving the broadcast packet, when a forwarding node sees that in the forwarding list of the previous node, there are two nodes exist which have lower ids than itself, the node then start to calculate the *Q-set* to exclude from *U-set*. However, for fault-tolerance, if any node in *P* or *Q* exists in *S(u),* the node *v* must try to cover that node again. Then, the 2-hop neighbor set that needs to be covered by *v*'s forward node list *F* is increased to $U(u,v)= N(N(v))-N(u)-N(v)-(P+Q-S(u))$. Like MPDP and MTDP, the set *B(u,v)* will remain same, $B(u,v)=N(v)-N(u)$. The MIDP algorithm remains the same as the generic algorithm for enhanced multicover dominant pruning (Algorithm 1). The area within the circles of *v, z* and *w* which is not overlapped by each other in Fig. 4 represents the set *S(u)* and the blue colored area is *U(u,v)*.

The only extra cost of the MIDP algorithm is that the set of single-covered nodes of each sender is piggybacked in the broadcast packet. Therefore, like MPDP, it consumes less bandwidth than the MTDP algorithm.

IV. Simulation and Experiments

To evaluate the performance of the broadcasting algorithms DP, TDP, PDP and the fault tolerant broadcasting algorithm MDP and its three proposed enhancements MTDP, MPDP, MIDP, extensive simulation experiments have been conducted. We simulate all the broadcasting algorithms and perform a comparative analysis based on the simulation results. We also implement our proposed fault tolerant optimized broadcasting algorithms and compare the performance of the algorithms with the multicover dominant pruning algorithm. We build the simulation program using Java. The simulation is basically implemented in Network Layer. From Section II and III we can see that all the discussed broadcasting algorithms are functions of two network parameters, (1) transmission range and (2) node population. We can observe their impacts on the protocols, assuming ideal Medium Access Control (MAC) or Physical Layers. If the functions of the protocols would depend on the Physical Layer or Data Link Layer parameters such as, collision probability, bit error rate etc. then a simulation with more realistic environments (like ns-3) could be justified. However considering those event's effects on our proposed broadcasting algorithms are beyond the scope of this work.

TABLE II. DESCRIPTION OF SCENARIOS

Number of Nodes (N)	**Transmission Range (m) of each node**					**Number of scenarios**	**Total Scena-rios**
100	125	150	175	200	225	10	50
150	125		175		225	10	30
200	125		175		225	10	30
250	125		175		225	10	30
300	125	150	175	200	225	10	50
350	125		175		225	10	30
400	125		175		225	10	30
450	125		175		225	10	30
500	125	150	175	200	225	10	50

We simulate a network where all nodes are randomly distributed in a 625m X 625m square area. The population ofnode is varied by varying the number of nodes from 100 to 500 in the stated deployment area. Moreover, we vary the transmission range from 125m to 225m to analyze the effect of transmission range. Here, we only consider the connected networks for the simulation results. Table II describes the scenarios we have generated for the simulation. We take 10 random scenarios for each case. Performance measures are calculated as an average of these random samples. An example of a case can be a scenario of 100 nodes and the transmission range of each node is 125m.

In order to evaluate the performance of the proposed heuristics, we consider three kinds of performance aspects: *number of forwarding nodes, redundancy and reachability*. We use the first metric "number of forwarding nodes" to evaluate the proposed algorithms through the variation of number of total nodes and their transmission range. Then, to see the effect of node failures on dominant pruning and on our proposed enhancements, we consider the latter two metrics. Number of forwarding nodes can be defined as the total number of nodes who forward or rebroadcast the broadcast packet. The average ratio of the number of well-behaved nodes receiving the broadcast packets to the total number of well-behaved nodes in the network is defined as reachability. Higher reachability ensures high coverage in the network that means more well-behaved nodes in the network will receive the packet. The goal of every broadcasting algorithm is to ensure high reachability. This measure of reachability can also be defined as *Coverage*. Redundancy is formally defined as the percentage of all well-behaved nodes participating in rebroadcasting averaged over all broadcast packets. The main goal of every broadcasting algorithm is to reduce the redundancy. The less redundancy in the network the better it is.

Now we present the simulation results and comment on various aspects of performance measures. To validate the effectiveness of our proposed enhanced multicover broadcasting algorithms MPDP, MTDP, MIDP, we compare simulation results of single-covered broadcasting protocols DP, PDP, TDP and multicover broadcasting protocol MDP with the simulation results of our proposed algorithms. Measurements of all these algorithms are plotted in the same graph. At first, we will see the performance in terms of percentage of forwarding nodes by varying both total number of nodes and the transmission ranges. The simulation is conducted under the static environment defined earlier.

To determine the effect of various transmission ranges on the percentage of forward nodes, we compare our proposed three algorithms MPDP, MTDP, MIDP with DP, PDP, TDP and MDP. Starting from 125m to 225m, the transmission range of each node is varied by increasing 25m per step in each type of networks and we verify the result for node population n = 100, n = 300 and n = 500. From Fig. 5(a), (b) and (c) it can be clearly seen that TDP and PDP both substantially outperform DP. But in the case of double covered broadcasting algorithms, undoubtedly MTDP, MPDP and MIDP show better performance than MDP to a great extent. The percentages of forwarding nodes of these three enhanced broadcasting algorithms are very close. Yet, multicover improved dominant pruning algorithm shows better performance than the MTDP and MPDP in all the three graphs. What is more striking in all these three graphs is that the trends of all heuristics are obviously downwards with the increase of the transmission range. So, we can come to a conclusion that as the average node degree increases (the increase of transmission range of a node leads to the increase of the number of nodes that are within that node's transmitter range. The number of neighbors (m) indicates the node degree, $d = m-1$), the percentage of forwarding nodes drops gradually.

In Fig. 6(a), (b) and (c), we present the effect of node density on the percentage of forwarding nodes with various transmission ranges (125m, 175m and 225m) while keeping the deployment area constant. Starting from 100 to 500, the node density is varied by increasing 50 nodes in the network at each step and we examine the result for transmission ranges 125m, 175m and 225m. Like the previous experiment, MTDP, MPDP and MIDP show much better performance than MDP. Here, the average difference between MIDP and our proposed algorithms is around 20%. The average percentage of forwarding nodes of these three enhanced broadcasting algorithms are almost similar. Yet, MIDP overshoots the performance of both the MTDP and MIDP in all the three graphs. In sum, from Fig. 6 it can be said that as the network gets denser, the percentage of forwarding nodes in our proposed algorithms rises steadily even though erratic sometimes.

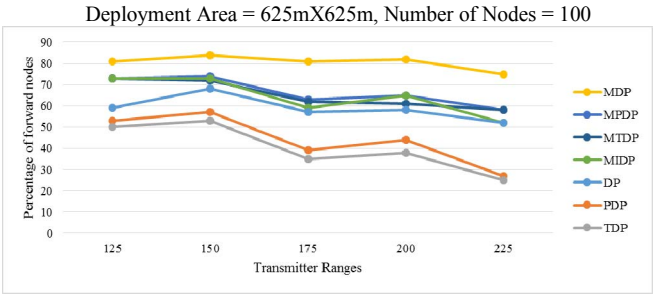


Fig. 5(a). Effect of transmitter ranges in sparse network.

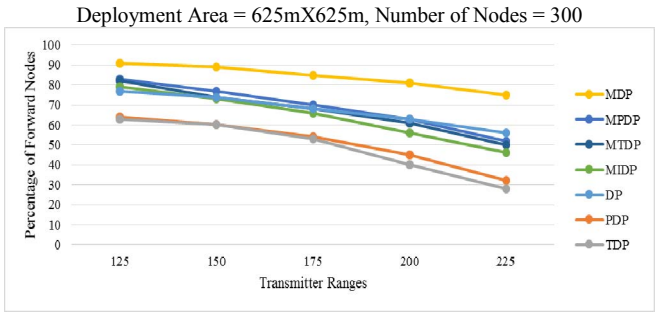


Fig. 5(b). Effect of transmitter ranges in moderaely dense network.

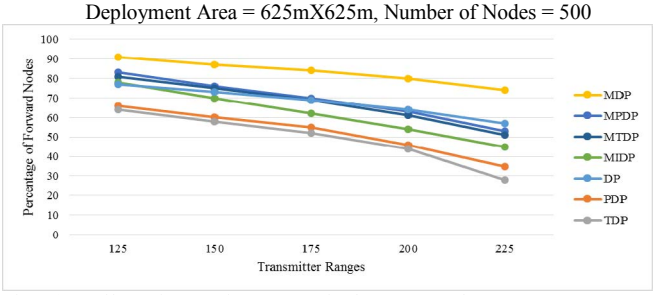


Fig. 5(c). Effect of transmitter ranges in dense network.

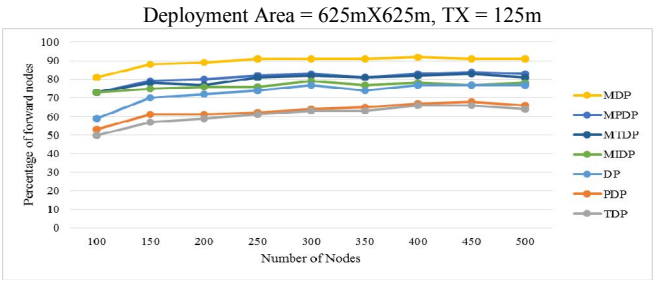


Fig. 6(a). Effect of node density when transmitter range is 125 m.

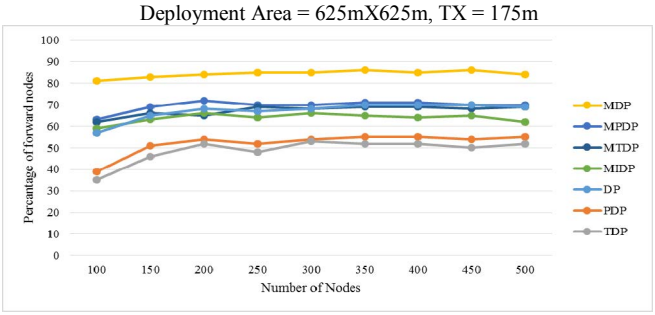


Fig. 6(b). Effect of node density when transmitter range is 175m .

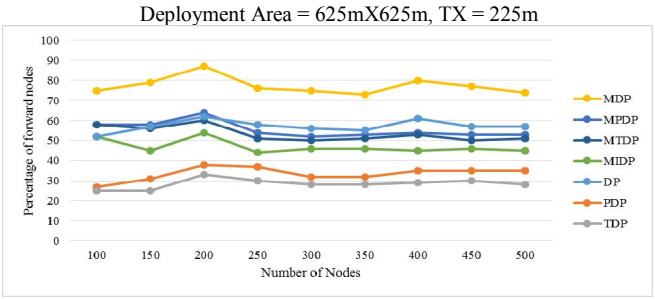


Fig. 6(c). Effect of node density when transmitter range is 225m.

In this section, we also present simulation results of reachability and ratio of redundancy-coverage while the environment becomes unreliable. To make the environment untrusted, we took an example network of 100 nodes in a 625m X 625m deployment area. The transmission range was set to 125m. We intentionally select a certain fraction of nodes randomly that misbehave by dropping packets without rebroadcasting them. For a particular scenario file, we first fix the number of well behaved nodes N and the number of misbehaving nodes M, such that $N + M = 100$. Thus, as we increase the number of misbehaving nodes, the larger set includes more misbehaving nodes from the previous scenario. This ensures a consistent evaluation of the impact of the enlarged population of misbehaving nodes by retaining all the features of the previous environment and simply making it worse. For the simulation experiments of our proposed algorithms, the results are averaged considering each node as a source node and for each node in the network rebroadcast decision is calculated. Performance measures are reported as an average of these random samples.

Fig. 7 presents the effect of percentage of misbehaving nodes on the reachability. Here, we compare single-covered broadcasting algorithms DP, PDP, TDP and IDP with the fault-tolerant broadcasting algorithms MDP, MPDP, MTDP and MIDP from the viewpoint of reachability. The environment of the network is varied by varying the number of misbehaving nodes from 10% to 100% while keeping the total number of nodes constant. We took the average reachability and redundancy/coverage of the scenarios where percentage of misbehaving nodes was fixed but we took each and every well-behaved nodes as a source node. For example, when we intentionally failed 10% nodes, then we calculated the reachability and redundancy/coverage by considering each and every 90 well-behaved nodes as source node which indicates source node can be anywhere in the network. Then, we took the average of the results of 90 individual well-behaved source nodes' reachability and redundancy/coverage. As we can see from the graph, the multicovered broadcasting algorithms appear to significantly outperform their single-covered counterparts. As we keep increasing the misbehaving nodes until 60%, the difference between the single-covered and multicover broadcasting algorithms rises gradually. Ultimately, the fault-tolerant broadcasting algorithms turn up the clear winner when we look at this simulation result, which is expected: these algorithms intentionally incur redundancy to improve reachability. To properly interpret the graphs in Fig. 7, we should realize that for a large number of misbehaving nodes (40%) the reachability of the single-covered broadcasting algorithms is so poor that they drops to below 80% and when 70% of the nodes are misbehaving, the reachability of all the heuristics plunged sharply, hitting a low of 0%. The ratio of redundancy and coverage is shown in Fig. 8. Here, the ratios of redundancy and coverage of the proposed enhanced multicover broadcasting algorithms are clearly less than that of MDP. From Fig. 8, we should realize that the spectacularly greater ratio of redundancy and coverage of the multicover algorithms with values in the range of 15%-36% results from the fact that the coverage of these algorithms are not very poor like their single-covered counterparts.

Lastly, we compare the reachability of the algorithms by varying the misbehaving nodes level-wise where we deactivated the source nodes' neighbors (first hop and second hop). We compared the performances regarding reachability in two kinds of network- dense and sparse. For dense network,

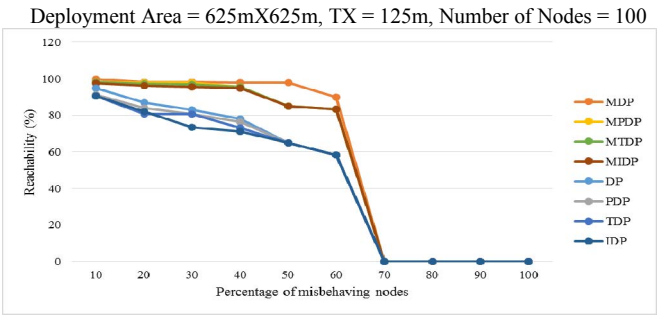


Fig. 7. Impact of misbehaving nodes in measurement of reachability.

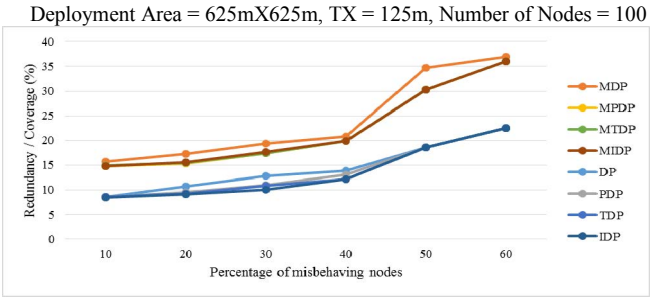


Fig. 8. Impact of misbehaving nodes in measurement of ratio of redundancy and coverage.

we took an example network of 100 nodes in the deployment area of 625m X 625m. The nodes' transmission range is fixed and that is 225m. For sparse network, we took an example network of 25 nodes in the deployment area of 1000m X 1000m. The nodes' transmission range is fixed and that is 300m. Like before, we calculated reachability considering each well-behaving node as source node. For each source node, we varied the misbehaving first hop and second hop neighbors from 10% to 100% and took the average of the reachability of each and every node as source nodes.

Fig. 9 and Fig. 10 compares DP, PDP, TDP and IDP with the fault-tolerant broadcasting algorithms MDP, MPDP, MTDP and MIDP from the viewpoint of reachability when misbehaving nodes are source nodes' first hop neighbors and the network is dense and sparse respectively. The environment of the network is varied by varying the number of misbehaving first hop neighbors from 10% to 100% while keeping the total number of nodes constant. As we can see from the first graph (Fig. 9), the multicovered broadcasting algorithms appear to significantly outperform their single-covered counterparts. As we keep increasing the misbehaving first hop neighbors until 50%, the reachability of all algorithms was 100%. From 60%, the trend experienced a gradual fall. Yet, fault-tolerant algorithms showed a much better performance (around 80% rechability) than others even when misbehaving nodes are 80% of the first hop neighbors. Inevitably, when the misbehaving nodes reached to 100%, the reachability of all the heuristics dropped to a low of 0%. Ultimately, the fault-tolerant broadcasting algorithms turn up the clear winner when we look at this simulation result, which is expected: these algorithms intentionally incur redundancy to improve reachability. To sum up, we can say that if at least 20% nodes are behaving well in one hop for a source node, we can achieve the reachability of around 80% by using the fault-tolerant broadcasting algorithms whereas the single covered algorithms achieved only 50% reachability. In Fig. 10, when

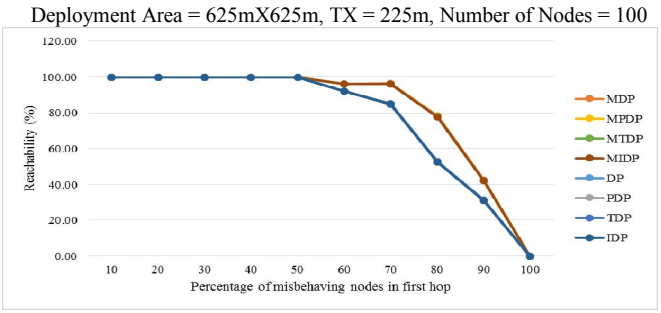


Fig. 9. Measurement of reachability in dense network when first hop neighbors are misbehaving.

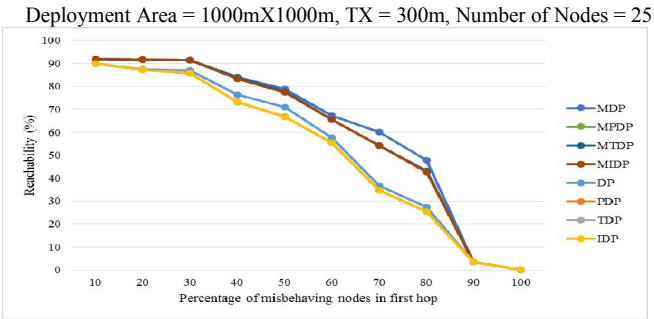


Fig. 10. Measurement of reachability in sparse network when first hop neighbors are misbehaving.

the network is sparse, unlike in the dense network the algorithms never gains 100% reachability. From 30% misbehaving nodes in first hop neighbors, the reachability starts falling dramatically from 90% reachability and reached to 0% when all the first hop neighbors are misbehaving. Like the previous graph, the fault-tolerant algorithms showed better performance than the fault-ignorant algorithms. When 70% nodes of first hop became uncooperative, the reachability of our proposed algorithms could not be up to the mark like MDP. Until 50% nodes are well-behaved in first hop neighbors, our proposed algorithms showed 80% reachability.

Fig. 11 and Fig. 12 illustrate the comparison in the measurement of reachability of DP, PDP, TDP and IDP with the fault-tolerant broadcasting algorithms MDP, MPDP, MTDP in both dense and sparse network while the second hop neighbors are misbehaving. From the graph of the Fig. 11, we can sum up that if at least10% nodes are behaving well in two hop for a source node, we can achieve the reachability of around 70- 80%. In Fig. 12 when the network is sparse, if at least 30% nodes are active in 2-hop neighbor-list, the reachability of our proposed broadcasting algorithms is above 80%, which is a marked improvement.

V. Conclusion and Future Work

In this paper, we have studied the broadcast process in ad hoc wireless networks with an objective to minimize the number of forwarding nodes of fault-tolerant broadcasting protocols. We have pointed out the deficiencies of the multicover dominant pruning (MDP) algorithm and proposed more optimized (three) novel efficient algorithms: multicover total dominant pruning (MTDP) algorithm, multicover partial dominant pruning (MPDP) algorithm and multicover improved dominant pruning (MIDP) algorithm. From the methodologies of these algorithms it has been clear that the size of the two-hop neighbor set that to be covered in MIDP is

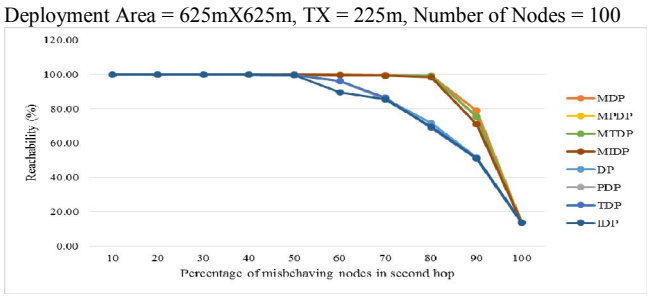


Fig. 11. Measurement of reachability in dense network when second hop neighbors are misbehaving.

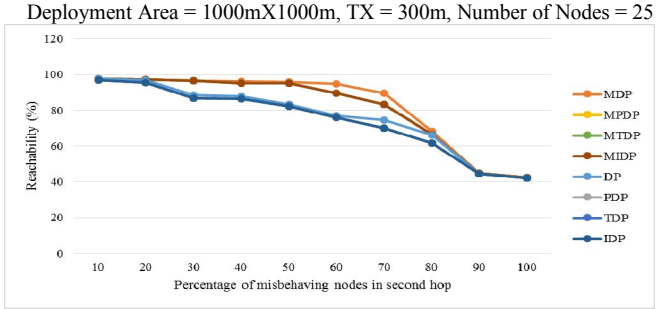


Fig. 12. Measurement of reachability in sparse network when second hop neighbors are misbehaving.

the smallest among all broadcasting protocols, $U_{MIDP} \leq U_{MTDP} \leq U_{MPDP} \leq U_{MDP}$. Simulation results have shown that MTDP, MPDP and MIDP all these three proposed algorithms have shown better performance than the MDP algorithm irrespective of density of the networks. The difference among MIDP, MTDP and MPDP is almost insignificant but MIDP showed the best especially in dense network. Moreover, irrespective of the dense and sparse network, these three algorithms showed better coverage like MDP in comparison with the single covered algorithms, although their coverage is not up to the mark like MDP sometimes especially when the majority of the population becomes misbehaving. In terms of redundancy - coverage ratio, our proposed algorithms again showed better performance than MDP as their ratio is less than that of MDP. Therefore, we can say that in practice, transmission ranges of nodes in a network may often be highly irregular and dynamic and due to this irregular transmission ranges, the network may become denser or sparser in some area, yet as long as the network is connected, our proposed algorithms will perform better. In future, we plan to design a fault-tolerant broadcasting algorithm which will be able to achieve not only more saved broadcast but higher reachability than MDP algorithm. Developing analytical models of MTDP, MPDP and MIDP to determine expected number of forwarding nodes required to complete a broadcast is another direction of future work.

REFERENCES

[1] W. Lou and J. Wu,"On Reducing Broadcast Redundancy in Ad Hoc Wireless Networks, " *IEEE Transactions on Mobile Computing,* vol. 1, pp. 111-122, April 2002.

[2] A. Rahman, P. Gburzynski and B. Kaminska, "Enhanced Dominant Pruning-based Broadcasting in Untrusted Ad-Hoc Wireless Networks," in *Proc. International Conference on Communication,* (Glagow, Scotland), pp. 3389-3394, June 2007.

[3] K. M. Alzoubi, P. J. Wan, and O. Frieder, "New distributed algorithm for connected dominating set in wireless ad hoc networks," in *Proc. 35th Hawaii International Conference on System Sciences,*vol. 9, (Hawaii, U.S.A), pp. 297-303, January 2002.

[4] G. Calinescu, I. Mandoiu, P. J. Wan, and A. Zelikovsky, "Selecting forwarding neighbors in wireless ad hoc networks," in *Proc. 5th International Workshop on Discrete Algorithms and Methods for MOBILE Computing and Communications,* (Rome, Italy), pp. 34–43, December 2001.

[5] H. Lim and C. Kim, "Flooding in wireless ad hoc networks," *Computer Communications Journal*, vol. 24, pp. 353–363, February 2001.

[6] S. Ni, Y. Tseng, Y. Chen, and J. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *Proc. The 5th Annual International Conference on Mobile Computing and Networking*, (Washington, U.S.A.), pp. 151–162, August 1999.

[7] W. Peng and X.C. Lu, "On the reduction of broadcast redundancy in mobile ad hoc networks," in *Proc. 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing*, (Boston, U.S.A), pp. 129–130, August 2000.

[8] A. Qayyum, L. Viennot, and A. Laouiti, "Multipoint relaying for flooding broadcast message in mobile wireless networks," in *Proc. 35th Hawaii International Conference on System Sciences*, (Hawaii, U.S.A.), pp. 3866 – 3875, January 2002.

[9] I. Stojmenovic, S. Seddigh, and J. Zunic, "Dominating sets and neighbor elimination based broadcasting algorithms in wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, vol.13, pp.14–25, January 2002.

[10] A. Rahman, M. E. Hoque, F. Rahman, S. K. Kundu and P. Gburzynski, "Enhanced Partial Dominant Pruning (EPDP) Based Broadcasting in Ad Hoc Wireless Networks," *Journal of Networks*, vol. 4, pp. 895-904, 2009.

[11] B. Saeed, C. Lung, T. Kunz, A. Srinivasan, "Multimedia Streaming for Ad Hoc Wireless Mesh Networks Using Network Coding," *International Journal of Communications*, *Network and System Sciences*, vol. 6, pp. 204-220, 2013.

[12] B. Das, R. Sivakumar, and V. Bharghavan,"Routing in ad-hoc networks using a virtual backbone," in *Proc.* 6th international conference on computer communications and networks**,** (Las Vegas, U.S.A.), pp. 1–20, September 1997.

[13] L. Lovasz, "On the ratio of optimal integral and fractional covers," *Discrete mathematics*, pp. 383–390, 1975.

[14] S. Bansal and M. Baker,"Observation based cooperation enforcement inad hoc networks," technical report, Computer Science Department, Stanford University, 2003.

[15] S.Buchegger and J. L. Boudec,"Performance analysis ofthe CONFIDANT protocol: Cooperation of nodes—fairness in dynamic ad-hoc networks," in *Proc. IEEE/ACM Symposium on Mobile Ad Hoc Networking and Computing*, (Lausanne, Switzerland), pp. 226–236, 2002.

[16] L. Buttyan and J. Hubaux,"Enforcing service availabilityin mobile ad-hoc WANs," in *Proc. 1st ACM International Symposium on Mobile Ad Hoc Networking and Computing,* (Boston, U.S.A.), pp. 87-96, 2000

[17] Y. Hu, D. Johnson, and A. Perrig,"SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks."*Ad Hoc Networks*, vol. 1, pp. 175–192, July 2003.

[18] Y.Hu, A.Perrig, and D. B. Johnson, "Ariadne: A secureon-demand routing protocol for ad-hoc networks," *Wireless Networks,* vol. 11, pp. 21-38,January 2005.

[19] P. Michiardi and R. Molva, "Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks," in *Proc. IFIP TC6/TC11 6th Joint Working Conference on Communications and Multimedia Security: Advanced Communications and Multimedia Security,* pp. 107-121, September 2001.

[20] S. Yi, P.Naldurg, and R. Kravets,"Security-aware ad hoc routing for wireless networks," in *Proc. 2nd ACM Symposiumon Mobile Ad Hoc Networking and Computing*, (LongBeach, CA, U.S.A.), pp. 299-302, October 2001.

[21] W. Yu, Y. Sun, and K.J.R. Liu,"Hadof: Defence against routing disruptions in mobile ad hoc networks," in *Proc. 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Informatics and Communications,* (Miami, U.S.A.), pp. 1252-1261, March 2005.

[22] S. Zhong, Y. Yang, and J. Chen,"Sprite: A simple, cheat-proof, credit basedsystem for mobile ad hoc networks," in *Proc. 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, Informatics and Communications,* (San Francisco, U.S.A.), pp. 1987-1997, March2002.

[23] C.J. Colbourn, L.K. Stewart, Permutation graphs: Connected Domination and Steiner Trees, *Discrete Mathematics*. pp.179-189, 1990.

[24] K. Bharath-Kumar and J. M. Jaffe, "Routing to multiple destinations in computer networks," *IEEE Transactions in Communications,* vol. 31, pp. 343-351, 1983.

A Graph Coloring Based Dynamic Channel Assignment Algorithm For Cognitive Radio Vehicular Ad Hoc Networks

Tareq Anwer Sohan*, Hasib Hamidul Haque†, Md. Asif Hasan‡, Md. Jahidul Islam§ & A. B. M. Alim Al Islam¶

*†‡ Department of Electrical and Electronic Engineering
§¶ Department of Computer Science and Engineering
*†‡§ United International University, Dhaka-1209, Bangladesh
§ University of Minnesota-Twin Cities, Minneapolis, United States
¶ Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

E-mail: *sohantas@gmail.com, †hasibhamidul@gmail.com, ‡asif.shm@gmail.com, §islam034@umn.edu and ¶alim_razi@cse.buet.ac.bd

***Abstract*—In this paper, we propose an intuitive graph coloring-based algorithm for Dynamic Spectrum Access (DSA) in Cognitive Radio-enabled Vehicular Ad Hoc Networks (CR-VANETs). Besides, we thoroughly investigate the performance of our algorithm through numerical and discrete event simulation. For performance evaluation, we design and simulate single-radio multi-channel CR-VANETs to investigate the operational challenges and performance of our proposed algorithm. In addition, we elaborately discuss the procedure of simulating CR-VANETs using a widely used discrete event simulator called `ns-2`, which is quite challenging and not properly documented in the literature.**

***Index Terms*—Vehicular Ad Hoc Network, Dynamic Spectrum Access, Cognitive Radio Network, CR-VANETs.**

I. Introduction

Vehicular Ad Hoc Network (VANET) is a promising technology that provides several safety and non-safety applications ([1], [2], [3], [4], [5]) to moving vehicles and pedestrians. VANETs have been envisioned to improve road safety and efficiency, and provide internet access on the move, by incorporating wireless communication and information technologies into the transportation system. We illustrate a possible communication scenario of VANETs in Figure 1.

However, data communication in VANETs is challenging due to its challenging features such as dynamic connectivity, high mobility, short link lifetime, and frequent network fragmentation. Therefore, efficient spectrum utilization and avoiding co-channel interference for ensuring reasonable network performance is extremely challenging in VANETs. Using *Cognitive Radio (CR)* [6] for VANET communication can facilitate efficient spectrum utilization by *Dynamic Spectrum Access (DSA)* ([6], [7]) feature.

VANETs that use CRs for communication are called CR-VANETs [8]. In CR-VANETs, road-side infrastructures provide frequency spectrum for communication to its users, i.e., Primary Users (PUs). On the other hand, the moving vehicles act as the Secondary Users (SUs), who use the unused portions of the frequency without interfering the PUs, using a Dynamic Spectrum Access (DSA) mechanism.

978-1-5090-0203-0/16/$31.00

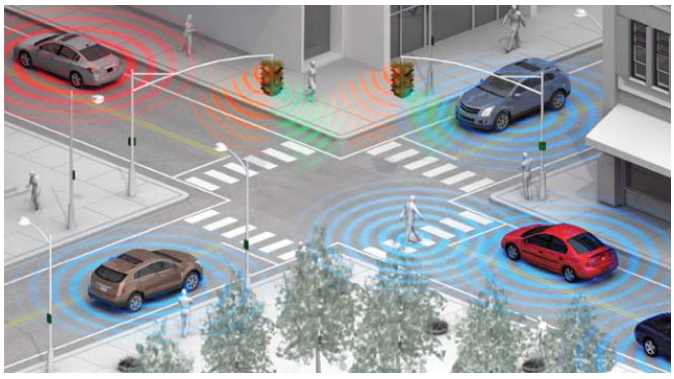


Fig. 1: Communication scenario in VANETs (Source: [9])

The basic feature of CR based networks is DSA. Previously, we investigate the challenges of designing efficient DSA techniques for CR-VANETs and provide numerical simulation results to delineate the possible challenges of DSA in CR-VANETs. Besides, we also highlighted the features that a DSA algorithm must have to consider to achieve good network performance in CR-VANETs [10]. In literature, for efficient DSA mechanism, we follow a number of algorithms based on game theory, agent based learning, heuristics evolutionary strategy, fuzzy logic and so on ([6], [11]). Most of these algorithms are proposed for static wireless network like Mobile Ad Hoc Network (MANET) and CR based mesh networks. However, the applicability of these algorithms for VANET is limited due to its challenging features. Therefore, in this paper we focus on these algorithm issues and making the following contributions:

- We design and simulate a CR-VANET scenario including road-side PU infrastructures, PUs, and moving vehicles

(i.e., SUs). Here, we use *iGraphics*[1] to mimic our CR-VANET scenario.

- Then, we numerically simulate our CR-VANET scenario and investigate different aspects of our network in terms of a number of performance metrics (percentage of successful communication, number of conflict between the channels and channel utilization).
- Subsequently, we propose a graph coloring-based Channel Assignment (CA) algorithm for single radio multi channel CR-VANETs [13]. In our simulation, we implement this algorithm and investigate its performance while operating in the network.
- In addition to numerical simulation, we perform discrete event simulation of our CR-VANET scenario as well. We use the most widely used discrete event simulator `ns-2` [14] for our simulation. Here, first, we point out the detailed procedure of simulating CR-VANETs using `ns-2`. Then, we present our simulation results in terms of several standard Quality Of Service (QoS) parameters.

II. Related Work

There are a number of research studies which propose efficient DSA algorithms based on graph theory, game theory, heuristics, evolutionary strategy, agent-based learning, fuzzy logic, etc ([6], [11], [15], [16]). However, most of these algorithms are proposed for CR-based mesh networks, MANETs, or other static wireless networks. Therefore, their applicability to CR-VANETs ([8], [17]) is limited due to the challenging features of VANETs. These limitations have not been addressed thoroughly in the literature. To the best of our knowledge, designing efficient dynamic channel assignment algorithms particularly for CR-VANETs and evaluating their performance using discrete event simulation is yet to be addressed in the literature.

III. CR-VANET Scenario

In this paper, we consider a roadside infrastructure-based single-radio multi-channel primary network having PUs and moving vehicles along the roads as SUs. These PUs and SUs share a set of orthogonal wireless channels. We illustrate our design in Fig. 2. Here, the roadside Base Stations (BS) provide wireless communication channels to the PUs within its transmission area. The vehicles are only allowed to use the channels that are not being used by any PUs within their transmission range.

For simplicity, we consider two entry points for vehicles in our simulation area shown in Fig. 2. Here, we allow vehicles to enter through the top and bottom horizontal roads (from right to left). On the other hand, vehicles can leave the simulation area through the exit points. The entry points are marked with brown-colored arrow and the exit directions are marked with blue-colored arrow. In case of a junction, vehicles take a probabilistic decision. We consider each direction equally likely for the vehicles. These junctions are marked with red-colored star signs in Fig. 2.

[1] *iGraphics* [12] is a graphical library containing some openGL drawing functions that can be used to draw basic graphical shapes in Visual C++.

TABLE I: Parameters used in Visual Studio 2012

Parameter	Value
# of maximum cars	50
# of channels	5
# of Primary users	8
# of Base stations	5
Maximum simulation time	40000 milliseconds
Transmission probability	20%
Radius of PUs	200 pixels
Radius of SUs	100 pixels

We start by generating a set of vehicles moving through the roads. Then, at each time epoch, we track each vehicle's position, number of neighbour, neighbouring PUs, etc. Based on these information, we evaluate our simulation parameters (percentage of successful communication, number of conflict between the channels and channel utilization). We discuss the simulation procedures and results in Section V. Next, we present the model parameters.

A. Model Parameters

We show our CR-VANET scenario in scale using Dia diagram editor [18].

Here, we consider a $5000\text{m} \times 5000\text{m}$ simulation area with 1 unit as 50m. The road-side lines are marked by purple color. The areas covered by green, ash, light blue colors or trees mean that there are no PUs in that area.

We place several primary BSs & Secondary BSs respectively for PUs & SUs. Besides, we mark the PUs by computers and SUs by vehicles. We assume that the PUs have no mobility. The arrival rates of the vehicles follow a Poisson distribution [19] with mean 5.

In our scenario, the maroon circles represent the coverage areas of PU BSs and the blue circles represent the coverage area of moving SUs. We take the coverage area of BSs as 2500m in radius. Besides, we consider the coverage area of moving vehicles as 1000m in radius. We also take the width of the roads to be 80-150 feet and speed of each vehicles to be 50m/sec.

However, when we plot the road scenario in visual studio 2012 [20], we have to change some parameters for our calculation. The measurements we follow in visual studio 2012 are illustrated in Table I.

Next, we discuss about our proposed channel assignment technique.

IV. Proposed Channel Assignment Technique

We adopt a centralized CA technique based on graph coloring [13]. We name our algorithm *LOSA*, denoting *Leave One Serve All* CA algorithm. We present the CA procedure of LOSA in Fig. 3.

As we already mentioned, we adopt a centralized network model. Therefore, the CA decision is centralised, i.e., channels are assigned to vehicles by the base stations in the network. Each vehicle communicates with its nearest base station and

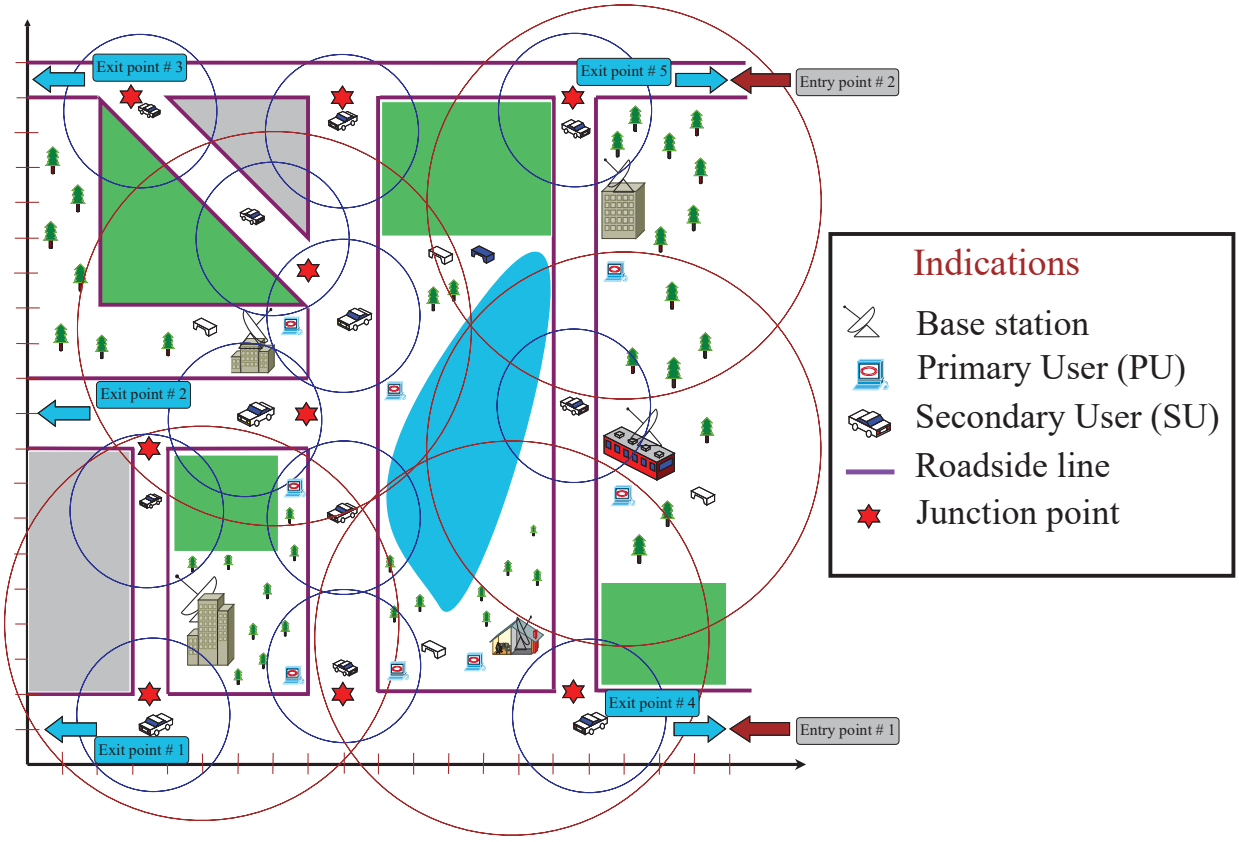


Fig. 2: CR-VANET model

share information regarding its channel access. Information include its current position, speed, direction, etc. Each base station assign channels to vehicles based on these information from all vehicles within its transmission range. .

Therefore, each base station maintains data structure to accommodate the following bookkeepings:

- Position, speed, and direction of each vehicle within its transmission range.
- Active vehicles (i.e., vehicles which are in transmission) and the corresponding channels being used for their communication.
- Active PUs and the corresponding channels being used for their communication.

At each epoch, all base stations in the network assign channels to moving vehicles within their transmission range as follows:

Step 1: First, the base station finds out the moving vehicles that are currently within its transmission range. Then, it takes the vehicles that are interested in accessing a channel. We denote these interested vehicles as $V = \{V_1, V_2, \ldots, V_n\}$. Besides, it finds the currently available channels, which we denote as $C = \{C_1, C_2, \ldots, C_m\}$. Here, m is the number of channel that are currently available for communication. We say a particular channel is not available if it is currently being used by any PU within the transmission range of the base station.

Step 2: Then, we generate the *network conflict graph* ([6], [21], [22]) based on the network topology graph of vehicles which represents the connectivity and communication among them. The vertices of a conflict graph correspond to the links (i.e., connections) between two vehicles, and edges are drawn between links (vertices) that can interfere with each other when assigned the same channels.
An example of conflict graph generation from topology graph is shown in Fig. 4. It also shows a vertex coloring of the conflict graph (which we denote as G) that we perform for assigning channels. Here, different colors represent different channels. Hence, we perform an *m-vertex coloring* of the conflict graph. For vertex coloring, we adopt the most commonly used greedy graph coloring approach ([21], [23]) in state-of-the-art DSA techniques designed for static CRNs.
Here, if there exist an $m-$coloring of G, we go to step 4, otherwise we go to step 3.

Step 3: If m colors are not sufficient for coloring all vertices of G, it means that we do not have enough channels to accommodate all the vehicle transmissions.

Therefore, we cannot serve all the vehicles leaving few vehicles unable to access a channel. However, we want this number to be minimized. Consequently, to serve maximum number of vehicles, we remove the vehicle (V_i) from V, which causes maximum number of conflicts in G while coloring with m colors. Hence, we prune all vertices and edges from G with are accountable for V_i and go back to step 2.

In this step, we deny access to one vehicle and look forward to serve the other vehicles. We assume all vehicles have equal priority and therefore, we choose to serve the maximum number of vehicles with currently available channels. Consequently, we leave the most conflicting vehicle unassigned. This *leave one serve all* strategy in this step of our algorithm has led to the naming of our algorithm.

Step 4: Eventually, G will be m-colorable. We map each color to a different channel and assign to the corresponding vehicle transmissions.

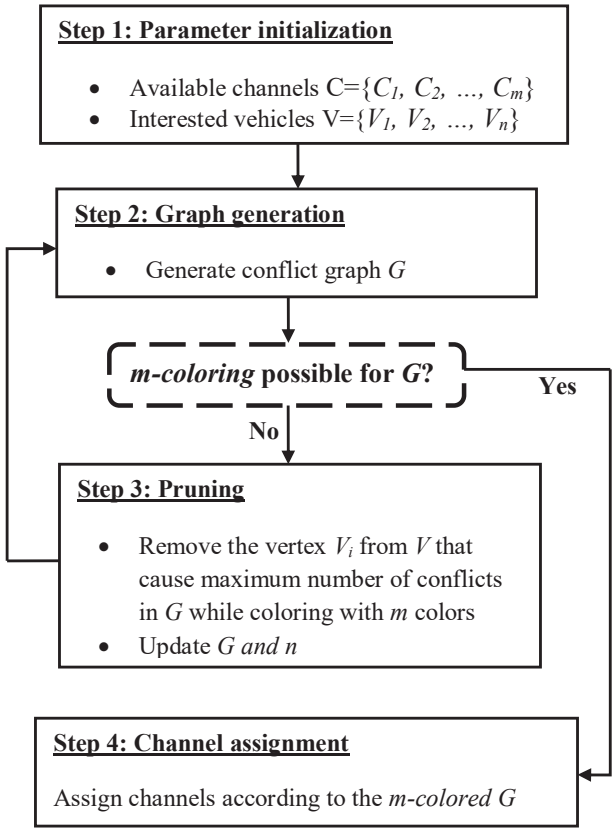


Fig. 3: Channel assignment procedure of our proposed algorithm

Here, we use a graph theory based approach as it is the most commonly used one ([6], [21], [22], [23]) for static CRNs. However, their applicability in CR-VANETs is subject to further investigation and future research work. Besides, investigating other class of algorithms such as heuristics, game theory based approaches to design DSA algorithms dedicated for CR-VANETs is also an open research problem in this regard. Next, we present numerical evaluation.

V. Numerical Evaluation

As we mention already, we designed our road scenario in DIA diagram editor [18]. To perform numerical simulation, we use iGraphics [12] tool. iGraphics is a graphical library containing some drawing functions that can be used to draw basic graphical shapes in Visual C++. These functions are implemented in openGL. It can be used to draw different shapes, display text in different fonts, change colors and many more.

While we simulate our road scenario, first we map the coordinates of DIA to our new interface in iGraphics. Then, we draw the complete road scenario. We allow moving vehicles to enter through entry point in the scenario, through the top and bottom horizontal roads (from right to left). On the other hand, vehicles can leave the simulation area not only through the exit points but also through the entry points. In case of a junction, vehicles take a probabilistic decision. We consider each direction equally likely for the vehicles. At each time epoch, we track each vehicle's position, number of assigned channels to the vehicles etc. Next, We will discuss the simulation settings.

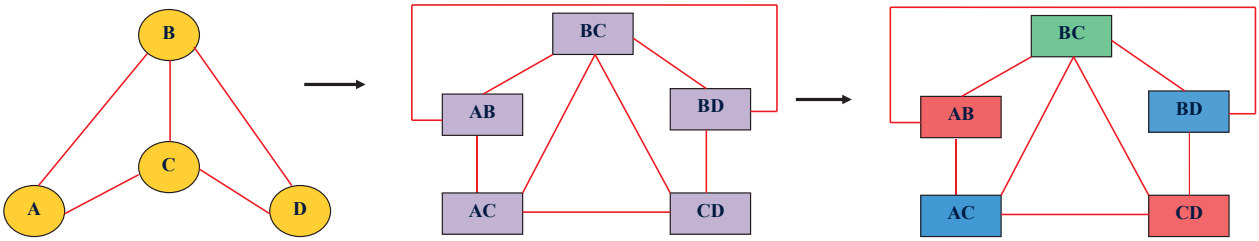
A. Simulation Settings

We need to use the extension $iGraphics$ in visual studio 2012 [20] before we start to write the code. After installing $iGraphics$, we choose $C++$ as coding language to plot the road scenario.

We place five BSs in our scenario. Besides, we mark the PUs by computers and SUs by vehicles. We assume that the PUs have no mobility. Here, the vehicles are moving by following Poisson distribution [19]. Besides, we consider the coverage area of moving vehicles as 100 pixels in radius. The coverage area of PUs are around 200 pixels. We also take the width of the roads to be $80-150$ feet and speed of each vehicles to be 50msec. We set transmission probability as 20%. We list these measurements (of iGraphics) in Table I.

In our simulation, we generate up to 50 vehicles. These vehicles arrive from two entry points maintaining two lanes. We control the flow of arrival rate of vehicles by following Poisson distribution (with mean 5). Afterwards, we track each vehicle for a simulation time of 40000ms. At each epoch, all BSs in the network assign channels to moving vehicles within their transmission range. We have already discussed the channel assignment procedure in the previous Section IV.

Through the simulation time, we track route of each vehicle whether it is under any coverage area of base stations or not. Alongside, we also find that if any vehicles are assigned with channel or not. If they are assigned with any channel, we also find that under which base stations it is assigned. It is worthy to mention that here we apply a dynamic VANET topology where these moving vehicles have different traveling paths, arrival times, etc. These vehicles not only arrive at different



(a) Network topology graph (b) Network conflict graph (c) Vertex coloring with 3 colors

Fig. 4: Use of graph-coloring in for assigning channels based on network conflict graph

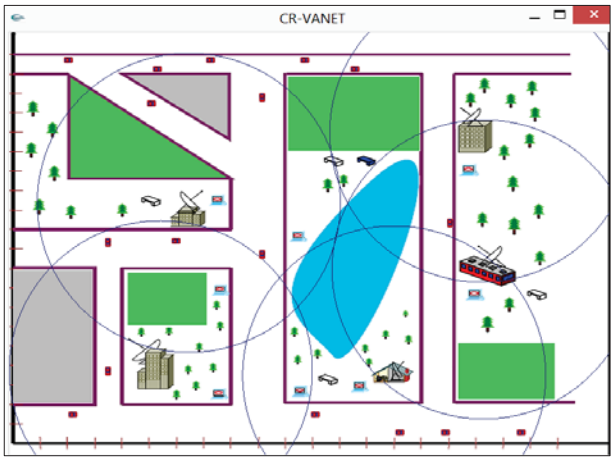


Fig. 5: Simulation environment

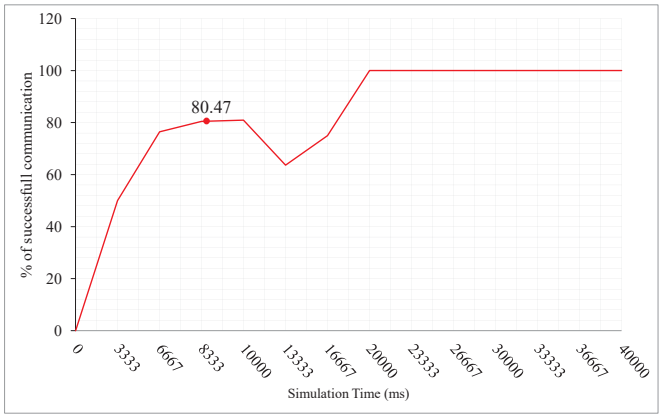


Fig. 6: Percentage of successful communication for each vehicles throughout the simulation time

times, they differ in their traveling paths due to the presence of junctions. Next, we present simulation results.

B. Simulation Results

Throughout the simulation time, each vehicle can successfully establish communication if they are assigned an unused channel. In our simulation, we find the percentage of successful communication for each vehicle and illustrate these values in Fig. 6. Here, in our simulation the average percentage of successful communications are about 80.47%. The rate of successful communication gradually increases from first few seconds. However, the rate of increase slows down from 6667ms and continuous to 10000ms. Then this rate falls from 10000ms and continuous to 13337ms. Afterwards, the rate increases with the time and after 13333ms the percentage of successful communication stables at 100 after 20000ms. This illustration depicts the overall availability of channels for each vehicle, which in our case 80.47% in an average. That is, according to our network model, each vehicle can access channels 80.47% of the times without interfering any PUs in the network. Besides, we find that after a certain period (i.e., 20000ms), our CA algorithm can assign channels to every vehicles without interfering any PUs.

In addition, we find the number of conflicts during assigning channels throughout the simulation time. Likewise, we find the maximum number of conflicts during simulation. We illustrated these values in Fig. 7. Here, we find that the of number of conflicts gradually increases between 0ms to 5000ms and 18333ms to 21667ms. On the other hand, it gradually decreases between 11667ms to 18333ms and 28333ms to 31667ms. The red line in this figure depicts the maximum number of possible in the whole network at each time epoch. Whereas, the blue one represents the total number of conflicts produced by our CA algorithm. Therefore, this figure demonstrates that our algorithm decreases the number of possible conflicts in the network significantly.

Finally, we find out the channel utilization for each channel, which we illustrate in Fig. 8. Here, we find that channel 1 has been assigned to SUs for 37% of the times, whereas, channel 2 and 3 are utilized in around 20% to 22% cases. However, channel 0 and 4 is utilized far less number of times (around 5% to 10%). Therefore, this figure indicates that the load of each channel may be improved further with more fair CA operation. That is, for best case scenario, each channel would be utilized for about 20% each, as there are 5 channels. However, as our objective is to minimize interference, the channel utilization fairness is not close to optimal. This is actually an open research problem [6] to design efficient CA algorithms for CRNs that can achieve optimal fairness in

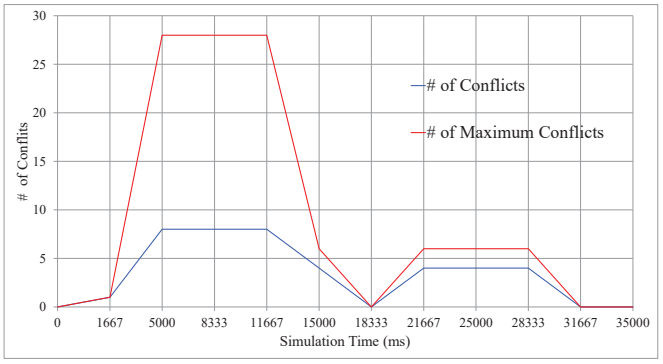


Fig. 7: Number of conflicts during simulation

addition to minimizing interference in the network, which is extremely challenging for CR-VANETs in particular. Next, we discuss about our simulation findings.

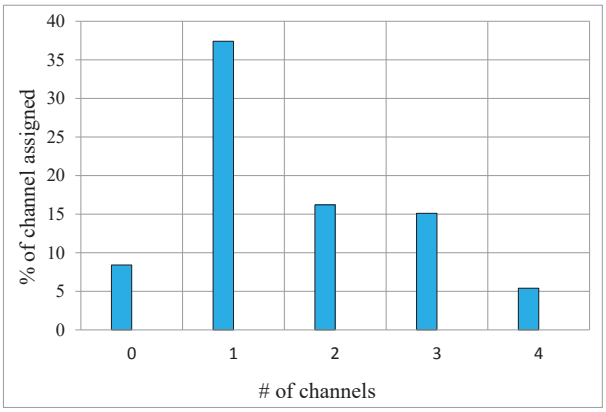


Fig. 8: Percentages of channel assignment

C. Simulation Findings

In numeric simulation, we find out percentage of successful communication for each vehicles throughout the simulation time, number of conflicts during simulation and percentages of channel assignment. We are able to get around 80.47% percentage of percentage of successful communication for each vehicles throughout the simulation time. By applying graph coloring algorithm, we minimize the number of conflicts during simulation and we are able to assign channels to SUs successfully. Next, we discuss about the experimental evaluation in Section VI.

VI. Experimental Evaluation

In this section, we discuss the detailed procedure of simulating CR-VANETs using discrete-event simulator `ns-2` [14]. Then, simulate our CR-VANET scenario following this procedure and present the simulation results.

A. Procedure of Simulating CR-VANETs in Discrete-event Simulator

Discrete event simulation is always preferable to get reliable results and to mimic real network environment. However simulating CR-VANETs in discrete-event simulator such as `ns-2` is more complicated things than other simulation ([24], [25], [26]). Therefore, we discuss discrete event simulation procedure of CR-VANETs.

Firstly, we need a `ns-2` based simulator for Cognitive Radio Cognitive Network (CRCN) [27]. This basic CRCN simulator provides all the basic CRN features and also enables to implement user defined DSA algorithms. Now, we need SUMO (Simulation of Urban MObility) to simulate VANET like mobility model generation [28]. The main features of SUMO include single vehicle routing, multi-lane streets, hierarchy of junction types, dynamic routing, etc. The main disadvantage of SUMO is that the traces generated by SUMO cannot be directly used by the available network simulator `ns-2`. Therefore, we need MOVE (MObility model generator for Vehicular networks) which is built on top of SUMO. It rapidly generates realistic mobility models for VANET simulations. With the help of MOVE, we convert SUMO traffic traces into `ns-2` compatible traces. The trace file generated by MOVE contains information of realistic vehicle movements which can be immediately used by popular network simulation tools such as `ns-2`. The procedure we follow in discrete event simulation are illustrated in Fig. 9

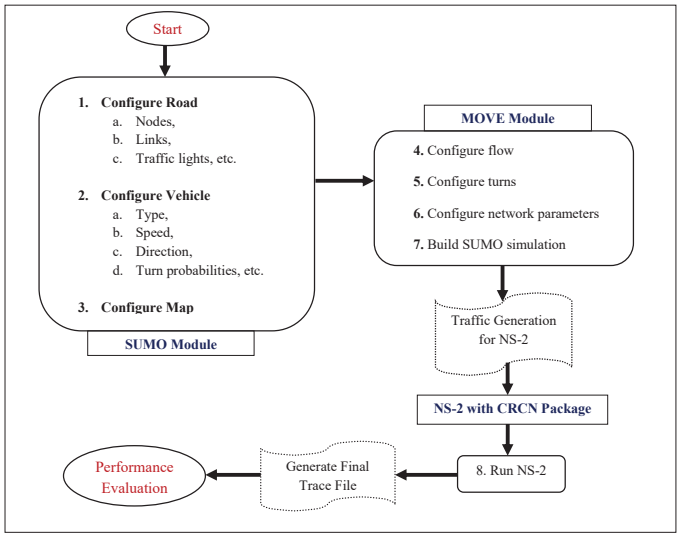


Fig. 9: Procedure of simulating CR-VANETs `ns-2`

- Firstly, we design a road topology in SUMO. Here we have to configure the nodes, multi-lane streets, traffic lights. Additionally, we configure vehicles types, their speeds, initial directions etc.
- Next to that, we need to convert SUMO traffic traces into `ns-2` compatible traces by using MOVE. Here, we also need to configure the network parameters.
- Finally, for evaluating for performance, we have to use these traces to `ns-2` (with CRCN simulator package integrated in it)

Next, we present simulation settings.

B. Simulation Settings

We consider the same network scenario that we used for our numerical simulation (Fig. 5). Here, we only consider Vehicle to Vehicle (V2V) communication. Two adjacent vehicles can communicate only when they both tune to the same channel. Besides, they operate in a half-duplex manner, i.e., they cannot receive while transmitting, and vice versa. In addition, we consider the communication range equal to the interference range for convenience and we adopt a binary interference model (i.e., two vehicles either interfere or do not interfere). In our simulation, we vary the number of vehicles (n) from 10 to 50 (i.e., $n \in [10, 50]$), and plot the corresponding QoS parameter values considering number of channels (m) to be 5 (i.e., $m = 5$).

In this performance evaluation, we consider the following QoS parameters:

- **Network Throughput:** It refers to the amount of data transmitted across the network in a given time, usually measured in kilo-bytes per second ($Kbps$). At each run, we calculated the average network throughput by averaging the individual data rates of all CR users for a given topology.
- **End-to-end delay:** The time taken for a packet to be transmitted across the network from source to destination is referred to as end-to-end delay. It is usually measured in $milliseconds(ms)$. In our simulation, we traced the sending and receiving times of each packet to find its end-to-end delay. Then, we averaged the end-to-end delays of all successful transmissions.
- **Packet delivery ratio:** The ratio of number of packets that are successfully delivered to a destination compared to the number of packets that have been sent out by the sender, is referred to as packet delivery ratio.

C. Simulation Results

Here, we illustrate average network throughput over varied number of vehicles. Then, we show average end-to-end delay and average packet delivery ratio.

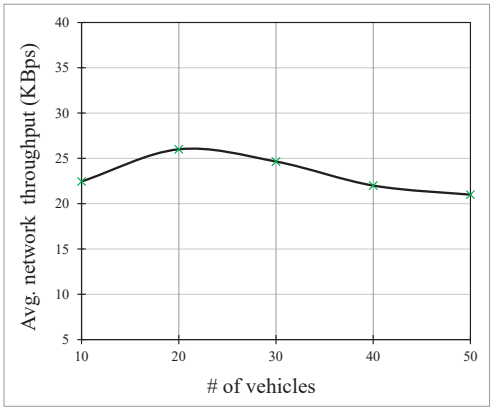


Fig. 10: Average network throughput

In our discrete event simulation, in a given time, we find out that the rate of data transmission gradually increases for 10 to 20 vehicles. Here, the data transmission rate varies between $22Kbps$ to $26Kbps$. However, this rate of data transmission decreases when the number of vehicles exceed 20 vehicles. This transmission rate is minimum when it reaches to 50 vehicles and the rate is about $20Kbps$. Here in our simulation the average network throughput is about $22Kbps$. We illustrated this average network throughput in Fig. 10.

Next to that, we find the average time taken for a packet to be transmitted across the network from source to destination Here we see that, the end to end delay is lower for 10 to 30 vehicles which is about 25ms to 20ms. However, after that this delay gradually increases from 21ms to 24ms for 31 to 50 vehicles. Here in our simulation the average end-to-end delay is about 23ms. In Fig. 11, we illustrated this average end-to-end delay.

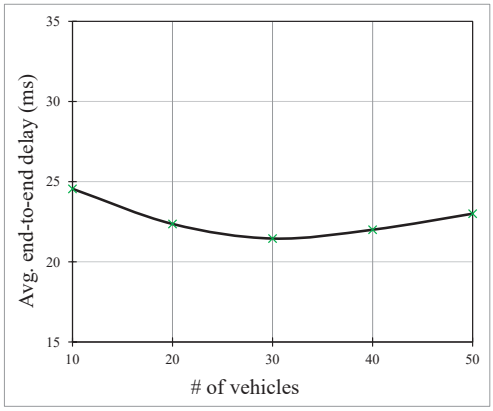


Fig. 11: Average end-to-end delay

Finally, we figure out the ratio of successful number of delivered packets to a destination. We illustrated this ratio in Fig. 12. Here, we find that the ratio increases 0.75 to 0.79 for first 20 vehicles. However, for 30 to 50 vehicles, this ratio of number of delivered packets decreases from 0.78 to 0.70. Here in our simulation the average packet delivery ratio is about 0.75. Next, we discuss about simulation results.

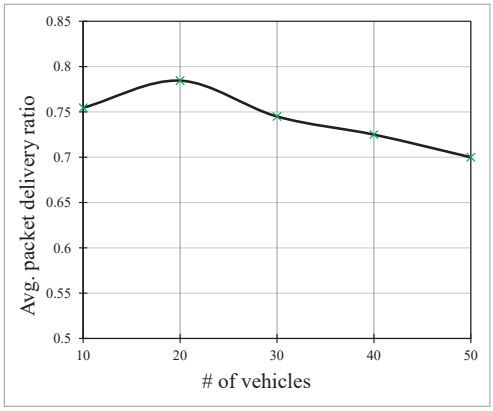


Fig. 12: Average packet delivery ratio

D. Simulation Findings

In experimental evaluation, we find out average network throughput over varied number of vehicles, average end-to-end delay and average packet delivery ratio. We are able to get around $22Kbps$ average network throughput for 50 vehicles. Then, we find about 23ms average end-to-end delay and around 0.75 average packet delivery ratio. Next, we discuss about our future work.

VII. Future Work

In future, we plan to extend our algorithm for multi-radio CR-VANETs and explore the energy considerations of CR-VANETs. Our goal is to investigating performances of other state-of-the art DSA algorithms for comparison and simulating over different type of CR-VANET scenario.

VIII. Conclusions

In this paper, we proposed graph coloring algorithm for channel assignment. We also design and simulate a VANET scenario and provide numerical simulation results to delineate the possible challenges of DSA in CR-VANETs. We do so by investigating the percentage of successful communication for each vehicle throughout the simulation time, number of conflicts during simulation time. Next to that, we find the percentage of channel assigned to the SUs. In addition, we point out the detailed procedure of evaluating the performance of DSA algorithms in CR-VANETs using discrete-event simulation. Finally, we perform discrete-event simulation with our scenario considering the complete presence and interference of PU transmissions.

IX. Acknowledgment

This work has been partially supported by the Ministry of Education, Government of the People's Republic of Bangladesh.

References

[1] A. K. Saha and D. B. Johnson, "Modeling mobility for vehicular ad-hoc networks," in *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*. ACM, 2004, pp. 91–92.

[2] H. Hartenstein and K. P. Laberteaux, "A tutorial survey on vehicular ad hoc networks," *Communications Magazine, IEEE*, vol. 46, no. 6, pp. 164–171, 2008.

[3] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions," *Communications Surveys & Tutorials, IEEE*, vol. 13, no. 4, pp. 584–616, 2011.

[4] T. H. Luan, L. X. Cai, J. Chen, X. Shen, and F. Bai, "Engineering a distributed infrastructure for large-scale cost-effective content dissemination over urban vehicular networks," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 3, pp. 1419–1435, 2014.

[5] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular ad hoc network," *Journal of network and computer applications*, vol. 37, pp. 380–392, 2014.

[6] E. Z. Tragos, S. Zeadally, A. G. Fragkiadakis, and V. A. Siris, "Spectrum assignment in cognitive radio networks: A comprehensive survey." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1108–1135, 2013.

[7] N. Cheng, N. Zhang, N. Lu, X. Shen, J. W. Mark, and F. Liu, "Opportunistic spectrum access for cr-vanets: A game-theoretic approach," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 1, pp. 237–251, 2014.

[8] N. Meghanathan, *Cognitive Radio Technology Applications for Wireless and Mobile Ad Hoc Networks*. IGI Global, 2013.

[9] "GM leverages WiFi Direct to give pedestrians a better chance of jumping out of the way of cars," http://www.extremetech.com/category/computing, [Online; last accessed September-2015].

[10] T. A. Sohan, H. H. Haque, M. A. Hasan, and M. J. Islam, "Investigating the challenges of dynamic spectrum access in cognitive radio-enabled vehicular ad hoc networks (cr-vanets)," *2nd International Conference on Electrical Engineering and Information Communication Technology*, 2015.

[11] P. F. Marshall, "Dynamic spectrum access as a mechanism for transition to interference tolerant systems," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*. IEEE, 2010, pp. 1–12.

[12] S. L. Delp and J. P. Loan, "A graphics-based software system to develop and analyze models of musculoskeletal structures," *Computers in biology and medicine*, vol. 25, no. 1, pp. 21–34, 1995.

[13] J. L. Gross and J. Yellen, *Handbook of graph theory*. CRC press, 2004.

[14] "The Network Simulator - ns-2," http://www.isi.edu/nsnam/ns/, [Online; last accessed September-2015].

[15] M. J. Islam, M. M. Islam, and A. A. Al Islam, "Evolutionary spectrum management in cognitive radio networks," in *In Proceedings of Workshop on Wireless Network and Communication*. Department of CSE, BUET, 2013, pp. 27–28.

[16] M. J. Islam, "Intelligent dynamic spectrum access exploiting a synergy between genetic algorithm and local search," Master's thesis, Department of Computer Science & Engineering, Bangladesh University of Engineering & Technology (BUET), 2015.

[17] T. K. Mak, K. P. Laberteaux, R. Sengupta, and M. Ergen, "Multichannel medium access control for dedicated short-range communications," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 1, pp. 349–366, 2009.

[18] "Dia Diagram Editor," http://dia-installer.de/, [Online; last accessed September-2015].

[19] W. J. Thompson, "Poisson distributions," *Computing in Science & Engineering*, no. 3, pp. 78–82, 2001.

[20] "Visual Studio," https://www.visualstudio.com//, [Online; last accessed September-2015].

[21] A. T. Hoang and Y.-C. Liang, "Downlink channel assignment and power control for cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 8, pp. 3106–3117, 2008.

[22] M. Bkassiny and S. K. Jayaweera, "Optimal channel and power allocation for secondary users in cooperative cognitive radio networks," in *Mobile Lightweight Wireless Systems*. Springer, 2010, pp. 180–191.

[23] B. Wang and K. R. Liu, "Advances in cognitive radio networks: A survey," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 1, pp. 5–23, 2011.

[24] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks: a survey and taxonomy," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 4, pp. 19–41, 2009.

[25] V. D. Khairnar and S. Pradhan, "Mobility models for vehicular ad-hoc network simulation," in *Computers & Informatics (ISCI), 2011 IEEE Symposium on*. IEEE, 2011, pp. 460–465.

[26] F. J. Martinez, C. K. Toh, J.-C. Cano, C. T. Calafate, and P. Manzoni, "A survey and comparative study of simulators for vehicular ad hoc networks (vanets)," *Wireless Communications and Mobile Computing*, vol. 11, no. 7, pp. 813–828, 2011.

[27] "Cognitive Radio Cognitive Radio (CRCN) simulator," http://faculty.uml.edu/Tricia_Chigan/Research/installation.htm, [Online; last accessed September-2015].

[28] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "Sumo (simulation of urban mobility)," in *Proc. of the 4th middle east symposium on simulation and modelling*, 2002, pp. 183–187.

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Short Papers

Poster & Demo

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

A Comparison between RSA and ElGamal based Untraceable Blind Signature Schemes

Khairul Alam[*1], Kazi Md. Rokibul Alam[*2], Md. Omar Faruq[*3], and Yasuhiko Morimoto[+]
[*]Department of Computer Science and Engineering, Khulna University of Engineering and Technology
Khulna-9203, Bangladesh
[+]Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima 739-8521, Japan
Email: Khairulcse0026@gmail.com[*1], rokib@cse.kuet.ac.bd[+], faruq2k10@gmail.com[*3], morimoto@mis.hiroshima-u.ac.jp[+]

***Abstract*—This paper presents a comparison between RSA and ElGamal based untraceable blind signature (BS) schemes through simulation. The objective is to provide a guideline while selecting either of them to develop an application. A BS scheme is a cryptographic protocol that can be used in cryptographic applications like electronic voting systems, electronic payment systems etc to conduct their privacy-related transactions anonymously but securely. While a user operates her electronic transactions employing a BS scheme over the internet, the BS scheme ensures the confidentiality of the secret message of the user. Besides, untraceability is a crucial criterion for any BS scheme because thereby the signer of this scheme is unable to link the message-signature pair after the BS has been revealed to the public. Two untraceable BS schemes: one is proposed by Hwang et al. and is based on RSA cryptosystem whereas the other is proposed by Lee et al. and is based on ElGamal cryptosystem have been chosen here for simulation. The outcome of the simulation model is the comparison of computation time requirement of blinding, singing, unblinding and verification phases of the chosen BS schemes.**

Keywords—Blind signature; RSA; ElGamal; Untraceability; Cryptography.

I. INTRODUCTION

A blind signature (BS) scheme can be used to conduct privacy-related applications like electronic voting systems, electronic payment systems etc to authenticate the identity of a user and / or her message anonymously [1, 11]. Normally digital signature (DS) and BS, both schemes employ two different parties *i.e.* a signer and a user (who is the message owner), where the user requests the signer to sign her message. But in a BS scheme, first the user blinds her message and then the signer signs on it by using his private signing key. Thus herein, the signer knows nothing about the content of the message [2]. Later on, anyone can publicly verify the legitimacy of the signature using the signer's public key. Herein if the employed BS scheme is untraceable, the signer is unable to link the message-signature pair even when the signature is revealed to the public [6]. Thus a BS scheme ensures the authenticity of a message [5].

In privacy-related applications, to ensure the secrecy of the message, usually a user demands a BS from the signer. For this, first she blinds her message. Then the signer sings on the blinded message, and from which later on the user can generate the unblinded signed message. Besides, a BS scheme assures that a user is unable to create a signed message by herself. Also the execution of a BS scheme can create at most one unblinded signed message [2]. Moreover, the signed message generated by a BS scheme ensures that neither it can be forged nor can be traced. For this reason, when a BS scheme is exploited, the authenticity of the signed message can be verified but the origin of the signed message cannot be traced [5].

An ideal BS scheme is supposed to satisfy the following requirements [1, 7, 9, 10].

- ***Correctness:*** the correctness of the signature of a message signed by a BS scheme can be checked by anyone using the signer's public key.
- ***Blindness:*** the content of the message should be blind to the signer; the signer of the BS scheme is unable to see the content of the message.
- ***Unforgeability:*** the signature is the proof of the signer, and no one else can derive any forged signature and pass verification.
- ***Untraceability:*** the signer of the BS scheme is unable to link the message-signature pair even when the signature is revealed to the public

Intuitively, existing BS schemes can be categorized as traceable [4, 9, 10] and untraceable [1, 2]. Although lots of BS schemes are available in the domain of cryptography, many of them can achieve only the above first three requirements. But untraceability is an essential criterion because thereby the secrecy of the message as well as the anonymity of its user is maintained. In this paper a comparison between two untraceable BS schemes has been proposed through simulation. One is proposed by Hwang et al. [1] and is based on RSA cryptosystem and the other is proposed by Lee et al. [2] and is based on ElGamal cryptosystem. The security of the first scheme depends on the difficulty of solving factoring problem (FP) whereas for the second scheme it depends on discrete logarithm problem (DLP). The outcome of the simulation model is the evaluation of computation time requirement of blinding, singing, unblinding and verification phases of the chosen schemes.

The outline of this paper is as follows: Section II explains the chosen BS schemes *i.e.* schemes proposed by Hwang et al. [1] and Lee et al. [2]. The experimental analysis of simulation

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

has been discussed in Section III. Finally Section IV explains the concluding remarks.

II. Untraceable Blind Signature Schemes

A. RSA based Hwang et al.'s Scheme

Hwang et al.'s BS scheme [1] is based on RSA cryptosystem and consists of five phases, these are: initialization, blinding, signing, unblinding, and verification. The scheme is briefly described as follows.

1) Initialization phase: The signer randomly chooses two large primes p and q, and computes $n = p * q$ and $\varphi(n) = (p - 1) * (q - 1)$. The signer chooses two large numbers e and d such that $GCD(e, \varphi(n)) = 1$ and $ed \equiv 1 \bmod \varphi(n)$. Let (e, n) be the signer's public key and d be the signer's private key for signing. The signer keeps (p, q, d) secure and publishes (e, n).

2) Blinding phase: The user has a message m, and she wishes to have it signed by the signer. She randomly selects two distinct integers' r_1 and r_2 as the blinding factor. Now she randomly chooses two primes a_1 and a_2 such that $a_1 \neq a_2$ and $GCD(a_1, a_2) = 1$. Then, she computes the blinded messages $\alpha_1 = r_1^{\,e} * m^{a1} \bmod n$ and $\alpha_2 = r_2^{\,e} * m^{a2} \bmod n$ and sends (α_1, α_2) to the signer.

3) Signing phase: After receiving (α_1, α_2) from the user, the signer randomly chooses two primes b_1 and b_2 such that $b_1 \neq b_2$ and $GCD(b_1, b_2) = 1$ and signs the blinded message by computing $t_1 = \alpha_1^{\,b1*d} \bmod n$ and $t_2 = \alpha_2^{\,b2*d} \bmod n$ and then sends them back to the user along with (b_1, b_2). Now (t_1, t_2, b_1, b_2) denote the blind signature.

4) Unblinding phase: After receiving (t_1, t_2, b_1, b_2) from the signer, the user computes a_1b_1 and a_2b_2 and finds two integers w and t such that $a_1b_1w + a_2b_2t = 1$. The parameters (a_1, a_2, w, t) are kept secret by the user. Then the user computes $s_1 = t_1 * r_1^{-b1} = m^{a1b1d} \bmod n$ and $s_2 = t_2 * r_2^{-b2} = m^{a2b2d} \bmod n$. The user can derive the signature S by computing $S = s_1^{\,w} * s_2^{\,t} \bmod n$ and then publishes (m, S).

5) Verification phase: As a result, S is the signature on the message m. Now anyone can verify the legitimacy of the signature by checking whether $S^e \equiv m \bmod n$.

B. ElGamal based Lee et al's Scheme

Lee et al.'s scheme [2] is based on ElGamal cryptosystem and also consists of five phases. These are: initialization, blinding, signing, unblinding, and verification phases. The signer first publishes the public information in the initialization phase. In the blinding phase, the user blinds her message and sends it to the signer requesting his signature. Then the signer signs on the blinded message in the signing phase. In the unblinding phase, the user derives the signature from the blinded signature. Finally, anyone can verify the legitimacy of the signature in the verification phase. The detail of Lee et al.'s scheme is as follows.

1) Initialization phase: Consider that p and q be two large primes, where $q|(p - 1)$, and $g \in Z^*_p$ with order q. The signer's secret key is x and public key is $y = g^x \bmod p$.

2) Blinding phase: The signer randomly chooses $\hat{k} \in Z_q$ and computes $\hat{r} = g^{\hat{k}} \bmod p$, and sends $\hat{r}$ to the user. Now the user randomly chooses $a, b \in Z_q$ and computes $r = \hat{r}^a * g^b \bmod p$, and blinds her message m by computing $\hat{m} = am\hat{r}r^{-1} \bmod q$ and sends $\hat{m}$ to the signer.

3) Signing phase: (i) The signer randomly chooses $\hat{k}_1$, $\hat{k}_2$, b_1, $b_2 \in Z_q$, and computes $\hat{r}_1 = g^{\hat{k}_1} \bmod p$ and $\hat{r}_2 = g^{\hat{k}_2} \bmod p$. Here $\hat{r}_i$ must satisfy $GCD(\hat{r}_i, q) = 1$. Then he sends $(\hat{r}_1, \hat{r}_2, b_1, b_2)$ to the user.

(ii) After receiving the blinded messages $\hat{m}_1$ and $\hat{m}_2$ from the user, the signer computes $\hat{s}_1 = x\hat{r}_1 + \hat{k}_1 b_1 \hat{m}_1 \bmod q$ and $\hat{s}_2 = x\hat{r}_2 + \hat{k}_2 b_2 \hat{m}_2 \bmod q$ and forwards them to the user.

4) Unblinding phase: (i) First, the user chooses five random numbers (a, b, c, d, e) and keeps them secret. (ii) After receiving $\hat{r}_1$ and $\hat{r}_2$ from the signer, the user computes $r_1 = \hat{r}_1^{\,ab1}\, g^c \bmod p$ and $r_2 = \hat{r}_2^{\,bb2}\, g^e \bmod p$. Then she computes $r = (r_1 r_2)^d \bmod p$. (iii) Now the user blinds her message m by computing $\hat{m}_1 = m\hat{r}_1 \left(\frac{r^{-1}}{2}\right) ad \bmod q$ and $\hat{m}_2 = m\hat{r}_2 \left(\frac{r^{-1}}{2}\right) bd \bmod q$ and sends $\hat{m}_1$ and $\hat{m}_2$ to the singer. (iv) After receiving $\hat{s}_1$ and $\hat{s}_2$ from the signer, the user can derive s_1 and s_2 by computing $s_1 = \hat{s}_1 \hat{r}_1^{\,-1} \frac{r}{2} + cdm \bmod q$ and $s_2 = \hat{s}_2 \hat{r}_2^{\,-1} \frac{r}{2} + edm \bmod q$. Then she can compute $s = s_1 + s_2 \bmod q$. The user publishes (m, r, s) to the public.

5) Verification phase: To verify (m, r, s), anyone can check the equation $g^s = y^r r^m \bmod p$ as follows.

$$
\begin{aligned}
g^s &\equiv g^{s1+s2} \bmod p \\
&\equiv g^{\hat{s}_1 \hat{r}_1^{\,-1} \frac{r}{2} + \hat{s}_2 \hat{r}_2^{\,-1} \frac{r}{2} + cdm + edm} \bmod p \\
&\equiv g^{(x\hat{r}_1 + \hat{k}_1 b1 \hat{m}_1)\hat{r}_1^{\,-1} \frac{r}{2} + (x\hat{r}_2 + \hat{k}_2 b2 \hat{m}_2)\hat{r}_2^{\,-1} \frac{r}{2} + cdm + edm} \bmod p \\
&\equiv g^{(x\hat{r}_1 + \hat{k}_1 b1 \hat{m}_1)\hat{r}_1^{\,-1} \frac{r}{2} + (x\hat{r}_2 + \hat{k}_2 b2 \hat{m}_2)\hat{r}_2^{\,-1} \frac{r}{2} + cdm + edm} \bmod p \\
&\equiv g^{xr + \hat{k}_1 b1 mad + \hat{k}_2 b2 mbd + cdm + edm} \bmod p \\
&\equiv y^r r^m \bmod p
\end{aligned}
$$

C. Untraceability

1) Hwang et al.'s scheme: It is the most important property for a BS scheme. For any given valid signature (m_i, s_i), the signer is unable to link this signature to the message *i.e.* herein the signer can be kept away from tracing the BS. The demonstration is as follows. The signer keeps a set of records $(\alpha_{1i}, \alpha_{2i}, t_{1i}, t_{2i}, b_{1i}, b_{2i})$ for every blinded message. However, when the user reveals (m_i, s_i) to the public, the signer has no way to get any information $(r'_{1i}$ and $r'_{2i})$ from these records. He cannot trace the relation between r_{1i} and r_{2i}. In addition, S consists of s_1 and s_2, neither of which the signer knows. Furthermore, without the knowledge of the secure integers $(a_{1i}, a_{2i}, w_i, t_i, r_{1i}, r_{2i})$, the signer cannot trace the BS [1].

2) Lee et al.'s scheme: The signer will keep a set of record $(\hat{m}_1, \hat{m}_2, \hat{r}_1, \hat{r}_2, \hat{k}_1, \hat{k}_2, \hat{s}_1, \hat{s}_2, b_1, b_2)$ for every blinded message. When the user reveals (m, r, s) to the public, the signer will compute two values $a'd'$ and $b'd'$, where $(a'd' = \hat{m}_1 m^{-1} r_1^{\,-1} \frac{r}{2} \bmod q)$ and $(b'd' = \hat{m}_2 m^{-1} r_2^{\,-1} \frac{r}{2} \bmod q)$, corresponding to each stored value $(\hat{m}_1, \hat{m}_2, \hat{r}_1, \hat{r}_2, \hat{k}_1, \hat{k}_2, \hat{s}_1,$

$\hat{s}_2$, b_1, b_2). However, the signer cannot trace the BS by detecting the equation $r = g^{\hat{k}_1 a' d' b_1 + \hat{k}_2 b' d' b_2 + cd + ed} \bmod p$. Because he does not know cd and ed unless he knew s_1 and s_2. Furthermore, S consists of s_1 and s_2, neither of which the signer knows. Therefore, without the knowledge of the secure numbers (a, b, c, d, e), the signer cannot trace the BS [2].

III. Experimental Analysis

A. Experimental Setup

A simulation of the chosen untraceable BS schemes has been developed, and the computation time requirements for blinding, signing, unblinding and verification operations have been measured. The environment consists of a 3.07 GHz CPU with 2048 MB of RAM running on windows 7 operating system. Programming language C with CodeBlocks 2010 with GMP [12] with 1024 bit modulus has been used for coding purpose. Sender and receiver data receiving time and travelling time of message is assumed to be negligible *i.e.* all computation time do not include the communication time. Besides, it is assumed that all secret unknown random numbers are prepared in advance; therefore their generation time is not considered. Moreover the operations of BS schemes that are not related to cryptography are not considered.

B. Experimental Results

Fig. 1 shows the computation time requirement while handling various operations *i.e.* blinding, signing, unblinding and verification operations employing Hwang et al.'s scheme. The Fig. shows that to blind a message and it requires 4.74ms. Then to sign on the blinded message by the signer, it requires 26.20ms. Later on to unblind the signed message by the user, it takes 11.31ms. Finally the verification of the BS by the user or by any third party using the signer's public verification key, it requires 0.24ms.

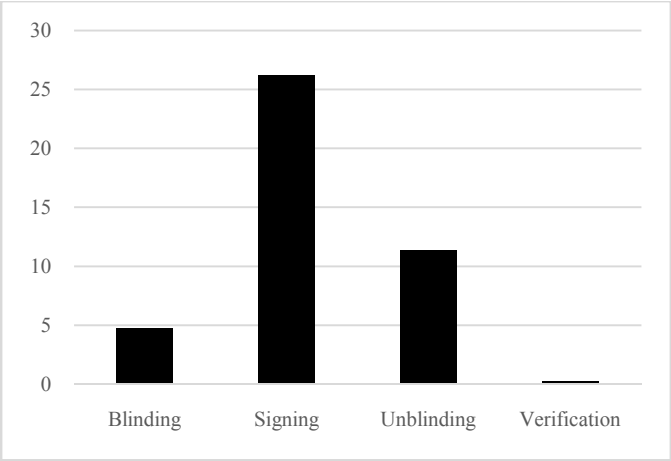


Fig. 1. Computation time requirement by Hwang et al.'s scheme.

Fig. 2 shows the computation time requirement to handle various operations *i.e.* blinding, signing, unblinding and verification operations employing Lee et al.'s scheme. The steps of this scheme are alike as the step of Hwang et al.'s scheme as mentioned in the above paragraph. However the manners of cryptographic techniques to conduct the operations are different. The Fig. shows that to blind the message, it needs 0.19ms. Then to sign on the blinded message by the signer, it needs 0.11ms. Later on to unblind the signed message by the user, it takes 0.17ms. Finally the verification of the BS by the user or by any third party using the signer's public key, it requires 0.09ms.

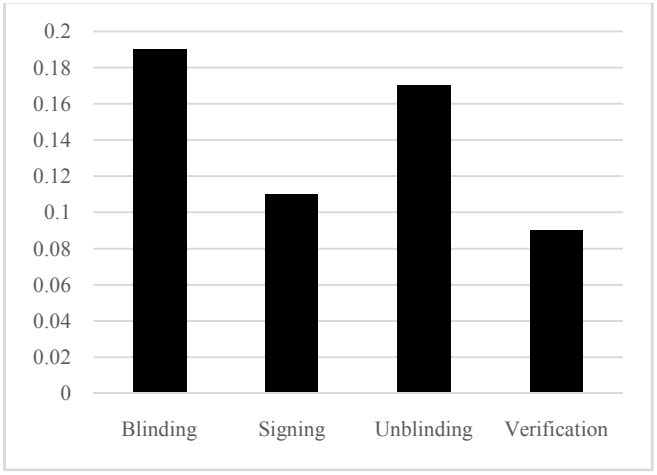


Fig. 2. Computation time requirement by Lee et al.'s scheme.

TABLE 1
Comparison between Hwang et al. and Lee et al.'s Schemes

Operation	Time requirement (ms) for the schemes	
	Hwang et al.	***Lee et al.***
Blinding Phase	4.74	0.19
Signing Phase	26.20	0.11
Unblinding Phase	11.31	0.17
Verification Phase	0.24	0.09

Table 1 shows the comparison of computation time requirement between Hwang et al.'s scheme and Lee et al.'s scheme. To blind a message, Hwang et al.'s scheme takes 4.74ms whereas Lee et al.'s scheme takes only 0.19ms. Later on while blinding a message, Hwang et al.'s scheme takes 26.20ms whereas Lee et al.'s scheme takes only 0.11ms. Here, Hwang et al.'s scheme uses another two primes with the secret signing key whereas in case of Lee et al.'s scheme it requires three random numbers. To unblind a signed message, Hwang et al.'s scheme takes 11.31ms whereas Lee et al.'s scheme takes only 0.17ms. Finally to verify the signature, Hwang et al.'s scheme takes 0.24ms and Lee et al.'s scheme takes 0.09ms. In this point, the comparison of computation shows that the time requirement of various operations by Hwang et al.'s scheme is greater than Lee et al.'s scheme.

IV. Conclusions

In this paper a simulation of two untraceable BS schemes *i.e.* Hwang et al.'s scheme and Lee et al.'s scheme has been performed that evaluates the computation time requirements of the cryptographic operations involved in different steps of the schemes. Although untraceability is a crucial but a tough requirement for any BS scheme, the literature review shows that both Hwang et al. and Lee et al.'s schemes are untraceable. Also both of them satisfy the other requirements of an ideal BS scheme and breaking their security is tough. But the comparison of computation time requirement shows that Hwang et al.'s scheme requires much time than Lee et al.'s

scheme to conduct the simulation. On the contrary, Hwang et al.'s scheme is based on RSA cryptosystem, and by far RSA based schemes are the easiest to understand and implement.

REFERENCES

[1] M. Hwang, C. Lee, and Y. Lai, "An untraceable blind signature scheme," in IEICE Trans. Fundamentals, Vol. E86−A, No. 7, pp. 1902−1906, 2003.

[2] C. Lee, M. Hwang and W. Yang, "A new blind signature based on discrete logarithm problem for untraceability" in Elsevier Applied Mathematics and Computation 164 (2005) 837-841.

[3] E. Mohammed, A. C. Emarah, and K. El−Shennawy, "A blind signature scheme based on ElGamal signature," IEEE/AFCEA EUROCOMM 2000 Information Systems for Enhanced Public Safety and Security, pp. 51–53, 2000.

[4] D. Chaum, "Blind signatures for untraceable payments," Advances in cryptology, CRYPTO'82, pp.199-203, 1982.

[5] A. K. Md. Rokibul, S. Mahmud and M. N. A. Khan "A Comparison between Traceable and Untraceable Blind Signature Schemes through Simulation," Proceedings of International Conference on Informatics, Electronics & Vision (ICIEV13), Dhaka, Bangladesh, May 2013.

[6] M. Kwon and Y. Cho, "Randomization enhanced blind signature schemes based on RSA," in IEICE Trans. Fundamentals, Vol. E82, No. 1, pp. 1−4, 1999.

[7] C.-I. Fan, W. K. Chen, and Y. S. Yeh, "Randomization enhanced Chaum's blind signature scheme," Comput. Commun, Vol. 23, pp. 1677-1680, 2000.

[8] F. Chiang, Y. Lin, and Y. Chang, "Comments on the Security Flaw of Hwang et al.'s Blind Signature Scheme" in International Journal of Network Security, Vol.6, No.2, pp.185–189, Mar. 2008

[9] Z. Shao, "Improved user efficient blind signatures," Electronics Letters, Vol. 36, No. 16, pp. 1372-1374, 2000.

[10] W. S. Juang and C. L. Lei, "Partially blind threshold signatures based on discrete logarithm," Comput. Commun, Vol.22, pp.73-86, Jan. 1999.

[11] A. K. M. Rokibul, S. Tamura, S. Taniguchi, and T. Yanase, "An Anonymous Voting Scheme based on Confirmation Numbers," IEEJ Transactions EIS, Vol. 130, No. 11, pp. 2065-2073, November 2010.

[12] T. Granlund. GMP Multiple Precision Arithmetic Library (GMP), Software available at http://gmplib.org/ (2014-10).

Short Paper: Enhancing Wi-Fi Security Using a Hybrid Algorithm of Blowfish and RC6

Nusrat Jahan Oishi, Md.Arafin Mahamud,Asaduzzaman
Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
cse_oishi@yahoo.com, arefin.cuet@gmail.com, asadcuet@gmail.com

Abstract— **Nowadays, Wi-Fi is a very popular Technology. Faster data transfer and Security both are very important for Wi-Fi. At present, Advanced Encryption Standard (AES) is used for Wi-Fi that is more secured than other encryption algorithms but it is not so fast because of it's complex functions. On the other hand, Blowfish is a very faster encryption algorithm but it cannot apply on Wi-Fi because of it's security problems. In this paper, a hybrid algorithm of Blowfish and Rivest Cipher 6 (RC6) is proposed that solves the security problems of Blowfish and also maintains the fastness of Blowfish and makes it able to use it in place of AES. It uses two innovative criteria. One is ingenious confusion process using two random numbers "a" and "w" that removes reflectively weak key attack and Known plaintext attack of Blowfish. The other is usage of one S-Box by overlapping process that eliminates the collision key attack of Blowfish. Sub key generation process of this algorithm also removes the Brute Force attack of AES. This paper tries to give an efficient algorithm that enhances the performance of Blowfish algorithm by adding a function of RC6 with it. The adding process is trickily handed here that makes the proposed algorithm as fast as Blowfish and also secured like existing AES. The proposed algorithm takes less encryption –decryption time like Blowfish and also secured like AES. Throughput, Average time for different data lengths and attack removal process are used to measure the efficiency of this proposed algorithm.**

Keywords—Blowfish;RC6;Collision key attack;Known plaintext attack;reflectively weak key attack (key words)

I. INTRODUCTION

Wi-Fi is a very popular technology but security is a matter of great concern for the field of Wi-Fi. Among many security processes, Cryptography is very popular network security process where the message of any formats is converted into a encrypted version that is unreadable by a human or computer. There are two types of cryptography Algorithms are found: one is Symmetric key cryptography where same key is used for both encryption and decryption(e.g. AES,Blowfish,RC6)and the another is Asymmetric key cryptography where different keys are used for encryption and decryption(e.g. RSA).

At present, AES [2] Symmetric key encryption algorithm is used for Wi-Fi network security but it is not so fast. On the other hand Blowfish algorithm is so fast but it has some security problems. In this paper, a 128 bit hybrid algorithm of Blowfish and RC6 is proposed that removes the security problems of Blowfish and also take less Encryption decryption time than AES.

A. Blowfish

In [1], Blowfish is 64 bit symmetric key algorithm which contains eighteen 32 bit sub keys and four 32bit S-boxes with 256 entries each. The main function of it is given below:
The Encryption process of Blowfish is , The input is a 64-bit data element, x. Divide x into two 32-bit halves: xL, xR.
Then, for i= 1 to 16.
XL = XL XOR Pi
XR = F(XL) XOR XR
Swap XL and XR
After the sixteenth round, swap XL and XR again to undo the last swap. Then, XR = XR XOR P17 and XL = XL XOR P18. Finally, recombine XL and XR to get the cipher text.
The F function is: F(XL) = ((S1,A + S2,B mod 2^32) XOR S3,C) + S4,D mod 2^32. Here 64 bit is divided among A,B,C,D registers where each of the register contains 8 bit.

B. RC6

In [3], RC6 is a 128 bit symmetric key encryption algorithm. The procedure is given below:

- Input: Plaintext is stored in four w-bit input registers A,B,C,D .Number r of rounds. w-bit round keys S[0,…….,2r+3].
- Output: cipher text is stored in A,B,C,D.
- Procedure: B=B+S[0]
  D=D+S[1]
  For i=1 to r do
  {
  t=(B+(2B+1)) <<< lgw
  u=(D+(2D+1))<<< lgw
  A=((A XOR t) <<< u)+S[2i]
  C=((C XOR u) <<< t)+S[2i+1]
  (A,B,C,D) = (B,C,D,A)
  }
  A = A+S[2r+2]
  C = C+S[2r+3]

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

II. Related Works

In[5],The B-R algorithm is also a mixer of Blowfish and RC6. It is also a 128 bit algorithm and the algorithm uses two S-boxes with 259 entries each. But this algorithm's time complexity is too large because in every iteration it uses two functions: one is Blowfish function and the other is RC6 function and it also contains the risk of Reflectively weak key attack and collision key attack for using the same function and two s-boxes. In[8],a comparison is made that proved Blowfish takes minimum encryption + decryption time that may be helpful for Wi-Fi but this paper did not discuss anything about it's security. In[7],four cases were shown by mixing and changing the number of the XOR and addition function of 'F' function but this does not able to remove the reflective and collision attack.

III. Proposed Algorithm

A possible faster and secure encryption algorithm is proposed here: In this proposed algorithm, 128 bit block of plaintext will be used as input. Here sub key generation of blowfish is used for making cipher text more powerful against brute force attack. The p-array consists of eighteen 64 bit sub keys from p1, p2……p18.Here will be used one 64bit s-box with 263 entries for substitution purpose. F(XL) that is used for Blowfish encryption function for finding next XR will be adjusted here with one s-box. Using one s-box can able to risk of collision attack between more than one S-boxes. 1-16 round of iteration is divided between Blowfish and RC6 modified using a variable 'a'.The value of 'a' is only known by sender and receiver. This type of variation will be able to reduce the risk of reflectively weak key attack[4]. The rotation number 'w' that is used at the portion of RC6 is also a variable that only known by sender and receiver.

The proposed algorithm is given below,

Input: Plaintext 128-bits P.
Output: Cipher text 128-bits C.
Begin:

```
Split plaintext into two 64-bit halves: XL, XR
For I= 1 to a Do
        XL=XL XOR P[I]
        XR = F (XL) XOR XR
        Swap XL and XR
For I= a to 16 Do
        t= (XL × (2XL +1)) <<<w
        u= (XR × (2XR +1)) <<<w
        XL= (XL XOR t) <<< u) + P [I] // first 32 bits of P[I]
        XR= (XR XOR u) <<< t) + P [I] // second 32 bits of P[I]
        Swap XL and XR
XR = XR XOR P17 and XL = XL XOR P18
```

End.

The new F – function with one S-box is given below,

Input: XL (64-bits)
Step1: Divide XL into eight 8-bits: a, b, c, d, e, f, g and h.
Step2: Store eight 8-bits in index [8], where index [1] =a, index [2] =b,..etc
Step3:

```
For i: =1 up to 8 Do
        X[i]=0
        For j:=0 up to 7 Do
                X[i]=X[i] OR (S1[index[i]+j])
        ROTATE_RIGHT (X[i], 8)
Next j : Next i
Y1=(((X[1] + X[2]) MOD 2^32)  XOR X[3]) + X[4]) MOD 2^32
Y2=(((X[5] + X[6]) MOD 2^32) XOR X[7]) + X[8]) MOD 2^32
        Combined Y1, Y2 into Y. Output: Y ( 64 bits).
```

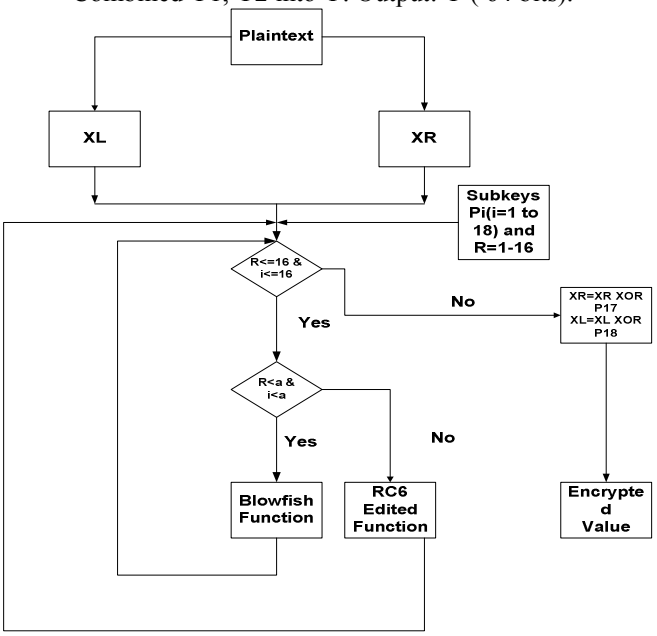
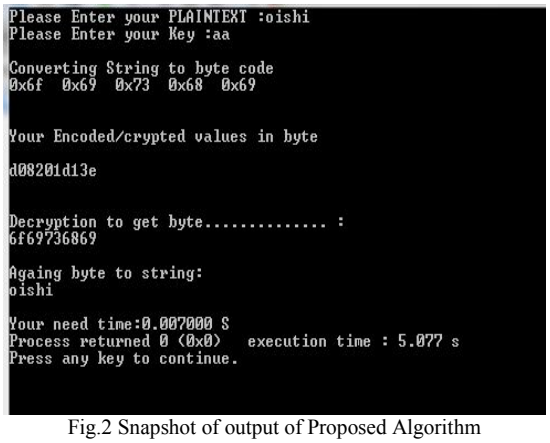


Fig.1 Flowchart of Proposed Algorithm

IV. Experimental Procedure and Input- Output

In Fig.2 and Fig.3 show the Encryption + Decryption time for any data length. The whole process is implemented in C.



```
Please Enter your PLAINTEXT :oishi
Please Enter your Key :aa

Converting String to byte code
0x6f  0x69  0x73  0x68  0x69


Your Encoded/crypted values in byte

d08201d13e


Decryption to get byte.............. :
6f69736869

Againg byte to string:
oishi

Your need time:0.007000 S
Process returned 0 (0x0)   execution time : 5.077 s
Press any key to continue.
```

Fig.2 Snapshot of output of Proposed Algorithm

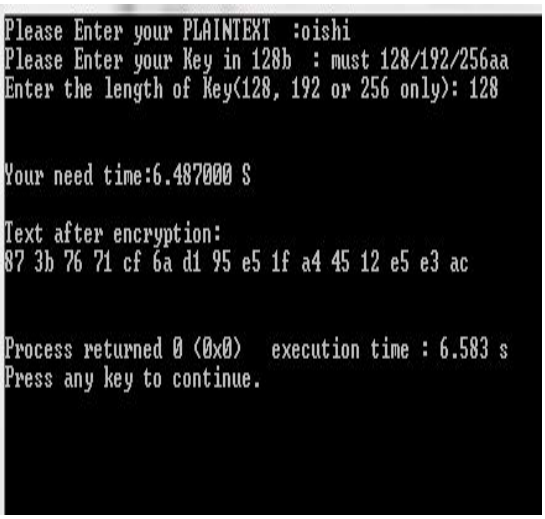


Fig.3 Snapshot of output of AES Algorithm

V. Performance Matrics

A. *Average Time:*

The average time of this experiment is calculated from total (Encryption + Decryption) time. Here take ten samples of total (Encryption + decryption) time and then calculate the average time of these samples.

B. *Throughput:*

Throughput is calculated from division of the total data size(Kbytes) by total time(Sec).

VI. Experimental Result

A. *Blowfish Vs Proposed Algorithm:*

In Blowfish , every iteration uses only a simple function of Blowfish. For these reason, Blowfish algorithm is not so strong in making confusion for the intruders. The proposed algorithm uses a variable "a" that's value is decided by sender and receiver to divide the 16 iterations into two different functions. So here exactly the number of iteration is not increased. Only the iterations are divided. So time complexity is not increased for this reason. But this algorithm uses one S-Box instead of 4 S-Boxes. So here a overlapping process is used for fitting the F–function with one S-Box that is happened by little iteration process. So in proposed algorithm time complexity is little larger than Blowfish algorithm but less than AES. Like[9], here also gives a table of time complexity comparison from where Throughput and Average time is found to show the comparison. and Fig.4 and Fig.5 show the graphical view of this comparison.

TABLE I. Comparative Throughput(Kbytes/Sec)and Average Encryption-Decryption Time of Blowfish and Proposed Algorithm with Different Data Length

Data length In Bytes	Blowfish Algorithm	Proposed Algorithm
22	15.447	19.018
50	31.705	35.337
64	38.368	40.875
79	43.454	48.966
90	48.250	51.379
110	52.920	56.779
128	64.856	68.986
190	70.559	72.109
200	73.936	76.802
220	75.355	78.900
Average Encryption-Decryption Time	51.485	54.9151
Throughput(Kbytes/Sec)	2.23	2.09

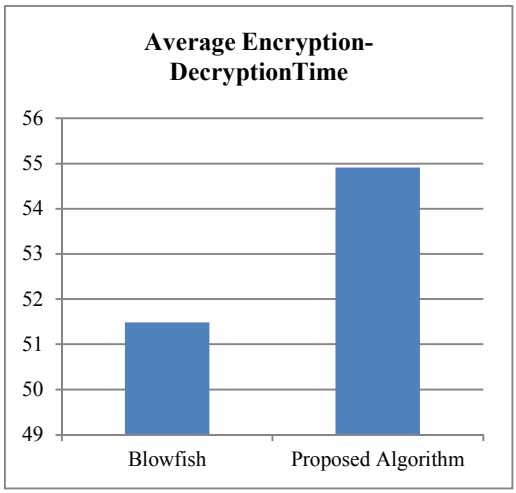


Fig.4 Comparative Graph of Average Encryption-Decryption time Between Blowfish and Proposed Algorithm

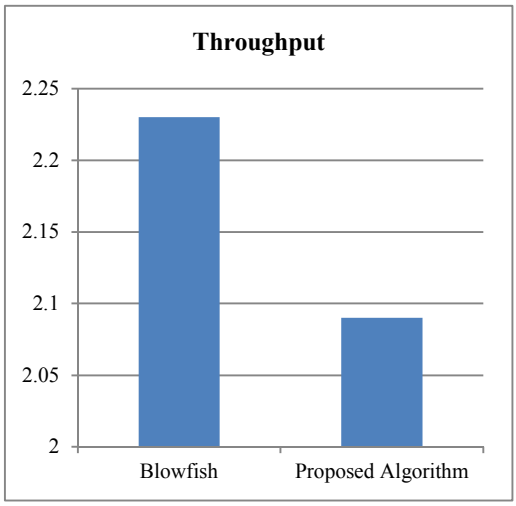


Fig.5 Comparative Graph of Throughput Between Blowfish and Proposed Algorithm

B. *Existing Wi-Fi Security System (AES) Versus Proposed Algorithm*

Nowadays, AES is used for WPA2 Wi-Fi protocol for ensuring Wi-Fi security. It is very secure but it's time complexity is high because it uses four large function in every iteration of it's encryption decryption process. But in proposed algorithm it uses two different functions those are divided into different iterations. So it's time complexity is smaller than AES. Like [9] ,here also gives a table of time complexity comparison from where Throughput and Average time is found to show the comparison. Table II shows the comparison and Fig.6 and Fig.7 show the graphical view of this comparison

TABLE II. COMPARATIVE THROUGHPUT(KBYTES/SEC) AND AVERAGE ENCRYPTION-DECRYPTION TIME OF AES AND PROPOSED ALGORITHM WITH DIFFERENT DATA LENGTH.

Data length In Bytes	**AES Algorithm**	**Proposed Algorithm**
22	41.592	19.018
50	72.816	35.337
64	80.932	40.875
79	87.971	48.966
90	91.135	51.379
110	98.679	56.779
128	102.506	68.986
190	115.695	72.109
200	118.805	76.802
220	120.225	78.900
Average Encrption-Decryption Time	93.0356	54.9151
Throughput(Kbytes/Sec)	1.23	2.09

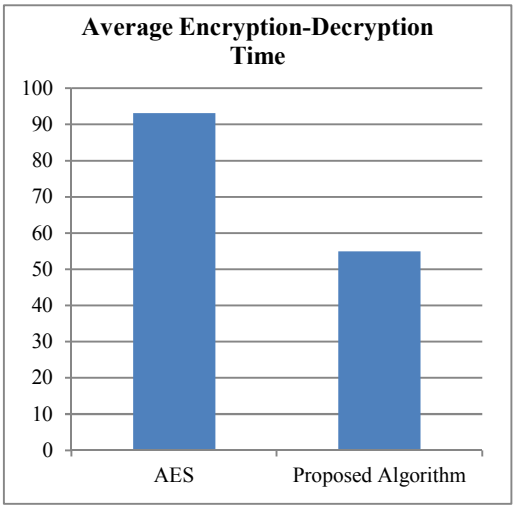


Fig.6 Comparative Graph of Average Encryption-Decryption Time Between AES and Proposed Algorithm

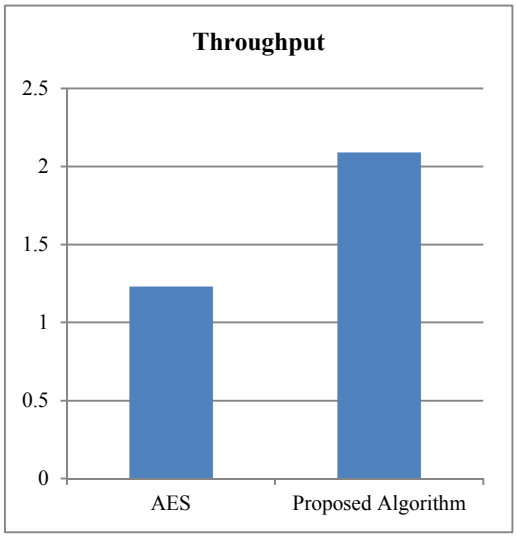


Fig.7 Comparative Graph of Throughput Between AES and Proposed Algorithm

VII. DIFFERENT TYPES OF ATTACK REMOVAL PROCESS

A. *Reflectively Weak Key Attack Removal Process*

In [4] , Self similarity attack is known as reflectively weak key attack. One function is reflectively used by Blowfish that creates this Reflectively weak key attack. Using "a" random variable this proposed algorithm divides the 16 iteration between two functions that removes the reflectivity of one function and the value of "a" is only known by sender and receiver.

B. *Collision Key Attack Removal Process*

In [6] ,describes that, This attack occurs when there is at least one collision into one of the four S-Boxes. Using one S-box with 263 entries this proposed algorithm removes this attack.

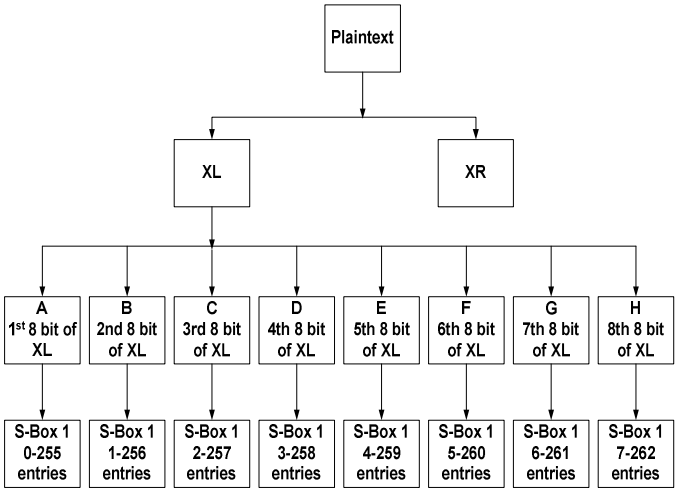


Fig.8 Flowchart of Removal process of Collision key Attack(128 bit)

C. Known Plaintext Attack Removal Process

The known-plaintext attack is an attack where the attacker has samples of both the plaintext and its encrypted version. In blowfish for one simple function there found fixed encrypted value for fixed plaintext according to a fixed key. But this algorithm makes different encrypted value for same key and same plaintext for different value of "a" and "w". Fig.9 and Fig.10 show the different encrypted values for same plaintext

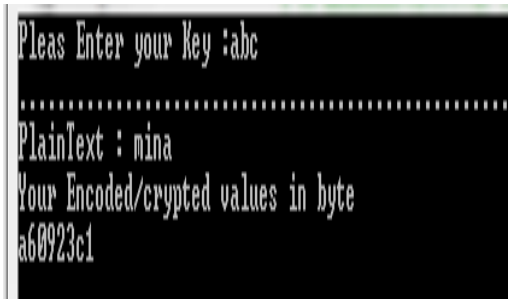


Fig.9 Known Plaintext Attack (1)

Fig.10 Known Plaintext Attack (2)

D. Brute Force and Dictionary Attack Removal Process

Brute force and Dictionary Attack consist of systematically checking all possible keys or passwords until the correct one is found. AES faces this problem because it has no sub key generation process to make strong enough it's key. In this algorithm here Sub key generation table is used that makes a key so much strong to remove these two attacks.

VIII. WHY PROPOSED ALGORITHM IS BETTER THAN EXISTING AES

AES is used for WPA2 Wi-Fi protocol for ensuring Wi-Fi security. But it's time complexity is large because it uses four complex functions. The proposed Algorithm is a enhanced version of Blowfish. Blowfish is faster than AES but it has some security problems. Proposed Algorithm solves this problems by adding a edited function of RC6 and also maintain less time complexity. The sub key generation process of this algorithm also removes the Brute force Attack of AES. By this way the proposed algorithm enhance the faster Blowfish algorithm and make it able to use it for Wi-Fi security in place of AES.

IX. CONCLUSION

Considering the concept of Faster and secured data transfer this paper tries to produce a faster algorithm like blowfish and secured like AES. The proposed algorithm improves the faster algorithm Blowfish by adding the edited function of RC6 and removing it's different attacks. It also uses "a" and "w" random variable to confuse the intruders by making different cipher text. It's one S-box criteria makes the time complexity little higher than Blowfish but it is not so high as the existing algorithm AES that is used in WPA,WPA2 web protocol. By this way, this paper is able to give a secured and faster algorithm than the existing AES. The whole process is implemented for text encryption. In future, it may be used for image file, PDF file and video file encryption.

REFERENCES

[1] B. Schneier, "Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)", *Fast Software Encryption, Cambridge Security Workshop Proceeedings*(December 1993),Springer-Verlag,1994, pp 191-204.

[2] Eric Conrad, **"**Advanced Encryption Standard", *GIAC Research in the Common Body of Knowledge*, California, 2007.

[3] R. L. Rivest , M.J.B. Robshaw , R. Sidney and Y.L. Yin , "The RC6 TM Block Cipher" ,M.I.T. Laboratory for Computer Science, 545 Technology Square, Cambridge,MA 02139 USA,RSA Laboratories, 2955 Campus Drive, Suite 400, San Mateo,CA 94403, USA, Version 1.1 ,August 20, 1998.

[4] Orhun Kara and Cevat Manap, "A New Class of Weak Keys for Blowfish", *TÄUB_ITAK UEKAE*, Gebze, Kocaeli, Turkey *f*orhun.

[5] Janan Ateya Mahdi*, "*Design and Implementation of Proposed B-R Encryption Algorithm" ,*IJCCCE,* VOL.9, NO.1, 2009.

[6]Evilcry,"Blowfish Study n' Reverse", http://evilcry.altervista.org

[7] Vaibhav Poonia, Dr. Narendra Singh Yadav, " Analysis of modified Blowfish Algorithm in different cases with various parameters", *International Journal of Engineering Research and General Science*, Volume 3, Issue 1, ISSN 2091-2730, January-February 2015.

[8] Gurjeevan Singh, Ashwani Kr. Singla, K.S. Sandha," Superiority of Blowfish Algorithm in Wireless Networks", *International Journal of Computer Applications,* Volume 44, No.11, pp-0975 – 8887, April 2012.

[9] Mar Preet Singh and Raman Maini ," Comparison Of Data Encryption Algorithms", *International Journal of Computer Science and Communication*, Vol. 2, No. 1,pp. 125-127, January-June 2011.

Chameleon: Defending Secret Information from Eavesdropping Over Physical Environment

Saiyma Sarmin[1], Saurabh Bagchi[2], A. B. M. Alim Al Islam[3]
[1,3]Dept. of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh
[2]Dept. of Computer Science
Purdue University
West Lafayette, Indiana
[1]smrity48@gmail.com, [2]sbagchi@purdue.edu, [3]alim_razi@cse.buet.ac.bd

***Abstract*—Eavesdropping of secret information is usually being prevented through using cryptography-based mechanisms. However, these mechanisms cannot guarantee protecting ones physical environments as information can be eavesdropped even before being encrypted, for example, by shoulder surfing. To address this problem, we propose a new technique called *Chameleon* to protect secret information from eavesdropping when the information is in transmission over physical medium. *Chameleon* is light-weight, easy-to-use, software-based solution to the solution. In *Chameleon*, we use a string, which is pre-mapped to a secret information, for transmitting over physical medium. After the string being received by the intended receiving device, a software module of *Chameleon* maps the string to the secret information and feed the secret information to the intended application. We perform real implementation of *Chameleon* to show its practical applicability.**

***Index Terms*—Eavesdropping, shoulder surfing, string mapping.**

I. INTRODUCTION

Spying on people has always been an effective way of obtaining information since the beginning of history. With the advent of technology, new ways of spying on somebody's communications have been devised, and consequently, new countermeasures have been developed as well. One of such most effective countermeasures to protect secret information while being in transmission through electronic communication is the use of cryptography-based mechanisms.

Encrypting secret information can guarantee protecting information during electric communication such as communication over a wireless network. However, this countermeasure cannot guarantee protecting one's secret information when the information is in transmission over physical environments. Here, the information can be eavesdropped even before being encrypted. Eavesdroppers can easily extract the secret information by monitoring somebody typing on a keyboard [4], observing the light reflected by the walls of a room [3], or even through analyzing the sound produced by keyboards [2]. For example, if a person types his password in a computer and someone else monitors him through shoulder surfing, then the password is no more guaranteed to be protected. Such eavesdropping becomes even more difficult to protect on recent advanced devices such as iPads, smartphones, and Google Glasses, which are expected to be pervasive all over the world in near future.

A number of research studies have experimentally demonstrated the potential of eavesdropping over physical environment. For example, a research study [1] conducted at the University of Massachusetts Lowell has showed that video from wearable devices such as Google Glasses and the Samsung smartwatch can be used for surreptitiously capturing four-digit PIN codes typed on an iPad from almost 10 feet away. Moreover, using a high-definition camcorder can perform similar capturing from nearly 150 feet away. The capturing tasks exploit a custom-coded video recognition algorithm, which is capable capturing without any image of the target devices' displays. These evidences clearly demonstrate significant threats of eavesdropping to secret information in transmission over physical environment.

To defend eavesdropping of secret information over physical environment, a few solutions [5], [6], [7], [9], [10] have already been proposed. These solutions are generally resource-hungry, complex, and difficult-to-use. Therefore, in this paper, we propose a new light-weight, simple and easy-to-use software based solution. Our proposed solution (named as ***Chameleon***) takes a string as input in place of a secret information such as a password. The string is not actually the secret information, however, is pre-mapped to the secret information within an electronic device before being utilized. Here, if an eavesdropper gets the string, he can never know the secret information only through the eavesdropping as the mapping is done within the device.

We make the following set of contributions in this paper:

- We propose a novel system for defending eavesdropping of secret information over physical environment.
- We implement our proposed system in real devices to demonstrate that how we can use the system in real life.

II. BACKGROUND AND RELATED WORK

Looking over someone's typing to get passwords, PINs, and other sensitive personal information has always been a problem that is difficult to overcome. When a user enters secret information using a keyboard, mouse, touch screen, or any traditional input device, a malicious observer may be able to acquire the secret information through eavesdropping the physical environment before the information gets entered in the devices. A few solutions have been proposed for this problem in recent times.

978-1-5090-0203-0/16/$31.00

Sobrado et al., [9] have proposed a scheme based on spatial relationships of geometrical shapes. Here, the password is a set of graphical symbols that are displayed in random order on a computer monitor. A user must choose the set of symbols through clicking inside a convex hull during the authentication period. To achieve sufficient confidence in the identity of the user, several correct clicks are required. This requirement makes the system complex and difficult-to-use.

Another authentication scheme, which claims to introduce resistance against shoulder surfing, is based on mathematical operations (e.g., requiring the user to perform modular arithmetic) [10]. Again, as the mathematical operations are not easy for all users, this approach remains difficult-to-use in many cases.

Beside, Hoanca et. al., [7] have proposed a password entry scheme that requires entering three keys for one key. Here, the positions of the three keys in a virtual keyboard create a parallelogram. The 4th place of the parallelogram represents the original key. This scheme is too complex for users to select right key, as any wrong selection of the first three keys eventually creates wrong key for the 4th place in the parallelogram.

Additionally, Kumar et. al., [6] have proposed another scheme named EyePassword to reduce shoulder surfing. Here, a user can enter sensitive input such as password, PIN, etc. through selecting keys of an on-screen keyboard using only the orientation of their pupils. This process demands high-resource hardware. Moreover, it offers a complex system to a user.

Another scheme [5] utilizes spy-resistant, security-sensitive onscreen virtual keyboard that allows users to enter private text without revealing it to observers. However, similar to the previous approach, changing the traditional key pattern in keyboard is not user friendly either as this requires considerably more visual search by a user when entering passwords.

Another research study, S3PAS investigates combining textual password with graphical without changing existing user password profiles [11]. However, the major drawback of S3PAS schemes is to offer longer login process. It is slightly more complicated scheme.

Nonetheless, another research study [12], demonstrates a stroke-based textual password entry system. It uses shapes of strokes on the grid as the origin passwords. It also changes the login interface of the system. It allows users to login with text passwords via traditional input devices. However, the major drawback of this system is relativity unfamiliar to the general people. It offers longer login process.

Another study named Web Password Filler [13] also stores the password for the purpose of automatic login. However, in this case anyone can get access the information only if he/she gets the access of the device without eavesdropping.

A common drawback of all the proposed schemes is that they require additional cognitive effort on the part of the user making them complex and difficult-to-use. Further elaboration of graphical password schemes is presented on the research study [14]. Which refer to the longer login process and difficult-to-use characteristics of the graphical password schemes. Besides, few approaches demand high-resource hardware. To the best of our knowledge a light-weight, simple, and easy-to-use scheme is yet to be proposed in the literature. Therefore, we propose such a scheme in this paper. we elaborate the proposed scheme in the next section.

III. PROPOSED SCHEME CHAMELEON

Our proposed scheme ***Chameleon*** utilizes a notion concept of string mapping to protect user's secure information. Here, ***Chameleon*** takes a string, which is not actually the secret information, however, will be eventually mapped to the secret information before being utilized. Fig. 1 shows block diagram and flow of actions of ***Chameleon***. Here, the known string is the user's generated random string, which user will enter in ***Chameleon***.

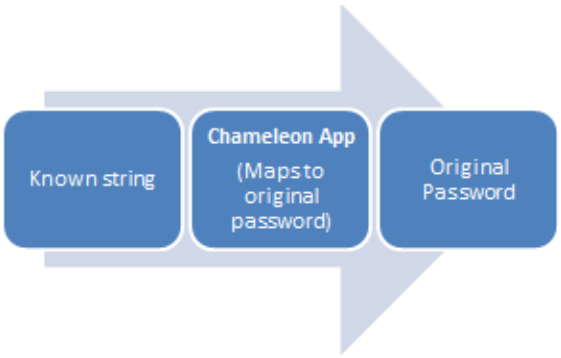


Fig. 1: **Block diagram and flow of actions in *Chameleon***

For example, if a user wants to sign-in to her Gmail account, then she will enter a known string to ***Chameleon***. ***Chameleon*** maps the entered string to the user's password pertinent for her Gmail account. Subsequently, it performs the task of logging in the Gmail account of the user through using the mapped password. In this way, even if an intruder eavesdrops and gets the input string, she can never know the original password as it is kept secret by ***Chameleon***. Consequently, capturing information does not suffice for making security threat.

In ***Chameleon***, a valid user has to provide both the input string and the secret information. It could be encrypted for further security. The real password is eventually mapped from the string before being used. The system directly feeds the secret information to the target application after performing the mapping. Therefore, in case an intruder knows the string being entered by the user without getting access to the system, she will not be able to retrieve the secret information as the mapping is done within the system and the secret information is directly fed to the target application without any visual presence to the intruder. To the best of our knowledge, we are the first to propose such a technique for defending eavesdropping of secret information over physical environment, which does not require any critical intelligence from the part of a user.

IV. IMPLEMENTATION DETAIL OF CHAMELEON

We have implemented a prototype of ***Chameleon***. Fig. 2 shows the Use-Case diagram of ***Chameleon***.

Use-Case narratives of *Chameleon*:

Name: Sign in
Priority: High
Actor: User
Precondition: User needs to add account in ***Chameleon***
Description: User gives email address, password for sign in.

Typical course of events:

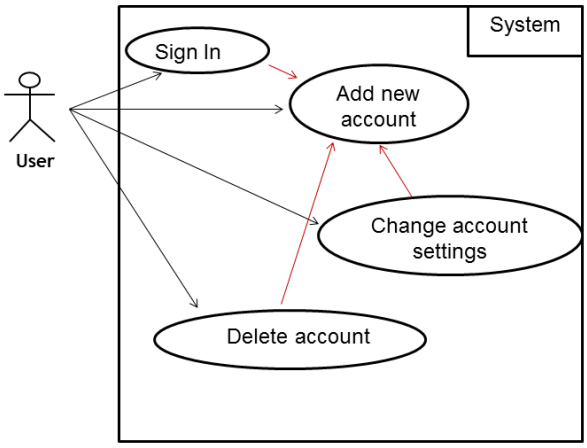


Fig. 2: **Use-case diagram of *Chameleon***

User	System Response
1. User needs to sign in	2. System asks for information
3. User gives information	4. System checks given information
	5.a System maps known string with real password for correct information 5.b System shows an error message for wrong information.

Name: Add new account
Priority: High
Actor: User
Description: User gives input string and real password, email address for add new account.

Typical course of events:

User	System Response
1. User needs to add new account	
2. User selects add new account	3. System asks for information.
4. User gives information	5. System adds new account.

Name: Change account settings
Priority: Low
Actor: User
Precondition: User needs to add account in ***Chameleon***
Description: User gives input string and real password, email address for change account settings.

Typical course of events:

User	System Response
1. User needs to change account settings	
2. User selects change account settings	3. System asks for information.
4. User gives information	5. System saves the modified account.

Name: Delete account
Priority: Low
Actor: User
Precondition: User needs to add account in ***Chameleon***
Description: User gives real password, email address for delete account.

Typical course of events:

User	System Response
1. User needs to delete account	
2. User selects delete account	3. System asks for information.
4. User gives information	5. System deletes the account.

Fig. 3 shows sequence diagrams pertinent for different modules of ***Chameleon***.

Subsequently, ***Chameleon*** maps the entered random password to original password and feeds to the target application. We developed the application for the windows and android platform. We present screenshots of our real implementation of ***Chameleon*** in Fig. 4. In Fig. 4(a), ***Chameleon*** Settings is shown. In Fig. 4(b), user enters input string and real password for storing in ***Chameleon*** database. In Fig. 4(c), user enters the input string to login to Gmail. In Fig. 4(d), ***Chameleon*** maps the input to original password and feeds it to the target application (Gmail in this case).

V. UNIQUENESS AND LIMITATIONS

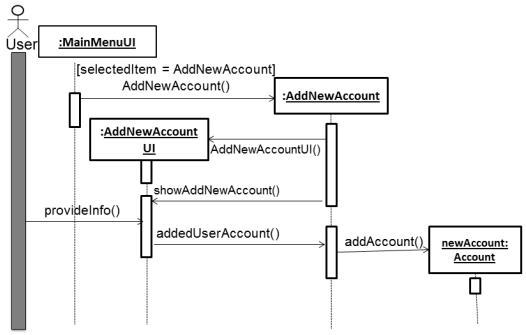
The uniqueness of ***Chameleon*** is its nature of being easy-to-use. It does not require any cognitive of memorization operation in the user's head to perform. Users can easily map their secret information such as passwords to known strings, which can be any easy word. Consequently, ***Chameleon*** reduces the potential of eavesdropping using a simple mapping and will help users to use their secret information in public places without worried about eavesdropping.

In ***Chameleon***, the input string exhibits three significances:

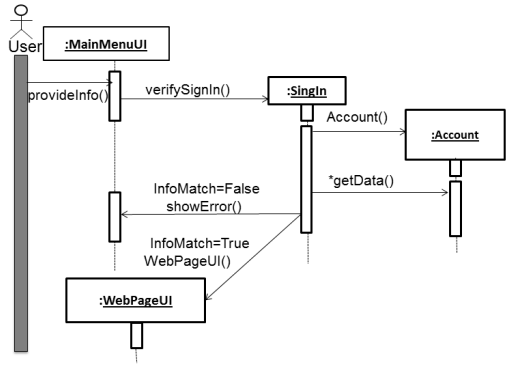
1) The use of input string enables to use a single device by multiple users. The application can be shared as long as one's password has not been eavesdropped by others. Here, the input strings distinguish different users. If we would omit the input string through using only a stand-alone application, only one user could use a device.

2) Devices such as cell phones frequently get lost [8] everywhere, even in developed countries such as USA. If we would omit the input string through using only a stand-alone applications, stolen devices will be a source of getting access to secret information as for getting access to the stand-alone module. However, in case of input string being used, access to secret information will demand the input string for permitting access to the secret information. Consequently, ***Chameleon*** exhibits its vulnerability only if both the device is stolen and the input string is known. We believe that occurring both events simultaneously posses a small probability to happen.

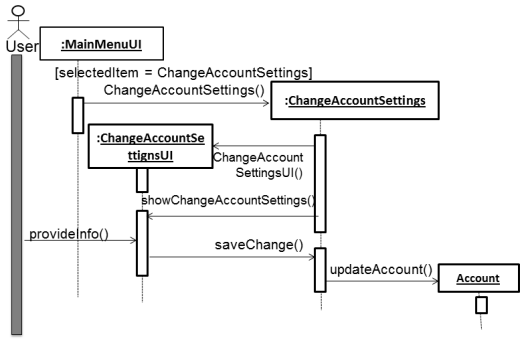
3) Note that, in ***Chameleon*** we use a mapping between an input string picked by the user and a secret information. The picking of input string by a user makes the mapping completely random. If we would use any state-of-the-art mapping technique, for example hashing or encryption of the



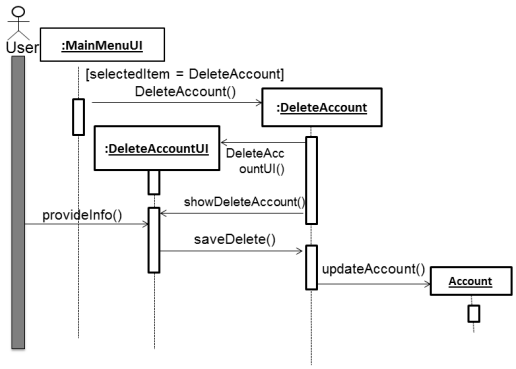
(a) **Add new account**



(b) **Sign in**

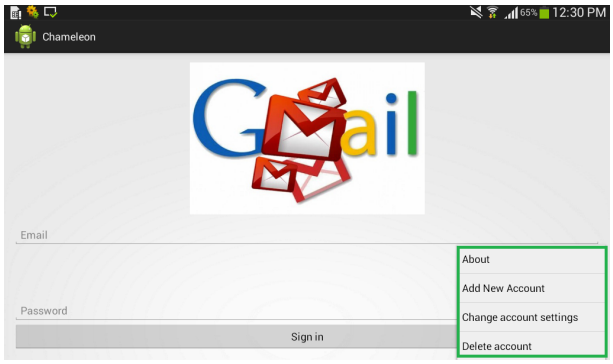


(c) **Change account settings**

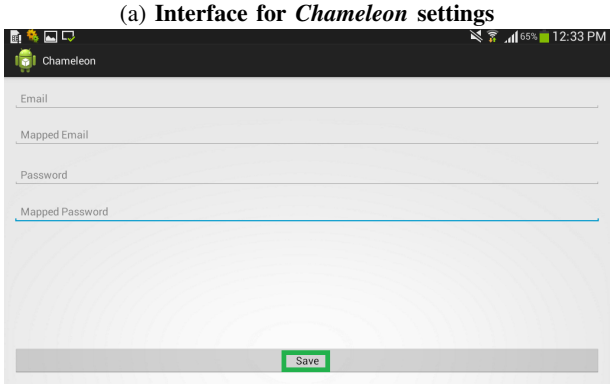


(d) **Delete account**

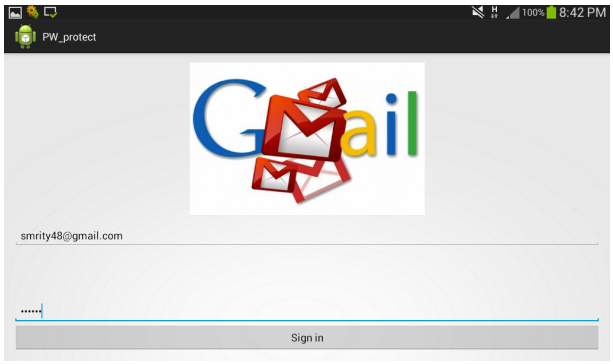
Fig. 3: **S**equence diagrams of *Chameleon*



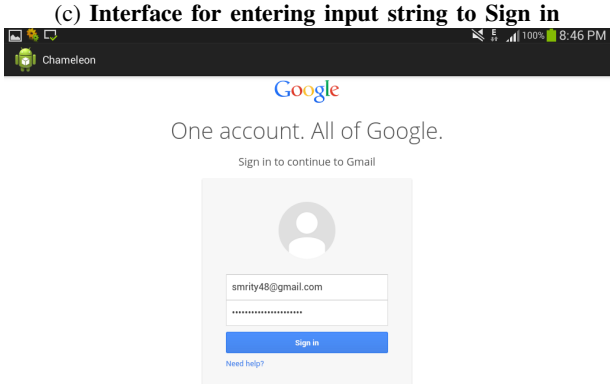
(a) **Interface for *Chameleon* settings**



(b) **Interface for entering input string and real password for stroing in database**



(c) **Interface for entering input string to Sign in**



(d) ***Chameleon* maps input string to real password and feeds to the target application**

Fig. 4: **Screenshoots of *Chameleon***

secret information, in place of the random mapping, the user will need to memorize the input string. This memorization would make the scheme complex and difficult-to-use.

VI. APPLICATIONS

Recent devices such as notebook, smartphone, etc., can exploit the notion of ***Chameleon*** for securing secret information. The application of ***Chameleon*** could be much more appealing for the systems that require voice for extracting secret information. Google Glass is one of such potential application.In such applications, secret information can easily be eavesdropped while being over air. ***Chameleon*** can greatly facilitate security in these cases.

VII. CONCLUSION

Eavesdropping of secret information, while being in transition over physical environment, always remains highly significant. Few schemes address this problem, however, mostly remain complex and difficult-to-use. Therefore, in this paper, we present a novel, simple and easy-to-use system to protect user's secret information from eavesdropping over physical environment. This approach is based on string mapping, where a user sets an input string that is mapped to the original secret information by the system.

We implement the system for android and windows platforms. We are planning to implement the system where users interact with the devices via voice commands. Besides, our application is not applicable for the disable person (e.g., blind); in future, we will propose a better implementation for the disable persons so that their secret information will also be protected from eavesdropping over physical environment.

ACKNOWLEDGMENT

This work has been conducted at and partially supported by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. Besides, this work has been partially supported by the Ministry of Education, Government of the People's Republic of Bangladesh.

REFERENCES

[1] Google Glass Snoopers Can Steal Your Passcode With a Glance, Retrieved on 30 August, 2015 from http://www.wired.com/2014/06/google-glass-snoopers-can-steal-your-passcode-with-a-glance/

[2] D. Asonov and R. Agrawal, Keyboard Acoustic Emanations. In Proceedings of the IEEE Symposium on Security and Privacy, pp. 3-11, 2004

[3] M. Kuhn, Time-Domain Eavesdropping Risks of CRT Displays. In Proceedings of the IEEE Symposium on Security and Privacy, pp. 3-18, 2002.

[4] D. Balzarotti, M. Cova, G. Vigna, ClearShot: Eavesdropping on Keyboard Input from Video, In Proceedings of the IEEE Symposium on Security and Privacy, pp. 170-183, 2008.

[5] D. Tan, P. Keyani and M. Czerwinski, Spy-resistant keyboard: more secure password entry on public touch screen displays, Proceeding OZCHI '05 Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, pp. 1-10, 2005.

[6] M. Kumar, T. Garfinkel, D. Boneh and T. Winograd. Reducing shoulder-surfing by using gaze-based password entry, Proceedings of the 3rd symposium on Usable privacy and security, pp. 13-19, 2007.

[7] B. Hoanca and K. Mock. Password Entry Scheme Resistant to Eavesdropping, Security and Management, Las Vegas, Nevada, 2008, pp. 119-125.

[8] Phone Theft In America. Retrieved on 17 October, 2015 from https://www.lookout.com/resources/reports/phone-theft-in-america

[9] L. Sobrado, J. C. Birget, "Graphical passwords", The Rutgers Scholar, An Electronic Bulletin for Undergraduate Research, vol. 4, 2002.

[10] N. Hopper and M. Blum. A Secure Human-Computer Authentication Scheme. Technical Report CMU-CS-00-139, Carnegie Mellon University, 2000.

[11] H. Zhao and X. Li, S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme. Advanced Information Networking and Applications Workshops, Vol. 2, pp. 467-472, 2007.

[12] Z. Zheng, X. Liu, L. Yin, and Z. LiuA, Stroke-Based Textual Password Authentication Scheme, Education Technology and Computer Science, Vol. 3, pp. 90-95, 2009.

[13] Web Password Filler, Retrieved on 22 October, 2015 from http://thycotic.com/products/secret-server/features/web-password-filler/

[14] A. H. Lashkari, O. B. Zakaria, S. Farmand, and R. Saleh, Shoulder surfing attack in graphical password authentication, International Journal of Computer Science and Information Security, Vol. 6(2), pp. 145-154, 2009.

UProve2: Privacy-Aware, Scalable, Ubiquitous Provenance to Enhance File Search

Annajiat Alim Rasel and Mohammed Eunus Ali
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
West Polashi, Dhaka-1000, Bangladesh
annajiat@gmail.com, eunus@cse.buet.ac.bd

***Abstract*—Rapid growth and highly scattered nature of files make efficient file management (e.g., recalling, identifying, locating, and retrieving files), a challenging problem. Traditional search requires the user to remember the file name, extension, path, storage device, etc. Existing research have shown that users often cannot remember these information accurately, which results in exhaustive, inconvenient, and repeated file search. We have observed that keeping the provenance of files aid in quick locating of desired/latest version of a file. We propose a provenance system that, (1) does not require migration to a modified/new storage/file system, (2) is unified, (3) considers ubiquitous nature of storage systems that are already in use, and (4) enhances file search, and (5) is privacy-aware.**

Keywords—File Search, File Provenance, File Tracking, Ubiquitous Provenance

I. INTRODUCTION

Capabilities of computing systems and storage devices are widely heterogeneous. Desktops and laptops (PCs) have more storage and computing power whereas tablets, smart-phones, and other devices have less. Flash/portable drives have only the storage part. Cloud drives and mailboxes can only store files but do not have computing capabilities from user's viewpoint. PCs may have multiple users but smart devices usually have one user. Portable drives may be shared among multiple users. As files travel [7], [8] through any of these devices, provenance data must be preserved. Existing research [5], [6], [9] use provenance to track files visible from within a single PC only.

Users cannot accurately recall name/location of files [3]. In a scenario where a user has multiple computing devices (e.g., home/office PCs, tablets, smart phones, etc.), files get scattered due to lack of a common storage space and network access. It results in situation like, "I remember emailing it to one of the co-workers but cannot seem to find it in email or usual locations where I save such files". When one or more files are found, user may try to guess the version from file name, location etc., and compare timestamp/content. Instead of the user having to wonder around, when user looks at a file, the file by itself should notify the user about its status, e.g., if it is the latest version or not, which files are related, and how they are related etc. After downloading an email attachment, computing systems should be able to detect its relation to existing other files, recommend name and location to save to. For example, it may suggest to save attached "latest.docx" as "2014 department budget v2.docx" in budget directory of the workspace. "Versionset" tried to provide a partial solution to these problems using a copy-aware computing ecosystem. It identifies and groups related files together and suggests new name for files. However, versionset fails to consider and provide provenance across multiple computing devices used by a user. Its view on provenance is greatly limited to a single user's point of view and constrained within a single computing system e.g. local mailboxes and local/network-shared drives accessible from a PC. Versionset loses provenance data when it travels across computing devices and storage spaces. Assume a scenario, when a user emails a file, "budget v1.xlsx" to own email address from office desktop using online email client or uploads to a cloud storage service. Later, the user resumes further work on that file from home resulting in "budget v3.xlsx". Then the file is brought to office and shared with co-worker both via flash drive. Co-worker returns "budget v6.xlsx" via flash drive. Final version, "budget v7.xlsx" is prepared from office laptop. In this scenario, provenance information is lost at multiple stages, e.g. when put to flash drive, when uploaded to cloud service, or emailed via web-mail client. As there is no unified provenance system, office laptop of the same user does not see prior provenance information. Even though some provenance information is available in office desktop, the system in laptop believes that "budget v7.xlsx" is the only version. Similarly, the system in office desktop believes "budget v1.xlsx" is the latest version! Ideally, provenance tracking should not stop while the file stays in non-local mailboxes / drives or crosses the boundary of a single users view. In this age of collaboration, files and their provenance are not limited to a single computing system or a single user. In order to account for usual scenarios such as the one discussed above, today's provenance systems must consider multiple computing devices, storage spaces with one or more users. Users store their files on heterogeneous devices such as portable or flash drives, online cloud storage, online mailboxes, multiple computing devices including smart devices, not just to local storage. Therefore, it is certain that there exists a scope for a privacy-aware unified provenance system encompassing: (1) multiple and heterogeneous computing systems, (2) multiple users, and (3) heterogeneous storage spaces.

We propose a ubiquitous provenance system that encompasses those. In short, we have following contributions:

- Integrating privacy awareness to our preliminary poster, UProve[10], a ubiquitous file provenance system that

978-1-5090-0203-0/16/$31.00

considers multiple computing systems and storage devices for interoperability.

- Formulate privacy-aware mechanism to retain, share, and aggregate provenance information in flash/portable drives, cloud drives, and from local/online mailboxes.
- Provide the user with an abstraction of a unified virtual large search space built on top of multiple scattered storage spaces and set privacy levels.
- Accommodate multi-user environment including collaborators with varying levels of privacy requirements.

II. Literature Survey

Many of the existing provenance aware systems go on from tracking user activities, file system activities to up to modifying file system, operating system kernel, etc. Some provenance systems consider one or a closed set of some computers where their designed provenance system will work. However, a five-year study of file system meta data by Microsoft says, "The fraction of file-system content created or modified locally has decreased over time. In the first year of our study, the median file system had 30% of its files created or modified locally, and four years later this percentage was 22%". This research outcome raises two concerns. First, as amount of locally modified file is gradually decreasing compared to overall number of files, the future of current provenance systems that mostly considers files within the closed system, becomes uncertain. Second, existing work fails to capture much of the provenance information associated with a file, when it is transferred to another system or do not consider how provenance information can live across boundaries of multiple computing systems.

Earlier systems such as “Connections” considered context of users work by taking file I/O and window focus into account. Provenance Aware Storage System (PASS) did significant work in showing how provenance awareness can be built into systems and proposes a storage system. It also identified some research challenges for provenance for data moving out of the system. However, PASS is more concerned about data flow and does not solve file provenance, file search enhancement, and specially interoperability of provenance systems.

Temporal locality based provenance systems tried to build provenance on the basis of work done within a time window. However, they erroneously group unrelated files together on which the user is working concurrently.

Causal provenance systems consider data flow among applications. They show erroneous result when the same software is used for multiple unrelated work.

TaskTracer tracks data flow among applications and provided detailed documentation of actions, flow of information, and lifetime of these information. It provided the users with provenance-based aid to recall and identify desired files.

Leyline enhanced TaskTracer by providing graphical search tool without requiring a file name. Files can be searched by actions/events e.g. copy/paste, download etc. It improved visual representation of both query and representation of provenance. However, it does not address interoperability of provenance systems.

A copy-aware computing ecosystem with Versionset captures provenance for PCs but fails to consider that provenance needs to be associated with files or storage devices instead of computing systems. It cannot retain provenance information outside locally mounted file systems. It does not consider multiple computers, users, online mailboxes, cloud drives, flash drives etc.

Existing solutions such are greatly limited to the scope of single user. They are constrained within a single computing system e.g. local mailboxes and local/network-shared drives accessible from a PC. Those lose provenance data (or do not consider at all!) when a file travels across computing devices and storage spaces. A system that works beyond local file system (and mailboxes), considering online mailboxes, other devices, stitching together provenance records across device/system to provide ubiquitous solution is absent.

III. UBIQUITOUS PROVENANCE TO ENHANCE FILE SEARCH

UProve2 obtains provenance information from file metadata as well as through monitoring user activities on files. For enhanced user experience, visualizer and search tools are integrated with OS GUI shells. UProve2 can assist authorized applications though APIs. For example, during insertion of a diagram, word processing application can show options to choose file from directories related to the file the user is working on. UProve2 consists of the following components:

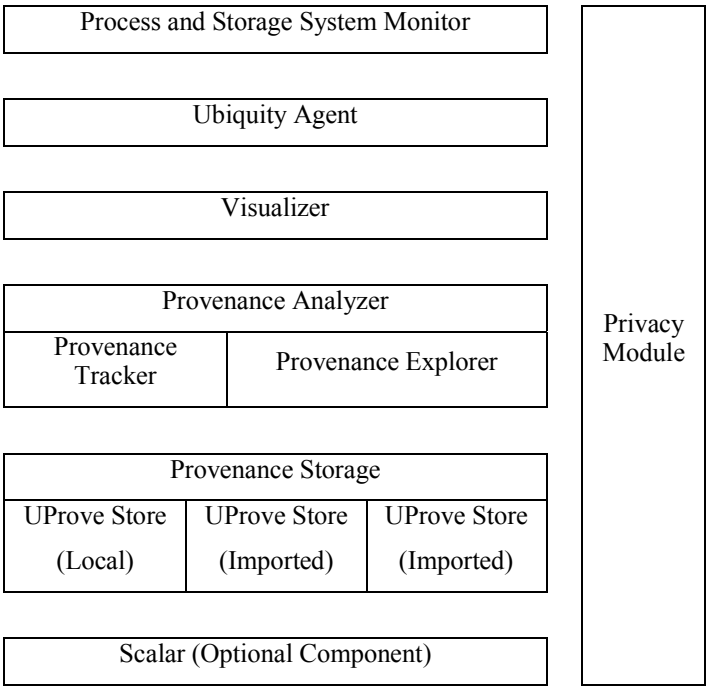


Fig. 1. Overview of UProve2 Architecture

A. Architecture

- Process and Storage System Monitor: For our prototype, we used "Process Explorer" by Mark Russinovich, a 3rd party tool to monitor File System. Different operating systems may have different monitoring tools/techniques available e.g., API hooks, process/file system monitoring, logs, etc. For monitoring other storage e.g., mail boxes, cloud drives, additional components may be added.
- Ubiquity Agent: It synchronizes provenance across different computing systems and storage devices with user's permission. The synchronization can be done for (1) all files, or (2) only files already available in either/both storage space. This synchronization takes place when a storage device becomes accessible (over the network or locally) from a computing system. For privacy and security, it features severing/cutting off desired links/edges in the provenance graph when synchronizing provenance data across devices.
- Visualizer: As provenance is a graph data, Visualizer provides navigational features somewhat like the ones provided by Google/Yahoo/Bing maps. Provenance can be explored in multiple dimensions: (a) time dimension (following a time line), (b) versions [1], [4], (c) by related directories, (d) based on a specific file, directory, or application, and (e) limiting or expanding visualization zone or scope to provenance data of other systems and storage devices.
- Provenance Analyzer: It consists of a tracker and an explorer module. "Provenance Tracker" keeps track of first class provenance data collected by process and file system monitoring. “Provenance Explorer" module tries to discover additional provenance by exploring metadata.
- Provenance Storage: It consists of one or more UProve Store. Provenance for files in a particular storage device is stored in the same device (e.g., local, flash, portable, and cloud drives) in an UProve Store. For online email accounts (with IMAP/POP3) and devices with limited storage capacity, provenance is stored at the local drive of the computing device that is accessing those resources. This results in multiple UProve Store. Synchronization also results in additional UProve Store. To reduce data duplication, provenance from each UProve Store is combined during query.
- Privacy Module: It assists the user in setting privacy policies and assists all other modules in functioning with specified privacy preferences.

B. Designing a System for File Provenance

File provenance is maintaining a record or history of files in as much details as possible. It may include various file information e.g., creation/ modified/ accessed date/timestamps, where a file came from, author information, who modified it and when etc. Provenance information may be obtained from file meta data as well as monitoring of user activities on files. This activities may include create / copy / move / download a file. Some of the actions such renaming a file can be done from both the Operating System (OS) shells or using "save as" from application software etc. Additionally, different file types have differences in both meta data types and its amount.

C. Tracking Provenance

There may be two modes of recording file provenance. First, monitoring user actions through running processes and file system activities and recording provenance from there. Second, provenance information may be discovered by examining existing data, albeit this may not be as accurate as capturing provenance information firsthand. To maximize amount of provenance information, we include provenance discovery as well to further enhance and complement first class provenance data collected by monitoring. For this purposes, two different agents are needed, one will try to discover provenance information from existing files, another agent continuously monitors user actions on the files and keep record by associating actions with both the file and the user. Discovery of provenance is done by examining meta data of files, emails and other information available from the file systems such as alternative data streams e.g. from NTFS streams, HFS forks. There are many specialized as well general purpose tools that can aid in document metadata extraction. When special purpose tools fails, general purpose tools often work but provide less information. Some of these include GNU libextractor, UNIX file program, hachoir-metadata, meta-extractor by National Library of New Zealand, Document Metadata Extraction, etc.

Some systems may store earlier versions of files that are periodically backed up by OS or other software e.g., TimeMachine in MacOS, System Restore in Windows, etc. To avoid provenance clutter, user may not be interested in information generated by such processors or automated backup systems. Hence, in addition to system processes, activities by backup software are excluded from monitoring as a user choice. Some files such as auto-generated temporary files or backup files may or may not need to be stored depending on the user. It is left to the user of the provenance system to determine how such files are to be treated.

D. Enhanced File Search

Enhancing file search is the major goal of this work. The provenance information being collected ubiquitously at different computers will aid in enhancing search for a latest (or desired) version of a file. There are several parts involved in providing enhanced file search experience. Provenance Visualizer facilitates the user with a multi-dimensional provenance exploration tool.

During execution of the file search, the system will lookup provenance information and try to show matching files along with their relation to other files so that. User is shown how the files are linked so that user can decide which one is the latest one or desired one. To avoid usability issues, traditional textual listing of search results is provided in addition to our display of search results as an inter-linked graph of files related according to the provenance information.

OS Shell Integration: For enhanced experience, Visualization and search tool needs to be all well integrated to OS GUI shells and provide necessary terminal tools.

Exposing API: A set of API is exposed so that OS or an authorized application can take benefit of UProve2. For example, in word processing application, during inserting a diagram or hyperlink to another file, the application can show options to choose file from related work directories specific to the file the user is currently working on.

To begin the file search, we need some input. Survey of existing literature shows that users cannot accurately recall information properly to search [1]. To aid in file search, we build on the existing work and provide both graphical input and textual input to properly capture search query. This flexible input system is very helpful in cases for example, user may not remember the name of the file, but the user remembers that s/he modified a text file that was originally received from someone as email attachment.

E. *Ubiquity*

Designing a provenance system to be a ubiquitous one requires many design considerations. These include adapting to the level of monitoring possible under OS of multiple heterogeneous computers. Furthermore, the amount of meta data available highly vary among different file systems and storage systems e.g. FAT32, cloud drives, email attachments do not support alternative data streams.

We provide ubiquity through storing provenance information specific to files of a storage device in that device. For example, local drives store provenance about local files, flash drives store provenance about files present in flash drive only. UProve2 provides synchronization options to import/export provenance information. It may be import/export information for (1) all files, (2) only files already available in either/both storage space. For non-computing devices such as portable drives, online email accounts (through IMAP or POP3), and cloud drives, provenance information will be maintained by respective computing device when it accesses one of those resources. A portable copy of UProve2 may be carried to run in systems where it is not already available.

For privacy and security, UProve2 also allows severing/cutting of desired provenance links in the provenance graph when synchronizing provenance data across devices. Where network access is available, synchronization is done over network as well.

F. *Privacy*

While merging provenance data from multiple source, if the owner is different, instead of showing file names, we can show the user who can be contacted for the latest files or additional details as shared by the user. Depending on need for individuals or organizations, varying levels of privacy may be set. Within an organization or among peers, more details may be shared such as file names, metadata, preview etc. When linking/sharing provenance data between multiple entities, the most restrictive privacy policy among privacy policies of all entities may be used.

G. *Scalar*

It is an optional component. It determines which PCs to utilize the most within the private cloud and as necessary, ensures the system scales up to public cloud by acquiring additional resources. Metrics mostly utilized for selecting resources to use first are (a) Throughput, (b) Free Storage Space, (c) Existing Utilization of Resources, and (d) Acquisition Cost. These metrics helps in minimizing costs and maximizing achievable benefits. This policy enables graceful degradation feature for the system as least-efficient or costliest resources (time or resource acquisition cost) are used at last. Users can calibrate these metrics based on their needs.

Cloud computing revolutionizes how software systems are developed and used as a utility-based model for delivering IT resources, e.g., infrastructure, software, and platform. This model treats any IT asset as a consumable-utility on a pay per use basis analogous to the electricity grid. It enables convenient, ubiquitous, and on-demand network access to a shared pool of easily configurable computing resources (e.g., networks, servers, storage, applications, and services). These resources can be rapidly provisioned and released, and requiring minimal management effort or interaction with the provider. Three major service offerings in cloud computing are Infrastructure- as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). IaaS provides processing, storage, networks, and other fundamental computing resources. Users can deploy and use required software stack for hosting applications. PaaS enables the consumer to configure hosting environment, and to develop/deploy applications onto the cloud infrastructure. SaaS allows the consumer to use the service provider's applications running on a cloud infrastructure.

Four types of cloud service ownership are public cloud, community cloud, hybrid cloud, and private cloud. Private cloud is used by a single organization managed by them or a third party. Community cloud is almost like a private cloud but may be shared across multiple organizations.

Public cloud is a cloud infrastructure provisioned for open use by the any individual or organizational entity. Hybrid cloud is a composition of two or more distinct (private, community, or public) cloud infrastructures that remain unique entities, but are coupled together to enable data and application portability (e.g., for load balancing between clouds). Cloud reduces resource acquisition cost and maintenance by acquiring and using additional IT resources only for the necessary duration, regardless of ownership.

Devices with less storage/compute capabilities e.g. smart, devices and in extreme cases where desktop/laptop runs out of the same capabilities, resources e.g. storage of other collaborating system may be requested or used as backup. For example, tablet PC may provision storage space from desktops, laptops or cloud drives to store provenance information in and thus establishing a hybrid storage cloud for optional scalability.

However, cloud is not required for our ubiquitous provenance system. Rather, it is an optional extension for scalability purposes and in case of limitation of local resources.

IV. Conclusions and Future Work

UProve2 provides a solution for provenance across multiple storage/computing devices. This ubiquity with privacy is achieved through opportunistic synchronization of provenance and distributed storage of provenance data.

We designed data structure and implemented a prototype that maintains provenance information. Provenance is stored in Apache DerbyDB (JavaDB) in embedded mode on a portable JVM per storage device.

In future, we will complete implementation and utilize provenance to further enhance user experience e.g., to auto-complete typing utilizing mostly used related files, naming folders, etc. We will explore other process monitoring tools, techniques, and kernel mode databases.

We will develop provenance plugins for cloud drives e.g., Google Drive, DropBox, etc.

We wish to explore cloud for scalability and integration with cloud services for transparent provenance [2].

Acknowledgment

We are thankful to plentiful members from Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET) for their valuable advice.

References

[1] Karlson, A.K., Smith, G., Lee, B.: Which Version is This?: Improving the Desktop Experience Within a Copy-aware Computing Ecosystem. In: Proc. of SIGCHI Conf. on Human Factors in Computing Systems, pp. 2669-2678. ACM, New York (2011)

[2] Dev, H., Ali, M.E., Sen, T., Basak, M.: AntiqueData: A Proxy to Maintain Computational Transparency in Cloud. DASFAA (2014) (to appear: Vol. 8505 of LNCS)

[3] Jensen, C., Lonsdale, H. Wynn, E., Cao, J., Slater, M., Dietterich, T.G.: The Life & Times of Files and Information: A Study of Desktop Provenance. In: Proc. of SIGCHI Conf. on Human Factors in Comp. Systems, pp.767-776. ACM, NY (2010)

[4] Muniswamy R., Kiran, K., Holland, D.A.: Causality-based Versioning. In: 7th Conf. on File and Storage Technologies, pp. 15{28. USENIX, California (2009)

[5] Shah, S., Soules, C.A.N., Ganger, G.R. and Noble, B.D.: Using Provenance to Aid in Personal File Search. In: Proc. of USENIX Annual Technical Conf., pp.13:1-13:14.USENIX, California (2007)

[6] Gyllstrom, K.A., Soules, C., Veitch, A.: Confluence: Enhancing Contextual Desktop Search. In: Proc. of the 30th Annual Intl. ACM SIGIR Conference on Research and Development in IR, pp.717-718. ACM, New York (2007)

[7] Agrawal, N., Bolosky, W.J., Douceur, J.R., Lorch, J.R.: A Five-year Study of File-system Metadata. ACM Trans. Storage 3. (2007)

[8] Muniswamy-Reddy, K.K., Holland, D.A.,Braun, U., Seltzer, M.: Provenance-aware Storage Systems. In: Proc. of Annual Conf., pp.4-4. USENIX, California (2006)

[9] Soules, C.A.N., Ganger, G.R.: Connections: Using Context to Enhance File Search. SIGOPS Oper. Syst. Rev. Journal vol 39., pp.119-132. ACM, New York (2005)

[10] Rasel, A.A., Ali, M.E.: Poster- UProve: Ubiquitous Provenance to Enhance File Search. Joint poster of International Provenance and Annotation Workshop (IPAW)/USENIX Workshop on the Theory and Practice of Provenance (TAPP). ProvenanceWeek, Germany (2014)

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Short Papers

Network Performance Analysis

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

The use of OMNET++ in the improvement of IVC Simulation Result

Trupti Nimje
Nagpur University,
Nagpur, INDIA
nimje09@gmail.com

Sachin Kohale
Asst. Prof., Dept. of EXTC,
SJCET, Palghar, Mumbai, INDIA
sdk_pz@yahoo.com

***Abstract*— OMNET++ is Similar to NS2 and NS3. OMNeT++ is also a public-source, component-based network simulator with GUI support. Its primary application area is communication networks. Like NS2 and NS3, OMNeT++ is also a discrete event simulator. But it is more advantages than NS2 as mentioned in Table III below. Till now OMNET++ is used only in IT systems, queuing networks, hardware architectures, or even business processes as well. In IVC, only a single OMNET++ simulator provides the simulation related to communication & it evaluates the protocols in its environment. But to evaluate protocols meaningfully or for proper evaluation of protocols it is not sufficient to use OMNET++ lonely. So to fulfill this requirement of evaluation it is necessary to use OMNET++ with road traffic simulator. Here, OMNET++ is used with road traffic simulator SUMO to not only fulfill the protocol evaluation result but also try to improve simulation result in case of IVC environment. One traffic scenario is developed here with the help of SUMO and according to this traffic information OMNET++ shows the result that will provide requirements of meaningful evaluation of protocol and more realistic result.**

Keywords*— *OMNET++, IVC, SUMO, NS-2, VANET

I. INTRODUCTION

Vehicular Ad-Hoc Network (VANET) [1] is one of the type of Mobile Ad-Hoc Network (MANET). It deploys the concept of continuously varying vehicular motion. The vehicles act as nodes in this network. The Vehicular Ad hoc Network technology is of interest in automobile industries to extend the comfort of commuters in terms of safety, transport efficiency, and information. In order achieve this automobile industries are keen on Vehicle to Vehicle (V2V) communication than Vehicle to Infrastructure(V2I) due to the high cost incurred in maintaining the road side infrastructure.

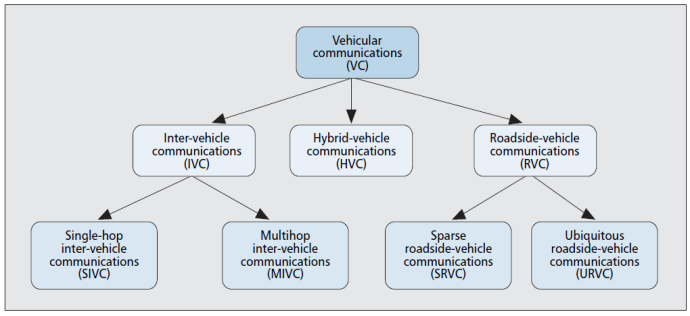


Fig. 1. A Taxonomy of Vehicular Communication Systems

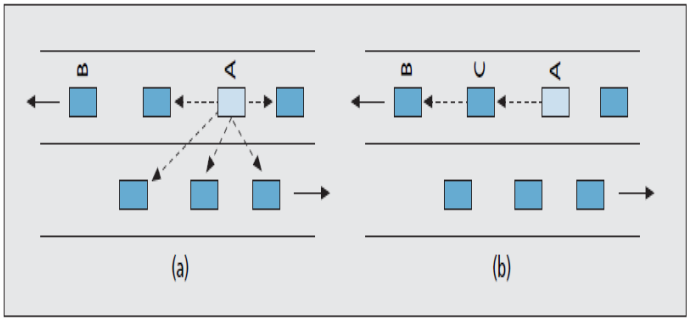


Fig. 2. a) Single-hop; b) multi-hop IVC systems

A. *Taxonomy of Vehicular Communication Systems*

Fig. 1 above shows a complete structure of taxonomy of vehicular communication systems. Whose, each system is explained in detail below.

1) Inter-Vehicular Communication Systems: IVC [2] systems are completely infrastructure-free; only onboard units (OBUs) sometimes also called in-vehicle equipment (IVE) are

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

needed. IVC systems are the main focus of this article. Depending on whether the information is retransmitted at intermediate hops or not, we can further distinguish between single-hop and multi-hop IVCs (SIVCs and MIVCs). SIVC systems are useful for applications requiring short-range communications (e.g., lane merging, automatic cruise control). MIVC systems are more complex than SIVCs but can also support applications that require long-range communications (e.g., traffic monitoring). The main difference between SIVC and MIVC systems is shown in Fig. 2. In an SIVC system vehicle A can send a message only to the cars that are in its transmission range (i.e., vehicle B never receives the message). On the other hand, in an MIVC system, vehicles not in the transmission range of vehicle A (e.g. vehicle B) can also receives the message through another vehicle (vehicle C in Fig. 2b). Here, vehicle C can relay the message. Therefore, an MIVC system requires a network layer capable of multi-hop routing.

2) Roadside to Vehicle Communication Systems: RVC[2] systems assume that all communications take place between roadside infrastructure (including roadside units [RSUs]) and OBUs. Depending on the application, two different types of infrastructure can be distinguished, sparse RVC (SRVC) and ubiquitous RVC (URVC) systems. SRVC systems are capable of providing communication services at hot spots. A busy intersection scheduling its traffic light, a gas station advertising its existence (and prices), and parking availability at an airport, are examples of applications requiring an SRVC system. SRVC system can be deployed gradually, therefore not required substantial investments before any available benefits. A URVC system is the holy grail of vehicular communication: providing all roads with high-speed communication would enable the application unavailable with any of the other system. Unfortunately, a URVC system may require considerable investments for providing full (even significant) coverage of existing roadways (especially in large countries like the United States).

3) Hybrid Vehicular Communication Systems: Hybrid Vehicular Communiction systems extend the range of RVC systems. In this, vehicles communicate with roadside infrastructure even when they are not in direct wireless range by using other vehicles as mobile routers. An HVC system enables as an RVC system with a larger transmission range. The main advantage is that it requires less roadside infrastructure. However, one disadvantage is that network connectivity may not be guaranteed in scenarios with low vehicle density.

B. VANET Simulation

It is practically not possible to do the experiments like testing on vehicular safety application because it is very difficult and costly [3]. So it is preferable to use the simulation tools for testing purpose before it is deployed in a real world to use. The use of simulator is cheap, simple and easy.

To evaluate VANET protocols and services, the first step is to perform an outdoor experiment. Many wireless technologies such as GPRS, IEEE 802.11p and IEEE 802.16 have been proposed for reliable traffic information. Before the technology hits the ground and can meet the expectations, a series of experiments should be performed to test it. These experiments could be expensive and highly complex to inherit all kinds of situation. For this purpose software simulations can play a vital role in imitating real world scenarios.

C. Problem

VANET relies on and is related to two other simulations for its smooth functioning, namely traffic simulation and network simulation. Network simulators are used to evaluate network protocols and application in a variety of conditions. The traffic simulators are used for transportation and traffic engineering. These simulations work independently but to satisfy the need of VANET, a solution is required to use these simulators together. Numerous traffic and network simulations have been tried to resolve the issues with VANET but every solution has had its shortcomings. There are a large number of traffic and network simulator and they need to be used together into what can be called VANET simulator [5][6]. There are few tools for VANET simulation but most of them have the problem of proper 'interaction'. Thus a proper selection of a simulator is also a question for simulation.

So to overcome the above problem of VANET simulation, network simulator 'OMNET++' is used here with road traffic simulator 'SUMO (Simulation in Urban Mobility)'.

II. RELATED WORK

A. OMNET++ Vs. NS2

TABLE I. COMPARATIVE VIEW OF OMNET++ AND NS2

Property of Simulator	**OMNET++**	**NS-2**

Effective Simulation Runtime	In a grid topology, the effective simulation runtime increases linearly with number of nodes (e.g., simulation time for 3025 nodes was 155 seconds) .	In a grid topology, the effective simulation runtime increases exponentially with number of nodes (e.g., simulation time for 3025 nodes was 508 seconds) .
Memory usage	The memory usage increases linearly e.g., for up to 2000 nodes the average memory usage was about 150 MB .	The memory usage increases linearly e.g., for up to 2000 nodes the average memory usage was about 500 MB .
Scalability	Highly scalable compared to NS-2	Low for large number of nodes
Learning curve	Fast	Steep
Reusability	Easier/faster to code and easily reusable in similar platforms with little changes due to its hierarchical structure.	No hierarchical structure .
IDE	Eclipse based IDE	No IDE support
Grid and cluster capability	Available	Available
Coupling between topology creation and protocol implementation	Loose coupling though high level Network Definition Language (NED).	Tight coupling due to C++ modules and OTCL interaction.
Dynamic topology creation	Supported	Not supported
Model management	Simulation kernel is independent of simulation models that enable model reusability without any patch.	Boundary between simulation core and models is blurred which requires patching to the simulation core in order to incorporate 3rd party models.
Module structure	Hierarchical module structure that facilitates easier implementation of a complex protocol.	Models are flat and tightly coupled which makes complex protocol implementation difficult.
Run time visualization of simulation	It can show packet transmissions while a simulation is running. It has run time environment as well as interactive execution environment, which allows one to examine the progres of simulation and change parameters	NAM (Network Animator) is a Tcl/TK based animation tool for viewing network simulation traces and real world packet traces. It support for visualizing the communication at the end of simulation run.
Availability of varieties of models	OMNET++ has a good variety of models for simulating computer systems and queuing systems, but lags behind the NS-2 simulator on availability of communication protocol models.	NS-2 has a rich set of communication protocol models (since it has been designed as a network protocol simulator, this is not surprising).
Documentation	Well organized and up to date	Fragmented.
Embeddability	OMNET++ simulation kernel can be embedded in other applications.	Not supported.
Physical layer signal model	Analog signal modeling is also possible. [MiXiM]	Not supported.
Signal reception models	SNR and BER based model	SNR based model

B. SUMO

SUMO (Simulation of Urban Mobility) [14] is an open source, highly portable, microscopic road traffic simulation package designed to handle large road networks. Its main features include collision free vehicle movement, different vehicle types, single-vehicle routing, multi-lane streets with lane changing, junction-based right-of-way rules, hierarchy of junction types, an openGL graphical user interface (GUI), and dynamic routing. SUMO can manage large environments, i.e., 10 000 streets, and it can import many network formats such as Visum, Vissim , ArcView, or XML Descriptions.

TABLE II. CPU AND MEMORY PERFORMANCE IN SUMO

Parameters	**SUMO**

CPU Usage	Between 5-17%, depending on the number of vehicles currently running on the traffic network.
Memory Usage	Between 12-16 MB, depending on the traffic network.

III. OMNET++ WITH SUMO

The network simulator OMNeT++ and the road traffic simulator SUMO used together to make a dedicated communication module. During simulation runs, these communication modules exchange commands, as well as mobility traces, via TCP connections. OMNeT++ is an event-based simulator, so it handles mobility by scheduling node movements at regular intervals. This fits well with the approach of SUMO, which also advances simulation time in discrete steps.

As shown in Fig. 3 below, OMNET++ is Network Simulator and SUMO is Road Traffic Simulator. Both OMNET++ and SUMO exchanges information with each other.

The Control modules integrated with OMNET++ and SUMO were able to buffer any commands arriving in between timesteps for its synchronous execution at a particular defined intervals.

At each timesteps , OMNET++ sends requests for vehicles mobility information to SUMO which in turn SUMO returns a position of all vehicles to OMNET++. OMNET++ reacts according to the received information and deletes particular node of vehicle that has reached destination, after that introducing new nodes in another timesteps.

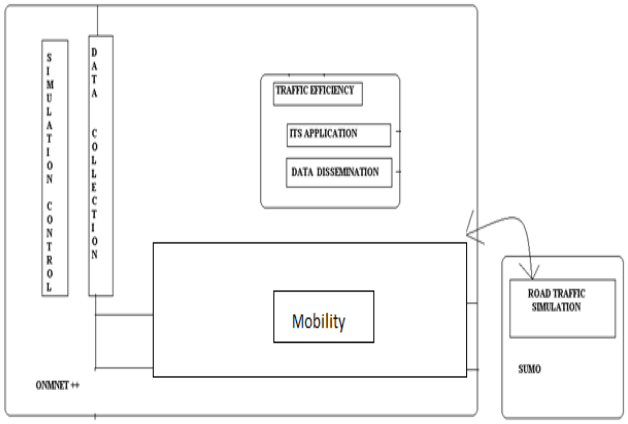


Fig. 3. Architecture of Bi-directional coupling of SUMO and OMNeT++

The network traffic and road traffic simulators, the interaction between OMNET++ and SUMO is simply based on request and response protocol as shown in Fig. 4 below. Road traffic in SUMO can be influenced by OMNET++ in a whole number of ways. Most importantly, time steps are generated to advance the simulation in SUMO.

Fig. 4 below shows Information exchange between OMNET++ and SUMO in one timestep.

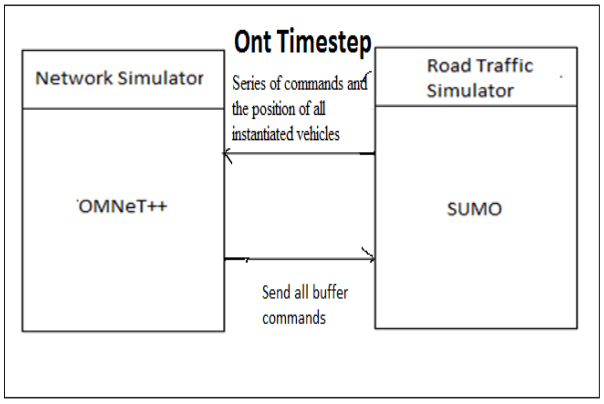


Fig. 4. Information exchange between OMNeT++ And SUMO in one timestep

IV. SIMULATION RESULT

A. *Road Traffic Scenario in SUMO*

The network includes four origins, four destinations as well as two un-signalized intersections. In the investigated area each road for outbound traffic has three lanes and the allowed traffic movements on each lane are restricted. U-turn behaviors are prohibited at all intersections. Moreover, a higher priority has been given to the eastbound and westbound traffic. For traffic demand there are four vehicle types (Cars A, B, C and D) in the network. All drivers are 50% perfect in driving. The corresponding information is listed in TABLE V below. One vehicle per vehicle type is going to leave from each origin to each destination every 30 seconds in the period under investigation from 15:00 to 15:15.

TABLE III. VEHICLE TYPES AND VEHICULAR CHARACTERISTICS

Vehicle Type	Max. Acceleration (m/s)	Max. Deceleration(m/s)	Length(m)	Max. Speed(m/s)
Car A	3.0	6.0	5.0	50.0
Car B	2.0	6.0	7.5	50.0
Car C	1.0	5.0	5.0	40.0
Car D	1.0	5.0	7.5	30.0

TABLE IV. COORDINATION DATA

Node name	x-coordinate	y-coordinate
91	-1000.0	+1000.0
92	-1000.0	0.0
93	+3000.0	0.0
94	+3000.0	+1000.0
911	-500.0	+1000.0
912	-500.0	0.0
913	+2500.0	0.0
914	+2500.0	+1000.0
1	0.0	+1000.0
2	0.0	0.0
3	+1000.0	0.0
4	+2000.0	0.0
5	+2000.0	+1000.0
6	+1000.0	+1000.0

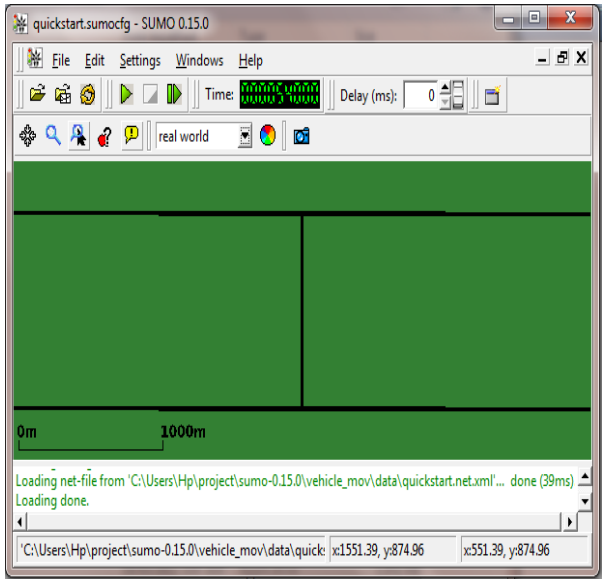


Fig. 5. Road Construction and Intersection of Traffic Scenario in SUMO

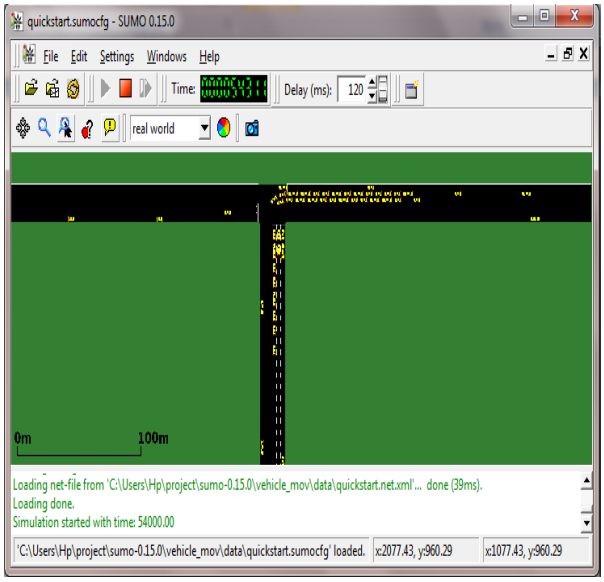


Fig. 6. Vehicle movements (Traffic Flow)

TABLE V. PARAMETERS AT THE END OF SIMULATION

Network Parameters		
Name	**Value**	**Dynamic**
Loaded Vehicles[#]	1440	Yes
Waiting Vehicles[#]	561	Yes
Departed Vehicles[#]	879	Yes
Running Vehicles[#]	610	Yes
Arrived Vehicles[#]	269	Yes
End Time[s]	54900	No
Begin Time[s]	54000	No
Step Duration[ms]	21	Yes
Simulation Duration[ms]	21	Yes
Idle Duration[ms]	0	Yes
Duration Factor[]	47.62	Yes
ups[#]	29047.62	Yes
Mean ups[#]	734266.29	Yes

B. Network Scenario in OMNET++

1) Information Exchange between OMNET++ and SUMO through command window:

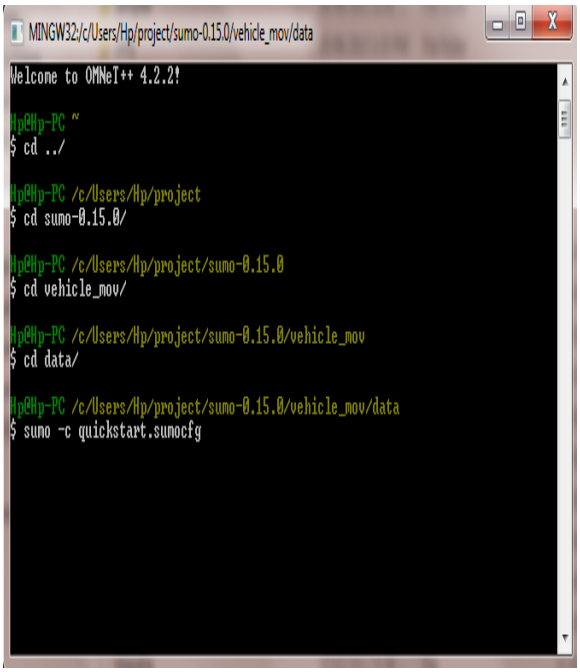


Fig. 7. Command window to open SUMO simulation in OMNET++ environment

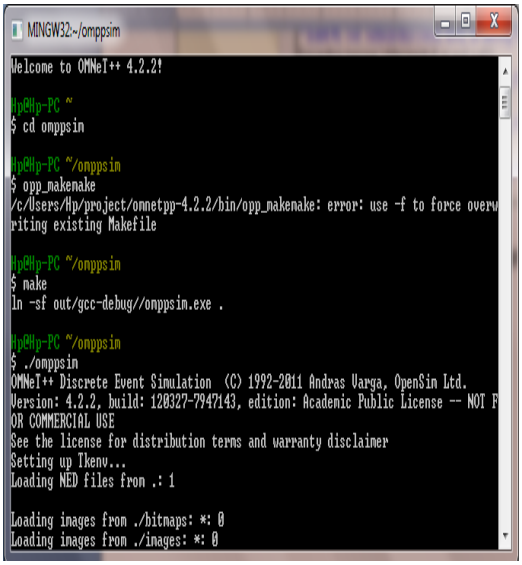


Fig. 8. Command window to open OMNET++ simulation

V. CONCLUSION

As Network Simulator OMNET++ indicates vehicles as nodes in the network. Also, Road Traffic Simulator SUMO gives the position of the vehicles. Thus by integrating , a better results in terms of locating vehicles and controlling its traffic both can be achieved. The simulation results shown above in Fig. 6 indicate vehicles movement. Also, Fig. 7 above indicates SUMO simulation in OMNET++ environment.

REFERENCES

[1] H. Hartenstein, K.P. Laberteaux, "A Tutorial Survey on Vehicular Ad Hoc Networks," IEEE Comm. Magazine, vol. 46, no. 6, pp. 164-171, June 2008.

[2] M.L.Sichitiu , M. Kihl, "Inter-Vehicle Communication Systems: A Survey", IEEE Communication Surveys and Tutorials, vol. 10, no. 2, pp. 88-105, 2008.

[3] Marco Fiore, Jerome Harri, Fethi Filali, Christian Bonnet, "Vehicular Mobility Simulation for VANETs Simulation", Symposium, 2007 ANSS apos 07. 40th Annual vol., issue 26-28, pg.301 – 309, March 2007.

[4] M. Raya, P. Papadimitratos, J.P. Hubaux, "Securing Vehicular Communications", IEEE Wireless Communication, Special Issue on Inter-Vehicular Communication, vol. 13, no. 5, pp. 8–15, Oct. 2006.

[5] J.J. Blum, A. Eskandarian ,L.J. Hoffman, "Challenges of Inter-vehicle Ad Hoc Networks," IEEE Trans. Intelligent Transportation Systems, vol. 5, no. 4, pp. 347-351, Dec. 2004.

[6] Tomoyuki Yashiro, Tempei Kondo, Hirotaka Yagome, Masafurru Higuchi, Yutaka Matsushita, "A Network based on Inter-Vehicle Communication", proc. Intelligent Vehicles Symposium, pp. 345–50, 1993.

[7] Francisco J. Martinez, Chai Keong Toh, Juan-Carlos Cano, Carlos T. Calafate ,Pietro Manzoni, "A survey and comparative study of simulators for vehicular ad hoc networks (VANETs)", Wirel. Commun. Mob. Comput. 2009 Published online in Wiley Inter Science, DOI: 10.1002/wcm.859.

[8] Fall K, Varadhan K. ns notes and documents. The VINT Project, UC Berkeley, LBL, USC/ISI, and Xerox PARC, February 2000.

[9] Martin J. GloMoSim. Global mobile information systems simulation library.UCLA Parallel Computing Laboratory, 2001.

[10] JiST/SWANS: Java in Simulation Time/Scalable Wireless Ad hoc Network Simulator, 2004.

[11] Walsh K, Sirer EG. A staged network simulator (SNS). Computer Science Department, Cornell University, 2003.

[12] Salem HH, Chrisoulakis J, Papageorgiou M, Elloumi N, Papadakos P. The use ofMETACORtool for integrated urban and interurban traffic control. evaluation in corridor peripherique, Paris. In Proceedings of Vehicle Navigation and Information Systems Conference, 1994; 645–650.

[13] Taylor NB. The CONTRAM dynamic traffic assignment model. Networks and Spatial Economics 2003; **3**: 297–322.

[14] D. Krajzewicz, G. Hertkorn, C. Rossel, P. Wagner, "SUMO (Simulation of Urban MObility):An Open-Source Traffic Simulation", Proc. Fourth Middle East Symp. Simulation and Modelling (MESM '02), pp. 183-187, Sept. 2002.

Impact of Mobile Nodes for Few Mobility Models on Delay-Tolerant Network Routing Protocols

Md. Sharif Hossen and Muhammad Sajjadur Rahim
Department of Information and Communication Engineering
University of Rajshahi
Rajshahi-6205, Bangladesh
sharif5613@gmail.com, sajid_ice@ru.ac.bd

***Abstract*—Delay-Tolerant Networks (DTNs) are sparse dynamic wireless networks, where most of the time a complete end-to-end path from the source to the destination does not exist. There are many real networks that follow this model, for example, military networks, vehicular ad-hoc networks (VANETs), wildlife tracking sensor networks, etc. In this context, conventional mobile ad-hoc routing schemes would fail, because they try to establish complete end-to-end paths, before any data sent. Therefore, performance analysis of different DTN routing mechanisms plays an important role in understanding the design of DTNs that encourages one to choose proper routing protocol for a particular scenario. This paper investigates the performance of replication-based DTN routing protocols, namely Epidemic, Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET), MaxProp, Resource Allocation Protocol for Intentional DTN (RAPID), Binary Spray-and-Wait (B-SNW), and Spray-and-Focus (SNF) in DTN scenario against varying number of mobile nodes for three mobility models, namely Random Walk (RW), Random Direction (RD) and Shortest Path Map Based (SPMB) movement. Performance has been evaluated and analyzed using Opportunistic Network Environment (ONE) simulator by considering three metrics: delivery probability, average latency and overhead ratio. Simulation results show that the investigated DTN routing protocols, in general, exhibit better performance in the SPMB movement model than other movement models, i.e., RW and RD, because they yield maximum message delivery probability, minimum overhead ratio (for B-SNW and SNF), and minimum average latency.**

Keywords—routing protocols; message replication; delivery probability; average latency; overhead ratio; simulation; delay-tolerant networks; opportunistic network environment simulator; random walk; random direction; shortest path map based movement model

I. INTRODUCTION

Personal communication devices, such as cellular phones, have enabled voice and data communications to mobile users, achieving global connectivity via infrastructure networks (cellular networks, WLANs). Local connectivity among the devices may additionally be obtained by forming ad-hoc networks since the mobile devices are virtually always turned on, and have the necessary radio interfaces, processing power, storage capacity, and battery lifetime to act as routers. However, such usually sparse ad-hoc networks generally cannot support the type of end-to-end connectivity required by the classic TCP/IP-based communications due to frequent topology changes, disruptions, and network partitions caused by the node movement. Instead, asynchronous message passing has been suggested to enable communication over the space-time paths that exist in these types of networks (e.g., DTNs). These networks were initially designed for deep-space communication, and one of the main characteristics is the adoption of an asynchronous communication mechanism called store-and-forward paradigm [1, 2].

There are many real-life networks, which follow this DTN paradigm, for example, satellite communication [3], wildlife tracking sensor networks [4], military networks, vehicular ad-hoc networks [5], etc. In addition, in north part of the Sweden, the communication between villages and the summer camps of the Saami population is provided when the nodes get connected [6]. Moreover, such environments can exist even when natural disaster or other effects destroy a stable infrastructure. A more interesting example of DTNs is the applications where sensors are attached to seals [7] and whales [8] to collect large number of sensor readings from the oceans.

This paper analyzes the impact of number of mobile nodes on different DTN routing protocols like Epidemic [9], PRoPHET [10], MaxProp [11], RAPID [12], Spray-and-Wait [13], and Spray-and-Focus [14]. These protocols have been analyzed under three mobility models, namely Random Walk (RW), Random Direction (RD), and Shortest Path Map Based (SPMB) movement model considering three performance metrics, namely delivery probability, average latency, and overhead ratio. The remainder of this paper is organized as follows: Section II briefly discusses about DTN routing protocols. Section III gives the overview of the mobility models. Section IV provides the introduction of simulator, reporting and visualization, simulation environment setup and performance metrics. Section V analyzes the simulation results. Section VI provides the concluding remarks and future works.

II. DTN ROUTING PROTOCOLS

This section gives a brief overview on the classification of DTN routing protocols, and summarizes the design of the replication-based DTN routing schemes, i.e., Epidemic, PRoPHET, MaxProp, RAPID, Spray-and-Wait, and Spray-and-Focus.

978-1-5090-0203-0/16/$31.00 ©2016 IEEE.

DTN routing protocols are classified into two basic schemes: single-copy and multi-copy. In single-copy schemes, a single copy of each message is forwarded through the network, which is called forwarding-based routing. Multi-copy schemes forward several copies of the same message to the network, i.e., replicate messages, and hence are called replication-based. There are several advantages and disadvantages to both schemes, and which scheme is desirable depends on the application scenario. Forwarding-based routing scheme is generally resource efficient as only a single copy of a message exists in the network at any given time, but it does not guarantee the best delivery ratio since the probability of finding the destination node is low, and usually exhibits high latency [15]. On the other hand, replication-based routing protocols obtain higher message delivery ratios [11], since several copies of the same message exist in the network, and at least one of those must reach the destination. Therefore, there is a typical tradeoff between the two schemes, whereby the former spends less resources but may provide low probability of successful delivery, whereas the latter tends to spend more resources but also provides better delivery ratios [14].

A. Epidemic

Epidemic routing is historically the first DTN routing protocol. It is flooding based in nature where every node continuously replicates a copy of message to all nodes it encounters that do not have the message copy in common. So, this algorithm is the best-effort approach to reach the destination compared to flooding which forwards a copy of every data packet to every node [13]. It has very high overhead ratio and has a large number of message copies in the network which results in network congestion.

B. Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET)

To improve the delivery probability and reduce the wastage of network resources in Epidemic routing, a new type of routing protocol called PRoPHET has been proposed in [10]. The basic operation of PRoPHET routing is similar to Epidemic, and it attempts to improve the delivery probability of messages. Message forwarding is based on the calculation of probability (also called delivery predictability) by each node to each destination node. When two nodes are encountered, messages are forwarded to a node that has higher delivery predictability. Delivery predictability $P_{(a,b)}$ is stored in internal delivery vector and gets updated whenever nodes meet each other. The delivery predictabilities used by each node are recalculated at each opportunistic encounter according to the three rules as follows:

(i) When node A encounters another node B, the predictability for B is increased. Equation (1) shows this calculation.

$$P_{(a,b)} = P_{(a,b)old} + (1 - P_{(a,b)old}) \times P_{init} \quad (1)$$

where P_{init} is an initialization constant.

(ii) The delivery predictability must age because if two nodes do not encounter each other in a while, then they are less likely to forward messages to each other. Equation (2) shows this ageing equation.

$$P_{(a,b)} = P_{(a,b)old} \times \gamma^{k} \quad (2)$$

where γ^{k} is an aging constant.

(iii) The delivery predictability also follows the transitive property, that is, if a node A frequently encounters node B and node B frequently encounters node D, then node D probably is a good node to forward message intended for node A. Equation (3) shows the effect of transitivity on delivery predictability.

$$P_{(a,d)} = P_{(a,d)old} + (1 - P_{(a,d)old}) \times P_{(\mathrm{a,b})} \times P_{(b,d)} \times \beta \quad (3)$$

where β is a scaling constant that decides how large impact the transitivity should have on the delivery predictability.

C. Resource Allocation Protocol for Intentional DTN (RAPID)

RAPID models DTN routing as a utility-driven resource allocation problem to intentionally optimize a single routing metric: average delay, missed deadlines, or maximum delay. The utility function assigns a utility value, U_i, to every packet depending on the routing metric to be optimized. U_i is defined as the expected contribution of the packet *i* to the given routing metric. RAPID replicates those packets first that locally result in the highest increase in the utility. For example, let us consider the metric to optimize is average delay. The utility function for average delay will be: $U_i = -(D_i)$, basically the negative of the average delay. Hence, the protocol replicates the packet that results in the greatest decrease in delay. Therefore, RAPID attempts to replicate all packets if network resources such as storage and bandwidth allow. The overall protocol is comprised of four steps as given below:

(i) Initialization: Metadata is exchanged between nodes in contact to help estimate the packet utilities.

(ii) Direct Delivery: Packets destined for immediate neighbors are transmitted.

(iii) Replication: Packets are replicated based on the marginal utility, i.e., the change in utility over the size of the packet.

(iv)Termination: Protocol ends when contacts break or all packets have been replicated.

D. MaxProp

MaxProp routing protocol uses several mechanisms to define a ranked list that determines which packets should be transmitted first and which packets should be dropped first. At the core of the MaxProp protocol is a ranked list of the peer's stored packets based on a cost assigned to each destination. The cost is an estimate of delivery probability. These probabilities are added to form a path score; the minimum score of all possible paths via the current peer to a destination is selected as the cost estimate. In addition, MaxProp uses acknowledgments sent to all peers to notify them of packet deliveries. These acknowledgements are 128-bit hashes of the message that are flooded into the network, and instruct nodes to delete extra copies of the message from their buffers. This helps free space so that outstanding messages are not dropped often. MaxProp assigns a higher priority to packets based on low hop-counts, and it also attempts to prevent reception of the same packet more than once. Besides each message maintains

a "hop list" indicating nodes it has previously visited to ensure that it does not revisit a node.

E. Spray-and-Wait

Spray-and-Wait (SNW) routing protocol is a class of replication-based schemes that attempts to find a good delivery ratio by limiting the number of replicas of a given message while keeping resource utilization low as in forwarding-based routing. SNW achieves resource efficiency by setting a fixed upper bound on the number of copies per message allowed in the network. The SNW protocol consists of the following two phases:

(i) Spray phase: For every message originating at a source node, L message copies are initially spread – forwarded by the source and possibly other nodes receiving a copy – to L distinct "relays".

(ii) Wait phase: If the destination is not found in the spraying phase, each of the L nodes carrying a message copy performs direct transmission (i.e., will forward the message only to its destination).

There are two main versions of the SNW routing protocol, known as Vanilla and Binary, respectively. The two versions differ in the mechanism employed to "spray" the L copies of a message. To achieve this, a simplest way called Vanilla is to transmit a single copy of the message from the source to the first L distinct nodes it encounters after the message is generated. The second version, referred to as Binary, works as follows: the source node starts with L copies of the message. It transfers $L/2$ of the copies to the first node it encounters. Each node then transfers half of its copies to future nodes they meet that have no copy of the message. When a node eventually gives away all of its copies, except for one, it switches to the wait phase where it looks for a direct transmission opportunity with the final destination of the message. The advantage of the Binary version is that messages are disseminated much faster than in the Vanilla version. Here, only the Binary Spray-and-Wait (B-SNW) is considered for simulation.

F. Spray-and-Focus

Spray-and-Focus (SNF) routing protocol overcomes the shortcomings of simple spraying algorithms. Existing spraying scheme, i.e., "Spray-and-Wait" scheme [13, 14], consists of two phases: in the first phase it distributes a fixed number of copies to the first few relays encountered, and in the second phase each of these relays waits until it encounters the destination itself (i.e., direct transmission). It is easy to see that, here, this scheme would spread all its copies quickly to the node's immediate neighborhood, but then few, if any, of the nodes carrying a copy might ever see the destination [16].

This problem could be solved if a sophisticated single-copy scheme is used to further route a copy after its handover to a relay, i.e., a scheme that takes advantage of transmissions (unlike direct transmission). Thus in second phase ("focus" phase) rather than waiting for the destination to be encountered, each relay can forward its copy to a potentially more appropriate relay, using a carefully designed utility-based scheme. Thus when a relay for a given message has only one forwarding token left for that message, it switches to the "focus phase". Unlike SNW, where in the wait phase messages are routed using direct transmission (i.e., forwarded only to their destinations) [13, 14, 17], in the focus phase of the SNF routing protocol a message can be forwarded to a different relay according to the given forwarding criteria as follows:

(i) Age of last encounter timers with transitivity: Let us assume that each node maintains a timer $\tau_i(j)$ for every other node j in the network, which records the time elapsed since the two nodes last encountered each other as follows: initially set $\tau_i(i) = 0$ and $\tau_i(j) = \infty$, $\forall\, i, j$; whenever i encounters j, set $\tau_i(j) = \tau_j(i) = 0$; at every clock tick, increase each timer by 1.

(ii) Single-copy utility-based routing: Let every node i maintain a utility value $U_i(j)$ for every other node j in the network. Then, a node A forwards to another node B a message destined to a node D, if and only if, $U_B(D) > U_A(D) + U_{th}$, where U_{th} (utility threshold) is a parameter of the algorithm

III. Mobility Models

This section introduces the mobility models that have been used in this research work.

A. Random Walk (RW) Mobility Model

RW mobility model is one of the simplest mobility models available. In this model, every node moves towards a new randomly chosen location. A random direction and speed is assigned to each node from a predefined range, and nodes of a network are independent from one another [18]. Whenever any node reaches the destination location, a new direction is again assigned from predefined ranges. Hence, the distributions of mobility parameters are a function of time. As the mobility parameters achieve the stable state of distributions, the simulation produces consistent results.

B. Random Direction (RD) Mobility Model

RD mobility model drives nodes up to the boundary of the simulation area before changing direction and speed. In this paper, the RD model proposed in [19] has been used for simulation. In this model, nodes will start at a random place on the simulation area, pick a random direction, and follow it to the edge of the simulation area. They will then pause and pick another direction to go in until they hit the edge again. There are other variants of this model. In the second variant of the RD mobility model as in [20], when a node reaches a boundary of the simulation area, it instantaneously re-enters into the simulation area from the opposite boundary. It then continues to move with the previous speed and direction. In the third variant, presented in [21], nodes would bounce off and continue to move with the previous speed in a new direction when a node reaches a boundary of the simulation area.

C. Shortest Path Map Based (SPMB) Mobility Model

A more realistic model is the SPMB mobility model, where instead of a completely random walk, the nodes choose a random point on the map and then follow Dijkstra's shortest path algorithm to discover the shortest path to that point from their current location. Points may be chosen completely random or from a list of Points of Interest (POI). These POIs may be chosen to match popular real-world destinations such

as tourist attractions, shops or restaurants [22]. SPMB is found easy to understand and competent to use in simulations, but it does not assure inter-contact time and contact time distributions that match real world traces when small number of nodes is used in the simulation [23].

IV. SIMULATION

This paper focuses on the performance analysis of Epidemic, PRoPHET, MaxProp, RAPID, Binary-SNW, and SNF routing protocols in a DTN scenario. All these routing protocols are simulated using Opportunistic Network Environment (ONE) with program version of 1.5.1. This section explains the ONE simulator, reporting and visualization tools with Graphical User Interface (GUI), the environment modeling parameters, and performance metrics.

A. The ONE Simulator

A series of simulations are carried out to evaluate the performance of above-mentioned protocols using the ONE. The ONE simulator could be run on Linux, Windows, or any other platform supporting Java. ONE is an agent-based discrete event simulation engine that is designed for evaluating the performance of DTN routing protocols. At each simulation step, the engine updates a number of modules that implement the main simulation functions. Unlike other DTN simulators which usually focus only on routing simulation, the ONE combines mobility modeling, inter-node contacts, DTN routing, message handling and visualization in one package that provides a rich set of reporting and analyzing modules. The elements and their interactions are shown in Fig. 1. A detailed description of the simulator is available in [24] and the ONE simulator project page [25], where the source code of the simulator is also available. Source codes are written in Java programming language.

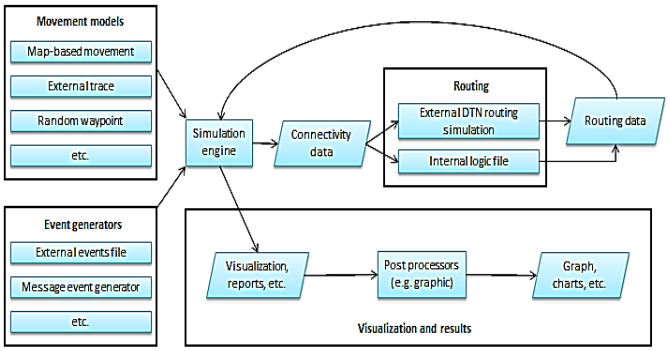


Fig. 1. Interaction of the ONE modules

B. Reporting and Visualization

ONE is able to visualize the results of the simulation in two ways: via an interactive Graphical User Interface (GUI) and by generating images from the information gathered during the simulation. Fig. 2 shows the GUI displaying the simulation in real-time. Node locations, current paths, connections between nodes, number of messages carried by a node, etc. are all visualized in the main window.

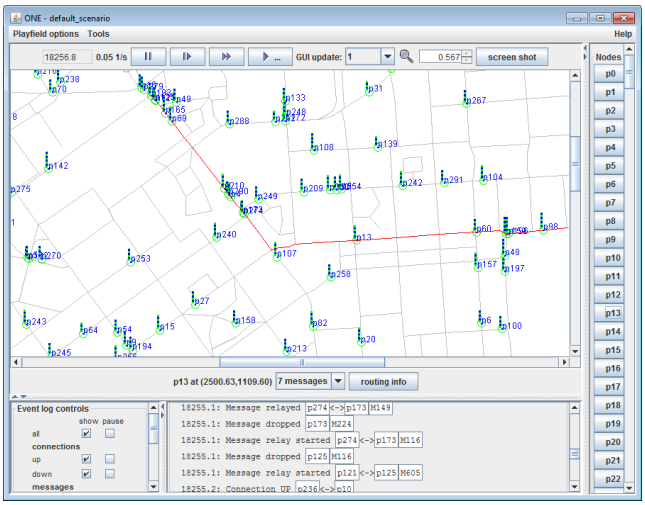


Fig. 2. Screenshot of the ONE simulator's GUI

Pressing the "routing info" button opens a new window where information about the routing module is displayed as shown in Fig. 3. When a node is chosen, the playfield view is centered on that node and current path is shown in red color.

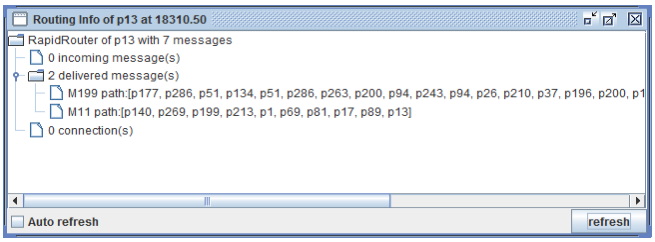


Fig. 3. Routing information at node 13

C. Simulation Environment Setup

The DTN network used for evaluation consists of a single group of pedestrians. Parameters of simulation setup and routing algorithms are specified in Table I and Table II, respectively. Table I shows the simulation configuration for analyzing the performance metrics by varying the number of mobile nodes, i.e., 100, 150, 200, 250, and 300 for each mobility model, i.e., RW, RD and SPMB, respectively. Table II summarizes simulation configuration for routing algorithms.

TABLE I. SIMULATION ENVIRONMENT PARAMETERS

Parameters	Values
Simulation Time	24 hours
Update Interval	1 second
Number of Nodes	100, 150, 200, 250, 300
Interface	Bluetooth Interface
Interface Type	Simple Broadcast Interface
Transmit Speed	250 kbps
Transmit Range	10 m
Routing Protocols	Epidemic, PRoPHET, MaxProp, RAPID, B-SNW, SNF
Buffer Size	5 MB
Message Generation Rate	2, i.e., one message in 25-35 seconds
Message TTL	300 (minutes)
Mobility Models	Random Walk, Random Direction, Shortest Path Map Based Movement
Message Size	500 KB – 1 MB
Simulation Area Size	4500 m × 3400 m

TABLE II. PARAMETERS FOR ROUTING ALGORITHMS

Routing Algorithms	Parameters	Values
Epidemic	N/A	N/A
PRoPHET	Seconds in Time Unit	30s
MaxProp	Max. Size of Probability	50
RAPID	Utility Algorithm	Average_Delay
B-SNW	No. of Copies (L)	6
SNF	No. of Copies (L)	6

D. Performance Metrics

Performance metrics used for evaluating the performance of DTN routing protocols are delivery probability and average latency, and overhead ratio. These metrics are defined as follows:

(i) Delivery Probability

It is defined as the ratio of the total number of messages delivered to the destination over the total number of messages created at the source.

$$\text{Delivery Probability} = \frac{\text{Number of messages delivered}}{\text{Number of messages created}}$$

(ii) Average Latency

It is defined as the measure of average time between messages generated and messages received by destination node.

$$\text{Average Latency} = \sum_{i=1}^{n} \frac{\text{Time when message received} - \text{Time when message produced}}{\text{Number of messages received}}$$

(iii) Overhead Ratio

The overhead ratio defines how many redundant packets are relayed to deliver one packet. It simply reflects the cost of transmission in a network.

$$\text{Overhead Ratio} = \frac{R-D}{D}$$

where R is the number of messages forwarded by relay nodes, and D is the number of messages delivered to their destination.

V. RESULTS AND DISCUSSION

The results presented here are obtained by running the simulations as per the parameters defined in Table I and II.

A. Performance Analysis on Delivery Probability

Delivery probabilities obtained for different mobility models for different routing schemes are depicted in Fig. 4. Hence, as the number of nodes increases, delivery ratio of packets also increases. The delivery probability is maximum in case of SPMB and minimum for RD movement. Therefore, all routing schemes provide good delivery ratios for SPMB movement model.

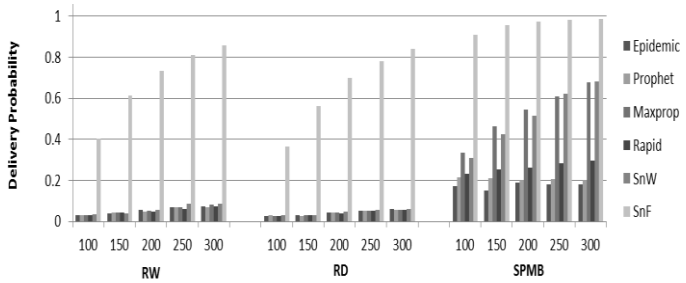


Fig. 4. Delivery probability vs. number of nodes

B. Performance Analysis on Average Latency

Average latencies obtained for different mobility models for different routing schemes are depicted in Fig. 5. As shown in Fig. 5, with increasing number of nodes, average latency increases in case of RW and RD for all routing schemes except SNW and SNF, where it decreases gradually. On the other hand, in case of SPMB movement, average latency increases for SNW and constant (approximately zero) for SNF routing. However, average latency is minimum in case of SPMB movement compared to RW and RD for all routing schemes. Therefore, SPMB movement model shows better performance than RD and RW in terms of average latency.

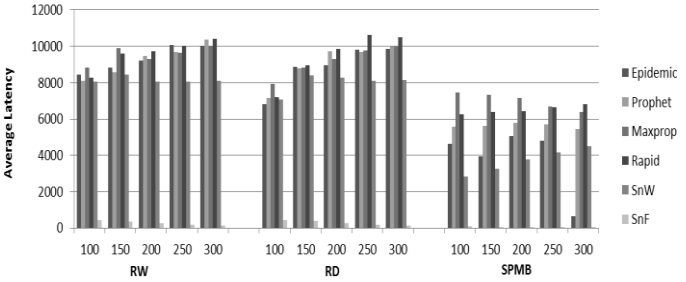


Fig. 5. Average latency vs. number of nodes

C. Performance Analysis on Overhead Ratio

Overhead Ratios obtained for different mobility models for different routing schemes are depicted in Fig. 6. In case of SPMB movement model, overhead ratio increases for Epidemic, Prophet, RAPID and MaxProp routing schemes with increase in the number of nodes, but decreases for B-SNW and almost constant for SNF routing. Hence, SPMB movement model exhibits better performance than RW and RD movement models, specially for B-SNW and SNF routing schemes.

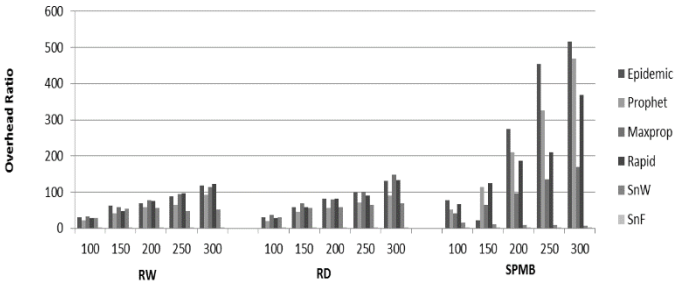


Fig. 6. Overhead ratio vs. number of nodes

VI. Conclusion and Future Works

In this paper, the performances of six routing protocols, namely Epidemic, PRoPHET, Binary Spray-and-Wait, MaxProp, RAPID, and Spray-and-Focus in a DTN scenario have been evaluated using the ONE simulator. Three different mobility models are used for the simulations. Simulation results show that Shortest Path Map Based (SPMB) movement model has provided with better performance than other movement models, i.e., Random Walk (RW) and Random Direction (RD) for the DTN routing protocols investigated here.

Future works may include investigation of routing protocols with some other mobility models, e.g., working day, office activity, map based, home activity, and evening activity movement models, etc. Further investigation can be done to derive different characteristics of the mobility models, such as distribution of contact durations, contact frequency, etc.

References

[1] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proc. of ACM SIGCOMM*, Karlsruhe, Germany, Aug. 2003, pp. 27–34.

[2] J. Scott, P. Hui, J. Crowcroft, and C. Diot, "Haggle: a networking architecture designed around mobile users," in *Proc. of IFIP WONS*, France, Jan. 2006, pp. 78–86.

[3] G. E. Prescott, S. A. Smith, and K. Moe, "Real-time information system technology challenges for NASA's earth science enterprise," in *Proc. of the 20th IEEE Real-Time Systems Symposium*, Phoenix, AZ, USA, Dec. 1999.

[4] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein, "Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet," in *Proc. of ACM ASPLOS*, San Jose, CA, USA, Dec. 2002, pp. 96–107.

[5] J. Ott and D. Kutscher, "A disconnection-tolerant transport for drive-thru internet environments," in *Proc. of IEEE INFOCOM*, Miami, FL, USA, Mar. 2005, vol. 3, pp. 1849–1862.

[6] A. Doria, M. Uden, and D. P. Pandey, "Providing connectivity to the saami nomadic community," in *Proc. of the 2nd International Conference on Open Collaborative Design for Sustainable Innovation*, Bangalore, India, Dec. 2002.

[7] G. W. Boehlert, D. P. Costa, D. E. Crocker, P. Green, T. O'Brien, S. Levitus, and B. J. L. Boeuf, "Autonomous pinniped environmental samplers: using instrumented animals as oceanographic data collectors," *Journal of Atmospheric and Oceanic Technology*, vol. 18, no. 11, pp. 1882–1893, 2001.

[8] T. Small and Z. J. Haas, "The shared wireless infostation model - a new ad hoc networking paradigm (or where there is a whale, there is a way)," in *Proc. of the 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Annapolis, MD, USA, Jun. 2003, pp. 233–244.

[9] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Department of Computer Science, Duke University, Tech. Rep. CS–2000–06, Apr. 2000.

[10] A. Lindgren, A. Doria, and O. Scheln, "Probabilistic routing in intermittently connected networks," *ACM Mobile Computing and Communication Review*, vol. 7, no. 3, pp. 19–20, Jul. 2003.

[11] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "MaxProp: routing for vehicle-based disruption-tolerant networks," in *Proc. of IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[12] A. Balasubramanian, B. N. Levine, and A. Venkataramani, "DTN routing as a resource allocation problem," in *Proc. of ACM SIGCOMM*, Kyoto, Japan, Aug. 2007, pp. 373–384.

[13] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *Proc. of ACM WDTN*, Philadelphia, PA, USA, Aug. 2005, pp. 252–259.

[14] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and focus: efficient mobility-assisted routing for heterogeneous and correlated mobility," in *Proc. of IEEE PerCom*, White Plains, NY, USA, Mar. 2007, pp. 79–85.

[15] S. Jain, K. Fall, and R. Patra, "Routing in a delay-tolerant network," in *Proc. of ACM SIGCOMM*, Portland, OR, USA, Oct. 2004, pp. 145–157.

[16] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms," in *Proc. of IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[17] T. Small and Z. J. Haas, "Resource and performance tradeoffs in delay-tolerant wireless networks," in *Proc. of ACM WDTN*, Philadelphia, PA, USA, Aug. 2005, pp. 260–267.

[18] B. D. Walker, T. C. Clancy, and J. K. Glenn, "Using localized random walks to model delay-tolerant networks," in *Proc. of IEEE Military Communications Conference,* San Diego, CA, USA, Nov. 2008.

[19] E. M. Royer, P. M. Melliar-Smith, and L. E. Moser, "An analysis of the optimum node density for ad hoc mobile networks," in *Proc. of IEEE ICC*, Helsinki, Finland, Jun. 2001, vol. 3, pp. 857–861.

[20] Z. J. Haas, "A new routing protocol for the reconfigurable wireless networks," in *Proc. of IEEE 6th International Conference on Universal Personal Communications Record*, San Diego, CA, USA, Oct. 1997, vol. 2, pp. 562–566.

[21] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 3, pp. 55–66, Jul. 2001.

[22] A. Keränen and J. Ott, "Increasing reality for DTN protocol simulations," Networking Laboratory, Helsinki University of Technology, Tech. Rep., Jul. 2007.

[23] T. Hossmann, T. Spyropoulos, and F. Legendre, "Putting contacts into context: mobility modeling beyond inter-contact times," in *Proc. of the 12th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Paris, France, May 2011.

[24] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE simulator for DTN protocol evaluation," in *Proc. of the 2nd International Conference on Simulation Tools and Techniques*, Rome, Italy, Mar. 2009.

[25] Project page of the ONE simulator, [Online], https://www.netlab.tkk.fi/tutkimus/dtn/theone.

Proceedings of
2016 International Conference on
Networking Systems and Security (NSysS)

7-9 January, 2016, Dhaka, Bangladesh

Short Papers

Application Specific Security

Organized by
Department of CSE, BUET
Dhaka, Bangladesh

Certificate Revocation in Vehicular Ad Hoc Networks: A Novel Approach

Nazmul Islam
Institute of Information & Communication Technology
Khulna University of Engineering & Technology
Khulna, Bangladesh.
Email: nazmul.kuet@gmail.com/nazmul_islam@kuet.ac.bd

***Abstract*—To manage the traffic load along with providing safety and updated information, vehicular communication networking is marching at a high pace in research area. But similar to other networks VANET is not free from vulnerabilities. This network has brought forward a number of sectors which demand security for continuing operations safely. Among them, this paper contributes to a burning issue– how network secrets are distributed confidentially among the legitimate members of the network. The recent mechanisms involve searching a database for identities of revoked entities which is time consuming. This paper proposes an efficient mechanism dropping the time consuming database searching and finally analyzes its security on known attacks while certificate revocation process.**

Keywords—Vehicular networks; communication security; message authentication; certificate revocation.

I. INTRODUCTION

Vehicular ad hoc networks (VANETs) have attracted extensive attentions recently as a promising technology for revolutionizing the transportation systems and providing broadband communication services to vehicles. Since vehicles communicate through wireless channels, a variety of attacks such as injecting false information, modifying and replaying the disseminated messages can be easily launched. A security attack on VANETs can bring severe harmful or fatal consequences to legitimate users. Consequently, ensuring secure vehicular communications is mandatory before any VANET application can be put into practice.

To maintain the revoked certificates, public Key Infrastructure (PKI) along with Certificate Revocation Lists (CRLs) is a well-recognized solution for securing VANETs. A CRL is generally issued by a Trusted Authority (TA) and contains all the certificates of revoked entities. In PKI, each entity in the network holds an authentic certificate, and every message should be digitally signed prior to its transmission. Authentication of any message is performed by first checking the sender's certificate in current CRL, then, verifying the sender's certificate and finally verifying the sender's signature on the received message. Checking the revocation status of the sender in a CRL may incur long delay depending on the CRL size and the employed mechanism for searching the CRL. The IEEE 1609 WAVE communication standards, which are also known as Dedicated Short Range Communications (DSRC) protocols, have emerged recently to enhance 802.11 to support wireless communications among vehicles for the roadside infrastructure [1]. According to the DSRC, each vehicle has to broadcast a message every 300 msec about its location, velocity, and other information. In such scenario, each OBU may receive a large number of messages in every 300 msec and it has to check the current CRL for all the received certificates which may incur long authentication delay depending on the CRL size and the number of received certificates. The ability to check a CRL for a large number of certificates in a timely manner leads an inevitable challenge to VANETs. This paper presents a mechanism that uses only a secret to prove the authenticity and don't require searching any database.

The remainder of the paper is organized as follows: The related works are discussed in Section II. Section III introduces the proposed mechanism. Security analysis is given in Sections IV while section V experimentally proves the gravity of the proposed mechanism and section VI concludes the paper indicating future research directions.

II. RELATED WORKS

Till now, a lot of research has been conducted to make the vehicular network risk free e.g. Laberteaux et al. [2] used car to car communication to speed up the CRL broadcasting, Zhu et al. adopted a probabilistic key distribution approach based on pre-deployed symmetric keys and introduced the GKMPAN protocol [3]. Papadimitratos et al. [4] proposed to partition the CRL into small pieces and distribute each piece independently. In [5], Raya and Hubaux ensured secure and privacy preserving communications to VANETs by using a classical PKI where each vehicle needs to preload a huge pool of anonymous certificates. They showed that PKI is the most viable technique to achieve the security requirements such as entity authentication, message integrity, non-repudiation, and privacy preservation. But, in their mechanism, revoking one vehicle implies revoking the huge number of certificates loaded in it. Studer et al. proposed an efficient authentication and revocation scheme called TACK [6] which adopts hierarchy system architecture with a central trusted authority and some regional authorities (RAs) for different regions of the network. Each vehicle, if enters a new region, must update its certificate from the RA dedicated for that region. The RAs require some processing time, e.g., 2 seconds, before sending

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

the new certificate to the requesting vehicle. Hence, the vehicle cannot send messages to neighboring vehicles within this period. But the WAVE standard requires each vehicle to beacon about its location, speed, and direction at every 100-300 msec. so TACK becomes unsuitable for the safety applications in VANETs.

In this paper, a mechanism is proposed to overcome the problem of long delay incurred in checking the revocation status of a certificate using a CRL. The mechanism requires all non-revoked vehicles to use a unique secret for communication. To achieve the goal, a protocol to serve legitimate vehicles with new secret is presented.

III. Certificate Revocation Mechanism

The configuration of the system and the proposed mechanism are described below.

A. System model

The system model of vehicular ad hoc network consists of the following components.

- A Trusted Authority (TA); responsible for providing anonymous certificates and distributing secret keys to all OBUs in the network.
- Roadside units (RSUs); network wide distributed fixed units that can communicate securely with the TA.
- OBUs; which are embedded in vehicles, can communicate either with other OBUs through vehicle to vehicle (V2V) communications or with RSUs through vehicle to infrastructure (V2I) communications. Any communication between TA and OBUs is established through RSUs. In order to establish a communication RSUs can be chosen based on distance or message load or considering the both [7].

According to the WAVE standard, each OBU should equipped with a tamper-resistant module named Hardware Security Module (HSM) [8] or Tamper Proof Device TPD [9] which is used to store the security materials, e.g., secret keys, certificates, etc. of the OBU. This module is responsible for performing all the cryptographic operations such as signing messages, verifying certificates, keys updating, etc. It is considered that valid OBUs cannot collude with the revoked OBUs as it is difficult to extort the security materials from their HSMs. Finally, it is considered that a compromised OBU is instantly detected by the TA.

B. System initialization

At the initial phase, a trusted authority (TA) has the followings– A secret key set $U_s=\{K_1^-, K_2^-, \ldots, K_n^-\}$and corresponding public key set $U_p=\{K_1^+, K_2^+, \ldots, K_n^+\}$, a selected initial secret S by TA which is shared with all non-revoked OBUs, a master secret key and corresponding public key pair $\{ p^-, p^+\}$ to generate and verify the signature of TA and finally a hash functions $H()$ e.g. MD5 to produce message digest. After TA has finished its initial phase, OBU_j gets the followings from TA– A secret key set $V_j^s \subset U_s$ i.e. n keys chosen randomly from U_s and corresponding public key set $V_j^p \subset U_p$, certificates with IDs for each elements of V_j^p e.g. $CERT_j^i(ID_{common} \| PK_j^i \| sig_{TA}(ID_{common} \| PK_j^i))$ where $PK_j^i \in V_j^p$ and ID_{common} contains all IDs of OBUs that have the common key PK_j^i. OBU_j also receives an initial secret S, TA's public key p^+ and the hash function $H()$ from TA.

To send a message M, OBU_j masks it by generating a random number R to hide the message from revoked OBUs. OBU_j concatenates S, own id ID_j and a time stamp together and generates hash of the result. It also puts signature on the concatenation of M and S^R. Finally all parts are sent concatenating a certificate which has corresponding public key of the private key used for signature i.e. $ID_j \| M^R \| T \| CERT_j^i(ID_{common} \| PK_j^i \| sig_{TA}(ID_{common} \| PK_j^i)) \| H(ID_j \| S \| T) \| sig_j(M \| S^R)$. After receiving, OBU_k checks the certificate's validity and gets proved ID of the sender. The signature part of TA in certificate makes the receiver ensured about sender's ID and his public key. A matched hash checking using receiver's own secret S proves a non-revoked sender. Finally the integrity is checked by the signature part of the message. Here, timestamp ensures the message is not a record of previous conversation.

C. Revocation

The revocation is triggered by the TA when there is an OBU_u to be revoked. The certificates of OBU_u must be revoked from TA's database. In addition, the key sets V_u^s and V_u^p of OBU_u and the current secret key S are considered revoked. Hence, a new secret key S' should be securely distributed among all the non-revoked OBUs. Also, each non-revoked OBU should securely update the compromised keys in its key sets V^s and V^p that are common with revoked OBUs. Hence the revocation process comprises secure distribution of S' among legitimate OBUs and updating their key sets V^s and V^p. Upon completion of the process no legitimate OBU contains any shared key with revoked OBUs and the revoked OBUs do not have current network secret S', hence they become isolated from the network. The revocation process is as follow:

TA picks a key pair $\{K_p^+, K_p^-\}$ from its key pool which has been loaded into majority number of OBUs but not into OBUs that are to revoke. TA also removes the common key pairs with revoking OBUs from its pool prior to the selection of new pairs. Then, TA generates a new random secret S' and encrypts with K_P^+ i.e. $E_{Kp+}(S')$. Besides, a New Key pair List (NKL) is also prepared and masked with S'. NKL is a table that consists of three columns– first of which contains all the public keys to be revoked i.e. public keys of revoked OBUs (NKL_{RevPK}) and the second one contains the corresponding public/private key pairs ($NKL_{PK/SK}$) by which the revoked keys will be replaced and finally the third column contains certificates (NKL_{CERT}) corresponding to $NKL_{PK/SK}$ i.e. i^{th} element of NKL_{CERT} corresponding to $NKL_{PKi/SKi}$ is $NKL_{CERTi}= CERT_i(ID_{common} \| PK_i \| sig_{TA}(ID_{common} \| PK_i))$ where ID_{common} contains all IDs that contains NKL_{RevPKi} and need to replaced with PK_i. So the final broadcast message of TA becomes: $REV_{msg} = E_{Kp+}(S') \| NKL^{S'} \| sig_{TA}(E_{Kp+}(S'))$. TA's signature ensures receivers about the authenticity of $E_{Kp+}(S')$ and encryption under K_p^+ enables only the legitimate entities who has the corresponding secret key K_p^- to decrypt and retrieve S'. Masking NLK with S' hides the new key pairs from revoked

entities and adversaries but only the entities who have retrieved S' successfully.

After receiving, OBU_j runs Algorithm 1 where S' is retrieved to get NKL; hence the revoked key pairs and certificates are replaced with their corresponding entities in NKL. S', then, re-encrypted with all aged key pairs distinctly and broadcast for other neighbor OBUs having matched key pairs with OBU_j. It is unnecessary to re-encrypt with newly placed keys as they are not familiar to other OBUs at current stage. Thus, all non-revoked OBUs get update about the new secret and key pairs with certificates. Note that, here, commutative cryptosystem is used so the order of encryptions or decryptions becomes discretionary e.g. in algorithm 2 encryption on $sig_{TA}(S')$ under the key K_i^+ can be written as $sig_{TA}(E_{Ki+}(S'))$. Moreover, OBU_j may get multiple instances of message for the same secret from distinct OBUs sharing valid key pairs with it. In that case, OBU_j only processes the first instance.

Algorithm 1. Revocation process at OBU_j
Require: $REV_{msg} = (E_{Kp+}(S')||NKL^{S'}||sig_{TA}(E_{Kp+}(S')))$ and p^+
1. Verify $sig_{TA}(E_{Kp+}(S'))$
2. **if** invalid **then**
3. Exit
4. **else**
5. decrypt $E_{Kp+}(S')$ for S'
6. retrieve NKL from $NKL^{S'}$
7. **for** each $K_i^+ \in V_j^p$
8. **do**
9. **if** $K_i^+ \in NKL_{RevPK}$ **then**
10. Replace $\{K_i^+, K_i^-\}$ with $\{K_i'^+, K_i'^-\} \in NKL_{PK/SK}$: $NKL_{RevPK} \rightarrow NKL_{PK/SK}$ is one to one mapping
11. Replace $CERT_i$ with $CERT_i' \in NKL_{CERT}$: $NKL_{PK/SK} \rightarrow NKL_{CERT}$ is one to one mapping
12. **end if**
13. **end for**
14. Run Algorithm 2
15. **end if**

Algorithm 2. Updates broadcast by OBU_j
Require S', $NKL^{S'}$, $sig_{TA}(S')$
1. **for** each $K_i^+ \in V_j^p \notin NKL_{PK/SK}$
2. **do**
3. $REV_{msg} = (E_{Ki+}(S')||NKL^{S'}||\ sig_{TA}(E_{Ki+}(S')))$
4. broadcast REV_{msg}
5. **end for**

IV. Security Analysis

In this section, the possible attacks on proposed model along with the security are analyzed.

A. Forging Attack

In a forging attack, an attacker tries to forge the secrets to communicate successfully within the network. It is assumed impossible for an attacker to get the keys and secrets from TA at beginning. Hence, eavesdropping new keys at key update procedure becomes the best option for an attacker. Here new key list NKL and S' are open in masked form i.e. $NKL^{S'}$ which is a discrete logarithmic problem to solve. Furthermore, an attacker may get new secret S' from $E_{Kp+}(S')$ which is encrypted with K_p^+ and an attacker or revoked OBU cannot decrypt as they don't have the corresponding secret key K_p^-. As all keys of the revoked OBU are replaced with new key set, there remains no way to decrypt $E_{Kp+}(S')$ to get S' and in turn, forge NKL from $NKL^{S'}$. Hence the model is resistant against forging attack.

B. Resistance to Replay Attack

Replay attacks are the network attacks in which an attacker spies the conversation between sender and receiver and takes the authenticated information e.g. shared keys and then contact to the receiver with them. The most common countermeasures of replay attack are session token, one-time password and time stamping. Since in each message an OBU includes the current time stamp, an attacker cannot record it at time T_i and replay it at T_{i+1}. Hence, the model is secure against replay attack.

C. Resistance to Colluding Attacks

For a colluding attack, a legitimate OBU colludes with a revoked OBU by releasing the current secret key S, such that the revoked OBU can use the secret for its need. But, note that, all security materials here of an OBU are stored in its tamper-resistant HSM. Moreover, all algorithms are executed in HSM which means it is difficult to export secrets to a revoked OBU.

V. Experimental Analysis

In order to prove the efficiency of the proposed technique some experiments that were conducted are described in this section. To perform the simulation MATLAB is used and all cryptographic properties, presented here, i.e. key pairs, encryption/decryption, signing etc are performed using ElGamal cryptosystem. For simplicity, the effect of vehicle density and their speed, size of area, transmission delay and data loss during propagation etc. are assumed negligible.

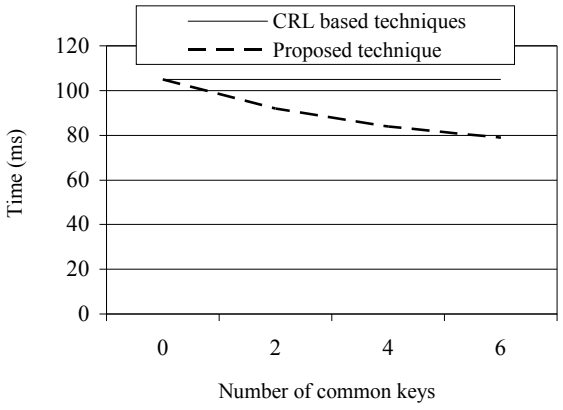


Fig. 1. Efficiency of the proposed technique in revoking the OBUs. Number of OBUs to be revoked is kept fixed.

The first experiment, result of which is shown in fig. 1, is conducted to prove the time efficiency of the proposed revocation technique. From the figure it is clear that the

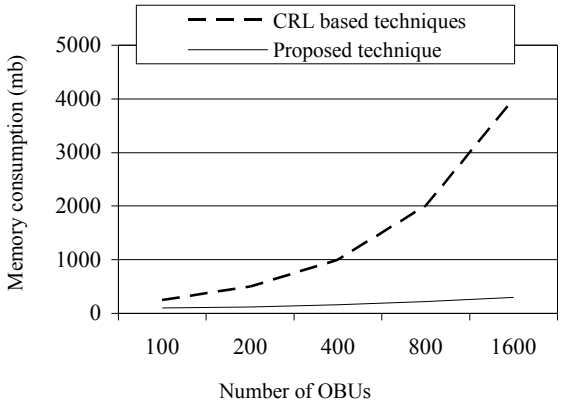


Fig. 2. Memory efficiency of the proposed technique.

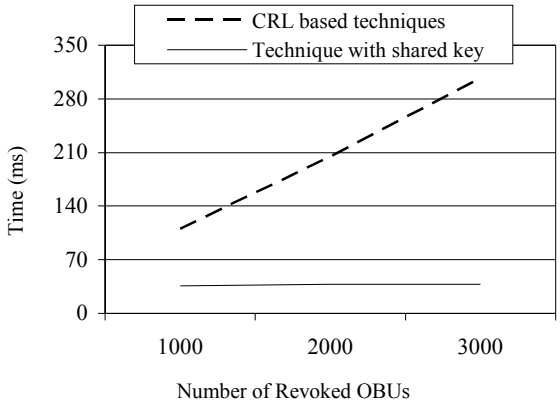


Fig. 3. Efficiency of shared secret in message authentication. For CRL based techniques time consumption to search the list increases with the growth of the list size.

proposed technique becomes successful in reducing the time consumption for establishing new secrets. The reason can be found analyzing the operation principle of the proposed technique. Unlike CRL based techniques, the proposed technique focuses on assigning certificate to keys rather than OBUs during the broadcast of revocation message which reduces the number of records of certificates. Namely, consider N OBUs with W key pairs each, among which Z pairs are common. To revoke all these OBUs CRL based techniques need to broadcast all NW certificates as each of them becomes distinct because of using individual IDs of OBUs in certificates. But in proposed techniques certificates of common keys becomes identical because of using ID_{common} for each of them. Hence, number of certificates needs to be broadcast reduces down to $N(W–Z)+Z$. Note that the proposed technique is not showing linear decrement which is a result of extra operation on receiver side due to increase of common keys.

In CRL based techniques [10], each vehicle may be stored with approximately 25000 certificates, and each certificate has 100 bytes, so, 2.5 mb is required to store the revocation data for one vehicle in CRL i.e. 250 mb for 100 adversary vehicles. Which is a very large size to be stored in OBUs, hence not memory efficient. On the other hand, the proposed mechanism does not need to remember the revoked entities but only a secret and some key pairs along with the corresponding certificates no matter how large the number of revoked entity is. This enables the proposed technique to be much more memory efficient. Fig. 2 shows the memory efficiency of the proposed technique. A small increase for the proposed technique is shown in the figure as increase of vehicles also enhances the need to store more number of keys to increase the probability of finding a neighbor OBU with a common key pair to establish a secure connection between them.

To prove the ultimate facility provided by the proposed technique the following experiment was conducted which is shown in fig. 3. Note that authentication of a message, if valid vehicles share a secret, becomes expedite than checking a CRL. This is because searching a list for a given entity requires time to complete depending on the size of the list and its organizing mechanism. Here, a binary search is performed on the list assuming the data is sorted. Though binary search reduces the time requirement to a large extent, it fails to make the process as faster as that of using a common secret. Using common secret to authenticate a message reduces the time consumption due to the fact that only legitimate entity can know about the secret.

VI. Conclusion and Future Work

This paper proposes an efficient certificate revocation mechanism for vehicular ad hoc network which replaces time consuming CRL checking by a secret that only belongs to the legitimate vehicles. Besides, the key sharing mechanism proposed in this paper abstain the revoked vehicle to have the new secret and enables other vehicles to update their secret information. Furthermore, it is resistant to common attacks during the revocation process. Therefore, the mechanism reduces time requirement to authenticate a message and hence decrease the message loss ratio due to message verification delay. The future work may focus on the certificate and message signature authentication acceleration. Moreover, how quickly the new secrets can be conveyed to legitimate entities and how to prohibit entities, that are still waiting for the updates, to communicate with revoked entities during the revocation process can be potential research issues.

References

[1] IEEE Std. 1609.2-2006, “IEEE Trial-Use Standard for Wireless Access in Vehicular Environments Security Services for Applications and Management Messages,” 2006.

[2] K.P. Laberteaux, J.J. Haas, and Y. Hu, “Security Certificate Revocation List Distribution for VANET,” Proc. 5th ACM int’l Workshop VehiculAr Inter-NETworking, pp. 88-89, 2008.

[3] S. Zhu, S. Setia, S. Xu, and S. Jajodia, “GKMPAN: An Efficient Group Rekeying Scheme for Secure Multicast in Ad-Hoc Networks,” J. Computer Security, vol. 14, pp. 301-325, 2006.

[4] P.P. Papadimitratos, G. Mezzour, and J. Hubaux, “Certificate Revocation List Distribution in Vehicular Communication Systems,” Proc. 5th ACM Int’l Workshop VehiculAr Inter-NETworking, pp. 86-87, 2008.

[5] M. Raya and J.-P. Hubaux, “Securing Vehicular Ad Hoc Networks, ”J. Computer Security, vol. 15, no. 1, pp. 39-68, 2007.

[6] A. Studer, E. Shi, F. Bai, and A. Perrig, “TACKing Together Efficient Authentication, Revocation, and Privacy in VANETs,” Proc. IEEE CS 6th Ann. Conf. Sensor, Mesh and Ad Hoc Comm. And Networks (SECON ’09), pp. 1-9, 2009.

[7] O. Faiza, B. Lila and G. Mourad, “The Involvement of RSUs in VANETs: Survey and Perspectives”, int’l J. Computational Engineering Research, vol. 03, no. 7, July, 2013

[8] P. Papadiamitratos, L. Buttyan, T. Holczer, E.Schoch, J.Freudiger, M. Raya, and J.P. Hubaux, “Secure Vehicular Communication Systems: Design and Architecture”, IEEE Communications Magazine, vol. 46, no. 11, pp. 100-109, November, 2008.

[9] P. Papadiamitratos, L. Buttyan, T. Holczer, E.Schoch, J.Freudiger, M. Raya, and J.P. Hubaux, “Secure Vehicular Communication Systems: Implementation, Performance, and Research Challenges”, IEEE Communication Magazine, vol. 46, no. 11, pp. 110-118, 2008.

[10] Haas, J.J., Y.C. Hu, and K.P. Laberteaux. “Design and analysis of a lightweight certificate revocation mechanism for VANET”, VANET '09: Proc. 6th ACM int’l workshop VehiculAr InterNETworking 2009.

Privacy and Security Problems of National Health Data Warehouse: A Convenient Solution for Developing Countries

Shahidul Islam Khan; Abu Sayed Md. Latiful Hoque
Department of Computer Science and Engineering (CSE)
Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh
nayeemkh@gmail.com; asmlatifulhoque@cse.buet.ac.bd

***Abstract*—Healthcare providers and researchers can discover the hidden knowledge from different health repositories if integration of data is performed by data warehousing. Integration of health records requires linkage of patients' data in different heterogeneous sources. Preserving record linkage in National Health Data Warehouse, by retaining identifiable attributes, is essential for effective data mining as well. In contrast identifiable health data have high risk to patient privacy and make the health information systems security vulnerable to hackers. In this paper, we have provided a practical solution: Global Patient Identification Technique (GPIT) that can anonymize identifiable private data of the patients while maintaining record linkage in integrated health repositories to facilitate knowledge discovery process. We have used encrypted mobile number, gender and NAMEVALUE of patients to produce Global Patient Identification Key. This system is being implemented in Bangladesh to develop National Health Data Warehouse. Our approach is also suitable for the developing countries where poverty and illiteracy rates are high among mass people.**

Keywords—Health / Medical Data; Data Privacy; Health Data Warehouse; Record Linkage; Data Mining;

I. INTRODUCTION

Data and Information has changed our lives and society. Data mining has made momentous advances over the past decades. Because of its potential power for solving complex problems, data mining has been successfully applied to diverse areas such as business, engineering, social media, and healthcare sectors. [1], [2]. For effective and efficient data mining, development of a data warehouse (DW) is required. A data warehouse is a subject-oriented, integrated, non-volatile, and time-variant collection of data to support management decisions [3]. It unites the data spread throughout an organization into a single central structure. Main advantages of warehouses are standardizing data across organizations and improved turnaround time for analysis and reporting. Long initial development time and associated high cost are treated as its major drawbacks [3].

A health data warehouse is a data store, different from hospitals' operational databases, used to analyze the consolidated historical health data [4], [5]. Development of a Health DW contains two key phases. Firstly, a conceptual view of the warehouse is specified according to the user requirements in the configuration phase. Secondly, the allied data sources and the Extraction-Transform-Load (ETL) process are determined. After the initial load, during DW operation, data must be refreshed on a regular basis such that data stored in the warehouse reflect the current state of the operational systems [3].

Patient records collected for diagnosis and prognosis typically encompass values of historical, clinical and laboratory parameters. Such datasets are characterized by their incompleteness i.e. missing values, incorrectness i.e. random noise in data and sparseness. The development of data mining tools for medical diagnosis and prediction is frequently motivated by the requirements for dealing with these characteristics of medical data sets. A huge amount of health records and related documents created by clinical diagnostic equipments are generated daily. These valuable data are stored in various medical information systems in various hospitals, departments and diagnostic laboratories. Data required to make proper medical decisions are trapped within fragmented and heterogeneous health systems that are not properly integrated. So integration of these health records into a single warehouse is necessary [6], [7].

Record linkage is the process of identifying record pairs from different information systems which belong to the same real world entity. Given two repository of records, the record-linkage process consists of determining all pairs that are similar to each other. Record linkage is essential when joining datasets based on entities that may or may not share a common identifier such as national id or social security number [8], [9]. For discovering effective knowledge such as correlations among diseases from medical dataset it is very essential to maintain record linkage. Health data have to be linkable in some way. But medical record linkage has inverse relation with patients' privacy. If a health dataset preserves record linkage means privacy of individuals are at risk. So, protecting privacy of patients, while maintaining effective record linkage, is an important research issue.

In the economically developed countries, literacy rates are high and citizens know their personal information such as date of birth and social security number correctly. Healthcare providers have also modern facilities such as IT, health

978-1-5090-0203-0/16/$31.00 ©2016 IEEE

Acknowledgments: This research is performed in the Department of CSE, BUET and funded by the ICT Division, Ministry of Posts, Telecommunication and Information Technology, Government of Bangladesh.

information management software, Internet etc. Each patient is assigned unique ID or reference number and all his medical information of different times are linked with that ID. But in developing countries poverty, population density and illiteracy are high. Many parsons even do not know their date of birth or full name. There is no unique ID available for all the citizens. Healthcare facilities are not modernized and also insufficient for highly dense population. People perform their diagnostic tests in different centers at different times and there is no linkage among them. So it is almost impossible to build efficient National Health Data Warehouse (NHDW) and perform data mining properly.

In this paper we have presented a brief overview of security and privacy risks of integrated healthcare information system. We have provided a practical solution namely Global Patient Identification Technique (GPIT) that can anonymize the identifiable private data of the patients while maintaining record linkage for doctors and researchers. Our approach is suitable for Bangladesh and other developing countries where poverty and illiteracy rates are high.

II. Security and privacy issues related to health data

Security of a Health Information System deals with protecting medical data from intruders, malwares, and frauds. It retains confidentiality and integrity of healthcare data. As medical systems are more interconnected and networked, security has become a huge challenge in healthcare sector.

A. Data security and privacy

Data or information security refers to protective digital security measures that are applied to prevent unauthorized access to computers, databases and websites. Examples of data security technologies include software/hardware disk encryption, backups, data masking and data erasure [10]. Data or information privacy deals with the ability an organization or individual has to determine what data in a computer system can be shared with others. It is considered an important aspect of information sharing. Wherever personally identifiable information is accumulated in any form. A major challenge in data privacy is to share information in a way that personally identifiable data is protected. Information privacy may be applied in numerous ways, including encryption, authentication and data masking.

Medical data are highly private. We do not want others to know about our medical or psychological conditions or treatments. It may affect one's insurance coverage or employment. Security and authentication systems are often required for individuals that process and store medical records. Many countries developed standards for doctor-patient relationships that preserve confidentiality. These standards protect patients' dignity and ensure the patients' comfort to reveal accurate information to receive the correct treatment.

B. National Health DW: Pros and Cons

There is no doubt that development of national health data warehouse is very much essential for every countries but it raise high risk to data security and privacy of citizens. After deployment of National Health DW, health service providers, and healthcare researchers can have access to private health data of millions of patients without bar. Prior integration to health data warehouse, these sensitive and private data reside to a single organization such as hospitals or diagnostic centers. Only the particular organization, where a patient's health data is generated, is responsible to protect the data privately. If any leakage of a patient data or any other kind of privacy violation occurs, that particular organization is liable for this. So every health service provider protects diagnostic test reports and other health data for its own interest.

The situation is far different in the case of NHDW. Here, if a patient's sensitive and private data leaks, who will be blamed? Against whom or which organization to file Defamation-suit? It is not possible to guarantee that all the doctors, health researchers, health service providers will execute their responsibility and no one will violate in any way. So proper measure has to be taken so that individual patient cannot be identified from health database or warehouse and their privacy is safeguarded.

C. Why health data are important to hackers?

There is a growing trend of hacking into medical records. Hackers' objective is to exploit personal information, which is a lucrative business to them. In U.S., a stolen Social Security Number might sell for 25 cents in the underground market, and a credit card number for $1. Whereas sell value of a comprehensive medical record may varies from 10$ to $1,000 in black market [11]. The data for sale includes names, birth dates, health policy numbers, diagnosis codes and billing information. Fraudsters use this data to create fake IDs to buy medical equipment or drugs. The hackers combine a patient number with a false provider number and file made-up claims with insurers, according to experts who have investigated cyber attacks on healthcare organizations.

Hospitals have low security, so it is relatively easy for the hackers to get large amount of medical data. Government sector like public health department systems that contain health-related data is also an increasing target of hackers for two main reasons. First, they are generally more vulnerable, as they are older systems running older, less secure software. Second, they are rich in data like personally identifiable information, healthcare, financial information [12].

III. Data Breaches of Health Information Systems

A data breach or leakage can be defined as any incident which involves loss or exposure of personal records digitally. Personal records means information about a person that cannot be obtained easily through other public means; and this information only known by an individual or by an organization under the terms of a confidentiality agreement. Cyber criminals recognize two critical facts about the healthcare industry:

- Healthcare organizations posses large and monetarily lucrative personal data
- Most of them do not have the resources and technologies to detect cyber attacks and effectively protect healthcare data.

According to 2015 Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data which covered 90 healthcare organizations in U.S., More than 90% of healthcare service providers had a data breach, and 40% had more than five data breaches over the past two years [13]. According to the Fifth Annual Study on 2014, medical identity theft nearly doubled in five years, from 1.4 million adult victims to over 2.3 million in 2014 [14]. In last ten years at least 18 health breach reported in Europe affected minimum 9,337,197 individual records [15]. The health records include details on the patients' conditions, names, home addresses and dates of birth. The main causes of data breach in healthcare sectors are illustrated in Fig. 1.

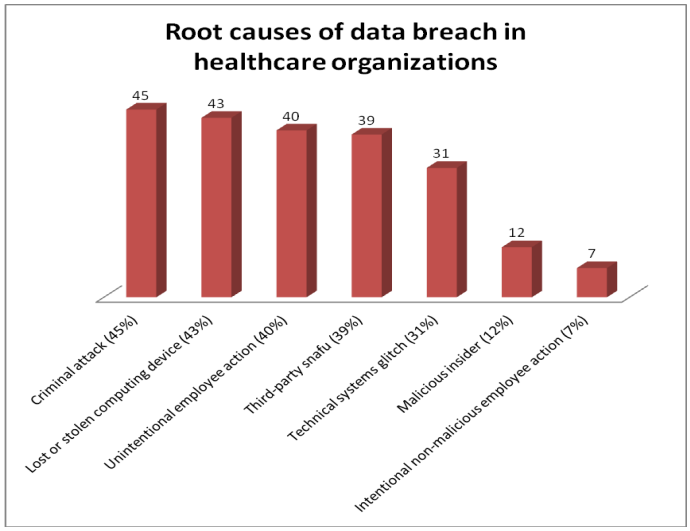


Fig. 1. Root cause of the healthcare organizations' data breach

Hacking is increasing in the *Healthcare Servers* at a shocking rate. From 21th October 2009 till data there are 1279 health data breach reported, among which only 190 is server attacks that is average 14.85%. But up to 1st August 2015, in last 12 months 58 of 255 are server attack, which is 22.74%. We have uncovered this insight by analyzing the data provided by U.S. Department of Health and Human Services [16]. Hackers are increasingly targeted to the health servers which are very alarming to national level health information system development. Fig. 2 presents a statistics of the main target of the hackers. We have collected the data from [13].

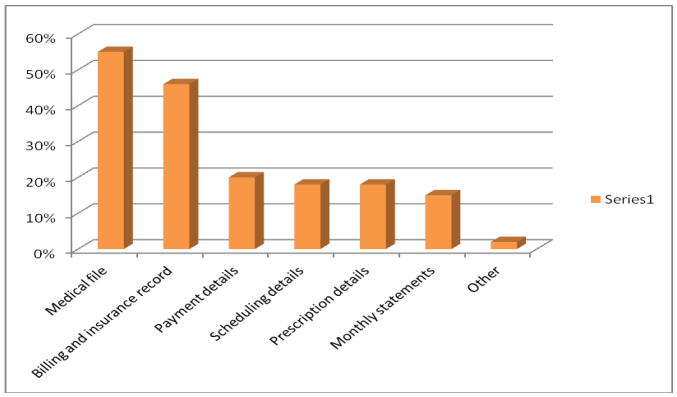


Fig. 2. Target of the Cybercriminals in Healthcare Systems

From Fig. 1 and Fig. 2 we can summarize that, currently criminal attacks are the main cause of healthcare information systems' data breach and the main target of the hackers are digital medical files or health records.

No information system can be assumed to be completely protected from all kind of criminal and cyber attacks. Security can be more vulnerable in the case of large scale, national level health information systems where Internet communication has to be maintained for the sake of easy data collection from far-most parts of the country. Integrated health information system should be designed in such a way that:

- There is enough data to maintain record linkage so that doctors, researchers can get useful insight from the system.
- If data breach occurs, individual patient's privacy will not be compromised.

We have provided a practical solution that is capable to provide above mentioned facilities. It is presented in the next section.

IV. A Solution: Global Patient Identification Technique

The block diagram of our system is shown in Fig. 3. To design the system we have followed a practical oriented approach suitable for Bangladesh, and other countries in the globe where rate of poverty and illiteracy among mass people is still high. Developing national scale information system is a major challenge is these countries due to large population and fewer resources.

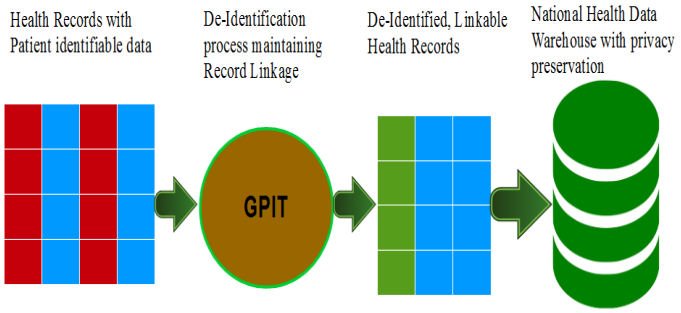


Fig. 3. Block diagram of global patient identification technique (GPIT) with privacy and record linkage preservation. Health Records with Patient identifiable attributes such as name, address, date of birth in heterogeneous format from various health service providers are inputted in the systems. These records are then de-identified preserving record linkage. These privacy preserved linkable health records are being stored into national health data warehouse as unified data format.

The input of GPIT system is health records provided by different health care organizations such as government and private hospitals, diagnostic centers, research centers, health NGOs. These data are is heterogeneous formats like Oracle, MS SQL or MySQL databases; CSV or MS Excel files etc. The detail architecture of the National Health Data Warehouse Bangladesh can be seen from our previous paper [17]. These raw health records contain attributes related to patient identification such as patient name, address, and mobile

number. Our global patient identification key (GPIK) algorithm works in two steps.

- In step 1, a GPIK is generated for each patient record using available patient identifiable data.
- In step 2, all identifiable data, capable to identify individual patients is removed from the health record

We have used three attributes to generate identification key; mobile number, name, and gender of a patient. Mobile number is stored in an encrypted format. Name is converted to NAMEVALUE. Age is used to generate year of birth and age group. Location is used to generate GEOCODE.

NAMEVALUE is the encrypted text string generated by our developed Name-Value Algorithm using significant and unambiguous characters contained in a patient's name. We have treated salutations and titles as insignificant. In the practical situations at most health centers or doctors' chamber, patients are asked and they tell their information i.e. name, age verbally. From pronounce to write, vowels are highly ambiguous and vowels can be written in many ways. This can be understood clearly from the following table. From Table 1 we can see that doctors or computer operator can write or entry following two patients in six or more ways. To remove ambiguity, vowels are discarded from significant portion of a name. Then the data is encrypted using simple encryption technique so that real name cannot be understood by the health warehouse users.

TABLE I. SELECTION OF SIGNIFICANT, UNAMBIGUOUS NAMEVALUE

Patient Name	Significant portion	Unambiguous significant portion	Encrypted NAMEVALUE
Mr. Abul Hosain	Abul Hosain	abl hsn	tagsemi
Mr. Md. Abul Hosen	Abul Hosen	abl hsn	tagsemi
Mohammad Abul Hosain	Abul Hosain	abl hsn	tagsemi
Mr. Subir Saha	Subir Saha	sbr sh	malsme
Sree Subir Saha	Subir Saha	sbr sh	malsme
Sree Subeer Saha	Subeer Saha	sbr sh	malsme

Mobile numbers of the patients are encrypted and concatenated before the NAMEVALUE. Gender information is also concatenated to get the Global Patient Identification Key (GPIK) that is shown in Fig. 4.

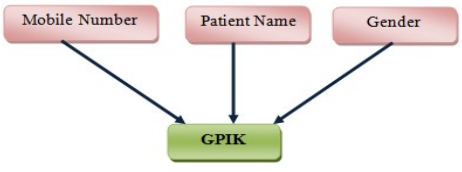


Fig. 4. Mobile no, name and gender of patients contribute to generate global patient identification key (GPIK)

GEOCODE is the 8 character de-identified address of a patient generated from his health record. We have used the *concept hierarchy* technique to produce GEOCODE from patients' address records [18]. This data is stored is the national health data warehouse to facilitate spatial data mining related to geographic location such as correlation of diseases and places. Generation of GEOCODE for Bangladeshi citizens can be understood by Fig. 5. This standard is used by Bangladesh government for passport and voter ID. It is also used by Directorate General of Health Services (DGHS), Ministry of Health & Family Welfare of Bangladesh Government [19], [20].

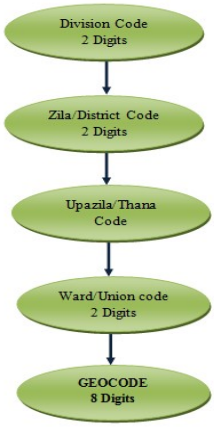


Fig. 5. GEOCODE sequence of 8 digits used in Bangladesh contains Division Code, Zila Code, Thana Code and Union Code of 2 digits each

Our algorithm of Global Patient Identification Technique (GPIT) is presented below:

Input: Health record set including patient identifiable data
Output: De-identified Linkable patient record
Steps:
Repeat

1. Encrypt mobile number
2. Convert patient name to NAMEVALUE
3. Convert address to GEOCODE
4. Convert Date of Birth or Age to BIRTHYEAR
5. Generate global patient identification key (GPIK) from encrypted mobile number, NAMEVALUE and Gender
6. Add GPIK and GEOCODE, BIRTHYEAR to record set
7. Delete Patient name, Mobile number, Address, Date of Birth, Credit Card number data.

Until last record

In Bangladesh many people do not know their date of birth. The case is most common to the aged rural people with less education. The major reasons are illiteracy, poverty, lack of social awareness etc. The situation of not knowing own birth date is more or less same among poor and uneducated people around the world. In many cultures and jurisdictions, if a person's real birthday is not known (for example, if he or she is an orphan), then their birthday may be considered to be January 1. [21]. The main reason of this, it is very easy to remember. So we have not considered date of birth rather considered birth year in our patient identification technique.

Another important reason is that in most hospitals and diagnostic centers, only patient's age is collected rather than date of birth. They have already millions of patients' health data scattered among their enormous health system without

patients' date of birth but with their ages. From patients' age data, their birth year can be easily calculated in the integrated warehouse data.

V. Motivation of GPIT in Developing Countries

We have used encrypted mobile phone numbers to distinguish individual patients because of the following reasons:

1. According to Bangladesh Bureau of Statistics total population of Bangladesh is 158,988,940 [19] and there are 124705000 active mobile connections [22].
2. Approximately 78.43 active mobile connections per 100 peoples are available.
3. Almost every family irrespective of rich or poor, urban or rural posses at least one mobile phone.
4. Every person must go through some security verification to purchase mobile SIM.
5. People already uses mobile for various identification and transaction purposes such as getting passport and national id card, performing financial transactions etc.
6. Mobile numbers are easy to remember and tell within shortest possible time.
7. Every mobile number is unique.
8. Most importantly, almost all health care centers collect mobile numbers of patients for communication and billing purposes. So mobile number is available with existing millions of health data.

No other identification number i.e. passport, national id, birth registration number has the above features. In all other SAARC countries the situation is more or less same which can be seen from Table II. In January 2015, there were 3.65 billion unique mobile users in the world with 7.09 billion active mobile subscriptions among 7.258 billion people in the world according to [23].

TABLE II. Statistics of mobile phone users in SAARC countries [24]-[27]

Country	Number of Mobile	Connection/ 100 citizens
India	960,579,472	77.58
Pakistan	140,000,000	77
Bangladesh	124,705,000	78.43
Sri Lanka	22,123,000	107
Nepal	18,240,670	86.82

In developing countries, population density and illiteracy are high. Many people do not know their date of birth or full name even. There is no unique ID available for all the citizens. People perform their diagnostic tests in different centers at different times and there is no linkage among them. Currently there is no option exist in Bangladesh Health DW to protect patients' privacy while providing efficient record linkage [28]. Using GPIT, privacy and record linkage problems can be solved for existing millions of health data available in different healthcare centers.

VI. Limitations and further research

A problem with mobile number based identification is many people use multiple mobile numbers. A parson with multiple mobile phones can provide one number in a health center and another one in other health center or the same center in different time. Thus the person's health data with two different mobile numbers will be treated as two different individual's data in warehouse. It will impact on mining results. A simple solution is to develop social awareness by the Government so that citizens provide only one number among their available numbers when taking health and other citizen services.

Another problem, though rarely, may occur due to change of mobile numbers by patients. For example a child after getting adult, own a mobile. He or she has already records in the health warehouse with his guardians' mobile number. This kind of problem can be addressed by writing simple DML query that will replace old patient identification key with new one generated with changed mobile number.

Though encrypted mobile numbers can uniquely distinguish patient data, we have also stored NAMEVALUE in the data warehouse because in some cases mobile number is insufficient for clustering patient data from health warehouse. For example, father and his child have same mobile number and GEOCODE but different NAMEVALUEs.

One of the many future research directions can be to design efficient data mining algorithms that can cluster all records of individual patients properly from National Health Data Warehouse.

VII. Related Works

A three step record linkage method is proposed in [29]. The first step is to standardize and indexing elementary identity fields using blocking variables. The second is to match similar pair records and finally in the third step clusters of coherent related records are created, using graph drawing technique, agglomerative clustering methods and partitioning methods.

In [30] for five institutions de-identified record with an exact match of patient first and last names and dates of birth were retrieved. Numbers of patient records existing for the topmost 250 commonly occurring first and last name pairs were determined. The authors also identified methods for managing duplicate records.

The authors in [8] synthesize this literature to formalize a new framework for privacy preserving interactive record linkage (PPIRL) with tractable privacy and utility properties and then analyze the literature using this framework.

Development, implementation and evaluation of a bespoke de-identification algorithm used to create the Mental Health register is discussed in [31]. The system is designed to create dictionaries using patient identifiers (PIs) entered into dedicated source fields and then identify, match and mask them (with ZZZZZ) when they appear in medical texts.

The authors of [32] developed a software application that performs data cleaning, preprocessing, and hashing of patient

identifiers to remove all protected health information. The application creates seeded hash code combinations of patient identifiers using an algorithm.

In summary, above works have their own merits and limitations but they are not effectively applicable in development of National Health Data Warehouse of Bangladesh or other economically developing countries. This is due to the unavailability of social security numbers or similar identification keys for the whole population. Another reason is the high illiteracy rates such that many people even do not aware of their date of births.

VIII. CONCLUSION

Health data warehouse development is a complex and time consuming process but is essential to deliver quality health services. Preserving record linkage by retaining identifiable attributes in national health data warehouse is essential for effective data mining. But identifiable health data have high risk to patients' privacy and also make the health information systems vulnerable to hackers. Disclosure of medical information about a patient may be harmful to his personal life and also for his career. Currently integrated health systems are in top hit list of cyber criminals as medical data worth 10 times higher than credit card numbers in the underground market.

In this paper, we have provided a practical solution: global patient identification technique (GPIT) that can anonymize the identifiable private data of the patients while maintaining record linkage to facilitate knowledge discovery by healthcare providers and researchers. We have used encrypted mobile number, gender and NAMEVALUE of patients to produce anonymize and linkable Global Patient Identification Key. Currently there is no option in Bangladesh Health DW to protect patients' privacy while providing efficient record linkage. Our system can be implemented in the national warehouse to protect patient privacy and to support efficient knowledge discovery. Using GPIT, patients' data can be shared and integrate among different government and private hospitals and diagnostic centers in Bangladesh. Our approach is also suitable for the economically developing countries where poverty and illiteracy rates are high among mass people.

REFERENCES

[1] U.M. Fayyad, G.P. and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," Advances in Knowledge Discovery and Data Mining, 1996, 1–36.

[2] A. Azzalini, and B.Scarpa, Data Analysis and Data Mining An Introduction, Oxford University Press, 2012.

[3] W. Inmon, Building the Data Warehouse, 4th edition, Wiley-New York 2005.

[4] T.R. Sahama, and P.R Croll, "A Data Warehouse Architecture for Clinical Data Warehousing," Australasian Workshop on Health Knowledge Management and Discovery, 2007.

[5] J.A. Lyman, K. Scully, and J.H. Harrison, "The development of health care data warehouses to support data mining," Clin Lab Med. 28,1 2008, pp. 55-71

[6] K. Cios, "Uniqueness of medical data mining," Artificial intelligence in medicine, Vol. 26, 2002, pp.1-24

[7] O. Maimon, and L. Rokach, Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer 2010.

[8] H.C. Kum, A. Krishnamurthy, A. Machanavajjhala et. at., "Privacy preserving interactive record linkage (PPIRL)," J Am Med Inform Assoc vol. 21, 2014, pp. 212–220.

[9] J. Liang, L. Chen, and S. Mehrotra, "Efficient Record Linkage in Large Data Sets," In Proc. of the Eighth International Conference on Database Systems for Advanced Applications, 2003.

[10] F. T. Harold, K. Micki, Information Security Management Handbook, Volume 2, CRC Press

[11] Why Hackers Are Targeting Health Data, http://www.databreachtoday.asia/hackers-are-targeting-health-data-a-7024

[12] Your medical record is worth more to hackers than your credit card, URL http://www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0HJ21I20140924

[13] Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data, Ponemon Institute Research Report, May 2015

[14] Fifth Annual Study on Medical Identity Theft 2014

[15] Central European University, Reported Breaches of Compromised Personal Records in Europe http://cmds.ceu.edu/sites/cmcs.ceu.hu/files/attachment/article/663/databreachesineurope.pdf

[16] U.S. Department of Health and Human Services https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

[17] S. I. Khan and A.S.M.L. Hoque, "Towards Development of National Health Data Warehouse for Knowledge Discovery", Intelligent Systems Technologies and Applications, Springer, Vol. 385 No.2, pp.413-421, 2016

[18] Y. Lu, Concept hierarchy in data mining: Specification, generation and implementation, Doctoral dissertation, Simon Fraser University, 1997.

[19] Bangladesh Bureau of Statistics, http://www.bbs.gov.bd/PageWebMenuContent.aspx?MenuKey=150

[20] DIRECTORATE GENERAL OF HEALTH SERVICES http://app.dghs.gov.bd/bbscode/

[21] http://www.syracuse.com/news/index.ssf/2011/01/on_new_years_day_wish_a_happy.html

[22] Bangladesh Telecommunication Regulatory Commission, http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-may-2015

[23] Digital, Social and Mobile in 2015, http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/

[24] http://www.trai.gov.in/WriteReadData/PressRealease/Document/PR-TSD-Feb-10042015.pdf

[25] http://www.pta.gov.pk/index.php?option=com_content&view=article&id=269&Itemid=599

[26] http://trc.gov.lk/2014-05-13-03-56-46/statistics.html

[27] http://thehimalayantimes.com/business/countrys-telephone-penetration-crossed-100pc-telecom-regulator/

[28] A Quiet Revolution: Strengthening the Routine Health Information System in Bangladesh, published by giz, accessed from http://health.bmz.de/good-practices/GHPC/A_Quiet_Revolution/HIS_Bangladesh_long_EN.pdf

[29] E.A. Sauleau, J. Paumier, and A. Buemi, "Medical record linkage in health information systems by approximate string matching and clustering," BMC Med Inform Decision Making, Vol. 5, 2005, pp.32–44.

[30] A.B. McCoy, A. Wright, Kahn M. et al., "Matching identifiers in electronic health records: implications for duplicate records and patient safety," BMJ Qual Saf Vol. 22, 2013, pp.219–24.

[31] C. F. Andrea, C. Danielle, T.M. Matthew et. al., "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records," BMC Medical Informatics and Decision Making , Vol. 13, 2013, pp.13:71.

[32] N. K. Abel, P. C. John, L. J. Kathryn et. al., "Design and implementation of a privacy preserving electronic health record linkage tool in Chicago," Journal of the American Medical Informatics Association, 2015, pp.1-9.

Author Index

S

T

U