IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Factors Influencing Infection Source Identification in Complex Networks: An Empirical Study

Syed Shafat Ali[1*], Syed Afzal Murtaza Rizvi[2]

[1,2]Department of Computer Science, Jamia Millia Islamia, New Delhi, India

[*]*Corresponding Author: shafat159074@st.jmi.ac.in, Tel.: +91-80766-26598*

DOI: https://doi.org/10.26438/ijcse/v7i5.17911804 | Available online at: www.ijcseonline.org

Accepted: 23/May/2019, Published: 31/May/2019

***Abstract*—** One of the important characteristics of the modern-day world is its high connectivity. While it has brought people closer and made lives easier, it has also paved way for harmful content, such as diseases, rumors, computer viruses, etc., to flow easily and spread even quicker. Therefore, finding the source of such unwanted diffusion processes becomes critical to mitigate the damages and avoid future threats. Consequently, infection source identification in complex networks has become an important problem with wide range of effective and meaningful applications. Researchers, over the years, have produced elegant and efficient solutions for the same. The main aim of this paper is to study the factors affecting locating a source of infection. This study largely focuses on four such factors: topology, graph density, infection probability and infection size. For performance analysis, three well known state-of-art source identification techniques, i.e., Dynamic Age (DA), Reverse Infection (RI) and Minimum Description Length (MDL), are employed. Largescale and extensive experiments conducted on various datasets indicate that all the four factors play critical roles in infection source identification, irrespective of the source identification technique employed.

***Keywords*—**Infection Source Identification, Information Diffusion, Social Networks, Complex Networks, SI Model

I. INTRODUCTION

The modern-day world is characterized by its high connectivity. Be it rapid urbanization, where people live in tightly knit societies, emergence of social networks where any person could connect with any other person, or technologies like computer networks or power-grid networks, connections are multi-dimensional and present everywhere. While this connectivity is highly desired as well as required, it has its own shadowy patches with the potential of great harm. One such unwanted scenario of such high connectivity is the diffusion or the spreading of harmful content, generally referred to as infection. For example, a disease breaking in a densely populated area could spread easily and do so at the speed of knots, resulting in human casualties (Zika Virus, Ebola Virus) [1,2]. Similarly, a rumor or false news diffusing through online social networks, like Facebook and Twitter, could potentially have drastic effects on human societies [3,4,5]. Besides this, a virus or a malicious program propagating through computer networks could end up destroying sensitive data and, thereby, compromising security [6,7,8]. Therefore, to mitigate the harm caused by such diffusion processes and avoiding them in future, locating their sources becomes essential. This is where the study of infection source identification comes into picture.

Given the importance of infection [1] source identification problem and its wide array of essential applications, researchers over the years have extensively studied it. Shah and Zaman, with their seminal work in the problem domain, studied infection source identification in tree-like networks under *Susceptible-Infected* (SI) model [9,10,11]. Following this, researchers approached the problem under similar conditions [12,13,14]. Afterwards, the problem was expanded and studied under more rigorous and testing conditions, i.e., under *Susceptible-Infected-Recovered* (SIR) and *Susceptible-Infected-Susceptible* (SIS) models [10,11,15,16,17]. Later on, researchers studied the problem in general graph networks by easing the constraints of tree-like networks [18,19,20,21].

This paper analyzes various graph factors and their impact on infection source identification. Mainly, four such factors, i.e., graph topology, density, infection probability and infection size, are considered. A brief overview of these factors will be presented in Section IV. The main aim of this paper is to shed light over the factors playing key role in locating the

[1] An infection could be a rumor, a computer virus, a disease, etc.

source of infection, which, in turn, would help researchers choose or develop proper techniques under different prevailing conditions, thereby, approaching the problem with prior knowledge. For the same, both real world (Facebook and US Power Grid (USPG)) and synthetic networks (Erdos-Renyi or ER-random and k-regular) are used, and classical *Susceptible-Infected* (SI) model (both homogeneous and heterogeneous) is employed to simulate infection diffusion over networks [22,23,24,25]. For performance analysis, three well-known state-of-art source identification techniques, i.e., Dynamic Age (DA), Reverse Infection (RI) and Minimum Description Length (MDL) are implemented in this paper [17,18,20,21]. Therefore, besides analyzing the impact the various graph factors, comparison of the performance of these methods are performed in various scenarios. The reason to choose these three techniques lies in the fact that all three techniques use contrasting approaches. While RI is Jordan-based, both DA and MDL are eigen-based, where DA works with eigen-vectors corresponding to largest eigen-value and MDL works with eigen-vectors corresponding to the smallest. Besides this, in addition to infected nodes in a graph, MDL utilizes the information provided by non-infected nodes as well. These methods will be discussed in detail in Section V.

The effect of topology is analysed on four datasets, i.e., ER-random, 4-regular, Facebook and US Power Grid (USPG), under heterogeneous SI model [22,23,24,25]. The results indicate that infection source identification is topology-dependent with different methods producing topology-specific results. Furthermore, it is observed that RI produces best results on ER-random graph and DA on Facebook. The analysis further shows that it is hard to detect infection sources on US Power Grid, a sparse graph with a very large diameter. Afterwards, analysis of the effect of graph density on source identification is performed on four ER-random and four k-regular graphs with different density. The results indicate that it is easier to find sources of infection if the underlying graph is dense as compared to sparse. Then, the third factor, i.e., the probability with which infection spreads across a network, is examined under homogeneous SI model on ER-random and Facebook networks, and it is found that higher infection probability leads to easier source detection. As for demonstrating the impact of infection size, two topologies, i.e., ER-random (with different properties) and Facebook are picked. The experimental results show that when the infection size is smaller, the source identification becomes easier. However, it is extremely hard to find infection sources when infection covers most of the underlying graph.

The contributions of this paper are three-fold:

1. A large-scale analysis of infection source identification is performed on different types of graphs.
2. Impact of various factors (topology, graph density, infection probability, infection size) on source identification is thoroughly examined.
3. The performance of three state-of-the-art source identification techniques, one Jordan-center based and two eigen based, is analyzed with respect to different factors on a multitude of graphs.

The rest of the paper is organized as follows. Section II provides a comprehensive literature on infection source identification problem. Section III presents a brief, technical overview of the source identification problem and Section IV explains various factors affecting source identification used in this study. In Section V, all the three source identification techniques are fully explained with technical aspects of their source estimation. Section VI presents a complete experimental layout and all the relevant aspects, i.e., experimental framework, theoretical description of SI model, datasets used, general experimental framework and evaluation measures used in this study. Results obtained and the relevant discussions can be found in Section VII and finally Section VIII provides a brief conclusion of this study and possible future directions.

II. RELATED WORK

Given its wide range of effective and important applications, infection source identification has gained a substantial amount of attention over the past decade. Early on, the problem was studied in its most ideal form under SI model while assuming the underlying topology is tree-like networks, and started off with the original work of Shah and Zaman [9]. They introduced the concept of rumor centrality of a node u in graph which is defined as the number of unique diffusion paths emanating from u. The node with the highest rumor centrality is called the rumor center and is considered to be the source. Later on, Shah and Zaman extended the concept of rumor centrality for locating the source of infection in general graphs by using BFS trees as a representation of the original graphs. Other works followed suit and studied the problem under the similar assumptions, i.e., the underlying graph topology is tree-like and infection spreads under SI model [10,11,12,13,14]. However, with SI model we get complete observation of infection graphs, i.e., we can easily tell which nodes are infected and which or not, but in real-world complete observations are hard to get. Therefore, researchers tackled the source identification problem under SIR (*Susceptible-Infected-Recovered*) and SIS (*Susceptible-Infected-Susceptible)* models, i.e., with partial observations [10,11]. Reference [17] proposed a sample path-based approach to identify infection source in tree graphs under SIR model and proved that the source associated with the optimal sample path is the Jordan center. Later on, Lou, et al., studied the sample path-based approach under SI and SIS models of infection and came to the same

conclusion [13,15]. Moving away from trees and on to generic graphs, to find the infection source, researchers relaxed the constraints of tree networks [18,19,20,21]. Reference [18] introduced the concept of dynamical age (DA) of a node inspired by its dynamical significance [26]. DA computes the amount of reduction in the largest eigenvalue of adjacent matrix, corresponding to a graph, after a node is removed. The node with the largest reduction is considered to be the source. Furthermore, researchers introduced the concept of minimum description length (MDL) for source identification [20,21]. They first compute the Laplacian matrix L corresponding to the graph and then identify the eigenvector corresponding to the smallest eigenvalue of L. A node with highest score in this eigenvector is considered to be the source. Reference [27] proposed K-Center to identify multiple sources of infection and works much like K-Means Distance heuristic. Picking the K initial centroids at random (where K is the number of sources), it partitions the graph into K partitions using Voronoi partitioning and updates the centers in each partition using effective distance [28,29]. When the convergence is reached, the centers in each partition are considered as sources. Furthermore, Label Propagation based Source Identification (LPSI) exploits the concept of source prominence to find the source using label propagation mechanism much like a Markov chain process [30]. Reference [31] studied the problem of multiple source detection under heterogeneous SIR model. To this end, they introduced the concept of Jordan cover - a set of nodes which covers all the observable infected nodes within minimum radius. This Jordan cover is considered to be the set of infection sources.

Infection source identification has also been studied under sensor-based observation. Sensors are special kind of nodes placed into a given network whose purpose is to provide information on their states (susceptible, infected or recovered), the infection direction and the time when they got infected. The advantage of using sensors is that they reduce the search for source to one specific part of the network. Pinto et al. provided a Gaussian method for single source estimation while assuming the infection propagation follows SI model in tree-like networks [32]. Later on, Louni and Subbalakshmi used high betweenness values of sensors to identify the bridges between communities in a network and showed that it was possible to reduce the number of sensors in a network and yet achieve better results [33]. In addition to this, Agaskar and Lu used Monte-Carlo method to identify sources in generic networks using sensors [34]. Reference [35] used Bayesian belief propagation model and Xie et al. employed moon-walk technique to identify sources under sensor observations [36].

III. INFECTION SOURCE IDENTIFICATION PROBLEM

Given an infection graph, $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$, with unknown source s^*, the goal is to estimate the source of infection, $\hat{s}$. For better accuracy, sometimes, researchers also make use of underlying graph, $G(V, E)$, where, besides infected nodes, non-infected nodes are utilized as well [20,21]. In addition to this, researchers also assume that infection probability or the time of infection is known, thereby, making it comparatively easier to locate the source of infection [27,37]. It is important to note that while under SIR and SIS models, only partial observations of infection are available, however, under SI model, nodes are distinguishable from infected to non-infected. Therefore, having the prior knowledge of the underlying model of infection is crucial for determining the applicable source identification technique.

IV. OVERVIEW OF THE FACTORS

Since this paper deals with understanding and demonstrating the effects of graph topology, density, infection probability and size on infection source identification, it becomes pertinent to briefly discuss them at first.

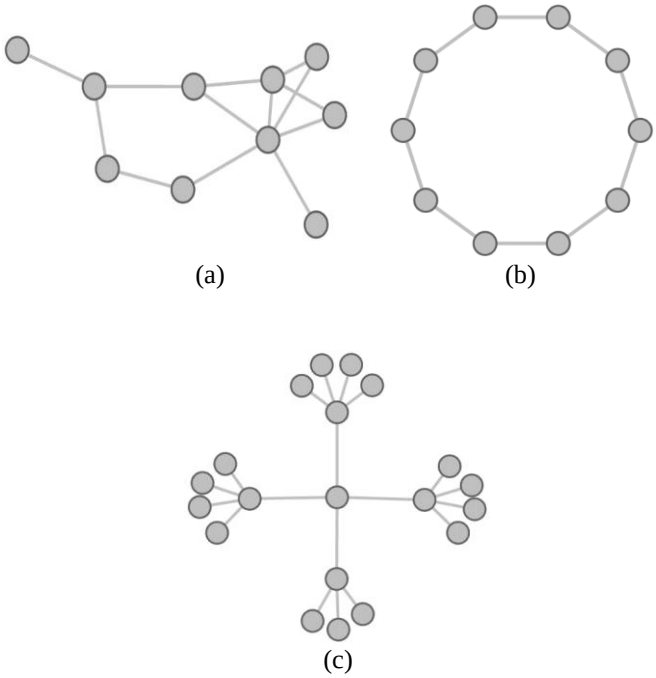


Figure 1. Graphs with different topologies: (a) ER-random, (b) 2-regular and (c) graph topology with scale-free property

Graph Topology: Infection may break in any type of network. It could be a real-world social network like Facebook where a person may start a rumor and with the help of friendship ties it diffuses through it. How a rumor may spread would depend on the type and the number of connections the user has with his/her friends and how, in turn, are they connected with other users. If a user doesn't have many active friends, the rumor may take time to spread

or may not catch at all. Therefore, this diffusion highly is governed by the scale-free property of a social network. Or, for example, in a human society, a human being with a contagious disease may spread it by being in contact with others. The spreading of a disease, therefore, would depend on the structure of the society. Therefore, the topology of a network over which an infection diffuses is imperative to the diffusion process itself. Figure 1 shows graphs with different topologies. In Figure 1(a), there is an ER-random graph topology of 10 nodes, while in Figure 1(b), there is a 2-regular graph topology of 10 nodes. Figure 1(c) shows a graph topology with scale-free property, prevalent in online social networks, like Facebook and Twitter. Node at the center in Figure 1(c) is more likely to easily diffuse content in the network compared to nodes at boundary.

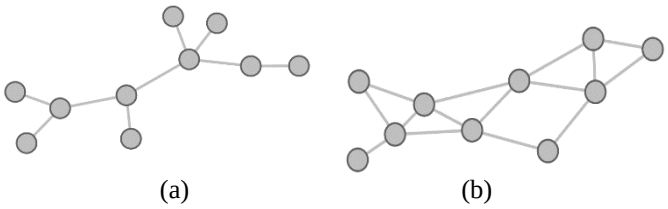


Figure 2. Graphs with different densities: (a) a sparse graph and (b) a dense graph.

Graph Density: Graph density is defined as the ratio of the number of edges in a graph to the total possible edges which it could have. How connected a node is to other nodes in a network may determine its spreading power. If the underlying network is very sparse, the infection would in turn become sparse and might be hard to localize. On the other hand, if the underlying topology is dense, infection would find it easier to spread and if caught on time, might be easier to localize and quarantine. The general rule of an infection spreading process is that there is higher density of infected nodes closer to the source than infected nodes farther away from it [30]. Therefore, if the underlying graph is sparse to begin with, it essentially becomes a hard task to distinguish the source node from a non-source node. Figure 2 shows two graphs with 10 nodes each. Due to the lesser number of edges, the graph in Figure 2(a) is comparatively sparser than the graph in Figure 2(b).

Infection Probability: The infection probability is defined as the intensity of infection diffusion between any two given nodes. It can be understood as the strength of friendship ties between two users in a social network. The stronger this tie, the higher the chances of each other sharing one another's content. Therefore, in any given network where nodes are actively tied with one another, there are higher chances of infection to spread quickly and have a far reach. Figure 3 shows a sample infection graph with node 5 as the source. The infection probability between node 5 and node 7 is 0.1 (10%) which can be considered as low infection probability, meaning it will be hard for infection to spread from node 5 to node 7. On the other hand, the infection probability between node 5 and node 8 is 0.7 (70%), a high infection probability, meaning it will be easier for infection to diffuse from node 5 to node 8.

Infection Size: Infection size is defined in terms of the number of infected nodes in a given infection graph. Generally, the longer the time given to an infection to spread, the larger the infection size, assuming the infection probability stays constant. Furthermore, the size of an infection intuitively depends on graph density factor as well. Again, by referring to Figure 3, it can be seen that, here, the infection size is 6 nodes.

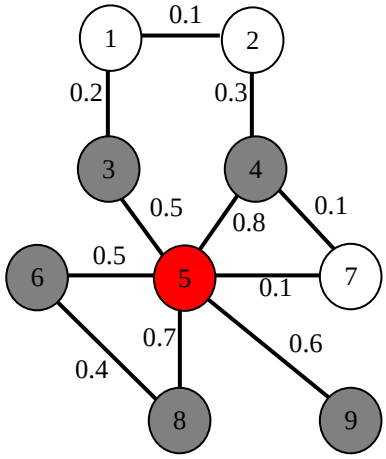


Figure 3. A sample graph of 9 nodes depicting infection of size 6. All the colored nodes are infected and nodes left uncolored are non-infected. Red node (node 5) is the source of infection. The weights on the edges indicate infection probability.

V. INFECTION SOURCE IDENTIFICATION METHODS

This study analyzes the impact of various factors on infection source identification using three well known state-of-the-art techniques: Dynamic Age (DA), Reverse Infection (RI) and Minimum Description Length (MDL) [17,18,20,21].

DA exploits the idea of dynamic age of a node [18]. The dynamic age of any node u in an infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$, where source s^* is unknown, is defined as the absolute difference between the largest eigenvalue of $\mathcal{G}$ and the largest eigenvalue of $\mathcal{G}$ when node u is removed from $\mathcal{G}$. Any node u in $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$ with the highest dynamic age is considered to be the source. Formally,

$$\hat{s}_{DA} = \underset{u \in \mathcal{V}}{argmax} \left(|\lambda_1 - \lambda_1^u|/\lambda_1\right) \quad (1)$$

where λ_1 is largest eigenvalue of the adjacency matrix corresponding to the infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$ and λ_1^u is the largest eigenvalue when u is removed from $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$.

Originally used under SIR model, RI uses reverse infection strategy [17]. Each infected node sends its information to its neighbors and the one receiving the information of all the

infected nodes first is considered to the source. The ties are broken using closeness centrality [38,39]. The source estimated by RI is the Jordan center of the graph. However, instead of SIR, this study uses RI under SI model where the complete information of the infection status of all the nodes is given, as has been previously done in [40]. Therefore, source estimation using RI under SI model could be formally defined as,

$$\hat{s}_{RI} = \underset{u \in V}{\operatorname{argmin}} (eccentricity(u)) \quad (2)$$

MDL uses Laplacian matrix L corresponding to the underlying graph $G(V, E)$ and extracts the sub-matrix L_s from L corresponding to the infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$ [20,21]. Then the node with the largest component of the eigenvector corresponding to the minimum eigenvalue of L_s is considered to be the source. Mathematically,

$$\hat{s}_{MDL} = \underset{u \in V}{argmax} (z_u \in \vec{Z}_{min}) \quad (3)$$

where $\vec{Z}_{min}$ is the eigenvector corresponding to the smallest eigenvalue of L_s and z_u is u^{th} component of this vector.

VI. Research Methodology

In this section, a detailed outlay of the experimentations used in this study is provided. Firstly, a general three-module experimental framework is presented and discussed. Secondly, the theoretical foundation of SI infection model - model with which infection process is simulated in this study - is explained. Thirdly, a brief overview of the datasets used in experiments is provided. Besides this, general experimentation set-up and evaluation measures are discussed in detail. Finally, all the three state-of-the-art source estimation techniques, i.e., DA, MDL and RI, used to study the impact of various factors on source identification, are explored with all the technical aspects thoroughly explained.

A. Experimental Framework

The experimental framework of this study is composed of three modules: (a) Infection Graph Generation Module, (b) Infection Source Estimation Module and (c) Evaluation Module.

Infection Graph Generation Module: In this module, firstly, a node is generated randomly from an input graph $G(V, E)$. This node is considered as the actual source of infection, s^*. Then together with s^*, input graph is fed to SI model, which starts infection spreading process originating from s^* and produces an infection graph, $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$. This infection graph serves as input to Infection Source Estimation Module.

Infection Source Estimation Module: In this module, an infection graph is provided as input to a source identification technique. A source identification technique then produces a node as output which is most likely to be the source of infection. This node is referred to as estimated source, $\hat{s}$ and serves as one of the inputs to the Evaluation Module.

Evaluation Module: This module takes input the infection graph and actual source s* from the Infection Graph Generation Module, and estimated source from the Evaluation Module. With the help of these inputs, this module evaluates the performance of source identification techniques by comparing s^* and $\hat{s}$ in the infection graph.

Figure 4 provides an illustration of all the modules and their individual components with intra-modular and inter-modular connections.

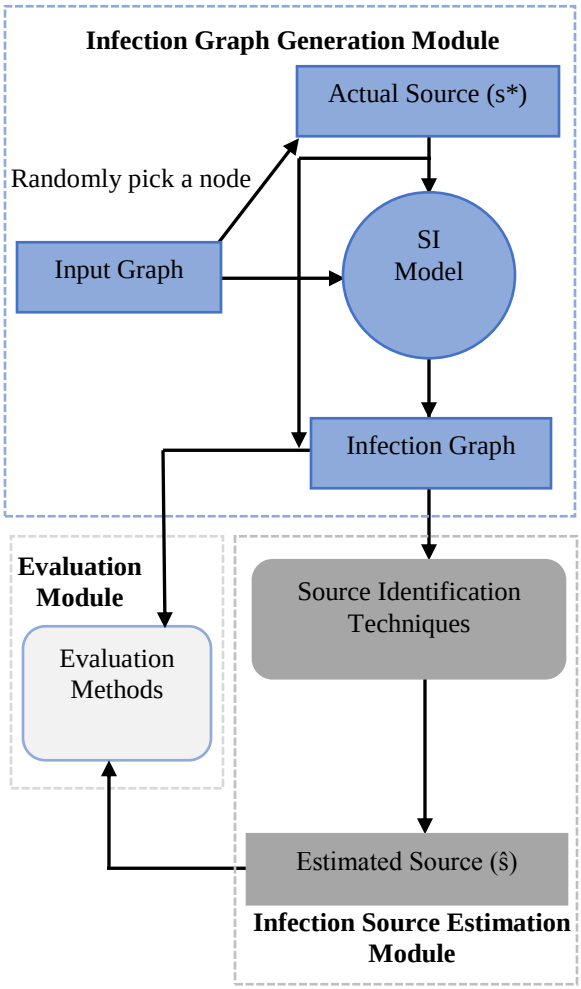


Figure 4. Three-module experimental framework used in this study.

B. Infection Model (SI)

Given a graph $G(V, E)$, where V is the set of nodes and E is the set of edges and an infection source s^*, this study employs the classical SI (*Susceptible-Infected*) model to simulate the process of infection spreading on *G* [3,4]. In his

model, there are two possible states for any node: susceptible (*S*) or Infected (*I*). A node may get infected by either receiving infection from its adjacent infected neighbors or simply by being the source. When it gets infected, it stays infected forever, i.e., it cannot change its state. The non-infected neighbors of an infected node are said to be susceptible to infection. In each time step (discrete in this study), each infected node tries to infect their susceptible neighbors with some probability p, where p depends on the strength of infection. The stronger the infection probability p between two nodes, the easier it is for infection to spread between them. In this study, two types of SI model are used, i.e., homogeneous and heterogeneous, depending on the factor to be analyzed. In homogeneous SI model, the infection probability is kept same across each edge in the graph. While in heterogeneous SI model, the infection probability varies across edges. In this study, when the infection model is heterogeneous SI, the infection propagation probability on each edge is uniformly distributed over (0,1) and it is assumed the infections are independent of each other [15,17,27,41]. If at any given time step *t*, a node *w* is susceptible to infection from any two of its neighbors, *u* and *v*, *w* gets infection with probability $p = 1 - (1 - p_{uw})(1 - q_{vw})$, where p_{uw} and q_{vw} are the infection probabilities (edge weights) between *u* and *w*, and *v* and *w*, respectively. This is illustrated in Figure 5. The infection is stopped when the desired number of nodes are infected (infection size), resulting in an infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$, where $\mathcal{V}$ is the set of infected nodes, $\mathcal{E}$ is the set of edges between them and s^* is the actual source of infection. It is important to note that any source identification technique, used to identify the source in a given infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, s^*)$, does not have any information on the location of s^*.

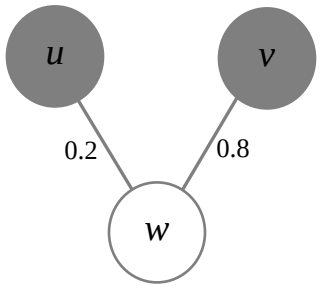


Figure 5. A sample example . Nodes u and v are infected and both are trying to infected w with probabilities 0.2 and 0.8, respectively. Therefore, the overall probability with which u and v both try to infect w is given as p=1-(1-0.2)(1-0.8)=.84.

C. Datasets

For each category of factors, i.e., graph topology, density, infection probability and size, different types of graphs are used. For analyzing the effect of topology, two synthetic networks, i.e., Erdos-Renyi or ER-random graph and 4-regular graph, and two real-world networks, i.e., Facebook and US Power Grid (USPG) [22,23,24,25]. For density, employ four ER-random and four k-regular graphs of four different types of densities are used. To analyze the behavior of source identification with respect to infection probability, one ER-random and Facebook is used and probability on graph edges is varied among {20%, 40%, 60%, 80%}. As for infection size, again ER-random graph and Facebook networks are used and four different infection graph sizes for each are taken. Table 1 summarizes the types of graphs used in studying these factors. The detailed datasets statistics will be provided in subsequent sections.

Table 1. Different types of graphs used in studying various factors.

	Graphs Used			
Factor	**k-regular**	**ER-random**	**Facebook**	**USPG**
Topology	☒	☒	☒	☒
Graph Density	☒	☒	☐	☐
Infection Probability	☐	☒	☒	☐
Infection Size	☐	☒	☒	☐

D. General Experimental Set-up

By employing the above defined SI model, infection spreading process is simulated on a given graph $G(V, E)$, where V is the number of nodes and E is the number of edges. For any given graph G, a source s^* is picked at random, and the infection spreading process, beginning from s^*, is simulated on G. This process is repeated over 100 independent runs while picking s^* randomly in each run. In each run, the infection process is continued until the desired number of nodes are infected. This results in 100 infection graphs, $\mathcal{G}_{i=1}^{100}(\mathcal{V}_i, \mathcal{E}_i, s_i^*)$ for any given graph. Note that in each infection graph, s^* is unknown. All the returned results are then averaged over these 100 runs and reported.

While this is a general experimental set-up and is applicable to all the analyses performed in this study, the type of SI model (homogeneous/heterogeneous) employed or the number of infected nodes (infection size) will depend on the type of factor being analyzed. Therefore, wherever necessary, factor-depended experimental parameters will be provided in the subsections of Section VII.

E. Evaluation Measures

To evaluate the performance of the source identification techniques used in this study (Section V), two traditional evaluation measures, i.e., average hop error and accuracy, are used.

Average Hop Error: Given an estimated source $\hat{s}$ and the actual source s^*, hop error, h, is calculated as the minimum distance d between $\hat{s}$ and s^*. Formally, $h = d(\hat{s}, s^*)$. Since, in this study, 100 infection graphs are generated for any given graph, therefore, to calculate the average hop error (AHE), the average of hop errors, h, between s* and s^, produced over all infection graphs, is taken.

Accuracy: Accuracy is defined as the number of correctly identified sources (i.e. h = 0) over all the infection graphs (100 in this study).

VII. RESULTS AND DISCUSSIONS

In this section, for each factor, factor-specific experiment outlay will be presented and the effects of each on infection source identification will be implored.

A. Impact of Topology

In order to understand the effect of topology on infection source identification, four different graphs are used - two synthetic, i.e., ER-random and 4-regular, and two real-world, i.e., Facebook and US Power Grid (USPG). The reason to pick these four graphs is embedded in varied characteristics of each and how each one of them is different from the other. The complete dataset statistics of each is given in Table 2. While the general experimental set-up is same as given in Section, here heterogeneous SI model is used and infection size is kept constant across all the topologies at 2-5%, i.e., infection spreading process is stopped when the 2-5% nodes of all the nodes are infected in any given topology. Afterwards, source identification techniques, defined in Section V, are employed on the resulting infection graphs. Upon investigating the results, it is found that source identification is highly topology dependent and, therefore, topology plays an important role in identifying the source of infection in a given graph. Figure 6 and Figure 7 show the average hop error and accuracy, respectively, as produced by different source identification techniques on various topologies. The results are discussed topology-wise.

Table 2. Dataset statistics of networks used to study impact of topology and infection size on source identification.

Topology	Nodes	Edges	Avg. deg.	Density
4-Regular	5,000	10,000	4.0	0.001
Random	5,000	24,943	9.98	0.002
Facebook	4,039	88,234	43.69	0.011
USPG	4,941	6,594	2.67	0.0005

1) 4-Regular: Figure 6(a) shows average hop error as produced by various source identification techniques on 4-Regular graph. It is understood that RI performs the best

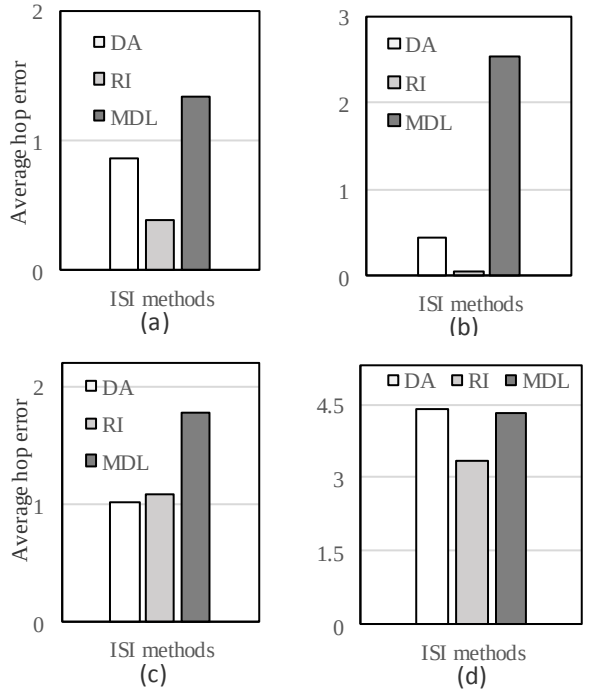


Figure 6: Average hop error (AHE) produced by various infection source identification (ISI) methods across different topologies: (a) 4-regular, (b) ER-random, (c) Facebook and (d) US Power Grid.

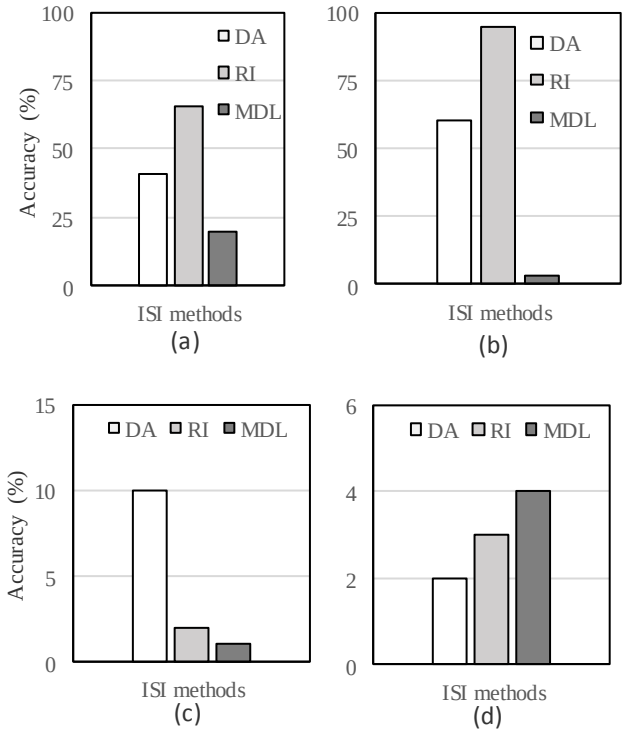


Figure 7: Accuracy produced by various infection source identification (ISI) methods across different topologies: (a) 4-regular, (b) ER-random, (c) Facebook and (d) US Power Grid.

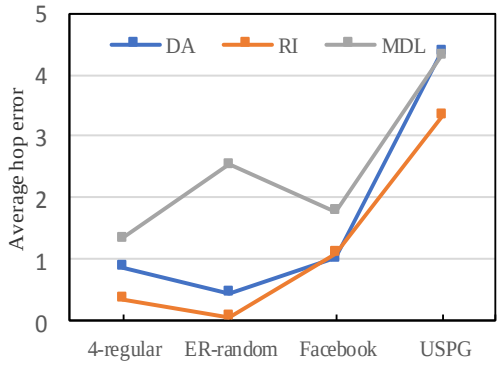
amongst the three methods with average hop error of 0.34, which means it, generally, estimates a source less than half a hop away from the actual source. RI is followed by DA with average hop error of 0.86 and MDL remains the worst performing source estimation method estimating a source 1.34 hops away from the actual. This is further supplemented by the accuracy of finding the source as shown in Figure 7(a). RI, with 66% accuracy, has the best accuracy among the employed methods, followed by DA with 41% and MDL with 20%. Therefore, here it can be concluded that on 4-Regular graph RI performs the best and has a good accuracy and average hop error.

2) ER-Random: As, can be seen from Figure 6(b) and Figure 7(b), both RI and DA work even better on ER-Random graph for both average hop error and accuracy. With the same infection size, RI generally finds a source within 0.05 hops with staggering accuracy of 95%, a clear improvement when compared against 4-Regular graph. DA as well improves its source identification with accuracy of 60% and average hop error of 0.44, again a clear improvement over 4-Regular. While RI and DA both improve their source identification on ER-Random graph, MDL performs worse when compared against its results on 4-Regular graph, with only 3% accuracy (Figure 7(b)) and 2.53 average hop error (Figure 6(b)).

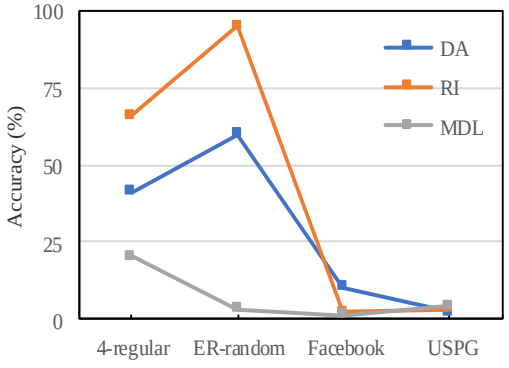
3) Facebook: While it was seen that RI and DA had improved performance on ER-Random graph, both equally take a dip when analyzed under Facebook topology. RI, whose accuracy was 95% in ER-Random graph, comes up with only 2% accuracy on Facebook (Figure 7(c)) and even worse average hop error of 1.09 (Figure 6(c)). DA, for a change, performs better than RI on Facebook, but when compared against its performance on 4-regular and ER-random graphs, it comes off worse with only 10% accuracy (Figure 7(c)) and 1.01 average hop error (Figure 6(c)). MDL continues to be the worst performing source identification method on Facebook as well. However, it has a better average hop error of 1.77 (Figure 6(c)) than was found in ER-Random graph.

4) US Power Grid (USPG): US Power Grid being a very sparse network, has a telling effect on all the three source identification methods. DA and RI, which produced good results on ER-random and 4-regular graphs, come up with meager 2% and 3% accuracy to locate a source as can be observed in Figure 7(d), respectively. Interestingly, MDL turns out to be the best for USPG with 4% accuracy. As for average hop error, RI estimates source closest to the actual source with 3.33 average hop error, followed by MDL and DA with 4.31 and 4.39 average hop error, respectively, as can be observed from Figure 6(d).

To supplement and summarize what is discussed above, Table 3 shows the best and worst performing source identification methods across the four employed topologies on average hop error and accuracy. It can be easily seen while DA performs best on Facebook, its performance isn't as good on other topologies. While MDL has the best accuracy on USPG network, RI, on the same network, produces estimated sources which are closest to the actual source than any other method. Besides this, Figure 8 shows the performance (average hop error in Figure 8(a) and accuracy in Figure 8(b)) of individual methods on different topologies. It is quite clear from this figure that RI performs better on synthetic networks, while its performance takes a dip in real-world networks, while MDL, somewhat, tends to work better on real-world networks. DA works best for Facebook topology, and therefore, should be more applicable to find sources in graphs with similar topology than any other comparing methods.



(a)



(b)

Figure 8. Performance of various source identification techniques on different topologies: (a) average hop error and (b) accuracy.

Therefore, from the discussion above, it can be understood that source identification is highly topology-dependent, with

the easiest source identification taking place in ER-Random graph, followed by 4-regular, Facebook and US Power Grid. RI works the best for ER-Random graph and 4-regular, while DA tends to perform better on Facebook. For a very sparse graph like US Power Grid, the detection becomes very hard, irrespective of the method used. Therefore, from this section it is established that source identification is topology-dependent and different kinds of source identification methods work best for different topologies.

Table 3. Best performing methods on average hop error (AHE) and accuracy (Acc.) across various topologies.

	Topology			
Evaluation Measure	**4-Regular**	**Random**	**Facebook**	**USPG**
AHE	RI	RI	DA	RI
Acc.	RI	RI	DA	MDL

B. Impact of Graph Density

In the above subsection, it is understood that it is very hard to detect a source on US Power Grid which is a very sparse graph. Therefore, to analyze the impact of graph density on infection source identification, four ER-random graphs with different densities are picked, i.e., Very Sparse (VS), Sparse (S), Dense (D) and Very Dense (VD). Besides ER-random, the impact of graph density is also analysed using four regular graphs of different densities: 3-regular, 5-regular, 7-regular and 10-regular. The statistics of these graphs have been presented in Table 4. Again, heterogeneous SI model is used and the infection graph size for each of the four ER-random and k-regular graphs is kept at the minimum of 5%. This, on an average, yields 50 nodes, 100 nodes, 150 nodes and 400 nodes for ER-random very sparse (VS), sparse (S), dense (D) and very dense (VD) graphs, respectively. For regular graphs, average infected nodes produced are 60 nodes, 75 nodes, 85 nodes and 140 nodes for 3-regular, 5-regular, 7-regular and 10-regular, respectively.

Table 4. Dataset statistics of various networks used to study the impact of graph density on source identification.

Topology	Nodes	Edges	Avg. degree	Density
3-Regular	1,000	1,500	3.0	0.003
5-Regular	1,000	2,500	5.0	0.005
7-Regular	1,000	3,500	7.0	0.007
10-Regular	1,000	5,000	10.0	0.01
Random (VS)	1,000	2,526	5.05	0.005
Random (S)	1,000	3,570	7.14	0.007
Random (D)	1,000	10,050	20.1	0.02
Random (VD)	1,000	25,037	50.1	0.05

The results indicate that graph density plays an important role in the performance of source identification methods. It is seen that, irrespective of the technique used, as the graph density is increased, source identification becomes easier. That is to say that source identification is easier on denser graphs than on comparatively lesser dense graphs. Figure 9 shows average hop error (Figure 9(a)) and accuracy (Figure 9(b)) of various source identification methods on ER-random graphs of different densities. As can be observed from this Figure 9(a), the average hop error produced by all the techniques is larger on sparse infection graphs as compared to the dense graphs. Average hop error is worst in ER-random very sparse (VS) graph and the best in very dense (VD). It is important to note that this observation holds true even when there are more infected nodes in denser graphs than in comparatively sparser graphs. This finding is further supplemented by noticing the accuracy of source identification techniques in Figure 9(b). Both MDL and RI show better performance as the graph density increases, incredibly, irrespective of the infection size. This shows the stability of these methods. While DA improves as well with the increase in graph density, however, it experiences comparatively lower performance in very dense (VD) graph as the number of nodes increase. Having said that, DA still performs better on very dense (VD) ER-random graph when compared against its performance on sparse (S) and very sparse (VS) graphs. As was seen in the precious subsection, RI continues to outperform every other state-of-the-art source identification technique on ER-random graph, finding

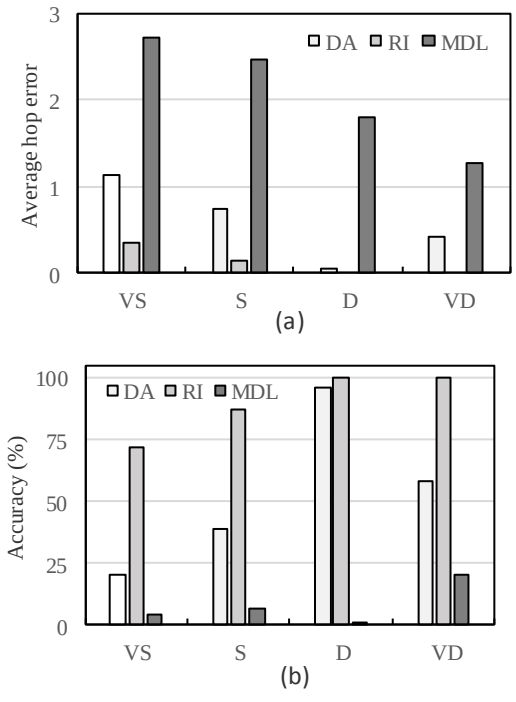


Figure 9. Performance evaluation of different infection source identification (ISI) methods, DA, RI and MDL, on ER-random graph having different densities, i.e., Very Sparse (VS), Sparse (S),

Dense (D) and Very Dense (VD) in terms of (a) average hop error and (b) accuracy.

the source with 100% accuracy on dense (D) and very dense (VD) graphs, irrespective of the infection size or graph density.

A similar trend can be observed on four k-regular with different densities. All the three employed techniques, i.e., DA, RI and MDL, find it relatively easier locate a source under 10-regular, a very dense k-regular graph, than under 3-regular, a very sparse k-regular graph. Figure 10 shows average hop error (Figure 10(a)) and accuracy (Figure 10(b)) of various source identification methods on k-regular graphs of different densities. As can be seen from this figure, both the average hop error and detecting accuracy of all the three methods are worst on 3-regular graph, but all three show continuous improvement on higher density k-regular graphs, i.e., 5, 7 and 10-regular. Besides this, as was seen in previous subsection, RI continues to be the best source identification method on k-regular graphs, followed by DA and MDL.

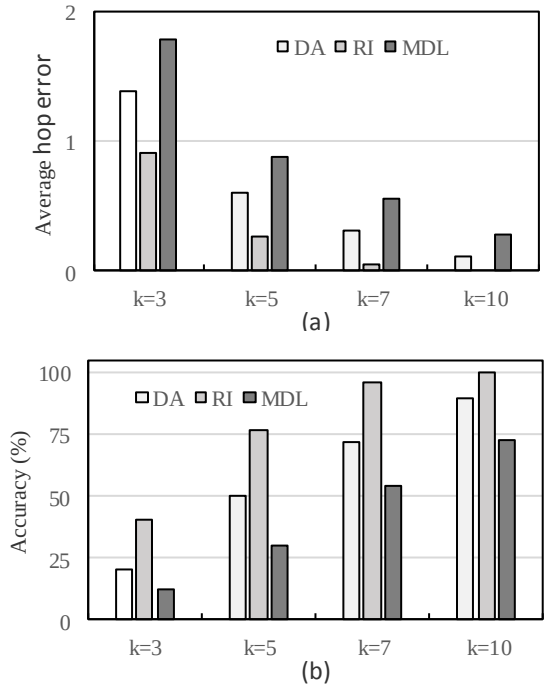


Figure 10. Performance evaluation of different infection source identification (ISI) methods, DA, RI and MDL, on k-regular graph with k=3,5,7 and 10 in terms of (a) average hop error and (b) accuracy.

C. Impact of Infection Probability

Since, how an infection would spread across a given network would depend on the infection probability between any two given nodes, the third factor analyzed in context to infection source identification is the infection probability. For the same, Facebook graph (used in Section VII(A)) an ER-random graph of 500 nodes and 1301 edges are used. Table 5 summarizes the statistics of the ER-random graph used in this part of study. The statistics of Facebook network can be found in Table 2. For this analysis, however, instead of heterogeneous SI model, homogeneous SI model, defined in Section VI(B), is employed. Furthermore, four different infection probabilities, 20%, 40%, 60% and 80% are used to simulate infection diffusion. Therefore, 100 infection graphs are generated for any given a graph (e.g., ER-Random) and infection probability (e.g., 40%), i.e., 4 sets of 100 infection graphs for each network. The infection size is kept between 100 and 150 nodes for ER-random graph and between 100 and 200 nodes for Facebook (due to higher density), and the performance of the source identification methods is analyzed.

Table 5. Dataset statistics of ER-random graph used to study the impact of infection probability on source identification.

Topology	Nodes	Edges	Avg. deg.	Density
ER-random	500	1,301	5.20	0.01

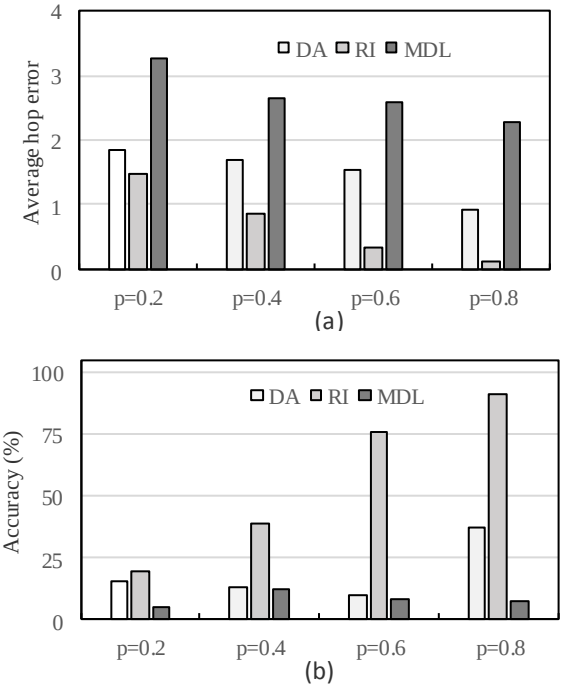


Figure 11. Performance evaluation of different infection source identification (ISI) methods, DA, RI and MDL, on ER-random graph with different infection probabilities, p=0.2 (20%), 0.4 (40%), 0.6 (60%) and 0.8 (80%) in terms of (a) average hop error and (b) accuracy.

The results indicate that source identification is highly dependent on the infection probability of the underlying graph. Figure 11 shows the average hop error (Figure 11(a)) and accuracy (Figure 11(b)) as produced by different source

identification techniques with different infection probabilities on ER-random graph. Each of the methods, i.e., DA, RI and MDL, perform the worst when the infection probability is the least, i.e., 20%. However, when the infection probability is increased, the performance of each consistently gets better. For example, from Figure 11(a) it can been seen that at 20% infection probability, the average hop error produced by RI is 1.47. However, at 80% infection probability, the average hop error becomes 0.12, a drastic improvement. Accuracy (Figure 11(b)), generally, as well tends to improve for both RI and DA. MDL performs the worst as far as the accuracy is concerned and doesn't show much improvement there as the infection probability increases.

A similar pattern could again be observed on Facebook. Figure 12 shows the average hop error (Figure 12(a)) and accuracy (Figure 12(b)) as produced by different source identification techniques with different infection probabilities on Facebook graph. As far as average hop error is concerned, as indicated in Figure 12(a), DA and RI perform almost in equal terms, while MDL continues to perform worse than the former two, as shown in fig. However, the improvement in all the three methods as the infection probability increases can be observed. Barring MDL, whose accuracy is almost negligible, both DA and RI improve their performance of accurately finding the infection source as the infection probability increases from 20% to 80% as shown in Figure 12(b).

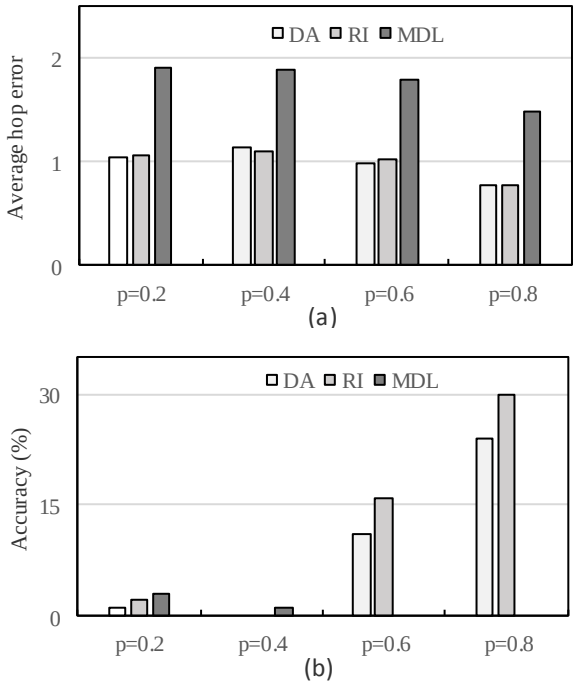


Figure 12. Performance evaluation of different infection source identification (ISI) methods, DA, RI and MDL, on Facebook graph with different infection probabilities, p=0.2 (20%), 0.4 (40%), 0.6 (60%) and 0.8 (80%) in terms of (a) average hop error and (b) accuracy.

From the above discussion, it can be understood that high infection probability tends to make it easier to locate a source, irrespective of the method used. The reason for the same comes down to the density of the infection graph. When the infection probability is low, it is intuitive to think that it gets hard to spread an infection. This very fact makes tends to generate infection graphs which very sparsely connected, i.e., the diameter of such graphs tends to be relatively larger, consequently, producing graphs with longer average path distances. This is similar to spreading infection on a graph like US Power Grid where it was hard to locate an infection source as was seen in Section. On the other hand, when the infection probability is high, it gets easier to spread infection, thereby, the chances of denser infection graphs increase. Therefore, in a nutshell, it can be argued that low infection probability translates to generating sparse infection graphs and high infection probability translates to generating dense infection graphs. And, as was seen in Section, it is easier to locate a source on dense graphs than sparse ones.

D. Impact of Infection Size

The fourth and the last factor analyzed in this study in context to infection source identification is infection size. For the same, ER-random and Facebook graphs are used, as were used in Section VII(A), whose statistics can be found in Table 2. Besides the general experimental set-up (Section), the model of infection used is heterogeneous SI. For each network, again four sets of 100 infection graphs of four different sizes 2-5%, 20-25%, 40-60% and ≥80% are generated. Figure 13 and Figure 14 show the performance of various source identification techniques on ER-random and Facebook.

As can be seen from Figure 13 and Figure 14, source identification is greatly dependent on infection size, irrespective of the topology. When the infection size is smaller, all the three techniques tend to find sources at relatively closer distances. However, as the infection size increases, it becomes hard to detect a source as average hop error drastically increases and accuracy decreases. For example, on ER-random graph, RI, the best performing source identification method on ER-random graphs (Section VII(A)), estimates sources 0.05 hops away from the actual source on average when the infection size is 2-5% as can be observed from Figure 13(a). However, when the infection size becomes ≥ 80% of the original graph, its performance gets reduced to a whopping 3.73 average hop error, which is even worse than DA (3.57). Accuracy of finding a source, as well, takes a dip, as can be seen in Figure 13(b). At 2-5% infection size, RI has an accuracy of 95%, but as the infection size grows, the accuracy of RI comes down to 0% when the infection size is ≥80%. This reduction in

performance, as the infection size increases, is reported by MDL as well as DA. On Facebook (Figure 14), a similar pattern can be seen, further testifying the fact that increasing infection sizes corresponds to worse source detection, regardless of the topology or the method employed.

From the discussion above, it can be concluded that if infection in a graph is somewhat localized, i.e., considerably smaller in size than the underlying graph, it is easier to distinguish a source node from non-source. As infection spreads and engulfs most of the underlying graph, the detection becomes almost impossible.

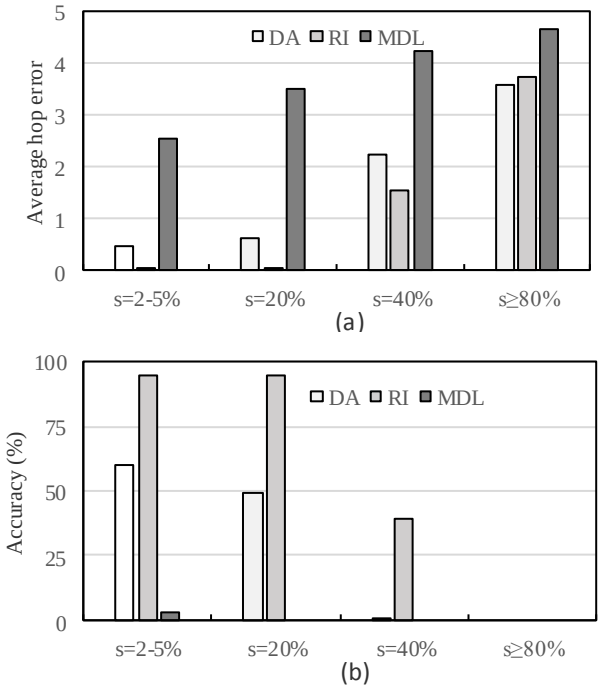


Figure 13. Performance evaluation of different infection source identification (ISI) methods, DA, RI and MDL, on ER-random graph with different infection sizes, s=2-5%, 20%, 40% and ≥80% in terms of (a) average hop error and (b) accuracy.

VIII. CONCLUSION AND FUTURE WORK

This paper aimed to analyse and understand the impact of various graph factors, i.e., graph topology, density, infection probability and infection size, on infection source identification, which, in turn, would prove helpful to researchers choose or develop proper techniques under different prevailing conditions. For the same, various networks were used and for analyses different types of source identification techniques were employed. Therefore, besides analyzing the impact the various graph factors, comparison of the performance of these methods in various scenarios were performed. The results showed that infection source identification is topology-dependent with different methods producing topology-specific results. Furthermore, it was observed that RI produced best results on ER-random graph and DA on Facebook. This analysis further went onto indicate that it is hard to detect infection sources on US Power Grid, a sparse graph with a very large diameter. Afterwards, analysis of the effect of graph density on source identification indicated that it is easier to find sources of infection if the underlying graph is dense as compared to sparse. Then, the third factor, i.e., infection probability, examined under homogeneous SI model showed that higher infection probability leads to easier source detection. Furthermore, analysing the impact of infection size demonstrated that when the infection size is smaller, the source identification becomes easier. However, it is extremely hard to find infection sources when infection covers most of the underlying graph. In future, researchers may take this work further and provide a theoretical explanation of why such an impact of the analysed factors is observed. Furthermore, based on the evidences, researchers maybe able to develop a general framework which, upon inspecting the type of graph under consideration, may be able to identify the technique best suited to identify an infection source in that graph.

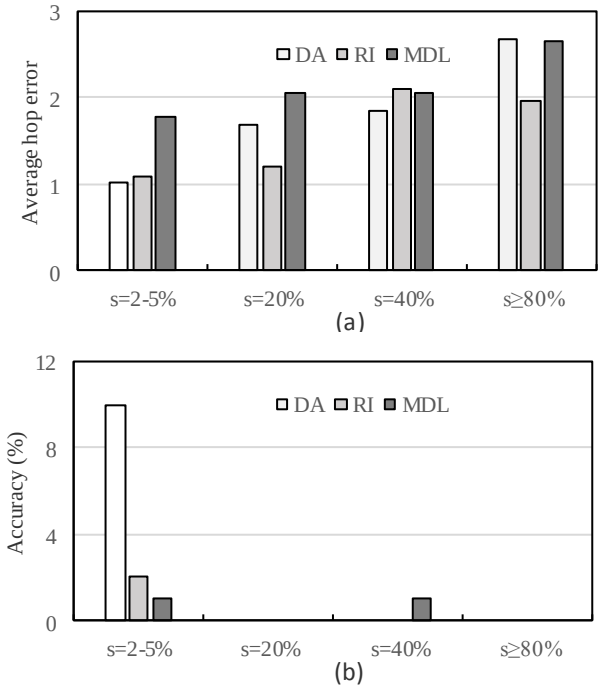


Figure 14: Performance evaluation of different ISI methods DA, RI and MDL on Facebook graph with different infection size s=2-5%, 20%, 40% and ≥80% in terms of (a) average hop error and (b) accuracy.

REFERENCES

[1] G. S. Campos, A. C. Bandeira, S. I. Sardi, "*Zika virus outbreak, Bahia, Brazil*", Emerging Infectious Diseases, Vol. **21**, No. **10**, pp. **1885**, **2015**.

[2] WHO Ebola Response Team, "*Ebola virus disease in West Africa—The first 9 months of the epidemic and forward projections*", New England J. Med., Vol. **371**, No. **16**, pp. **1481–1495**, **2014**.

[3] H. Allcott, M. Gentzkow, "*Social media and fake news in the 2016 election*", Journal of Economic Perspectives, Vol. **31**, No. **2**, pp. **211-36**, **2017**.

[4] B. Doerr, M. Fouz, T. Friedrich, "*Why rumors spread so quickly in social networks*", Commun. ACM, Vol. **55**, No. **6**, pp. **70–75**, **Jun. 2012**.

[5] C. Pash, "*The lure of naked hollywood star photos sent the Internet into meltdown in New Zealand*", Business Insider Australia, Vol. **4**, **Sep. 2014**.

[6] D. MacRae, "*5 viruses to be on the alert for in 2014*", Computer Business Review, Tech. Rep., **Feb. 2014**.

[7] A. Singla, K. Jain, A. Gairola, "*Delving into Security of Networks - Times Need*", International Journal of Scientific Research in Network Security and Communication, Vol. **2**, Issue. **3**, pp. **1-8**, **2014.**

[8] U.K. Singh, C. Joshi, S.K. Singh, "*Zero day Attacks Defense Technique for Protecting System against Unknown Vulnerabilities*", International Journal of Scientific Research in Computer Science and Engineering, Vol. **5**, Issue. **1**, pp. **13-18**, **2017.**

[9] D. Shah, T. Zaman, "*Detecting sources of computer viruses in networks: Theory and experiment*", In the Proc. ACM SIGMETRICS Int. Conf. Measur. Model. Comput. Syst. (SIGMETRICS), New York, NY, **USA**, pp. **203–214**, **Dec. 2010**.

[10] L. J. Allen, "*Some discrete-time si, sir, and sis epidemic models*", Mathematical Biosciences, Vol. **124**, Issue. **1**, pp. **83–105**, **1994**.

[11] R. M. Anderson, R. M. May, B. Anderson, "*Infectious diseases of humans: dynamics and control*", Vol. **28**, Wiley Online Library, **1992**.

[12] N. Karamchandani, M. Franceschetti, "*Rumor source detection under probabilistic sampling*", In the Proc. IEEE Int. Symp. Inf. Theory (ISIT), Istanbul, **Turkey**, pp. **2184–2188**, **2013**.

[13] W. Luo, W. P. Tay, M. Leng, "*Identifying infection sources and regions in large networks*", IEEE Trans. Signal Process., Vol. **61**, No. **11**, pp. **2850–2865**, Jun. **2013**.

[14] D. T. Nguyen, N. P. Nguyen, M. T. Thai, "*Sources of misinformation in online social networks: Who to suspect?*", In the Proc. of 2012 IEEE Milit. Commun. Conf. (MILCOM), Orlando, FL, **USA**, pp. **1–6**, **2012**.

[15] W. Luo, W. P. Tay, M. Leng, "*How to identify an infection source with limited observations*", IEEE J. Sel. Topics Signal Process., Vol. **8**, No. **4**, pp. **586–597**, Aug. **2014**.

[16] W. Luo, W. P. Tay, "*Identifying infection sources in large tree networks*", In the Proc. of 9th Annu. 2012 IEEE Commun. Soc. Conf. Sensor Mesh Ad Hoc Commun. Netw. (SECON), Seoul, **South Korea**, pp. **281–289**, **2012**.

[17] K. Zhu, L. Ying, "*Information source detection in the SIR model: A sample-path-based approach*", IEEE/ACM Transactions on Networking (TON), Vol. **24**, Issue **1**, pp. **408-421**, **2016**.

[18] V. Fioriti, M. Chinnici, J. Palomo, "*Predicting the sources of an outbreak with a spectral technique*", Appl. Math. Sci., Vol. **8**, No. **135**, pp. **6775–6782**, **2014**.

[19] W. Luo, W. P. Tay, "*Identifying multiple infection sources in a network*", In the Proc. of 2012 Conf. Rec. 46th Asilomar Conf. Signals Syst. Comput. (ASILOMAR), Pacific Grove, CA, **USA**, pp. **1483–1489**, **2012**.

[20] B. A. Prakash, J. Vreeken, C. Faloutsos, "*Spotting culprits in epidemics: How many and which ones?*", In the Proc. of 2012 IEEE 12th Int. Conf. Data Min. (ICDM), Brussels, **Belgium**, pp. **11–20**, **2012**.

[21] B. A. Prakash, J. Vreeken, C. Faloutsos, "*Efficiently spotting the starting points of an epidemic in a large graph*", Knowl. Inf. Syst., Vol. **38**, No. **1**, pp. **35–59**, **2014**.

[22] J. Leskovec, J. J. Mcauley, "*Learning to discover social circles in ego networks*, In Advances in Neural Information Processing Systems, Curran Associates, Inc., pp. **539–547**, **2012**.

[23] D. J. Watts, S. H. Strogatz, "*Collective dynamics of 'small-world' networks*", Nature, Vol. **393**, No. **6684**, pp. **440–442**, **1998**.

[24] P. Erdos, A. Rényi, "*On random graphs I*", Publ. Math. Debrecen, 6, pp. **290-297**, **1959**.

[25] P. Erdos, A. Rényi, "*On the evolution of random graphs***"**, Publ. Math. Inst. Hung. Acad. Sci, Vol. **5**, No. **1**, pp. **17-60**, **1960**.

[26] J. G. Restrepo, E. Ott, and B. R. Hunt, "*Characterizing the dynamical importance of network nodes and links*," Phys. Rev. Lett., Vol. **97**, No. **9**, Art. No. **094102**, **Sep. 2006**.

[27] J. Jiang, S. Wen, S. Yu, Y. Xiang, W. Zhou, "*K-center: An approach on the multi-source identification of information diffusion*", IEEE Transactions on Information Forensics and Security, Vol. **10**, Issue. **12**, pp. **2616-2626**, **2015**.

[28] K. Yang, A. H. Shekhar, D. Oliver, and S. Shekhar, "*Capacity constrained network-Voronoi diagram: A summary of results*", In Advances in Spatial and Temporal Databases, Berlin, **Germany**: Springer-Verlag, pp. **56–73**, **2013**.

[29] D. Brockmann and D. Helbing, "*The hidden geometry of complex, network-driven contagion phenomena*", Science, Vol. **342**, No. **6164**, pp. **1337–1342**, **2013**.

[30] Z. Wang, C. Wang, J. Pei, X. Ye, "*Multiple Source Detection without Knowing the Underlying Propagation Model*", In the Proceedings of the Thirty-First 2017 AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, California, **USA**, pp. **217-223**, **2017**.

[31] K. Zhu, Z. Chen, L. Ying, "*Catch 'Em All: Locating Multiple Diffusion Sources in Networks with Partial Observations*", In the Proceedings of the Thirty-First 2017 AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, California, **USA**, pp. **1676-1683**, **2017**.

[32] P. C. Pinto, P. Thiran, and M. Vetterli, "*Locating the source of diffusion in large-scale networks*", Phys. Rev. Lett., Vol. **109**, No. **6**, Art. No. **068702, Aug. 2012**.

[33] A. Louni, K. P. Subbalakshmi, "*Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks With Probabilistically Varying Internode Relationship Strengths*", IEEE Transactions on Computational Social Systems, Vol. **5**, Issue. **2**, pp. **335-343**, **2018**.

[34] A. Agaskar and Y. M. Lu, "*A fast Monte Carlo algorithm for source localization on graphs*", In the Proc. of 2013 SPIE Opt. Eng. Appl., San Diego, CA, **USA**, Art. No. **88581N**, **2013**.

[35] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, "*Bayesian inference of epidemics on networks via belief propagation*", Phys. Rev. Lett., Vol. **112**, No. **11**, Art. No. **118701, 2014**.

[36] Y. Xie, V. Sekar, D. A. Maltz, M. K. Reiter, and H. Zhang, "*Worm origin identification using random moonwalks*", in Proc. IEEE Symp. Security Privacy, Oakland, CA, **USA**, pp. **242–256**, **2005**.

[37] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "*Inferring the origin of an epidemic with a dynamic message-passing algorithm*", Phys. Rev. E, Vol. **90**, No. **1**, Art. No. **012801**, **2013**.

[38] L. C. Freeman, "*Centrality in social networks conceptual clarification*", Soc. Netw., Vol. **1**, No. **3**, pp. **215–239**, **1978**.

[39] M. E. J. Newman, "*Epidemics on networks, in Networks: An Introduction*", Oxford, **U.K.**: Oxford Univ. Press, Ch. **17**, pp. **700–750**, **2010**.

[40] B. Chang, E. Chen, F. Zhu, Q. Liu, T. Xu, Z. Wang, "*Maximum a Posteriori Estimation for Information Source Detection*", IEEE Trans. on Systems, Man, and Cybernetics: Systems (Early Access), pp. **1-15**, **May**, **2018**.

[41] Z. Chen, K. Zhu, and L. Ying, "*Detecting multiple information sources in networks under the SIR model*," in *Proc. 48th Annu. Conf. Inf. Sci.Syst. (CISS)*, Mar. 2014, pp. 1–4.

Authors Profile

Syed Shafat Ali completed his bachelor's and master's degrees in computer science from University of Kashmir, Jammu and Kashmir, India, in the years 2011 and 2014, respectively. From year 2014 to 2015, he worked as a software engineer in Rooman Technologies, Bangalore, India. Following this, he joined Jamia Millia Islamia (A Central University) to pursue PhD after qualifying the entrance examination while securing the 2nd rank. He also qualified the State Eligibility Test (SET) in 2018 for the eligibility for assistant professor post. His research interests include theoretical computer science, graph theory, data mining, machine learning and analysis of online social networks. Presently, he is studying infection source identification in complex graphs.

With over 33 years of research and academic experience in the institutions in India, Australia, the UAE and the USA, presently, Dr. S.A.M. Rizvi is a Professor in the Department of Computer Science at Jamia Millia Islamia (A Central University), New Delhi, where he, recently, completed his second term in office as Head of the Department. He has more than 170 research publications and has authored 6 text books as well. So far in his academic career, he has supervised 19 PhDs in different areas of computer science. Besides various academic positions held by him and being the Founder-Director of many institutions, he has also served as Chief Manager (IT) in Goa Shipyard under the Ministry of Defense, Government of India. In addition to this, he has served as a Consultant-Programme Director (MIS) at Abu Dhabi University in UAE and Jawaharlal Institute of Technology (JIT) in India. Prof. Rizvi has a rich experience of academic management, evaluation/accreditation process and administration through the roles such as founder-director/dean/head of institutions/departments in various institutions.

.

