EPA: Exoneration and Prominence based Age for Infection Source Identification

Syed Shafat Ali[†], Tarique Anwar[‡◇], Ajay Rastogi[†], Syed Afzal Murtaza Rizvi[†]
[†]Jamia Millia Islamia, New Delhi, India
shafat159074@st.jmi.ac.in,ajay148115@st.jmi.ac.in,sarizvi@jmi.ac.in
[‡]Indian Institute of Technology Ropar, Punjab, India
[◇]Macquarie University, Sydney, Australia
tarique@iitrpr.ac.in

ABSTRACT

Infection source identification is a well-established problem, having gained a substantial scale of research attention over the years. In this paper, we study the problem by exploiting the idea of the source being the oldest node. For the same, we propose a novel algorithm called Exoneration and Prominence based Age (EPA), which calculates the age of an infected node by considering its prominence in terms of its both infected and non-infected neighbors. These non-infected neighbors hold the key in exonerating an infected node from being the infection source. We also propose a computationally inexpensive variant of EPA, called EPA-LW. Extensive experiments are performed on seven datasets, including 5 real-world and 2 synthetic, of different topologies and varying sizes to demonstrate the effectiveness of the proposed algorithms. We consistently outperform the state-of-the-art single source identification methods in terms of average error distance. To the best of our knowledge, this is the largest scale performance evaluation of the considered problem till date. We also extend EPA to identify multiple sources by developing two new algorithms - one based on K-Means, called EPA_K-Means, and another based on successive identification of sources, called EPA_SSI. Our results show that both EPA_K-Means and EPA_SSI outperform the other multi-source heuristic approaches.

KEYWORDS

Infection source identification, Rumor detection, Exoneration and Prominence, Complex networks, Information diffusion

ACM Reference Format:
Syed Shafat Ali, Tarique Anwar, Ajay Rastogi, Syed Afzal Murtaza Rizvi. 2019. EPA: Exoneration and Prominence based Age for Infection Source Identification. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3358035

1 INTRODUCTION

Infection source identification deals with the localization or detection of the source of a diffusion or infection process in a given network. An infection could be a rumor or false news spreading through online social networks, a virus propagating on computer networks, infectious diseases spreading in the networks of human societies, or a harmful high-voltage power surge in a power grid network. Due to its wide range of applications and the significance of mitigating the damages, the infection source identification problem has attracted researchers from varied domains.

Researchers, over the years, have studied the infection/information source identification problem, beginning with the work of Shah and Zaman [18]. They studied the problem on tree-like networks and assumed the information diffusion follows *Susceptible-Infected* (SI) model. Following that, [7, 12] also approached the problem in the same vein under the similar assumptions. Later on, the problem was expanded and explored under more difficult and real-world-prevalent assumptions, i.e., the underlying models of diffusion are *Susceptible-Infected-Recovered* (SIR) and *Susceptible-Infected-Susceptible* (SIS) [1, 13, 23]. In reality, however, trees-like networks are uncommon. Therefore, researchers eased the constraints of tree-like networks to identify infection sources in general graphs [4, 6, 11, 15, 19, 22].

Generally, the existing works only consider infected nodes to locate the source. For example, Dynamic Age (DA) [4] and Reverse Infection (RI) [23] (implemented under SI model [2]) utilize the information provided by infected nodes while discarding their non-infected neighbors. However, these non-infected neighbors hold the key to exonerate an infected node from being the source of infection (Example 1). While there are some studies (Minimum Description Length (MDL) [15]) which do consider the effect of non-infected nodes, yet their performance is not any better. Furthermore, RI only considers one of the Jordan centers of an infection graph to be the source. However, since every node has equal probability to be the source, there are chances that the actual source may not be a Jordan center. Furthermore, some studies make use of infection probabilities between two nodes [6] and others assume that the time of infection is given [10]. However, in real-world, both infection probability and time of infection are extremely hard to capture.

Example 1: Consider a hypothetical network of friends given in Figure 1, where red nodes indicate people having information pertaining (infected) to the rumor and blue nodes, those unaware

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11...$15.00
https://doi.org/10.1145/3357384.3358035

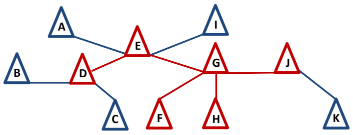


Figure 1: A hypothetical rumor-infected friendship network

of it[1] (non-infected). Links between any two nodes indicate a direct friendship tie. The aim is to identify the person amongst the red nodes that spread the rumor in the given network. For a person with rumor, it would be logical to investigate its neighborhood (friends). For example, if person D is assumed to be source of rumor, checking the state of its friends, we see that B and C are unaware of the rumor. Therefore, it is quite unlikely that D spread the rumor, because, intuitively, if D were the source, its friends should be privy to it. This shows the importance of non-infected neighbors of an infected node. However, if we look at person G, all of its friends are aware of the rumor, thereby, making G highly likely to be the source of rumor. ■

In this paper, we study the infection source identification problem under heterogeneous SI model of infection. We exploit the idea of the age of a node - the older the node, the longer time it gets to infect its neighbors, and therefore, the higher its chances of being the source. Since, this study assumes that the time of diffusion is unknown, we determine the age of an infected node by considering its prominence in terms of its both infected and non-infected neighbors. For a node to be the source, it should have the following three properties. 较小的免责效应 未感染邻居比例小、 *i*) *Lesser exoneration effect*: If a node is the source, then it should have lesser proportion of non-infected neighbors as 较高的节点重要性 感染邻居比例大 compared to any other node. *ii*) *Higher node prominence*: If a node is the source, it should have higher proportion of infected neighbors 较高的本地重要性 as compared to any other node. *iii*) *Higher Local Prominence*: Due 距源较近的节点会有更多的时间去感染邻居，表现出更高感染程度 to the fact that a source node receives maximum time to infect its neighbors, nodes at closer distances to the source will also receive sufficient amount of time to infect their neighbors, in turn, thus, exhibiting higher degree of node prominence as well. Therefore, if a node is to be considered the source, then it should have higher local prominence than that of any other node. The local prominence, in turn, determines the age of a node. These properties are discussed in detail in Section 4.

Based on the same idea, we propose a novel algorithm called Exoneration and Prominence based Age (EPA) for infection source identification in general graphs. To determine the age of a node u, EPA, essentially, calculates the local prominence of u by taking the sum of the prominence of all infected nodes within $r - 1$ distance (argued in Section 4.1) from u. Finally, the node with highest age is considered to be the source. In addition to EPA, we propose EPA-LW, a lightweight variant of EPA, in which, instead of calculating the age of every node in the graph, we only consider the nodes with minimum eccentricity to be the candidates for source. The rationale behind this is discussed in Section 4.2. It is worth noting that both EPA and EPA-LW take the effect of non-infected nodes, carrying essential signatures for source detection, into account, which is generally seen to be lacking in the existing literature, as discussed above. In addition to this, besides infection graph, we only consider the underlying graph and assume infection probabilities [6] and infection time [10] are unknown. We compare EPA and EPA-LW against three well-known state-of-the-art techniques: Dynamic Age (DA) [4], Minimum Description Length (MDL) [15] and Reverse Infection (RI) [23] (implemented under SI model [2]) on seven networks, two synthetic and five real-world. The results show that both EPA and EPA-LW outperform all the three compared methods. Primarily being the single source detection algorithm, we show that EPA can also be extended and used to find multiple infection sources. To this end, we propose two multi-source identification algorithms - one based on K-Means, EPA_K-Means, and the other based on successive identification of sources, EPA_SSI. We evaluate the performance of these two methods against two heuristic algorithms based on K-Means, i.e., Distance Centrality (DC) and Closeness Centrality (CC), previously used in [14, 22]. Our results indicate the superiority of EPA_K-Means and EPA_SSI over DC and CC in finding multiple sources of infection. The main contributions of this paper are summarised below.

(1) We propose EPA, a novel algorithm to identify the infection source in general graphs by exploiting the concepts of exoneration effect and local prominence. In addition, a lightweight variant of EPA, called EPA-LW, is also proposed, having substantially lesser computational cost.
(2) We conduct a large-scale performance evaluation on seven datasets, including 5 real-world and 2 synthetic, of different topologies and varying sizes to demonstrate the efficacy of both the proposed algorithms. To the best of our knowledge, this is the largest scale performance evaluation of the considered problem till date.
(3) We further extend EPA to multi-source infection scenario and provide two algorithms - EPA_K-Means and EPA_SSI. Both the proposed multi-source identification methods outperform the Distance and Closeness heuristics.

The rest of the paper is organized as follows. Section 2 provides a brief literature on infection source identification. In Section 3, preliminaries pertaining to this paper are discussed. The theoretical foundation of the proposed approaches together with their algorithms are presented in Section 4. Section 5 covers all the single source and multi-source experimental evaluation and relevant results. Finally, the paper is concluded in Section 6 with a summary of our findings.

2 RELATED WORK

Given the importance of the problem, infection source identification has attracted a significant attention and has been extensively studied over the past decade. In the early years, the problem was studied in its most ideal form, i.e., under SI model on tree-like networks, and began with the seminal work of Shah and Zaman [18]. They introduced the concept of rumor centrality of a node u in the network, defined as the number of unique diffusion paths originating from u. The node with the highest rumor centrality is called the rumor center and is considered as the source. Later, they extended rumor centrality to find source in general graphs. Other works [7, 12] followed and worked under the same assumptions.

Later, researchers tackled this problem under more difficult assumptions, i.e., the underlying models of infection are SIR and SIS

[1] Some of the figures in this paper are best viewed in color.

[1]. [23] proposed a sample path-based approach to identify infection source in tree graphs under SIR model and provided reverse infection (RI) algorithm to find the source in general graphs which is also its Jordan center. In RI, messages, containing the node IDs, are passed between nodes and the node which gets messages from all the infected nodes first is considered the source. [12, 13] examined the sample path-based approach under SI and SIS models.

Moving away from trees, [4, 11, 15] relaxed the constraints of tree networks to find infection source on general graphs. [4] introduced the concept of dynamical age DA of a node inspired by its dynamical significance [17]. DA computes the amount of reduction in the largest eigenvalue of adjacent matrix, corresponding to a graph, after a node is removed. The larger this reduction the higher the chances of a node to be the source. Besides finding the single source of infection, DA can also be used to find multiple sources by considering top m ranked nodes as m sources. [15, 16] introduced the concept of minimum description length (MDL) for source identification. They first compute the Laplacian matrix L corresponding to the infection graph and find the eigenvector corresponding to the smallest eigenvalue of L. The node with highest score in this eigenvector is considered to be the source. Akin to DA, MDL can also be used to find multiple sources. Jiang, et al. [6] introduced K-Center to identify multiple sources. The technique works much like K-Means Distance heuristic while assuming that the infection probabilities are known, which in real-world are hard to get. [19] proposed Label Propagation based Source Identification (LPSI) algorithm, which exploits the idea of source prominence to find multiple sources using label propagation mechanism much like a Markov chain process. [22] studied the problem of multiple source detection under heterogeneous SIR model. To this end, they introduced the concept of Jordan cover which is considered to be the set of infection sources.

3 PRELIMINARIES

Notations: Table 1 provides a brief description of the commonly used notations in this paper.

Table 1: Notations used in this study

Notation	Description
$G(V,E)$	Underlying graph
$\mathcal{G}(\mathcal{V},\mathcal{E})$	Infection graph
N_u	Total neighbors of node u
η_u	Infected neighbors of u
t_v^l	Time when v at level l got infected
t_v	Total time v received to infect its neighbors
P_u	Node prominence of node u
P_u^l	Node prominence of node u at level l
$P^l(u)$	Level prominence of level l when traversing the graph from node u
P_u^{local}	Local prominence of node u

Infection Model: Given a graph $G(V,E)$, where V and E denote the sets of nodes and edges respectively, and an infection source s^* , this study employs the classical SI (Susceptible-Infected) model [1] to simulate the process of infection spreading on G. In this model, a node is in either of the two states: susceptible (S) or Infected (I). Once a node gets infected by either receiving infection from its adjacent infected neighbors or simply by being the source, 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, we employ heterogeneous SI model in which the infection probability varies amongst edges. If at a given time-step t, a node w is susceptible to infection from any two of its neighbors, u and v, at time-step $t+1$, w may get infected 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.

Example 2: Figure 2(a) presents an example of the heterogeneous SI model. In the current time-step, nodes u and v are infected, and both are trying to infect w with probabilities 0.2 and 0.8, respectively. Therefore, the overall probability with which w may get infected at the next time-step is $p = 1-(1-0.2)(1-0.8) = 0.84$. ■

Definition 1: (**Underlying Graph**) Underlying graph $G(V,E)$ is the original topology over which the infection spreads, where V is the set of nodes and E is the set of edges. Basically, starting from a random node $u \in V$ as infected, using SI model, infection spreads over G when u infects its neighbors and they in turn infect their neighbors and so on. ■

Definition 2: (**Infection Graph**) Infection graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ is a connected subgraph of underlying graph G containing infected nodes $\mathcal{V} \in V$, and edges $\mathcal{E} \in E$, connecting $\mathcal{V}$. ■

Definition 3: (**Source**) A source s^* is the node from which the infection starts on the underlying graph G. ■

Definition 4: (**Fringe Node**) A fringe node u is an infected node at the boundary of the infection graph $\mathcal{G}$, farthest from the source s^*. ■

Definition 5: (**Infection Source Detection Problem**) Given an infection graph $\mathcal{G}$ and its corresponding underlying graph G, the problem undertaken in this study is to identify the infection source s^* in $\mathcal{G}$, assuming that $\mathcal{G}$ has been infected by a single source under the heterogeneous SI infection model. ■

4 SOURCE IDENTIFICATION IN GENERAL GRAPHS

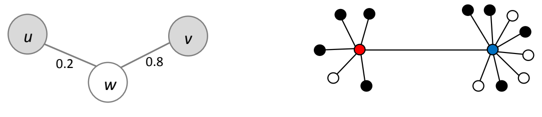
Our overall idea to solve the source identification problem is based on the age of a node. The older the node in a network, the more time it receives to infect its neighbors, and therefore, the higher its chances of being the source. Frioriti et. al [4] have earlier worked along the same lines, while exploiting the concept of dynamical importance of nodes [17]. However, we undertake the problem of finding the age of a node by exploiting the concepts of exoneration and prominence.

免责效应

Exoneration Effect: Exoneration effect on a node u is defined in terms of its surrounding neighbors. Essentially, surrounding nodes could be imagined as an alibi, holding the power to substantially contribute to the vindication of a node assumed to be a culprit of a rumor or could be instrumental for the conviction of the same. If u is infected but at the same time is surrounded by a larger proportion of non-infected neighbors, there is lesser probability of the same to be the culprit - it is exonerated by its neighbors. Formally, the exoneration effect of u denoted by ξ_u is defined as $\xi_u \propto \frac{1}{\eta_u}$, where, η_u is number of infected neighbors of u.

节点重要性

Node Prominence: We define the prominence of an infected node u in terms of the exoneration effect it experiences from its non-infected neighbors. The lesser the exoneration effect on u, the higher its prominence. Formally, prominence of u is denoted by P_u and defined as $P_u \propto \frac{1}{\mathcal{S}_u}$.



(a) Heterogeneous SI model (b) Exoneration Effect

Figure 2: Illustrations

Example 3: Figure 2(b) shows the exoneration effect experienced by the two centers (red and blue nodes) of this graph. All colored nodes are infected nodes. We can see that both the centers have the same number of infected nodes, but the blue center has more non-infected neighbors than the red one. This higher proportion of non-infected nodes exonerate the blue center from being the culprit (source). ■

本地重要性

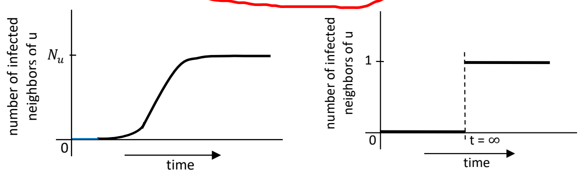
Local prominence: We can understand that exoneration effect measures a node's neighboring prominence. However, this prominence is not enough to convict a node accused of being the culprit. Therefore, to increase the chances of convicting the actual culprit, we need to analyze a node's non-immediate neighbors as well. If a node u is to be considered the source or culprit of a rumor, intuitively, not only should it have its own prominence, but, its non-immediate, closer neighbors (relatively at smaller hop distances from u or, intuitively, at earlier levels l) should also show some degree of such prominence. This prominence is what we refer to as local prominence, i.e., if a node is to be considered the source of an infection, it should have self prominence and subsequently its non-immediate closer neighbors should have some degree of prominence as well. This points us towards the age of a node. Intuitively, the higher the local prominence of a node, the older it is in terms of infection time, and higher its chances of being the source. Formally, local prominence of a node u, denoted by P_u^{local}, is defined as $P_u^{local} = \sum_{l=0}^{k} P^l(u)$, where k is the level *closer* to u, and $P^l(u)$ is the level prominence (defined in Sections 4.1 and 4.2) of all the nodes l hops away from u. We have mathematically determined the best value of k in Section 4.1. Also, $P_u^{local} \propto A_u$, where A_u is the age of node u. Figure 4 graphically shows the local prominence.

4.1 Mathematical Foundation

We firstly show the technicalities latent within our central idea which we theorize on a k-regular tree, and extend it later for generic graphs.

As discussed earlier, if a node is old, it should have a very small or negligible exoneration effect from its neighbors, since it will have had enough time to infect all its neighbors. The oldest node, intuitively, should be considered as the source/culprit of an infection process. Therefore, if a source u has total of N_u neighbors, at time $t_u = 0$ (total infection time u received to infect its neighbors), trivially, it would have infected none of its neighbors. At some point in time $t_u = t$, it would infect some fraction of its neighbors and when t_u is large enough, it would infect all its N_u neighbors. Formally, $\eta_u = \begin{cases} 0 & \text{when } t_u = 0 \\ N_u \times f(t) & \text{when } t_u = t \end{cases}$ where η_u is infected neighbors of u, $f(t_u)$ is a function of time and $N_u \times f(t_u)$ is some fraction of N_u. Also, as $t \to \infty$, $\eta_u \to N_u$. This is illustrated in Figure 3(a). Observe that the change in η_u with respect to time t_u is defined by a positive sigmoid function ($t_u \geq 0$). Therefore, $f(t_u)$ can we written as Eqn. 1, where $\alpha \in R \ll \infty$ is an arbitrarily large real number and significantly smaller than ∞. This leads us to Theorem 1.

$$f(t_u) = \frac{1}{1 + e^{-(t_u - \alpha)}} \tag{1}$$



(a) Source node u (b) Fringe node u

Figure 3: Rate of change of infected neighbors

感染时间趋于无穷时，感染邻居数趋于全部邻居数

Theorem 1: *Consider a k-regular tree with infinitely many levels and source node u as the root. As time $t_u \to \infty$, the number of its infected immediate neighbors $\eta_u \to N_u$.*

Proof: For a source node u as the root (represented as red node in Figure 4), at time $t_u = t$, its infected neighbors η_u is $N_u \times f(t)$ (found above). Using Eqn. 1, $\eta_u = N_u \times \frac{1}{1+e^{-(t_u-\alpha)}}$. At time $t_u = 0$, trivially η_u is 0. As $t_u \to \infty$, we have to show, $\lim_{t_u \to \infty} \eta_u = N_u$.

$$\begin{aligned} \lim_{t_u \to \infty} \eta_u &= \lim_{t_u \to \infty} N_u \times \frac{1}{1 + e^{-(t_u-\alpha)}} \\ &= N_u \times \lim_{t_u \to \infty} \frac{1}{1 + e^{-(t_u-\alpha)}} = N_u \times 1 = N_u \end{aligned} \tag{2}$$

Hence, Theorem 1 holds. ■

Conversely, for a fringe node u, an opposite effect could be observed. At time $t_u = t$, trivially, the number of infected neighbors will be 0. But when t_u approaches infinity (level infinity), the exoneration effect from its surrounding neighbors will be maximum. In the worst-case scenario, it will have only one infected neighbor (from which it received the infection). Therefore, the number of infected neighbors of u at $t_u = \infty$ will be 1, which is a constant. Formally, $\eta_u = \begin{cases} \chi_\infty(t) & \text{when } t_u = t \\ 1 & \text{when } t_u = \infty \end{cases}$, where η_u is infected neighbors of u and $\chi_\infty(t)$ is a function of time. This is illustrated in Figure 3(b).

Observe that the change in η_u with respect to time is defined by a Heaviside function. Therefore, $\chi_\infty(t)$ can we written as Eqn 3. This leads us to Theorem 2.

$$\chi_\infty(t) = \begin{cases} 1 & \text{if } t \geq \infty \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

时间趋于无穷，边缘节点感染邻居数趋于1

Theorem 2: *Consider a k-regular tree with infinitely many levels and fringe node u (node at the boundary of infection) as the root. As time $t_u \to \infty$, the number of its infected neighbors $\eta_u \to 1$.*

Proof: For a fringe node u as the root (represented as blue node in Figure 4), at time $t_u = t$, its infected neighbors η_u is $\chi_\infty(t)$, as shown above. From Eqn. 3, we can easily see that at time $t = 0$, $\chi_\infty(t) = 0$, and $\chi_\infty(t) = 1$ when time $t \to \infty$. Thus, Theorem 2 holds as well. ■

So far, we have mathematically shown the exoneration effect and the resulting prominence of a node. Next, we shall discuss the technical aspects of local prominence as discussed above. Consider the tree in the Figure 4. Let the node at the root be source (red node at level 0). We aim at showing that if a node is source then, besides having node prominence it should have some degree of local prominence as well. In addition to this, as the levels reach infinity, the source should exhibit no or negligible prominence. This leads us to Theorem 3.

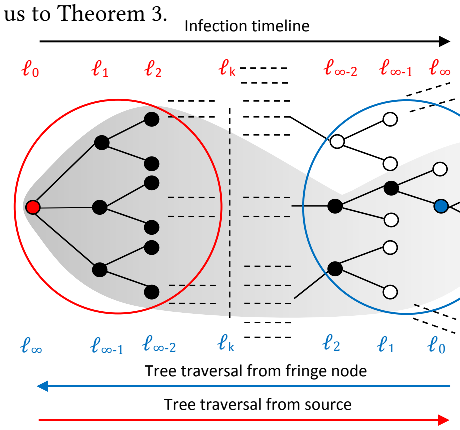

Infection timeline
ℓ_0 ℓ_1 ℓ_2 ℓ_k $\ell_{\infty-2}$ $\ell_{\infty-1}$ ℓ_∞
ℓ_∞ $\ell_{\infty-1}$ $\ell_{\infty-2}$ ℓ_k ℓ_2 ℓ_1 ℓ_0
Tree traversal from fringe node
Tree traversal from source

Figure 4: Sample k-regular graph (k=3) showing local prominence. Red node is the source of the infection and blue node is the corresponding fringe node. The shaded region shows the relevant infected (colored) and non- infected nodes. The red circle indicates the local prominence of source when traversing the tree from the source itself, illustrated by the direction of red arrow and red level labels. Conversely, the blue circle indicates the local prominence of fringe node when traversing the tree from the fringe itself, illustrated by the direction of blue arrow and blue level labels. The color intensity of shaded region corresponds to level prominence.

随着层数的增加，节点v在l层的感染邻居数趋于0

Theorem 3: Consider a k-regular tree with infinitely many levels and source node u as the root. Then the number of infected neighbors of a node v at level l, η^l_v, decreases and approaches 0, as the levels increase.

Proof: Let source node u be the root (red node in Figure 4), t^l_v be the infection time of a node v at level l (i.e., v is l hops away from u), and t_v be the total time v received to further infect its neighbors. Then, $t_v = \tau - t^l_v$, where τ is the total infection time (τ is a significantly large number close to ∞ and $\tau \gg \alpha$) (t_v and t^l_v are illustrated in Figure 5), $\eta^l_v = N_v \times f(t_v)$, and $f(t_v) = \frac{1}{1+e^{-(t_v-\alpha)}} = \frac{1}{1+e^{-(\tau-t^l_v-\alpha)}}$. Therefore, $\eta^l_v = N_v \times \frac{1}{1+e^{-(\tau-t^l_v-\alpha)}}$. At level 0 (source level), if v is the source, then $t^l_v = 0$, so $t_v = \tau$ and therefore, $\eta^l_v = N_v \times \frac{1}{1+e^{-(\tau-0-\alpha)}} = N_v$ as was found in Theorem 1. As $t^l_v \to \tau$, we have to show that at a very large level τ, $\eta^l_v = 0$.

$$\begin{aligned}
\lim_{t^l_v \to \tau} \eta^l_v &= \lim_{t^l_v \to \tau} N_v \times \frac{1}{1+e^{-(\tau-t^l_v-\alpha)}} \text{ (from above)} \\
&= \lim_{t^l_v \to \tau} N_v \times \frac{1}{1+e^{-(\tau-\tau-\alpha)}} \\
&= N_v \times \epsilon \approx 0 \qquad (4)
\end{aligned}$$

where ϵ is an extremely small number. Hence, it holds. ■

Conversely, for a fringe node as the root of the tree (given in Figure 4 as blue node), there should be no prominence at earlier levels, but this prominence should be at levels far from the fringe node itself (closer to source), as shown in Figure 4.

Theorem 4: Consider a k-regular tree with infinitely many levels and a fringe node u as the root. Then the number of infected neighbors, η^l_v, of a node v at level l, increases, as the levels increase and approaches N_v.

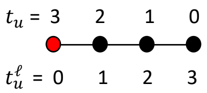


Figure 5: A sample graph showing the time of infection of node u at level l (t^l_u) and total infection time u received to infect its neighbor (t_u)

边缘节点u作为根的话，随着层数的增加，节点v在l层的感染邻居数趋于总邻居数

Proof: For a fringe node as the root, its infinite levels will correspond to levels closer to the source (Figure 4). Theorem 3 shows that at levels closer to the source, the number of infected neighbors of v approaches N_v. The proof of this theorem is converse to that, and thus it trivially holds. ■

From Theorems 3 and 4, we deduce two corollaries.

随着层数的增加，节点重要性减小

Corollary 1: If we have a k-regular tree with infinitely many levels, and root node u is the infection source with $t_u \to \infty$, the prominence of nodes at different levels decrease as levels increase. ■

Corollary 2: If we have a k-regular tree with infinitely many levels, and root node v is a fringe node with $t_v \to \infty$, the prominence of nodes at different levels increase as levels increase. ■

From Corollary 1, traversing the tree from source node u,

$$P^{l_0}_{w_0}(u) > P^{l_1}_{w_1}(u) > P^{l_2}_{w_2}(u) > \cdots > P^{l_k}_{w_k}(u) > \cdots > P^{l_\infty}_{w_\infty}(u) \qquad (5)$$

where $P^{l_i}_{wi}(u)$ is the node prominence of w_i at level l_i when traversing the tree from the source node u as the root (red node in Figure 4). From Corollary 2, traversing the tree from a fringe node as the root (blue node in Figure 4), we have,

$$P^{l_0}_{w_0}(v) < P^{l_1}_{w_1}(v) < P^{l_2}_{w_2}(v) < \cdots < P^{l_k}_{w_k}(v) < \cdots < P^{l_\infty}_{w_\infty}(v) \qquad (6)$$

where $P^{l_i}_{wi}(v)$ is the node prominence of w_i at level l_i when traversing the tree from the fringe node v.

源节点和边缘节点的重要性是相同的

Intuitively, for a k-regular tree, $\sum_{i=0}^{\infty} P^{l_i}_{wi}(u) \approx \sum_{i=0}^{\infty} P^{l_i}_{wi}(v)$. It indicates that the overall prominence of source and fringe node is equivalent. Therefore, in this way source identification becomes indeterminate. However, from Eqns. 5 and 6, we have,

$$P^{l_0}_{w_0}(u) + P^{l_1}_{w_1}(u) + \cdots + P^{l_k}_{w_k}(u) \gg P^{l_0}_{w_0}(v) + P^{l_1}_{w_1}(v) + \cdots + P^{l_k}_{w_k}(v) \qquad (7)$$

Intuitively,

$$\begin{aligned}
P^{l_0}(u) + P^{l_1}(u) + \cdots + P^{l_k}(u) &\gg P^{l_0}(v) + P^{l_1}(v) + \cdots + P^{l_k}(v) \qquad (8) \\
\text{or, } P^{local}_u &\gg P^{local}_v \qquad (9)
\end{aligned}$$

where $P^{l_i}(u)$ and $P^{l_i}(v)$ are the sum of the prominence of all the nodes at level l_i when traversing the tree from source node u and fringe node v, respectively, and P^{local}_u and P^{local}_v are the local prominence of nodes u and v, respectively. Note that at level l_0 there will be only one node, i.e., root of the tree.

Therefore, in order to find the source, we have to find the best l_k. For the same, let us consider a tree network (Figure 4). We know that the longest path from a node u to any other node in a tree is the eccentricity of u. And if u is the root of the tree (red node), the longest path from u to any other node is the minimum eccentricity of tree (Jordan centre). Again, by definition, minimum eccentricity of a graph (tree or otherwise) is the radius of that graph. If the source is the root of the tree (in Figure 4 denoted by red node), then to cover any node in the tree from the source, covering the radius will be sufficient. Therefore, we have to examine prominence of nodes till one level lesser than the radius, since a node at level l_{k-1}

will experience the exoneration effect from the nodes at level l_k. Therefore, $l_k = l_{r-1}$, where r is the radius. Hence, we have shown, in order to find the source, the best l_k is one lesser than the radius.

4.2 EPA: Exoneration and Prominence based Age

To find the infection source in general graphs, we firstly calculate the age of all nodes from the given infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Starting from any node u, we apply BFS to traverse the infection graph. Since the nodes generated by BFS in iteration l are l hops away from u, we treat such nodes belonging to level l. To calculate the exoneration effect on infected nodes as generated by BFS, and thereby their prominence, we take the corresponding non-infected neighbors as well, by considering the underlying graph $G(V, E)$. Therefore, at level l, we take the sum of prominence of all the infected nodes present in l. We refer to this as *level prominence* (the sum of the prominence of all nodes present in l), shown in Figure 6(a). Formally, traversing from node u, the prominence of level l is defined as $P^l(u) = \sum_{v \in \mathcal{V}_l} P^l_v$, where $\mathcal{V}_l \in \mathcal{G}.\mathcal{V}$ is set of nodes at level l. For a node v at level l, we calculate the prominence using Eqn. 10, where I_v and O_v are the corresponding infection degree (from infection graph) and the original degree (from underlying graph) of v.

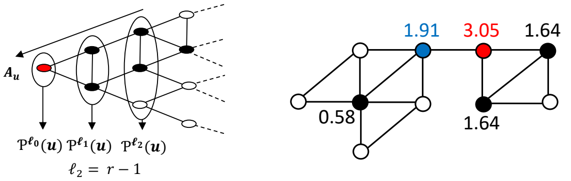
$$P^l_v = \left(\frac{I_v}{O_v}\right) / \left(\frac{1}{1 + \ln O_v}\right) \tag{10}$$

We stop the BFS traversal when we reach the radius, i.e., it processes nodes till $l = r - 1$ (for prominence) and generates nodes till $l = r$ (for exoneration effect). Every node at level $r - 1$, will be exactly $r - 1$ hops away from u. Therefore, the age of node u will be the sum of the prominence of each level starting from level 0 (level of u) to $r - 1$. Formally, $A_u = \sum_{l=0}^{r-1} P^l(u)$, which is also our local prominence (P^{local}_u). From the above definition of $P^l(u)$, $A_u = \sum_{l=0}^{r-1} \sum_{v \in \mathcal{V}_l} P^l_v$. Finally, A_u is defined in Eqn. 11 by using Eqn. 10,

节点年龄

$$A_u = \sum_{l=0}^{r-1} \sum_{v \in \mathcal{V}_l} \left(\frac{I_v}{O_v}\right) / \left(\frac{1}{1 + \ln O_v}\right) \tag{11}$$

Figure 6(a) shows the prominence of different levels and the resulting age of source (sum of prominence at different levels), where black and white circles represent infected and non-infected nodes, respectively.


A_u
$\mathcal{P}^{\ell_0}(u)$
$\mathcal{P}^{\ell_1}(u)$
$\mathcal{P}^{\ell_2}(u)$
$\ell_2 = r - 1$
1.91
3.05
1.64
0.58
1.64

(a) Level prominence (b) Age score by EPA

Figure 6: Illustrations on sample graphs

Algorithm 1: Exoneration and prominence based age for source identification, EPA (Underlying graph (UG) G, infection graph (IG) $\mathcal{G}$)

```
1  $D_{UG} \leftarrow$ degree matrix of $G$ ;
2  $D_{IG} \leftarrow$ degree matrix of $\mathcal{G}$;
3  $D_{UI} \leftarrow$ submatrix of $D_{UG}$ corresponding to $D_{IG}$;
4  $d_{IG} \leftarrow$ diagonal of $D_{IG}$ ; // list containing infection degrees
     of infected nodes
5  $d_{UI} \leftarrow$ diagonal of $D_{UI}$ ;  // list containing original degrees
     of infected nodes
6  $\vec{P} \leftarrow \left(\frac{d_{IG}}{d_{UI}}\right) / \left(\frac{1}{1+\ln d_{UI}}\right)$ ; // element wise operation resulting a
     vector of size $|\mathcal{G}.\mathcal{V}|$
7  $radius \leftarrow$ radius of $\mathcal{G}$;
8  foreach $node\ u \in \mathcal{G}.\mathcal{V}$ do
9      $level \leftarrow 0, U \leftarrow \{u\}$;
10     $\mathcal{N} \leftarrow$ initialize empty set;
11     $A_u \leftarrow age(U, \vec{P}, level, \mathcal{N}, \mathcal{G}, radius)$ ;  // age calculation
12     $A'_u \leftarrow \frac{A_u}{ecc(u)}$;                              // penalized age
13 $\hat{s} \leftarrow \arg\max_{u \in \mathcal{G}} A'_u$;
14 return Estimated source $\hat{s}$;
```

Algorithm 2: Age calculation, *age*(Set of nodes U, Prominence Vector $\vec{P}$, Integer *level*, Set of nodes $\mathcal{N}$, Infection Graph $\mathcal{G}$, Integer *radius*)

```
1  $\mathcal{N} \leftarrow \mathcal{N} \cup U$;
2  if $level > radius - 1$ then
3      return (0);
4  $lage \leftarrow 0$, $\mathcal{K} \leftarrow$ initialize empty set;       // $lage$ gets the level
     prominence of each $level$
5  foreach $u \in U$ do
6      $lage \leftarrow lage + \vec{P}(u)$;
7      $\mathcal{K} \leftarrow \mathcal{K} \cup$ neighbors of $u$ in $\mathcal{G}$ ;
8  $U \leftarrow \mathcal{K} - \mathcal{N}$;
9  return $(lage + age(U, \vec{P}, level + 1, \mathcal{N}, \mathcal{G}, radius))$;
```

We are calculating the level prominence till level $l = r - 1$, and we know that for a source to be the center, it should, intuitively, have minimum largest path distance among all the nodes from itself to any other node in graph. Conversely, a fringe node should have the maximum largest path distance among all the nodes. Therefore, to further discriminate the two, we introduce a concept of penalty. We penalize the age A_u of node u by dividing it with its largest path distance, which is nothing but the eccentricity of u, i.e., $A'_u = \frac{A_u}{ecc(u)}$, where $ecc(u)$ and A'_u are the eccentricity of node u and the penalized age score of node u, respectively. Finally, we consider the node with the oldest age (highest age value) as the source, $source(\hat{s}) = \arg\max_{u \in \mathcal{G}(\mathcal{V}, \mathcal{E})} A'_u$.

Now, we present our source identification algorithm EPA (Algorithm 1). It consists of three main steps. *Step 1*: Firstly, we extract original and infection degrees of infected nodes using underlying graph G and original graph $\mathcal{G}$ (Lines 1-5). Then, we calculate the prominence of every node in the infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ represented by the vector $\vec{P}$, defined in Line 6. *Step 2*: This step is the main part of our algorithm, where the actual age of every node is calculated. Starting from an infected node $u \in G.V$, BFS technique is employed to traverse the infection graph as shown in Algorithm 2. At each level $l = 0$ to $r - 1$, the prominence of l (level prominence) is computed and recursively added to give the final age score of u while new nodes are generated for level $l + 1$ (Lines 1-9, Algorithm 2). *Step 3*: Finally, when the age of a node u is returned, a penalty in terms of the eccentricity of u is given to the age of u (Line 12, Algorithm 1). This results in a penalized age score of u, and the

Algorithm 3: Lightweight source identification, EPA-LW (Underlying graph G, infection graph $\mathcal{G}$)

```
1 Lines 1-7 of Algorithm 1;
2 $\Gamma \leftarrow \arg\min_{u \in \mathcal{G}.\mathcal{V}} ecc(u)$;        // a set of nodes with minimum eccentricity
3 foreach node $u \in \Gamma$ do
4     $level \leftarrow 0, U \leftarrow u$;
5     $\mathcal{N} \leftarrow$ initialize empty set;
6     $A_u \leftarrow age(U, \vec{P}, level, \mathcal{N}, \mathcal{G}, radius)$
7 $\hat{s} \leftarrow \arg\max_{u \in \Gamma} A_u$;
8 return Estimated source $\hat{s}$;
```

node with the highest penalized age source (A'_u) is considered to be the source. Ties are broken at random.

Example 4: In Figure 6(b), both blue and red nodes are the centers of the infection graph (sub-graph of the underlying graph with colored nodes). The red node with the highest age score is considered as the source (estimated by EPA). Also, the node with the lowest age score (0.58) has been exonerated by a large number of non-infected neighbors, therefore showing the higher exoneration effect on the same. This comparatively small score also makes sense in that it is farthest away from the estimated source. In addition to this, infected nodes closer to the estimated source both have age scores of 1.64 which again makes sense because both have small exoneration effect. ■

Earlier, we showed that starting from a node u, level l_{r-1} is the best level until the infection graph is to be traversed (Section 4.1) to calculate the level prominence (exoneration effect on infected nodes at level l_{r-1} will be from non-infected nodes at level l_r). Therefore, essentially, we are considering the nodes of an infection graph till its radius. But we know that the radius of a graph is also its minimum eccentricity. Therefore, instead of analyzing all the nodes in the infection graph, we pick the nodes with minimum eccentricity as the candidates for source. Thus, the candidate set $\Gamma = \arg\min_{u \in \mathcal{G}(\mathcal{V},\mathcal{E})} ecc(u)$. This results into a modified lightweight version of EPA, called EPA-LW, given in Algorithm 3. Line 1 of this algorithm corresponds to Lines 1-7, Algorithm 1. Then, instead of calculating the age of all the nodes in the graph, we pick only those with minimum eccentricity as source candidates, represented by a set Γ (Line 2) and compute the age of nodes in Γ (Lines 3-6) using Algorithm 2. Since we are only picking the minimum eccentricity nodes as candidates, penalty is trivial here. Therefore, for EPA-LW, *source*$(\hat{s}) = \arg\max_{u \in \Gamma} A_u$ (Line 7).

4.3 Computational Complexity

Theorem 5: Given an infection graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ of n nodes and e edges, with radius r, the computational complexity of EPA is $O(n^2)$ in best case and $O(n^3)$ in worst case.

Proof: Starting from a node $u \in n$, EPA uses BFS technique to traverse the infection graph till radius r. Let $n_r \in n$ denote the number of nodes covered within the radius of u and $e_r \in e$ denote the edges amongst n_r. Then, the computational complexity of EPA to calculate the age of u is $O(n_r + e_r)$. If the infection graph is very sparse, i.e., domination of the number of nodes over egdes (case of a ring graph), this complexity becomes $O(n_r + n_r) = O(n_r)$ (best case). Similarly, if the infection graph is very dense (case of a complete graph), i.e., domination of the number of edges over nodes, $n_r = n$ and $e_r = e = \frac{n \times (n-1)}{2}$. Therefore, in this case, the complexity of EPA to calculate the age of node u will be $O(n + n^2) = O(n^2)$ (worst case). Therefore, given an infection graph with n nodes, the overall complexity of EPA to calculate the ages of n nodes will be $O\left(n^2\right)$ in best case and $O\left(n^3\right)$ in worst case. ■

Note that the computational complexity of EPA-LW, in worst case, is also $O(n_e^3)$, where n_e is the number of nodes having minimum eccentricity in infection graph, $\mathcal{G}(\mathcal{V},\mathcal{E})$. In average case scenario, $n_e \ll n$, where n is the total number of nodes in $\mathcal{G}(\mathcal{V},\mathcal{E})$.

4.4 Relation between EPA and EPA-LW

Since EPA initially considers every node in the infection graph equally probable of being the source, if there are n infected nodes, the probability of a node being the source is $1/n$. EPA-LW on the other hand, considers only Jordan centers (minimum eccentricity nodes) as candidates for source. So, the probability of a Jordan center J being the source is $\mathbb{P}(J = s^*) = \frac{1}{n} \times \frac{n_e}{n} = \frac{n_e}{n^2}$. where n_e is the number of minimum eccentricity nodes in the graph.

Theorem 6: For an infection graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ with n nodes and n_e Jordan centers, the initial probability of a Jordan center being the source under EPA-LW lies in $\left[\frac{1}{n^2}, \frac{1}{n}\right]$.

Proof: Consider the following two cases.

Case 1 (Worst Case): Consider a chain graph with n nodes. The value of n_e will be either 1 or 2 depending on whether n is odd or even, respectively. As shown above, $\mathbb{P}(J = s^*) = \frac{n_e}{n^2}$, which equals $\frac{1}{n^2}$ or $\frac{2}{n^2} \geq \frac{1}{n^2}$.

Case 2 (Best Case): Consider a complete graph with n nodes. The value of n_e will be same as the number of nodes n, i.e., $n_e = n$. Therefore, $\mathbb{P}(J = s^*) = \frac{n_e}{n^2} = \frac{n}{n^2} = \frac{1}{n}$, which is same as the initial probability of a node being the infection source under EPA.

Considering the two extreme cases above (minimum and maximum number of Jordan centers), the probability of a Jordan center being the source $\mathbb{P}(J = s^*) \in \left[\frac{1}{n^2}, \frac{1}{n}\right]$ ■

Note that these probabilities are initial probabilities associated only with the infection graph. They will be affected when we consider the underlying graph (exoneration effect from non-infected neighbors).

4.5 Extension of EPA for Multiple Sources

EPA, being the single source identification technique, can be easily extended to identify multiple sources of infection in a complex graph with a good detection rate and error distance. We propose two techniques, one based on K-means and the other based on successive identification of sources, resulting into two algorithms-EPA_K-Means and EPA_SSI (Successive Source Identification).

EPA_K-Means: Shown in Algorithm 4, this technique starts with randomly picking m initial centroids, where m is the known number of sources (Line 1). In the clustering step, using these initial centroids, we make m partitions of a given graph using Voronoi graph partitioning scheme as used in [6] (Line 3). Then, we compute the single infection source in each partition by using EPA (Lines 5-8), and iteratively update them until convergence is achieved (Lines 2-10). We then consider the final m centroids as the set of sources (Line 11). A similar technique can also be seen in [12] [6].

EPA_SSI: Shown in Algorithm 5, this technique starts with estimating the first source $\hat{s}$ by using EPA (Line 3). Thereafter, we

```
Algorithm 4: EPA_K-Means ( Underlying graph $G$, infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, number of sources $m$)
1  Randomly pick $m$ nodes $\hat{S} \leftarrow \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m\}$, from $\mathcal{G}.\mathcal{V}$;
2  repeat
3      Partition $\mathcal{G}$ into a set of $m$ partitions $C$ using k-means, by considering $\hat{S}$ as initial centroids;
4      $newc$ ← initialize empty set;
5      foreach $c \in C$ do
6          Construct a connected sub-graph $\mathcal{G}_c \subseteq \mathcal{G}$ by using $c$;
7          Find estimated source $\hat{s}$ in $\mathcal{G}_c$ by using Algorithm 1;
8          $newc \leftarrow newc \cup \hat{s}$ ; // $\hat{s}$ is the new center of $c$
9      $\hat{S} \leftarrow newc$ ;                    // update centres
10 until convergence;
11 return Estimated set of sources $\hat{S}$;
```

```
Algorithm 5: EPA_SSI (Underlying graph $G(V, E)$, infection graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, number of sources $m$)
1  $i \leftarrow 1$, $\hat{S}$ ← initialize empty set;
2  while $i \leq m$ do
3      Find the source $\hat{s}$ using Algorithm 1;
4      Set $\hat{s}$ unifected by updating the adjacency matrix of $\mathcal{G}$;
5      $\hat{S} \leftarrow \hat{S} \cup \hat{s}$, $i \leftarrow i + 1$;
6  return Estimated set of sources $\hat{S}$;
```

set the previous source to *uninfected* by setting the $\hat{s}$-th row and column entries to 0 in the adjacency matrix of the given infection graph (Line 4). To find the next source, EPA is applied on the updated adjacency matrix. This process of finding the sources, one at a time, is repeated successively until m number of sources are retrieved (Lines 2-6), where m is the known number of sources.

5 EXPERIMENTAL EVALUATION

Sections 5.1 – 5.3 present our experimental results of the primary algorithms EPA and EPA-LW proposed in this paper, whereas Section 5.4 presents the results of EPA_K-Means and EPA_SSI proposed as extensions. It is found that EPA and EPA-LW outperform state-of-the-art single source detection methods, and EPA_K-Means and EPA_SSI generally outperform the standard heuristics.

5.1 Datasets

We evaluate the performance of our two proposed methods for single source identification on seven network topologies (2 synthetic i.e., 4-regular and ER (Erdős-Rényi)-random [3], and 5 real-world i.e., Facebook [9], US Power Grid (USPG) [20], Karate [21], Les Misérables (Lesmis) [8] and Jazz [5]) whose statistics are given in Table 2 (first two columns).

5.2 Experimental Setup

First, we pick a source, s^*, at random and then use SI heterogeneous model (described in Section 3) to spread infection on each topology, $G(V, E)$. We keep the edge weight (infection probability) uniformly distributed over (0,1). The infection spreading simulation is stopped when the desired number of nodes are infected resulting in infection graph, $\mathcal{G}(\mathcal{V}, \mathcal{E})$. To analyze the scalability, we consider two infection graph sizes - small and large. The statistics of both the infection graph sizes for each topology are provided in 2 (last two columns). Since Karate, Les Misérables (Lesmis) and Jazz are relatively small graphs, therefore, for large infection graphs we only considered Facebook, USPG, ER-random and 4-regular. All the reported results are averaged over 100 independent runs. We have implemented[2] all the proposed and compared methods in R.

Table 2: Dataset statistics

Network	# Nodes	# Edges	Small	Large
Facebook [9]	4,039	88,234	2-5%	40-60%
US Power Grid (USPG) [20]	4,941	6,594	2-5%	40-60%
4-regular	5,000	10,000	2-5%	40-60%
ER-random [3]	5,000	24,943	2-5%	40-60%
Karate [21]	34	98	60%	-
Les Misérables (Lesmis) [8]	77	254	65%	-
Jazz [5]	198	2,497	70%	-

Evaluation measures: For evaluation, we use the error distance (*ED*) between the actual and estimated source. Formally, error distance is defined as $ED = h(s^*, \hat{s})$ where $h(s^*, \hat{s})$ is the distance between actual source s^* and estimated source $\hat{s}$ in hops. Besides this, we also use average error distance (AED), which is the average of all *ED* over 100 independent runs.

Compared methods: We compare the performance of EPA and EPA-LW, against the following existing methods, *i*) Dynamic Age (DA) [4], *ii*) Minimum Description Length (MDL) [15], and *iii*) Reverse Infection (RI) [23] (described in Section 2). RI, originally, was devised for SIR model of infection, but since SI is a special case of SIR, this method can also be used for source identification when the propagation model is SI [2].

5.3 Experimental Results

In this section, we present the performance of the two proposed single source identification algorithms. For small infection graph sizes, we use both EPA and EPA-LW, and for the large we only test the performance of EPA-LW due to the considerably smaller computational cost associated with it.

Results on Small Infection Graphs: Figure 7 shows the frequency of error distances across all the seven topologies. Our results show that both the proposed single source identification approaches, EPA and EPA-LW, mostly and consistently find the actual source mostly within distances 0 and 1, which is generally better than all the three state-of-the-art methods. The best performance of EPA and EPA-LW is found on ER-random graph, where both the methods find a source with 95% accuracy which is only matched by RI amongst the comparing methods, while DA and MDL even estimate a source 3 hops away from the actual source. The performance of both the proposed methods is even more pronounced when contrasted against the other methods on 4-Regular network with better accuracy than all the comparing methods. This performance of EPA and EPA-LW against the comparing methods can also be observed on Facebook network. On USPG network, source detection becomes quite hard for all the methods. The reason for the same is the high sparsity of USPG network (density = 0.0005).

[2] The source codes of the proposed as well as compared methods, along with the datasets, are publicly available at GitHub (https://github.com/tesla121/EPA-Data-Code).

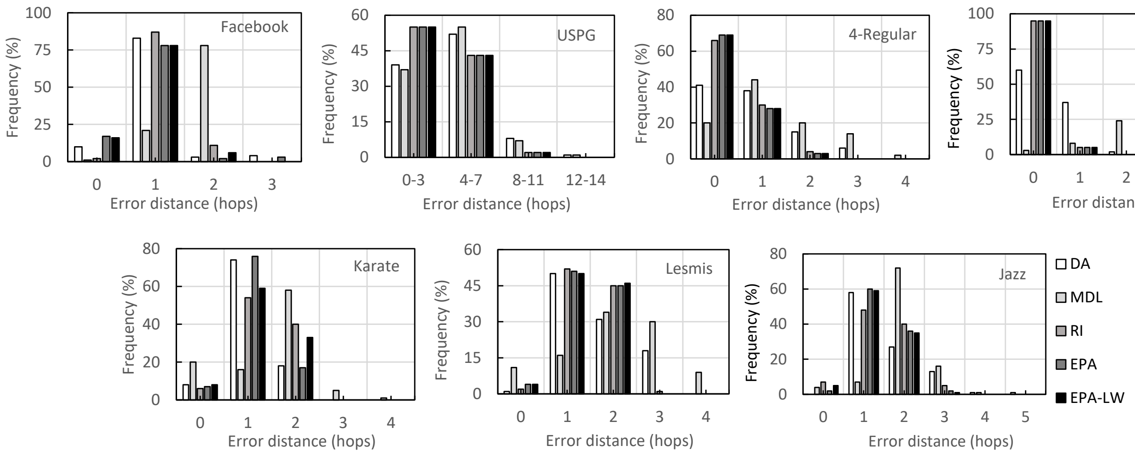


Figure 7: Frequency of error distances across different topologies (small infection graphs)

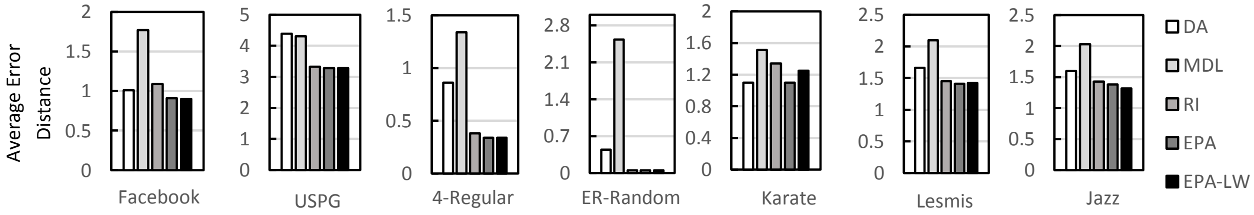


Figure 8: Average error distances across different topologies (small infection graphs)

Even then, EPA and EPA-LW finds the source within 0-3 hops from the actual source. On smaller networks (Karate, Lesmis and Jazz), again EPA and EPA-LW mostly find sources within 0 and 1 hops, which is better than other methods.

The above findings are further substantiated when we analyze the average error distance. As shown in Figure 8, EPA outperforms all the other methods on Facebook, USPG, 4-regular, Lesmis and Jazz. On Karate, EPA produces average error distances of 1.1 similar to DA and, on ER-random, 0.05, similar to RI. EPA-LW produces similar performance to EPA on Facebook, USPG, ER-random, 4-Regular and Lesmis.

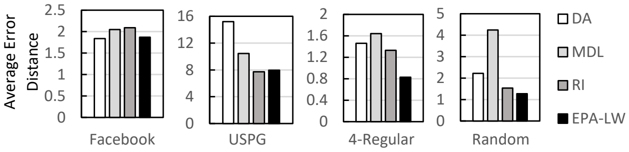


Figure 9: Average error distances in large infection graphs

Results on Large Infection Graphs: To check the scalability, we analyze the performance of EPA-LW on average error distance. The reason we pick only EPA-LW for comparison on large graphs is the lesser computational time associated with it. As can be seen in Figure 9, EPA-LW significantly outperforms all the three methods on two synthetic networks, i.e., ER-random and 4-regular, with average hop errors of 1.27 (37% accuracy) and 0.83 (41% accuracy). This finding is important, because given EPA-LW takes only a fraction of nodes as input, as compared to DA and MDL, and yet produces better results. On Facebook, EPA-LW produces results similar to DA while outperforming MDL and RI. On USPG, EPA-LW has a similar average error distance to RI, while beating the other two competing methods significantly. Note that while other methods keep fluctuating their performance as the networks change, EPA-LW, continues to give consistent performance, proving its consistency.

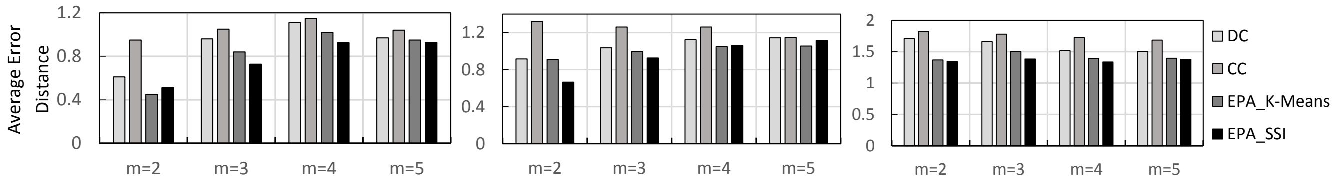
5.4 Multiple Source Detection Results

For multiple sources ($m > 1$), we evaluate the performance of EPA_K-Means and EPA_SSI on three datasets - Karate, Les Misérables and Jazz. For a given number of sources m, we randomly pick m nodes as seeds and use SI heterogeneous model to spread infection on each topology. We stop the simulation when at least 30%, 50% and 80% of the nodes are infected in Lesmis, Karate and Jazz networks, respectively. All the reported results are averaged over 100 independent runs.

For evaluation, we use two widely used measures, i.e., error distance and detection rate as used in [22]. The error distance is defined as $\min_{\Omega \in permutation(\hat{S})} \sum_{i=1}^{m} \frac{d(s_i,p_i)}{m}$, where $S = \{s_1, s_2, \ldots, s_m\}$ is the set of m actual sources, $\hat{S}$ is the set of m estimated sources, and $\Omega = (\omega_1, \omega_2, \ldots, \omega_m)$ is a permutation of $\hat{S}$. Detection rate is defined as $DR = \frac{|S \cap \hat{S}|}{m}$. All the reported resulted are averaged over 100 independent runs.

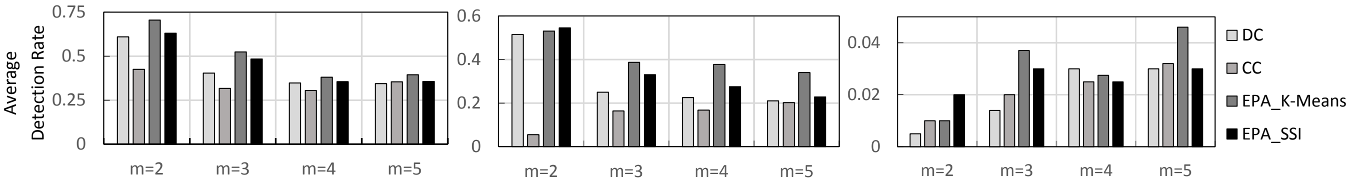
We compare the performance of EPA_K-Means and EPA_SSI against two K-Means based heuristic algorithms, i.e., Distance Centrality (DC) and Closeness Centrality (CC), previously used in [22] and [14]. In both these heuristics, the initial centroids are picked at random. Using K-Means, during the clustering step for each iteration, distance and closeness centroids become new centers of each partition for DC and CC, respectively. This process is repeated until convergence is achieved.

As can be seen in Figure 10, for any number of sources ($m =$ 2, 3, 4, 5), both the proposed algorithms outperform the CC and DC heuristics across all the networks as far as average error distance



(a) Karate (b) Lesmis (c) Jazz

Figure 10: Average error distance (AED) for number of sources $m = 2, 3, 4, 5$ across three topologies on multi source detection



(a) Karate (b) Lesmis (c) Jazz

Figure 11: Average detection rate (ADR) for number of sources $m = 2, 3, 4, 5$ on multi source detection

is concerned. On Karate (Figure 10(a)), EPA_SSI performs better than EPA_K-Means for sources, $m = 3, 4, 5$, while both the proposed methods are better than DC and CC. On Lesmis (Figure 10(b)), again both proposed approaches outperform the heuristics, with EPA_SSI slightly showing the edge over EPA_K-Means for lesser number of sources. As for Jazz network (Figure 10(c)), we can see a more significant outperformance of both EPA_K-Means and EP_SSI over DC and CC for any number of sources, with EPA_SSI continuing being slightly better than EPA_K-Means as far as average error distance is concerned.

As for the average detection rate (Figure 11), EPA_K-Means significantly outperforms both CC and DC for any number of sources on Karate, as shown in Figure 11(a). EPA_SSI, albeit not as significantly as EPA_K-Means, also outperforms both CC and DC. On Lesmis (Figure 11(b)), while both the proposed methods perform better than the heuristics, we observe that EPA_K-Means has a significantly better performance than DC and CC for $m = 3, 4, 5$, but for $m = 2$ EPA_SSI slightly performs better than EPA_K-Means. As for Jazz (Figure 11(c)), again EPA_K-Means comes up with better average detection rate than the other heuristics. From the discussion above, we understand that both the proposed approaches have better average error distance and average detection rate than DC and CC heuristics, irrespective of the network topology.

6 CONCLUSION

In this paper, we studied the infection source identification problem by exploiting the idea of the source being the oldest node. To this end, we proposed a novel algorithm called Exoneration and Prominence based Age (EPA), which calculates the age of an infected node by considering its prominence in terms of its both infected and non-infected neighbors. We also proposed a computationally inexpensive variant of EPA, called EPA-LW. The proposed methods consistently outperform the state-of-the-art methods from several perspectives on seven standard datasets including both small and large infection graphs of different topologies. We further extended EPA to identify sources in multi-source infection scenario, resulting into two algorithms called EPA_K-Means and EPA_SSI. Our results validate the superiority of EPA_K-Means and EPA_SSI over other multi-source heuristic approaches.

REFERENCES

[1] Linda J. Allen. 1994. Some discrete-time SI, SIR, and SIS epidemic models. *Mathematical Biosciences* 124 (1994), 83–105. Issue 1.

[2] Biao Chang, Enhong Chen, Feida Zhu, Qi Liu, Tong Xu, and Zhefeng Wang. 2018. Maximum a Posteriori Estimation for Information Source Detection. *IEEE Trans. on Systems, Man, and Cybernetics: Systems* (May 2018), 1–15.

[3] Paul Erdös and Alfréd Rényi. 1959. On random graphs I. *Publ. Math. Debrecen* 6 (1959), 290–297.

[4] Vincenzo Fioriti, Marta Chinnici, and Jesus Palomo. 2014. Predicting the sources of an outbreak with a spectral technique. *Appl. Math. Sci.* 8, 135 (2014), 6775–6782.

[5] Pablo M. Gleiser and Leon Danon. 2003. Community structure in jazz. *Advances in complex systems* 6, 4 (2003), 565–573.

[6] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. 2015. K-center: An approach on the multi-source identification of information diffusion. *IEEE Trans. on Inf. Forensics and Security* 10 (2015), 2616–2626. Issue 12.

[7] Nikhil Karamchandani and Massimo Franceschetti. 2013. Rumor source detection under probabilistic sampling. In *IEEE ISIT*. 2184–2188.

[8] Donald E. Knuth. 1993. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley.

[9] Jure Leskovec and Julian McAuley. 2012. Learning to discover social circles in ego networks. In *NIPS*. 539–547.

[10] A.Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová. 2013. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E* 90, 1 (2013).

[11] Wuqiong Luo and Wee Peng Tay. 2012. Identifying multiple infection sources in a network. In *ASILOMAR*. 1483–1489.

[12] Wuqiong Luo, Wee Peng Tay, and Mei Leng. 2013. Identifying infection sources and regions in large networks. *IEEE Trans. Signal Process.* 61, 11 (2013), 2850–2865.

[13] Wuqiong Luo, Wee Peng Tay, and Mei Leng. 2014. How to identify an infection source with limited observations. *IEEE J. Sel. Topics Signal Process.* 8 (2014), 586–597. Issue 4.

[14] Wuqiong Luo, Wee Peng Tay, and Mei Leng. 2017. On the universality of Jordan centers for estimating infection sources in tree networks. *IEEE Trans. on Systems, Man, and Cybernetics: Systems* 63 (2017), 4634–4657. Issue 7.

[15] B. Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. 2012. Spotting culprits in epidemics: How many and which ones?. In *IEEE ICDM*. 11–20.

[16] B. Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. 2014. Efficiently spotting the starting points of an epidemic in a large graph. *Knowl. Inf. Syst.* 38, 1 (2014), 35–59.

[17] Juan G. Restrepo, Edward Ott, and Brian R. Hunt. 2006. Characterizing the dynamical importance of network nodes and links. *Phys. Rev. Lett.* 97, 9 (2006).

[18] Devavrat Shah and Tauhid Zaman. 2010. Detecting sources of computer viruses in networks: Theory and experiment. In *ACM SIGMETRICS*. 203–214.

[19] Zheng Wang, Chaokun Wang, Jisheng Pei, and Xiaojun Ye. 2017. Multiple Source Detection without Knowing the Underlying Propagation Model. In *AAAI*. 217–223.

[20] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440–442.

[21] Wayne W. Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33, 4 (1977), 452–473.

[22] Kai Zhu, Zhen Chen, and Lei Ying. 2017. Catch 'Em All: Locating Multiple Diffusion Sources in Networks with Partial Observations. In *AAAI*. 1676–1683.

[23] Kai Zhu and Lei Ying. 2016. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Trans. on Networking* 24 (2016), 408–421. Issue 1.