K-Coverage: A Monitor Node Selection Algorithm for Diffusion Source Localizations

Yuexin Zhang[1]* and Jianjun Zhang[2]
[1]China University of Mining and Technology, Xuzhou, Jiangsu, China
zhengyuexin121@live.cn
[2]Shanxi Hongning Railway Co. LTD, Shenmu, Shanxi, China
office@sxhntl.com

Abstract

The method of selecting monitor nodes has a direct impact on the accuracy of the infection source localization, and infection source localization methods that use monitor observation tend to focus on the source localization itself and ignore the selection of monitor nodes. The biggest problem of using graph centrality to select monitor nodes is that the distribution of the selected monitor nodes may be too concentrated, thus affecting the effect of infection source localization. In order to solve the problem of centralized distribution of monitor nodes, a hierarchical method of selecting monitor nodes using K-shell is proposed. To further improve the effectiveness of the selection, the overlapping range of neighbors is introduced into the selection method. Through simulation experiments on various networks of various sizes, the monitor node selection method can effectively improve the accuracy of infection source localization.

Keywords: Source localization, Monitor observation, Monitor node, K-shell

1 Introduction

With the rapid development of the Internet, online social networking applications, such as Twitter, Weibo, WeChat, Reddit, forums, etc., affect everyone's life to a greater or lesser extent. Online social applications make people closer to each other, and these close ties make up the well-known social network. According to a social network analysis study[1], in 2020, the global social penetration rate reached 49 percent, about 4.14 bn active global social media population worldwide. While this close connection facilitates people's lives, it also facilitates the spread of negative information. Today, with the ubiquity of cell phones, everyone from the elderly to children is a node in a social network, and everyone receives a large amount of unverified information every day [13, 16]. These large amounts of information are likely to be mixed with things like misinformation and rumors, and the consequences of spreading them widely are often serious.

Problems not only in social networks, but also in the spread of COVID-19, computer viruses transmitted on the Internet, contaminants in the water network, high-voltage surges in the power grid, and failures in sensor networks, all require the localization of infection and failure sources in the corresponding networks [5, 7, 12]. These "faulty networks" are collectively called infection networks and "faulty sources" are collectively called infection sources, so this series of problems are finally abstracted as the problem of infection source localization in infection networks. Therefore, infection source localization has a wide range of applications in all walks of life, and rapid and accurate identification of infection sources is of considerable importance.

Research Briefs on Information & Communication Technology Evolution (ReBICTE), Vol. 6, Article No. 8 (December 1, 2020)
*Corresponding author

In the initial study[11], the infection source localization method was designed to be used in a tree-like network in the form of a complete observation. As the research progressed, the observation methods for infection source localization became more diverse, with snapshot observation being used for localization. In order to find the source of infection more accurately and quickly, localization methods using monitor to record the time of infection are becoming popular.

At present, most of the infection source localization methods using monitor observations focus on the localization itself and ignore the selection of monitor nodes. The distribution of monitor nodes in the graph is related to the level of access to the transmitted information, and the source localization method relies on the information obtained from the monitor nodes, so the method of selecting monitor nodes directly affects the accuracy of source localization. The source localization method using monitor observation often adopts the traditional graph centrality such as random[10, 18, 3, 9, 8], degree[10, 18, 19, 8], and betweenness[18, 19, 14, 8] in selecting monitor nodes. The biggest problem with using centrality for selection is that the distribution of the selected monitor nodes may be too concentrated. As shown in Figure 1, the degree centrality is used to select three monitor nodes on the Karate network, with the blue nodes being the monitor nodes. It can be seen that the nodes ”33” and ”34” are large degree values and meet the selection criteria, but these large nodes are often directly connected to each other, which results in a non-centralized distribution of the monitor nodes. There will be a large overlap in the range of information received by the monitor nodes.

In this paper, we propose K-Coverage, a monitor node selection method based on K-shell, to address

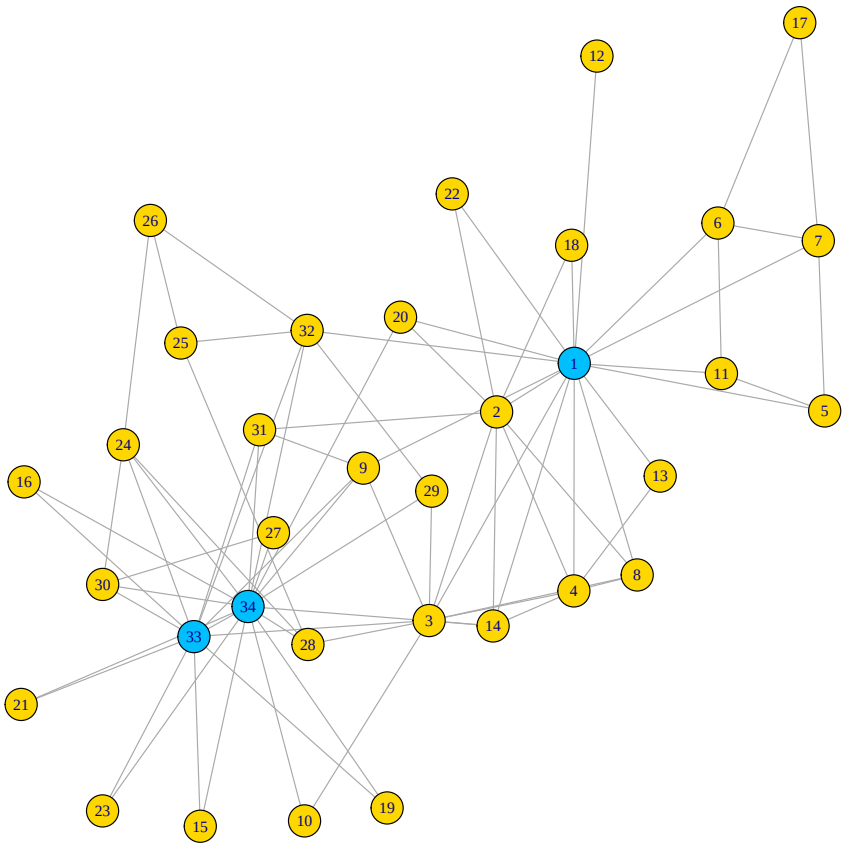


Figure 1: Select three monitor nodes in the Karate network.

the unbalanced distribution of monitoring nodes in the graph, and the experiments show that this method has a higher localization accuracy than the traditional centrality selection method.

2 Related Works

Current methods for locating the source of infection are mainly classified according to the method of observation of the infection. Observation methods refer to the methods used to capture the changing processes of the infection network over time. Current studies fall into three main categories: complete observation, snapshot observation, and monitor observation[4]. Shah and Zaman[11] were the first to propose the concept of rumor localization, and they studied it in a tree-like network, using the SI (Susceptible-Infected) model for information diffusion. They proposed the Rumor Centrality approach, which the number of propagation paths of a node is defined as the rumor centrality, and the larger the rumor centrality, the more likely the node is to be the source of infection. Compared with rumor centrality search for an infection source, assigning a group of nodes as the infection source greatly reduces the probability of missing the real infection source and reduces the scale of node search. Dong et al.[2] then proposed a Local Rumor Center approach, which designates a set of nodes as the source of infection. Zhu and Ying[20] proposed a method using Jordan Centrality in a tree-like network and SIR (Susceptible-Infected-Recovered) model. The Jordan centrality of a node is the maximum shortest distance from infected nodes and recovered nodes, and the greater the Jordan centrality, the more likely it is to be an infected source.

In this paper, we study the source localization method using monitor observation, also known as sensor observation. In this context, monitor refers to the selection of a certain number of nodes in a network using some sort of selection method, which is used to record the time at which they receive information. This is called static monitor selection if the nodes are selected before the infection starts, or dynamic selection if the nodes are selected during the infection. The difference between a monitor node and a normal infected node is that the monitor node records its own infection time.

Pinto et al.[10] proposed a single source localization method using Gaussian estimation, which was subsequently improved by Paluch et al.[9] and Shelke et al.[14]. Shi et al.[15] proposed a single source localization method using Markov random fields, and the experimental results demonstrate that this method is more effective than the Jordan Centrality, DMP, using snapshot observation. Wang[18] proposed a more general approach to locate by calculating the Spearman correlation coefficient between the distance from the node to the monitors and the time of monitors receiving the information. According to the general law of information diffusion, ideally the distance and time of information diffusion should be directly proportional, so that the closer the Spearman correlation coefficient of a node is to 1, the more likely it is to be the source of infection.

Spinelli et al.[17] proposed a general framework for locateing the source of infection using dynamic monitor node placement. The framework is divided into three steps: the first step is to select a small number of nodes in the network to obtain whether the infection has started. These nodes are called static monitor; the second step is to select some nodes in the uninfected position of the network by some method after the static monitor nodes get the infection start, which are called dynamic monitor; the third step is to use the information collected by the monitor nodes to make source localization. Localization can be done after the infection has stopped (Offline) or the infection is in progress (Online).

The source localization method used in this paper to evaluate the performance of monitor node selection is based on the Spearman correlation coefficient proposed by Wang.

3 Method

In this paper, we use K-shell to layer the network. By selecting monitor nodes in each layer of the network, the monitor nodes are evenly distributed in the graph, which not only avoids the problem of over-concentration of the selected monitor nodes, but also ensures the coverage of the monitor nodes. The rule in each layer is that the node in the current layer covers the node in the outer layer. If a node in the k layer is n, then the score S_n of node n is defined as.

$$S_n = | n_{nei} \cap (N_1 \cup N_2 \cup \ldots \cup N_{(k-1)}) \setminus (M \cup M_{nei}) | \tag{1}$$

where n_{nei} is the neighbors of node n, N_i is all nodes with K-shell value $k = i$, M is the already selected monitor nodes, and M_{nei} is the neighbors of the already selected monitor nodes. M_{nei} is actually the coverage area of the already selected monitor nodes, so the further selected monitor node should avoid these already covered nodes.

This method only takes into account the coverage of outer layer nodes, but ignores the coverage of inner layer nodes. In reality, the network is so diverse that it cannot be excluded that there may be cases where $k = i$ nodes are too few to select one node. Although these less than one case will be accumulated to the next layer for selection, but this layer between the formation of a blind area. Therefore, the overlap range of neighbors is taken into account when selecting monitor nodes. The smaller the coverage area between the neighbors of the selected node and the neighbors of the monitor node, the higher the ranking of the node as a potential monitor node candidate. The perfected score S_n of the node n is defined as.

$$S_n = | n_{nei} \cap M_{nei} | \tag{2}$$

The following is a simple demonstration of the implementation of the K-Coverage algorithm using the Karate network, as the K-shell layer is not well-displayed in the graph, we use the node's degree to roughly distinguish the K-shell value, the higher the degree, the larger the representation of the node in the graph. According to the rule, the outermost node, $k = 1$, is not selected for monitor. The interesting thing is that there is only one node with $k = 1$, and none of the $k = 1$ nodes is directly connected to it, so it can only be selected according to the degree value. As shown in Figure 2(a), node "10" is selected as the monitor node, the remaining $k = 2$ nodes are selected as the covered nodes for the next round. As shown in Figure 2(b), node "6" covers node "17" of $k = 2$, and its degree value is higher, so node "6" is selected. The remaining $k = 3$ nodes are the covered nodes for the next round of selection. Finally, $k = 4$ nodes are selected, as shown in Figure 2(c). Although nodes "34" and "1" cover more outer nodes, they also include the neighbors of nodes "10" and "6", so they can not be selected as monitor nodes. Finally, nodes "33" and "2" are selected, so the algorithm ends.

4 Experiments

4.1 Datasets and Evaluation Metrics

The Spearman correlation coefficient based infection source localization method[18] is used to evaluate the validity of the monitor node selection methods. The source localization method is based on calculating the Spearman correlation coefficient between the distance from the infected node to the monitor node and the time at which the monitor node received the information; the closer the correlation coefficient is to 1, the more likely it is to be an infected source. The Spearman correlation coefficient ρ_k for node k is defined as follows:

Algorithm 1: K-Coverage

```
Input: graph, rate
Output: picked

coreness = coreness(graph)  // K-shell value asc order
last_not_picked = coreness[1]  // outermost layer no monitor node selected
rate = coreness * rate / coreness[-1]  // calculate the picked number of each layer

for c in coreness[-1]
  c = degree(graph, c) // desc order
  count = length(c) * rate

  for node in c
    adjs = are_adjacent(graph, node, last_not_picked)
    scores[node] = length(adjs)
    last_not_picked = last_not_picked[-adj]  // remove nodes that have been covered
  end for
  scores = order(scores, desc)

  for i in count
    for pick in picked
      adj = are_adjacent(graph, scores[i], pick)  // make sure it is not covered by the monitor nodes
      if adj
        break
      end if
    end for
    if !adj
      picked = scores[i]
    end if
  end for
  last_not_picked = difference(c, picked)
end for
return picked
```

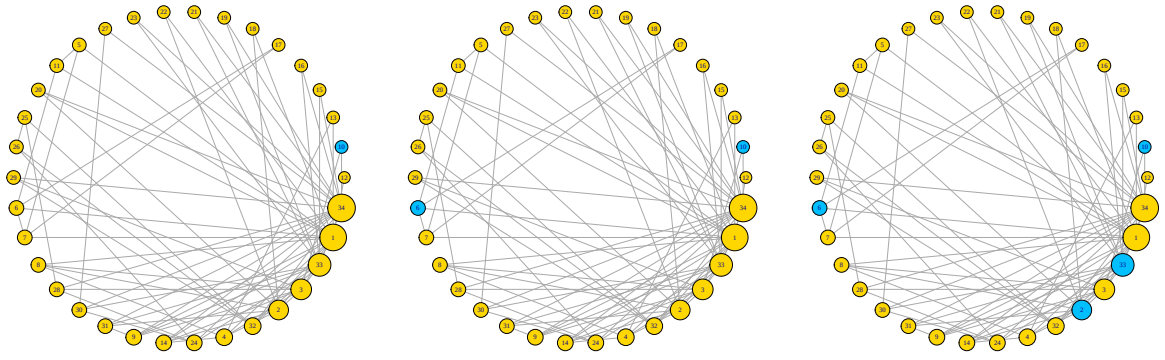
$$\rho_k = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (d_i^k - t_i)^2 \tag{3}$$

where n is the number of monitor nodes, d is the distance from node k to the monitor node, and t is the time when the monitor node receives information.

Classical graph centrality metrics: degree centrality, betweenness centrality, and random are used for the comparison of monitor node selections. The SI model was used for the information diffusion model. We performed simulations on 9 real networks, and each network has been simulated for 100 times: Zachary's Karate Club network, the network of characters in Les Miserables Lesmis, the RFID sensor network, the network of jazz musicians Jazz,, the network of U.S. airports in October 2010 USairports. immunoglobulin interaction network Immuno, Yeast interaction network, Facebook friend network, and US power grid USPG, the details of the datasets are shown in Table 1.

Algorithm 2: K-Coverage-Nei

```
Input: graph, rate
Output: picked

coreness = coreness(graph)  // K-shell value asc order
last_not_picked = coreness[1]  // outermost layer no monitor node selected
rate = coreness * rate / coreness[-1]  // calculate the picked number of each layer

for c in coreness[-1]
  c = degree(graph, c) // desc order
  count = length(c) * rate

  for pick in picked  // get picked neighbors
    picked_neighbors = neighbors(graph, pick)
  end for
  picked_neighbors = difference(unique(picked_neighbors), picked)

  for node in c  // calculate the number of the intersection of neighbors
    scores[node] = length(intersect(neighbors(graph, node), picked_neighbors))
  end for

  scores = order(scores, asc)
  picked = scores[1: count]
end for
return picked
```



(a) (b) (c)

Figure 2: Using K-Coverage to select monitor nodes in Karate network.

In order to quantify the effectiveness of monitor node selection, the area under the receiver operating characteristic curve (ROC), or AUC, is used as the evaluation metric to evaluate the accuracy of the source localization algorithm, and indirectly to evaluate the effectiveness of the monitor node selection method by the accuracy of the source localization algorithm. AUC value is calculated by true positive rate (TPR) and false positive rate (FPR). In order to calculate AUC value, Spearman correlation coefficient ρ

Table 1: Simulation experiment datasets.

Dataset	Node	Edge	Clustering Coefficient	Average Degree
Karate	34	78	0.2557	4.5882
Lesmis	77	254	0.4989	6.5974
RFID	75	32424	0.588	30.3733
Jazz	198	2742	0.5203	27.697
USairports	745	4618	0.3385	12.3973
Immuno	1316	6300	0.4851	9.5745
Yeast	2617	11855	0.4687	9.8467
Facebook	4039	88234	0.5192	43.691
USPG	4941	6594	0.1032	2.6691

should be ranked in descending order.

$$TPR = \frac{TP}{k} \tag{4}$$

where TP is the number of real sources in ρ and k is the number of real sources, and since this algorithm is a single source localization algorithm, TP can only be 0 or 1 and k is 1.

$$FPR = \frac{FP}{n-k} \tag{5}$$

where FP is the number of false positives in ρ, which is the number of error source nodes, and n is the number of Spearman correlation coefficient ρ. The horizontal coordinate of the receiver operating characteristic curve is FPR, the vertical coordinate is TPR, and AUC is the area under the curve.

4.2 Results

The AUC of the experimental results is the average AUC value of 100 simulations.

Figure 3 demonstrates the accuracy of using Spearman correlation coefficients for localization with 10% of the monitor nodes selected. Both K-Coverage and K-Coverage-Nei show a large improvement in accuracy when compared to the centrality selection methods. It can also be found that the effectiveness of using centrality to select monitor nodes is closely related to the network structure, such as in Lesmis and RFID networks, the use of betweenness centrality to select monitor nodes will have good results. It is also found that the effectiveness of K-Coverage and K-Coverage-Nei has a strong relationship with the average degree value of the network, the higher the average degree value of the network, the better the effectiveness of using K-Coverage and K-Coverage-Nei for monitor node selection.

In order to investigate the relationship between the select proportion of the monitor nodes and the source localization accuracy, we set three select proportion of 1%, 5%, and 10%, respectively. Figure 3 shows the relationship between the select proportion and the source localization accuracy of different monitor nodes. It can be seen that with the increase of the select proportion, the accuracy of source localization is also constantly improving. At the same time, in the case where the select proportion is small, such as 1%, both K-Coverage and K-Coverage-Nei achieve good results, much better than the method using centrality selected monitor nodes. There is also an interesting experimental phenomenon, a higher proportion of monitor nodes, random selection of monitor nodes on individual networks also achieved good results. This is because the random method used by us is not really random, but a kind of average sampling,

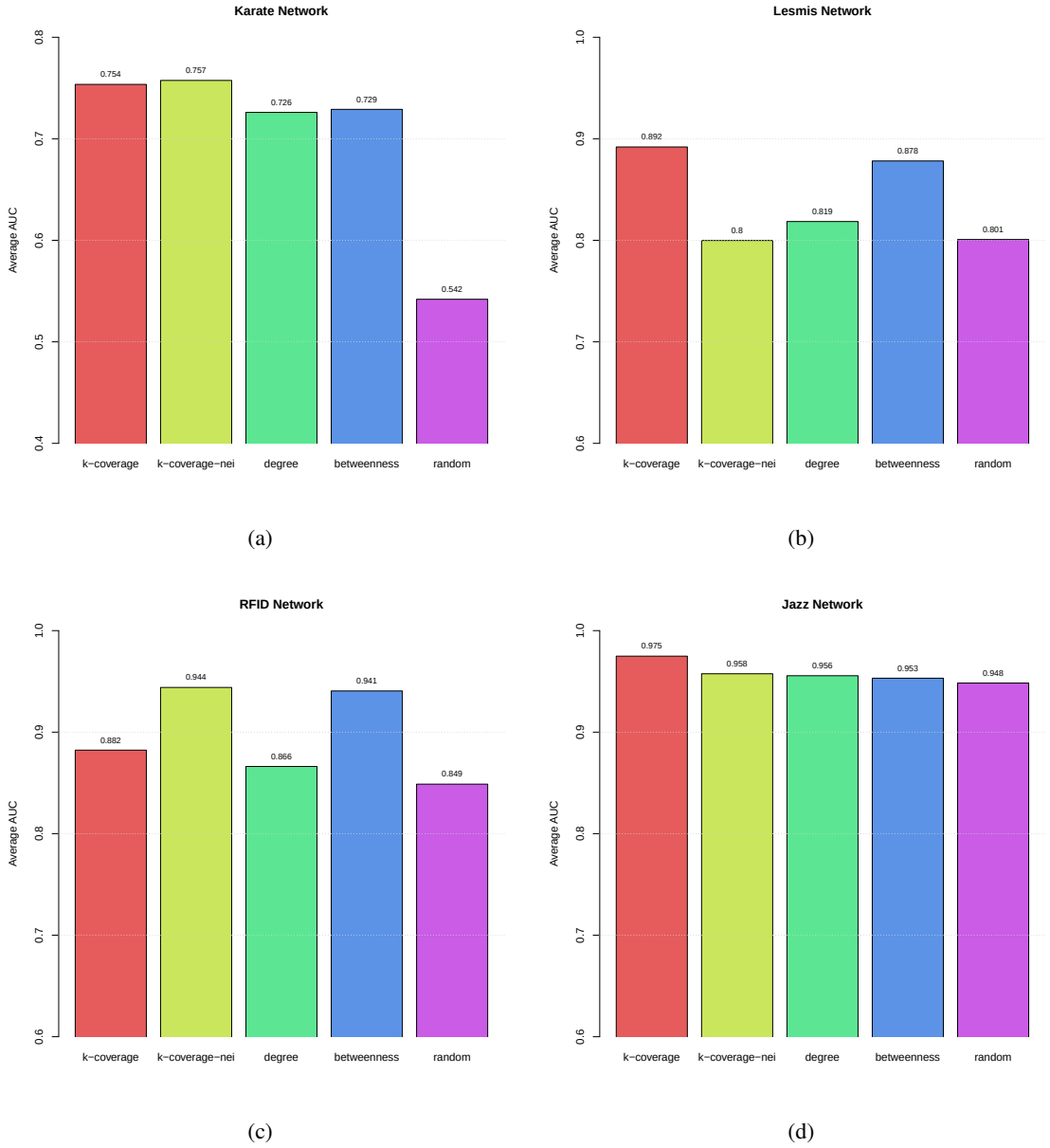


Figure 3: Influence of different selection methods on localization accuracy.

so that the selected monitor nodes are naturally evenly distributed in the whole network. Of course, in the case of a small percentage of selections, the results are not as good as using the centrality selection method.

Finally, Table (2) and Figure (5) summarize the experimental results. Except for individual networks, K-Coverage-Nei is better than K-Coverage.

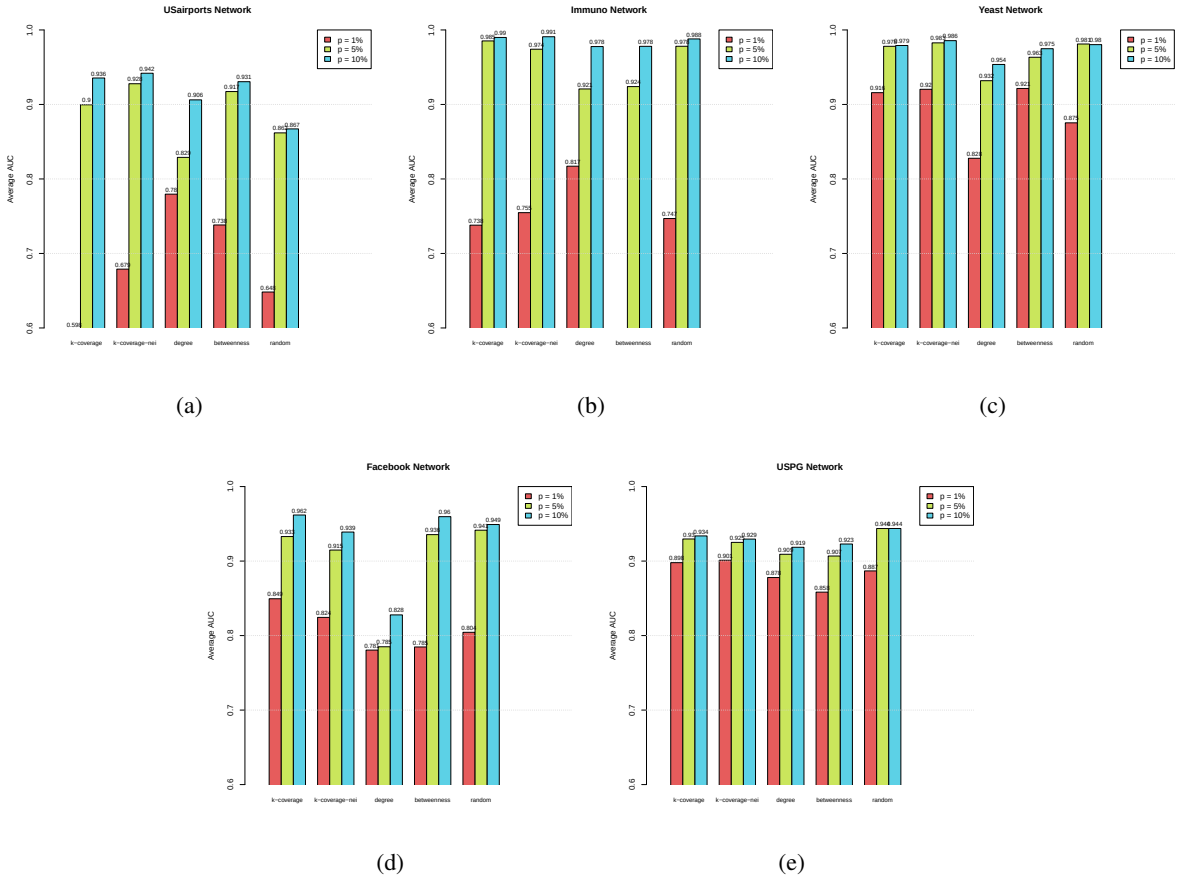


Figure 4: Influence of different select proportion on localization accuracy.

5 Conclusion

With the application of infection source localization problems in various industries, it is becoming more and more important to improve the accuracy of localization. An infection source localization method based on monitor observation has the advantage of being more accurate and faster than a probability-based method. In order to fully amplify this advantage, the placement of monitor nodes also needs to be studied in depth. In this paper, in order to solve the unevenness of the monitor nodes selected by the traditional graph centrality, the K-Coverage method is used to select the monitor nodes hierarchically, and the K-coverage method is proposed to distribute the monitor nodes in each region of the graph. At the same time, in order to solve the blind spots between layers, the overlapping range of neighbors is considered, and the K-Coverage-Nei method is proposed to further improve the effectiveness of K-Coverage.

References

[1] J. Clement. Social network statistics. `https://www.statista.com/topics/1164/socialnetworks/` [Online; Accessed on November 30, 2020].

[2] W. Dong, W. Zhang, and C. W. Tan. Rooting out the rumor culprit from suspects. In *Proc. of the 2013 IEEE International Symposium on Information Theory (ISIT'13), Istanbul, Turkey*, pages 2671–2675. IEEE, October 2013.

Table 2: Summary of simulation results.

Dataset	Select Proportion	K-Coverage	K-Coverage-Nei	Degree	Betweenness	Random
Karate	10%	0.7535615	**0.7574772**	0.7260000	0.7290000	0.5419299
Lesmis	10%	**0.8919413**	0.7998499	0.8185000	0.8781585	0.8007992
RFID	10%	0.8822952	**0.9440410**	0.8662069	0.9406034	0.8488013
Jazz	10%	**0.9748529**	0.9575157	0.9555732	0.9531526	0.9483753
	1%	0.5982787	0.6789572	**0.7795875**	0.7382372	0.6482445
USairports	5%	0.8995430	**0.9278056**	0.8289572	0.9173197	0.8618422
	10%	0.9355006	**0.9419053**	0.9061408	0.9305447	0.8671718
	1%	0.7379758	0.7547713	**0.8171271**	0.2584064	0.7468839
Immuno	5%	**0.9852690**	0.9741313	0.9207929	0.9239819	0.9780926
	10%	0.9899276	**0.9908897**	0.9777086	0.9780114	0.9879333
	1%	0.9158223	0.9203361	0.8277513	**0.9214603**	0.8753172
Yeast	5%	0.9780456	**0.9826621**	0.9318824	0.9633456	0.9811325
	10%	0.9791235	**0.9856307**	0.9536808	0.9748262	0.9802944
	1%	**0.8494117**	0.8243977	0.7805063	0.7846915	0.8043920
Facebook	5%	0.9329196	0.9147816	0.7849834	0.9355050	**0.9414378**
	10%	**0.9618247**	0.9389648	0.8277513	0.9596012	0.9490543
	1%	0.8978628	**0.9011372**	0.8779126	0.8583064	0.8867819
USPG	5%	0.9296420	0.9250907	0.9091270	0.9067817	**0.9436781**
	10%	0.9337446	0.9294991	0.9185217	0.9228880	**0.9437358**

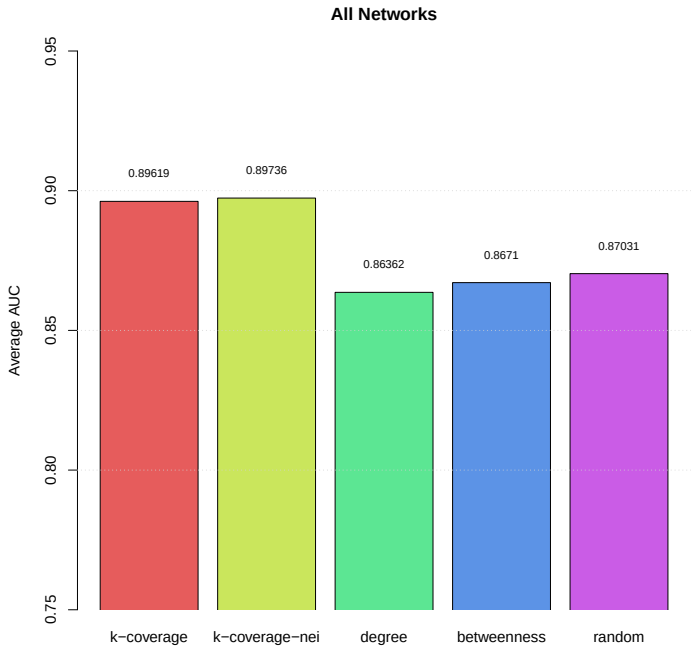


Figure 5: Average AUC values of different selection methods in all networks.

[3] Z.-L. Hu, Z. Shen, C.-B. Tang, B.-B. Xie, and J.-F. Lu. Localization of diffusion sources in complex networks with sparse observations. *Physics Letters A*, 382(14):931–937, April 2018.

[4] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, 19(1):465–481, 2016.

[5] Z. Li, C. Xia, T. Wang, and X. Liu. Which node properties identify the propagation source in networks? In *Proc. of the International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP'19), Melbourne, Australia*, volume 11944 of *Lecture Notes in Computer Science*, pages 256–270. Springer, Cham, December 2019.

[6] A. Louni and K. Subbalakshmi. A two-stage algorithm to estimate the source of information diffusion in social media networks. In *Proc. of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS'14), Toronto, ON, Canada*, pages 329–333. IEEE, July 2014.

[7] A. Louni and K. 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*, 5(2):335–343, February 2018.

[8] V. Melnyk and I. Styopochkina. Malicious information source detection in social networks. *Theoretical and Applied Cybersecurity*, 1(1):75–81, May 2019.

[9] R. Paluch, X. Lu, K. Suchecki, B. K. Szymański, and J. A. Hołyst. Fast and accurate detection of spread source in large complex networks. *Scientific reports*, 8(1):1–10, February 2018.

[10] P. C. Pinto, P. Thiran, and M. Vetterli. Locating the source of diffusion in large-scale networks. *Physical review letters*, 109(6):068702, August 2012.

[11] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8):5163–5181, August 2011.

[12] V. Sharma, I. You, and G. Kul. Socializing drones for inter-service operability in ultra-dense wireless networks using blockchain. In *Proc. of the 2017 ACM CCS International Workshop on Managing Insider Security Threats (ACM CCS MIST'17), Dallas, USA*, pages 81–84. ACM, October 2017.

[13] V. Sharma, I. You, F.-Y. Leu, and M. Atiquzzaman. Secure and efficient protocol for fast handover in 5g mobile xhaul networks. *Journal of Network and Computer Applications*, 102:38–57, January 2018.

[14] S. Shelke and V. Attar. Origin identification of a rumor in social network. In *Cybernetics, Cognition and Machine Learning Applications*, pages 89–96. Springer, 2020.

[15] C. Shi, Q. Zhang, and T. Chu. Source identification of network diffusion processes with partial observations. In *Proc. of the 36th Chinese Control Conference (CCC'17), Dalian, China*, pages 11296–11300. IEEE, July 2017.

[16] F. Song, Z. Ai, Y. Zhou, I. You, K.-K. R. Choo, and H. Zhang. Smart collaborative automation for receive buffer control in multipath industrial networks. *IEEE Transactions on Industrial Informatics*, 16(2):1385–1394, October 2019.

[17] B. Spinelli, L. E. Celis, and P. Thiran. A general framework for sensor placement in source localization. *IEEE Transactions on Network Science and Engineering*, 6(2):86–102, December 2017.

[18] H. Wang. An universal algorithm for source location in complex networks. *Physica A: Statistical Mechanics and its Applications*, 514:620–630, January 2019.

[19] S. Xu, C. Teng, Y. Zhou, J. Peng, Y. Zhang, and Z.-K. Zhang. Identifying the diffusion source in complex networks with limited observers. *Physica A: Statistical Mechanics and its Applications*, 527:121267, August 2019.

[20] K. Zhu and L. Ying. Information source detection in the sir model: A sample-path-based approach. *IEEE/ACM Transactions on Networking*, 24(1):408–421, November 2014.

Author Biography

Yuexin Zhang is a graduate student of China University of Mining and Technology, majoring in software engineering technology. His research interest lies at the rumor source localization of social network.

Jianjun Zhang graduated from Dalian University of Technology, majoring in civil engineering. He is an ITC railroad convergence expert with extensive railroad knowledge and abundunt railroad operation and management experiences.