Privacy-Preserving Vital Node Identification in Complex Networks: Evaluating Centrality Measures under Limited Network Information

1st Diallo, Diaoulé
Institute for Software Technology
German Aerospace Center (DLR)
Cologne, Germany
diaoule.diallo@dlr.de

2nd Hecking, Tobias
Institute for Software Technology
German Aerospace Center (DLR)
Cologne, Germany
tobias.hecking@dlr.de

***Abstract*—Identifying vital nodes in complex networks is pivotal across various research domains such as social network analysis, epidemiology, and physics, with centrality measures being commonly employed. Despite growing privacy concerns, its impact on vital node identification, especially in networks with sensitive data like Bluetooth-based contact networks, remains underexplored. Our study assesses centrality measures' efficacy under constrained privacy settings, where only limited neighbor information is accessible. Through simulations, we pinpoint algorithms optimal for privacy-sensitive node vitality estimation, emphasizing the influence of network characteristics and algorithmic traits like multi-aggregation. This work enhances the understanding of privacy-centric methods in complex network analysis.**

***Index Terms*—vital node identification, influential node ranking, network centrality measures, complex networks analysis, privacy-sensitive network analysis, epidemic modeling and analysis,**

I. Introduction

Vital nodes in complex networks, crucial for network function and structure, are central to numerous domains including information propagation [1], power grid analysis [2], economics [3], and especially infectious disease modeling, accentuated by the COVID-19 pandemic [4]–[6]. Determining these nodes aids in predicting and managing disease spread and individual infection risks.

Mobile Bluetooth-based contact tracing apps, such as the German "Corona-Warn-App" [7], offer personal infection risk estimates. However, privacy concerns have restricted data scope to immediate neighborhood contacts, complicating precise risk calculations. This constraint signals the need to understand node vitality estimations under such limited data scopes. Our research probes the balance between precision and privacy in node vitality estimations, evaluating recent methodologies across diverse network datasets under different privacy constraints. Results indicate that merely incorporating additional edge information often diminishes estimation accuracy while considering second-degree neighbors can approach the precision of full network analyses. We commence with a background on node vitality and privacy, subsequently detailing our experimental setup. The latter sections present our results and their broader implications.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://dx.doi.org/10.1145/3625007.3627513

II. Background

The following sections introduce the concept of node vitality as well as the role of privacy in node vitality estimation algorithms.

A. Estimating Node Vitality

In unweighted and undirected networks, the vitality of a node is determined by the network topology and represents the spreading potential as well as the receptive potential of a node. This potential refers to any kind of information (infection spreading, fake news spreading, etc.). The importance of a node is typically described by network centrality measures. Among the classical methods are degree and path aggregating measures such as betweenness, closeness, eigenvector centrality, and its relatives [8]. Different attempts have been made to categorize modern approaches to vital node identification [9], [10]. Given that this work investigates data demand and performance of node vitality estimation, we roughly follow the categorization of [10] into local, semi-local, global, and hybrid approaches. Section III, introduces several methods for each category.

B. Privacy in node vitality estimation

As stated in the introduction there is a trade-off between the accuracy of node vitality estimation and the requisite amount of potentially sensitive network information. For instance, social networks contain sensitive personal information, and

the analysis of these networks must adhere to strict privacy regulations. Similarly, Bluetooth-based contact networks for fighting pandemics may, as a negative side effect, reveal large parts of societal structures that do not comply with data protection regulations. Thus, privacy and data protection in such scenarios are paramount.

[11] presents a secure multiparty computation protocol for a scenario where some individuals in the network treat their data as private while others do not. The authors design protocols for popular methods such as K-shell decomposition, ensuring privacy-preserving computation. In contrast to removing nodes from the graph globally, this work investigates the unique challenges of vital node estimation within limited egocentric sub-graphs. [12] proposes the secure multiparty computation ranking (SMPC-ranking) protocol, which enables participants in a network to collaboratively identify influential nodes while preserving the privacy of each private network. While our study concentrates on vitality estimation within the confines of limited egocentric network information, whereas SMPC-ranking explores vital node identification within a context where multiple private sub-graphs collaboratively contribute to the estimation.

To the best of our knowledge, comprehensive evaluations of multiple vital node estimation algorithms within a node egocentric network visibility context are scarce. We try to fill this gap by examining the performance of these algorithms under stringent privacy conditions, thereby highlighting their strengths and limitations.

III. Experimental setup

The following section presents the setup for the simulation-based performance evaluation of multiple vitality estimation algorithms. Different privacy settings are defined and the typical SIR-based evaluation method for vitality estimation is described. Subsequently, evaluated algorithms are introduced.

A. Computing ground truth vitality

Consider a network represented by $G = (V, E)$, where V denotes the set of nodes, E represents the set of edges and N is the number of nodes. Within this network, let $\langle k \rangle$ denote the average degree of first-order neighboring nodes and $\langle k^2 \rangle$ is the average degree of second-degree neighbors. To obtain ground truth for evaluating node vitality estimation under varied privacy, we utilize the SIR spreading model [13], as it is commonly used for vitality estimation [14]–[16]. Initially, every network node is *susceptible* (S) with a sole index node as *infected* (I). Neighbors of the index risk infection at rate β, transitioning to *recovered* (R) post-infection. To assure statistical stability, each node is considered an index node 1000 times. A node i's vitality value $Vitality(i)$ is defined as $Vitality(i) = \frac{N_{R_i}}{N}$, where N_{R_i} is the number of recovered nodes. The transmission rate β typically aligns with or slightly exceeds a network's epidemic threshold β_{th} [9], [10], [17], computed via degree distributions as

$$\beta_{th} \approx \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}.$$

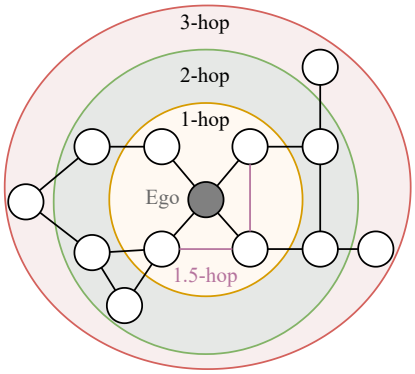

Fig. 1. Different configurations of a "horizon" of ego, i.e. its neighborhood degree.

In this research, we chose $\beta = 1.25 \times \beta_{th}$ to ensure sufficient spreading dynamics. Simulations were executed using the Epidemics on Networks Python library [18]. Algorithm performance was gauged by the correlation between the algorithm's rankings and the SIR-derived ground truth, employing Kendall's tau coefficient τ [19], with a range of -1 to $+1$ indicating correlation strength.

B. Vital node estimation under limited local network information

Traditional vital node identification studies often assume full network access. However, practical scenarios, driven by privacy concerns or platform restrictions, limit this. For instance, many social media platforms allow viewing only up to friends of friends. Similarly, decentralized contact tracing apps often grant users visibility only to their immediate neighbors [20]. In such contexts, a pressing question emerges: How does limiting data sharing to direct and indirect neighborhoods (e.g., 2-hop, 3-hop) impact the accuracy of vital node estimation?

To address this, we tested various node vitality estimation algorithms under different ego-centric network visibilities or ego horizons (see Fig. 1). Scenarios ranged from observing solely first-order neighbors (1-hop), extending to 1.5-hop, 2-hop, 3-hop, and full graph visibility (also referred to as privacy settings). In the 1.5 hop scenario neighbours of ego only share their links to nodes that ego already knows, which is acceptable in certain privacy scenarios. Even methods typically calculated on the entire graph, like eigenvector centrality, were adapted to these sub-graphs. Our exploration into limited network visibility's effects offers insights into node vitality estimation's reliability under privacy constraints. These insights are crucial for developing robust, privacy-aware contact tracing systems during times of infectious disease outbreaks.

C. Vitality estimation algorithms

Local vitality estimation algorithms focus solely on the 1-hop neighborhood of a node, making them efficient with low computational costs, but they often lack accuracy due to their limited scope. For instance, in undirected and unweighted

TABLE I
ALL EVALUATED VITALITY ESTIMATION ALGORITHMS.

Algorithm	degree	K-shell	K-shell iteration	entropy	neighb. aggreg.	others
LGC [14]	✓					
NINL [21]	✓				✓	
ERM [15]	✓			✓		
CLD [25]	✓					clust. coeff.
LH-Index [26]	✓				✓	
GC [16]		✓				
GC_+ [16]		✓			✓	
IGC [27]	✓	✓				
IGC_+ [27]	✓	✓			✓	
DKGC [22]	✓	✓	✓			
$DKGC_+$ [17], [22]	✓	✓	✓		✓	
MCGC [17]	✓	✓				eigenvec.
IC [28]	✓	✓	✓		✓	
GLSC [29]	✓	✓			✓	neighb. set
GLI [30]	✓	✓	✓			
MCDE [31]	✓	✓		✓		
MCDWE [31]	✓	✓		✓		
ECRM [23]	✓		✓	✓	✓	
LS [32]	✓	✓				neighb. set

networks, the node degree becomes the sole metric. Semi-local algorithms, on the other hand, utilize a predefined truncation threshold to determine the number of considered nodes. An exemplar is the NINL [21] approach, which incorporates additive neighbor layer information, aggregating nodes' influence iteratively, and offering a representation of the network's spreading dynamics. Global approaches encompass the full network, deriving vitality from its entire structure. This can increase computational costs, especially for expansive networks. A notable example is the improved K-shell decomposition [22], [23], which builds upon the well-known K-shell method [24] by including the iteration number of removement. Lastly, hybrid methodologies merge both global and local perspectives. They assimilate data from the entire network and from immediate surroundings. The Gravity Centrality (GC) is a hybrid approach, factoring in the K-shell value to compute node importance within a specified truncation radius. Table I displays all algorithms utilized in this study, along with the network measures and principles upon which they are founded. As explained in III-B, all algorithm types will be evaluated with limited neighborhood information to verify their ability to determine node vitality without full network access.

IV. RESULTS

To ensure a thorough assessment, we used diverse datasets of 12 networks, namely Jazz, Email, Power, USAir, Router, Dolphin [33], French-school [34], Network science, Infectious, Contiguous [35], Celegans [36] and Sfhh [37]) to evaluate the performance of the 19 state-of-the-art vitality estimation algorithms shown in Table I under different privacy conditions.

A. Overall ranking of algorithms

The average rankings of the evaluated algorithms for each privacy setting across all networks are depicted in Table II.

TABLE II
AVERAGE RANKS OF ALGORITHMS ACCORDING TO KENDALLS τ WITH GROUND TRUTH VITALITY OVER ALL NETWORKS. TOP 3 IN **BOLD**; '-' DENOTES INFEASIBILITY OF EXECUTION.

Method	1-hop	1.5-hop	2-hop	3-hop	full-graph
IGC+	9.0	8.0	4.0	**1.0**	**1.0**
ERM	-	17.0	**2.0**	**3.0**	**2.0**
NINL	9.0	**3.0**	**1.0**	**2.0**	**3.0**
GC+	9.0	13.0	7.0	5.0	4.0
ECRM	9.0	10.0	6.0	6.0	5.0
DKGC+	9.0	9.0	**3.0**	4.0	6.0
MCGC	9.0	19.0	19.0	19.0	7.0
IGC	9.0	12.0	5.0	7.0	8.0
DKGC	9.0	4.0	9.0	8.0	9.0
GC	9.0	14.0	8.0	9.0	10.0
LGC	9.0	16.0	11.0	12.0	11.0
LH-Index	9.0	6.0	13.0	10.0	12.0
CLD	9.0	15.0	10.0	11.0	13.0
GLI	9.0	5.0	12.0	13.0	14.0
GLSC	9.0	7.0	18.0	16.0	15.0
IC	9.0	11.0	14.0	18.0	16.0
MCDE	9.0	**2.0**	16.0	15.0	17.0
MCDWE	9.0	**1.0**	15.0	14.0	18.0
LS	-	18.0	17.0	17.0	19.0

A pattern emerges under the full-graph and 3-hop conditions, with IGC_+, ERM, and NINL consistently appearing in the top three algorithmic positions. When the privacy setting shifts to 2-hop, the rankings adjust, showcasing NINL, ERM, and $DKGC_+$ as the dominant three algorithms. This means that NINL seems to achieve the best balance between the amount of network information needed and performance. As all algorithms solely rely on degree information in the 1-hop setting, they produce identical estimations. When examining the transition from a full-graph to a 3-hop setting, the algorithmic rankings show minor fluctuations in the produced rankings. Surprisingly in the 1.5 setting, no method performs better than simple node degree. This will be discussed in more detail below.

B. Results on Single Networks

Fig. 2 shows Kendall's τ correlations with the SIR-based ground truth for the top five methods for each network, with privacy settings displayed on the x-axis. As already shown above overall ECRM and IGC_+ perform best when the 3-hop neighborhood or full graph is visible to the ego node, while the differences become more subtle in the more restrictive settings with NINL at the top rank in most 2-hop cases. The performance of all methods increases only marginally when transitioning from the 2-hop to the 3-hop and full-graph settings. In specific cases such as the French-school, Celegans, and USAir networks, this is due to their small network diameters, where the network diameter is 3 or smaller. However, for networks with larger path lengths, considering the 2-hop neighborhood of an ego node already provides enough information for robust vitality estimation. NINL performs best in 7 out of the 12 networks in the 2-hop setting. Interestingly, although ECRM emerges as a high-performing algorithm in numerous instances, it isn't featured in the top four rankings under any of the privacy conditions as seen in Table II. This

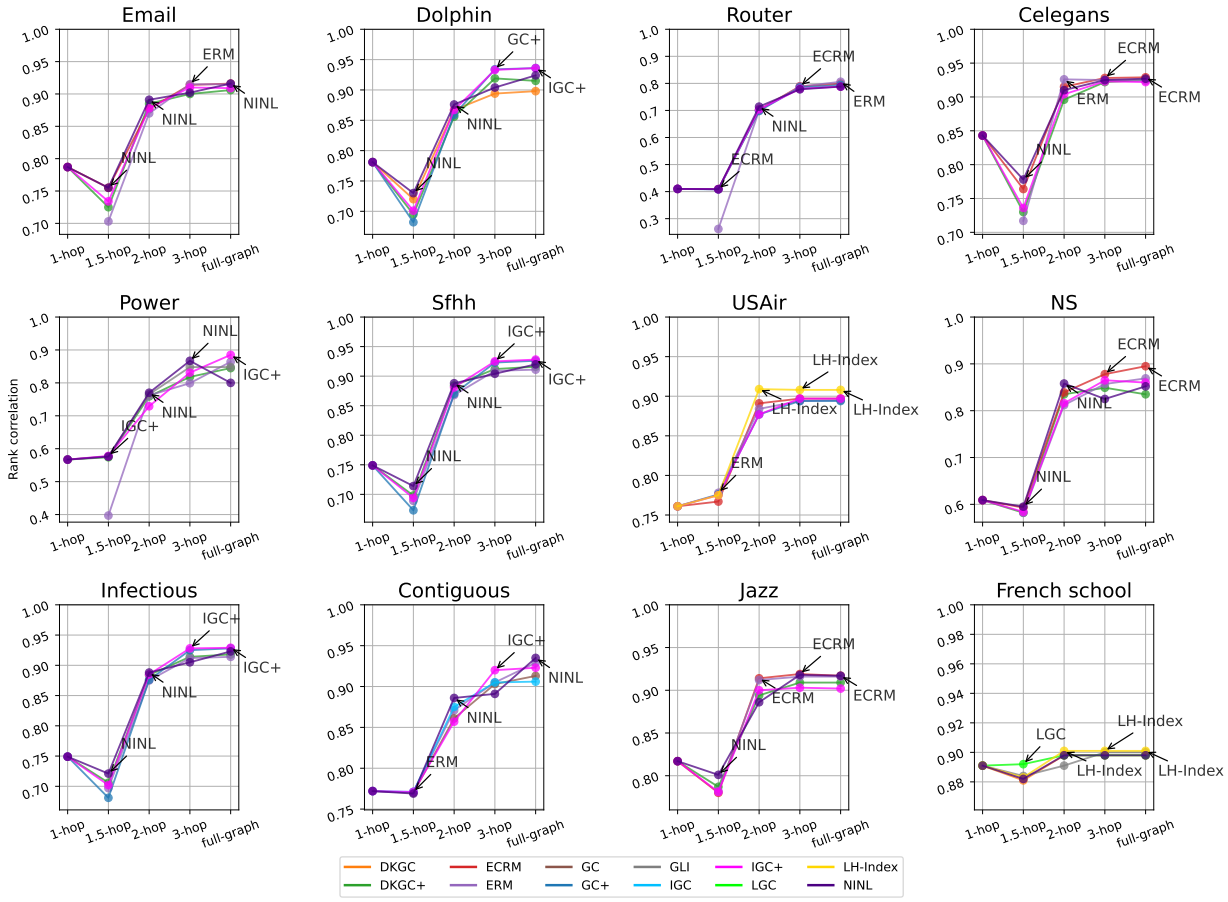

Fig. 2. Rank correlations of the top 5 vitality estimation algorithms across different networks. For each privacy setting (x-axis), the best-performing approach is annotated. The transmission rate is given by $\beta = \beta_{th} \times 1.25$.

underscores the observation that an algorithm's effectiveness can be highly dependent on the specific network, implying that what works well in one network may not necessarily replicate its performance in another. Another notable observation is that the rank correlations drop in almost all cases from the 1-hop to the 1.5-hop setting. This means that in the scenario where the edges between the neighbors of a node are known (which might be acceptable regarding privacy), these additional edges rather introduce less reliable information for node vitality estimation.

C. Impact of algorithm properties

When applying privacy constraints to networks, the available information for vitality estimation algorithms diminishes. The K-shell decomposition, when used in an algorithm within these constraints, sees a notable reduction in the K-shells assigned to nodes. In contrast, K-shell iteration retains more distinct values under privacy constraints. This affects algorithm rankings: while GC+ sees a drop as privacy settings tighten, DKGC+, using both K-shell values and iteration numbers, climbs from the 6th in full-graph to the 3rd position in 2-hop. Top algorithms in the 2-hop to full-graph settings generally adopt aggregation steps in a message-passing manner. For instance, NINL determines influence values by aggregating neighboring degrees iteratively, facilitating continuous information exchange between nodes. Given that frontrunners ERM and ECRM also use multi-aggregation, this method is effective across various networks and privacy conditions. Lastly, algorithms like ERM, ECRM, MCDE, and MCDWE utilize entropy principles for vitality estimation. While ECRM excels in specific networks, ERM typically surpasses others in an average context across all networks.

D. Implications for privacy-sensitive node vitality estimation

Our analysis reveals several critical findings with implications for privacy-sensitive node vitality estimation. For most networks, the supplementary edges provided in the 1.5-hop setting do not offer a significant advantage over the 1-hop setting. This suggests that in contexts such as estimating node vitality in contact networks, like in contact tracing applications, employing mechanisms that reveal edges between common contacts would not provide any notable advantages. Interestingly, access to just the 2-hop neighborhood produces results almost on par with having access to the entire graph. This balance between privacy and efficacy is essential, especially during scenarios demanding swift interventions like pandemics. Our

research also highlights the importance of neighbor aggregation, emphasizing the role of message-passing in estimating spreading dynamics, especially with restricted network visibility. Overall, our research illuminates a new and privacy-aware perspective on node vitality estimation, providing important insights for both academics and practitioners in the field.

V. Conclusion

Our investigation of privacy-sensitive node vitality estimation has revealed key insights for the domain. Notably, while extending to a 1.5-hop setting might not be as effective, the 2-hop neighborhood provides results almost equivalent to full graph access. This suggests a feasible balance between preserving privacy and ensuring accurate network analysis. Emphasizing the role of neighbor aggregation, our findings underscore the significance of the message-passing approach, particularly with limited network data. In conclusion, our research provides a valuable basis for future studies aiming to bridge the gap between efficient node vitality estimation and privacy preservation. The insights drawn can inform the design of applications that are both conscious of privacy concerns and capable of pinpointing vital nodes. Future research should delve deeper into understanding the interplay between network attributes and the required information for trustworthy vitality predictions, especially within privacy-focused algorithms.

References

[1] W. Xu, T. Li, W. Liang, J. X. Yu, N. Yang, and S. Gao, "Identifying structural hole spanners to maximally block information propagation," *Information Sciences*, vol. 505, pp. 100–126, 2019.

[2] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the north american power grid," *Physical Review E*, vol. 69, no. 2, p. 025103, 2004.

[3] A. Garas, P. Argyrakis, C. Rozenblat, M. Tomassini, and S. Havlin, "Worldwide spreading of economic crisis," *New Journal of Physics*, vol. 12, no. 11, p. 113043, 2010.

[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[5] S. S. Chaharborj, K. N. Nabi, K. L. Feng, and et al, "Controlling COVID-19 transmission with isolation of influential nodes," *Chaos, Solitons, and Fractals*, vol. 159, p. 112035, 2022.

[6] Z. Dong, Y. Chen, T. S. Tricco, C. Li, and T. Hu, "Hunting for vital nodes in complex networks using local information," *Scientific Reports*, vol. 11, p. 9190, 2021.

[7] B. Sowmiya, V. Abhijith, S. Sudersan, R. Sakthi Jaya Sundar, M. Thangavel, and P. Varalakshmi, "A survey on security and privacy issues in contact tracing application of covid-19," *SN computer science*, vol. 2, pp. 1–11, 2021.

[8] S. P. Borgatti, "Centrality and network flow," *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[9] A. Asgharian Rezaei, J. Munoz, M. Jalili, and H. Khayyam, "A machine learning-based approach for vital node identification in complex networks," *Expert Systems with Applications*, vol. 214, p. 119086, 2023.

[10] A. Zareie, A. Sheikhahmadi, and M. Jalili, "Influential node ranking in social networks based on neighborhood diversity," *Future Generation Computer Systems*, vol. 94, pp. 120–129, 2019.

[11] V. B. Kukkala and S. Iyengar, "Identifying influential spreaders in a social network (while preserving privacy)," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 2, pp. 537–557, 2020.

[12] W. Hu, X. Xia, X. Ding, X. Zhang, K. Zhong, and H.-F. Zhang, "SMPC-ranking: A privacy-preserving method on identifying influential nodes in multiple private networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2022.

[13] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[14] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, and T. Zhou, "Identifying influential spreaders by gravity model," *Scientific Reports*, vol. 9, no. 1, p. 8387, 2019.

[15] A. Zareie, A. Sheikhahmadi, and A. Fatemi, "Influential nodes ranking in complex networks: An entropy-based approach," *Chaos, Solitons & Fractals*, vol. 104, pp. 485–494, 2017.

[16] L.-L. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, "Identifying influential spreaders in complex networks based on gravity formula," *Physica A: Statistical Mechanics and its Applications*, vol. 451, pp. 205–212, 2016.

[17] Z. Li and X. Huang, "Identifying influential spreaders by gravity model considering multi-characteristics of nodes," *Scientific Reports*, vol. 12, no. 1, p. 9879, 2022.

[18] J. C. Miller and T. TIng, "EoN (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks," *Journal of Open Source Software*, vol. 4, no. 44, p. 1731, 2019.

[19] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1, pp. 81–93, 1938.

[20] J. Li and X. Guo, "COVID-19 contact-tracing apps: a survey on the global deployment and challenges," *arXiv preprint arXiv:2005.03599*, 2020.

[21] J. Zhu and L. Wang, "Identifying influential nodes in complex networks based on node itself and neighbor layer information," *Symmetry*, vol. 13, no. 9, p. 1570, 2021.

[22] Z. Li and X. Huang, "Identifying influential spreaders in complex networks by an improved gravity model," *Scientific Reports*, vol. 11, no. 1, p. 22194, 2021.

[23] A. Zareie, A. Sheikhahmadi, M. Jalili, and M. S. K. Fasaei, "Finding influential nodes in social networks based on neighborhood correlation coefficient," *Knowledge-Based Systems*, vol. 194, p. 105580, 2020.

[24] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[25] S. Gao, J. Ma, Z. Chen, and et al., "Ranking the spreading ability of nodes in complex networks based on local structure," *Physica A: Statistical Mechanics and its Applications*, vol. 403, pp. 130–147, 2014.

[26] Q. Liu, Y.-X. Zhu, Y. Jia, and et al., "Leveraging local h-index to identify and rank influential spreaders in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 512, pp. 379–391, 2018.

[27] J. Wang, C. Li, and C. Xia, "Improved centrality indicators to characterize the nodal spreading capability in complex networks," *Applied Mathematics and Computation*, vol. 334, pp. 388–400, 2018.

[28] Z. Wang, Y. Zhao, J. Xi, and C. Du, "Fast ranking influential nodes in complex networks using a k-shell iteration factor," *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 171–181, 2016.

[29] H. Hu, Z. Sun, F. Wang, L. Zhang, and G. Wang, "Exploring influential nodes using global and local information," *Scientific Reports*, vol. 12, no. 1, p. 22506, 2022.

[30] Y.-Z. Yang, M. Hu, and T.-Y. Huang, "Influential nodes identification in complex networks based on global and local information," *Chinese Physics B*, vol. 29, no. 8, p. 088903, 2020.

[31] A. Sheikhahmadi and M. A. Nematbakhsh, "Identification of multi-spreader users in social networks for viral marketing," *Journal of Information Science*, vol. 43, no. 3, pp. 412–423, 2017.

[32] Y.-P. Wan, J. Wang, D.-G. Zhang, H.-Y. Dong, and Q.-H. Ren, "Ranking the spreading capability of nodes in complex networks based on link significance," *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 929–937, 2018.

[33] R. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[34] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PLOS ONE*, vol. 10, no. 9, p. e0136497, 2015.

[35] J. Kunegis, "KONECT: the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13 Companion. Association for Computing Machinery, 2013, pp. 1343–1350.

[36] M. Kaiser and C. C. Hilgetag, "Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems," *PLOS Computational Biology*, vol. 2, no. 7, p. e95, 2006.

[37] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?" *EPJ Data Science*, vol. 7, no. 1, pp. 1–18, 2018.