Unravelling the Memetic Mysteries of Reddit: The Power of Eigen Central Nodes in Spreading the Laughs

Arju Yadav
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
a.yadav@se22.qmul.ac.uk

Arko Chatterjee
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
a.chatterjee@se22.qmul.ac.uk

Eleanor Prashamshini
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
e.prashamshini@se22.qmul.ac.uk

Riya Dodthi
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
r.dodthi@se22.qmul.ac.uk

***Abstract*—This paper investigates the role of eigenvector centrality in modelling the diffusion of memes over social media networks in comparison to betweenness centrality. Eigenvector centrality and betweenness centrality are widely used network metrics that measure the importance of a node. By conducting simulations, we demonstrate that centrality is crucial in the propagation of memes in social networks. Specifically, our results suggest that nodes exhibiting high eigenvector centrality are more prone to becoming "super-spreaders" of memes, as they are more likely to kickstart the dissemination process and play an influential part in its spread. We further explore the impact of change in network structures on the spread of memes and show that targeting nodes with high eigenvector centrality can significantly enhance the spread of memes in certain network structures. Our findings highlight the importance of eigenvector centrality as a tool for predicting and controlling the spread of memes over social media and have implications for marketers, advertisers, and social media platforms seeking to maximise the impact of their campaigns.**

***Index Terms*—Eigenvector Centrality, Community Detection, Network Analysis, Meme spreading, SI modeling**

I. Introduction

Memes have become a ubiquitous presence on the internet, often compared to viral infections that spread rapidly through social networks. One such platform where memes thrive is Reddit, a popular social network where users share content on a variety of topics called subreddits. The posts within a subreddit creates a hyperlink to other related subreddits, thereby creating an interconnected network. Within this network, users tend to form communities that share information and memes more closely than with others. This has significant implications on how cultures influence transmission of information among communities and ultimately shape the internet as we know it.

To understand the dynamics of the communities on Reddit and the role of central nodes in the spreading of memes, different centrality measures such as betweenness, closeness, and eigenvector centrality can be explored. In this paper, we focus on eigenvector centrality and betweennness centrality, and their role in the spreading of memes in the network. By analysing the centrality of subreddits and their connections, we aim to gain insights into the structure and dynamics of these communities and how they influence the spread of memes.

II. Related work

The phenomenon of transmission of information among communities has long been of interest in social network analysis. A fundamental aspect in the circulation of information through connections are various centrality measures, which refers to the importance of a node in a network. Through the literature review, we delved into the role of network centrality in information diffusion.

In the spreading of memes and other online content through social networks, there are two vital metrics that determine the *"structural virality"* of a piece of content: the intrinsic appeal of the content itself, and the network structure through which it spreads [1]. In particular, it was found that content shared through highly central nodes in a network tend to have greater virality, when controlling other factors. S Kumar et al [2] provides a data-driven view of intercommunity interactions and conflict on Reddit, examining cases where users of one community are mobilized by negative sentiment to comment in another community.

Paper [3] provides a more detailed analysis of the relationship between network centrality and information diffusion. The authors found that the relationship between centrality and diffusion is complex and depends on the type of content being diffused, as well as the underlying network structure. Wei and Wang's work focuses on the use of eigenvector centrality in modeling the spread of infectious diseases through networks [4]. The paper finds that nodes with high eigenvector centrality are particularly important in spreading the infection, and that targeting these nodes with interventions, such as vaccination campaigns, can be an effective way to control the spread of the disease.

III. Dataset and Network Presentation

A. Dataset and Network Statistics

In this paper, we analyse the Reddit Hyperlink (RH) Network [2], which is a directed network that captures the connections between different subreddits on the Reddit platform. The RH Network is constructed from publicly available Reddit data spanning from January 2014 to April 2017 [6]. Each node in the network represents a subreddit, while each edge represents a hyperlink or mention from one subreddit to another. Specifically, the source node in the network is the subreddit where the hyperlink was posted, while the target node is the subreddit the hyperlink directs to. This dataset provides valuable insights into the relationships between subreddits and how information and ideas spread across the Reddit platform, making it an ideal dataset for network analysis.

TABLE I
Network Statistics

Nodes	35776
Edges	137821
Average Degree	3.852
Density	0.000108
Average Path Length	4.39
Diameter	13
Average Clustering Coefficient	0.134

B. Community Detection and Analysis

Community detection algorithms are used in network analysis to identify groups of nodes within a network that are more densely connected to each other than to the rest of the network. These groups are known as *"communities"*. To identify communities in our network, we utilize the community detection algorithm available in Gephi [7]. This algorithm employs modularity-based community detection [8] and allows for adjustment of the 'Resolution' [9] value to optimize modularity. In order to determine the optimal resolution for our network, we ran the algorithm with various resolution values as detailed in Table II.

The average size of communities for various resolution values is 541.2 and the standard deviation on the size of communities is 17.24.

It is observed that smaller resolution values result in higher number of communities detected within the network as compared to larger resolutions values. Furthermore, selecting higher resolution values results in a single large community with a majority of the nodes while the rest of the communities have less than 1% nodes. Fig. 1 captures a visual representation of the community partitions for resolution 1 and 10.

TABLE II
Communities formed & Modularity for various resolutions

Resolution	Number of Communities	Modularity
0.5	563	0.477
0.8	549	0.499
1	537	0.507
3	541	0.134
10	516	0.031

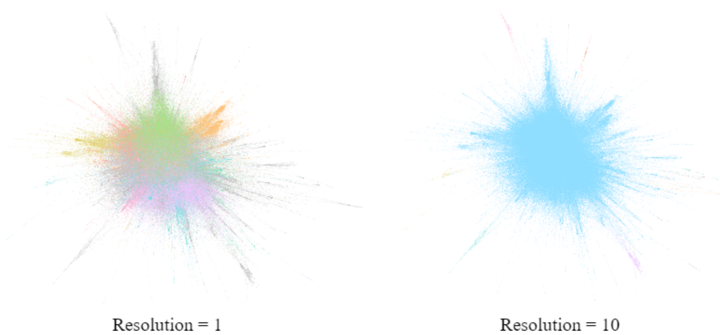

Fig. 1. Impact of Resolution value on Community Detection

After analysing the community structure for different resolutions, the value 1 emerges as the optimal resolution as it attains the highest modularity for our network. This indicates that the network is best organized into distinct communities with dense connections within and fewer connections across communities for this resolution.

1) Clustering Coefficient: Clustering coefficient of a node characterizes the density of connections in the neighbourhood of a node. In the case of the RH network, clustering coefficient is observed to be inversely proportional to the degree of subreddit. Tightly connected local clusters can be useful for rapid dissemination of information.

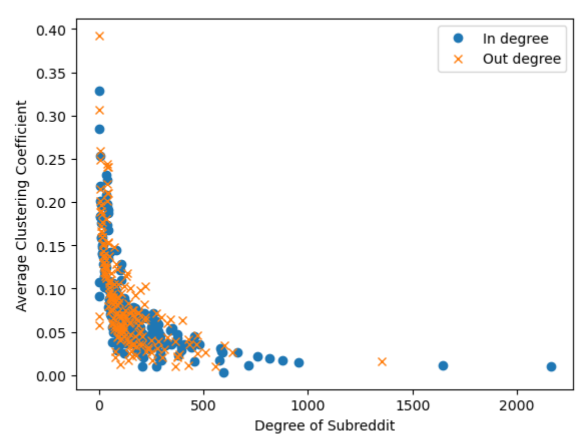

Fig. 2. Clustering coefficient vs Degree of Subreddit

2) Betweenness Centrality: Betweenness centrality measures the degree to which a node lies on the shortest paths between other nodes in a network. Nodes with higher betweenness centrality are crucial for maintaining communication and facilitating the spread of information throughout the network, as they serve as key intermediaries for the transmission of messages or ideas.

TABLE III
TOP 10 NODES HAVING HIGHEST BETWEENESS CENTRALITY

Node	Betweenness Centrality
59	0.052598
122	0.049402
41	0.044145
225	0.028852
57	0.017028
224	0.012314
0	0.012105
13	0.011317
233	0.010504
11	0.009267

3) Eigenvector Centrality: Eigenvector centrality measures the importance of a node in a network, by considering both the number of connections it has and the importance of those connections. This centrality measure is particularly relevant in the context of information spreading, as it can identify key players or influencers who play a critical role in disseminating information throughout the network. Nodes with high eigenvector centrality are likely to have a significant impact on the network's overall structure, so changes in their behavior or activity can lead to consequential changes in the network's dynamics.

TABLE IV
TOP 10 NODES HAVING HIGHEST EIGENVECTOR CENTRALITY

Node	Eigenvector Centrality
59	1.000000
41	0.883302
166	0.586556
36	0.580925
42	0.545450
55	0.468273
190	0.462987
97	0.432495
224	0.428874
225	0.393312

4) Link Symmetry: The distribution of indegree and outdegree across subreddits in the network is observed to be similar, indicating a higher degree of link symmetry [10]. This feature promotes increased network connectivity, which can facilitate the rapid spread of information throughout the network.

IV. NETWORK ANALYSIS

A. Simulating spread of memes using Susceptible-Infected (SI) modelling

Susceptible-Infected (SI) epidemic model [5] is a method of modelling the spread of infectious diseases which can be used to emulate the spread of memes through the RH network. In this type of modelling, each individual in the network can either be in one of two states: **susceptible (S)** or **infected (I)**. Once an individual is infected, it remains in the infected state and cannot change its state or vice versa.

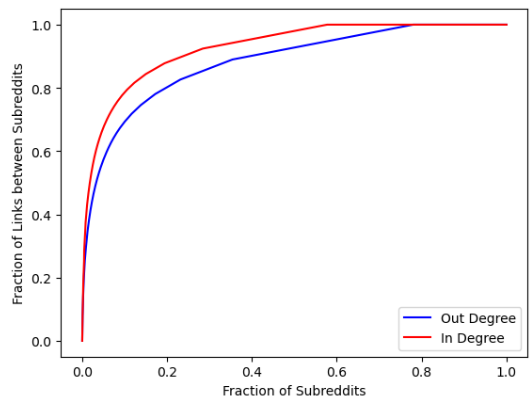

Fig. 3. Comparative link symmetry for in and out degree

The probability of infection is proportional to the number of infected individuals in the network in SI modelling. Therefore higher the infected nodes, higher the chance of susceptible nodes getting infected. Transmission rate is also one of the defining parameters in the SI model, wherein it determines the probability of a susceptible node becoming infected on contact with an infected node.

We will analyse how different nodes get infected in the networks by infecting the **most central** node and the **least central** node in the network by leveraging two centrality measures - *betweenness centrality* and *eigenvector centrality*

We are using *betweenness centrality* as it measures the number of shortest paths between pairs of nodes in the network that pass through a particular node. Therefore a node having a high betweenness centrality infers it interconnects between many different clusters. These are critical as they are responsible for facilitating the transmission across clusters within the network. Table V showcases the node and measure for the most and least central node in the network for betweenness centrality. We are choosing Node 14075 instead of other nodes having betweeness centrality measure as 0.0, because it is able to exhibit diffusion of memes through SI modelling.

TABLE V
NODES SELECTED FOR BETWEENESS CENTRALITY

Betweenness Centrality	Node	Measure
Most Central Node	59	0.052598
Least Central Node	14075	0.000

Leveraging the *most* and *least* central node as the starting infected node, we are computing the probable nodes infected in the network with SI epidemic modelling by running it for 10 iterations with transmission and recovery rates as **tau = 0.3**, **gamma = 0.1** respectively.

Using *EoN.fast_SIR()* [11] we obtain a list of nodes affected over time as it spreads through the network. The list obtained is further used to visualise the time steps taken to infect other nodes in the network from the initial infected node. The process is repeated for the least central node and the overall process is replicated using *eigenvector centrality*. Table VI showcases the measure for the most and least central node in the network for eigenvector centrality.

TABLE VI
NODES SELECTED FOR EIGENVECTOR CENTRALITY

Eigenvector Centrality	**Node**	**Measure**
Most Central Node	59	1.000
Least Central Node	21099	0.000

Along with betweenness centrality we are also choosing eigenvector centrality as it measures the importance of a node in a network based on the importance of its neighbors. In the context of SI modeling, eigenvector centrality can be utilised to find a node that, if infected, would have the highest impact on the spread of the infection. Eigenvector centrality is also a crucial centrality measure as it is less susceptible to manipulation since it is based on the overall structure of the network and that it prioritizes influence over connectivity.

B. Simulating spread of memes within and across communities

In this study, we employ a methodology to examine the spreading of memes within and across the five largest communities of the selected partition. Specifically, for the resolution value of 1, we identified the five largest communities in the given network, along with the percentage of nodes they hold as shown in Table VII.

TABLE VII
5 LARGEST COMMUNITIES

Community	**Percentage of Nodes**
4	16.45
15	14.71
19	7.75
6	7.54
1	5.01

To investigate the spread of memes, we randomly select a node from each community and run the SI model, where the infection starts with the selected node. We then plot a graph of the number of infected nodes, denoted as N_c, for each community c (c = 4, 15, 19, 6 and 1) at each time-step t, resulting in different plot lines for each community. This process is repeated for nodes selected from the other four communities, and the spread of memes is similarly observed.

Furthermore, we create a random directed graph with the same number of edges and nodes as the original network using the *'erdos_renyi_graph'* method. We then repeat the same process with this network to create a baseline and compare the results. Through this analysis, we aim to gain insight into the impact of the network structure on meme diffusion and identify the factors that play a crucial role in this process.

C. Simulating scenarios to analyse node participation in meme spreading

In order to gain further insight into the role of nodes in spreading information, we will conduct simulations to assess the impact of removing nodes from the network. Specifically, we will consider three different scenarios: removal of random nodes, removal of nodes with highest betweenness centrality, and removal of nodes with highest eigenvector centrality. In the first scenario, we will randomly remove a given percentage of nodes from the graph and monitor the number of infected nodes among the 5 largest communities, starting with a randomly infected node in the largest community. In the second scenario, we will remove the top percentage of betweenness central nodes from the graph and repeat the same process. Finally, in the third scenario, we will remove the top percentage of eigenvector central nodes and monitor the impact on the spread of information.

For each scenario, we will simulate the removal of 5%, 10%, 15%, 20%, and 25% of the nodes, and plot the evolution of the number of infected nodes in each of the 5 largest communities over time. These simulations will provide valuable insights into the role of different types of nodes in spreading information and may help identify critical nodes that play a key role in the dissemination of information within a network.

V. RESULTS AND DISCUSSION

A. Spread of memes using Susceptible-Infected (SI) modelling

On analyzing the spread of memes across two centrality measures : *betweeness centrality* and *eigenvector centrality* through SI epidemic modelling over 10 iterations as seen in Figure 4, we can observe for the most centrally connected node (Node 59) in betweeness centrality, the spread across nodes is coherent across larger time step, than for the least centrally connected node (Node 14075). The hops to a distance of 2, 3 and 4 for Node 59 showcases a steady diffusion of memes whereas for Node 14075 the diffusion of memes is short-lived but staggers to distant hops at 7 or 8.

In the scenario of eigenvector centrality measure, it is observed that the diffusion of memes across time steps and distance is much larger when compared to betweeness centrality, thereby making it the *super-spreaders* of memes. Figure 4 showcases the spread through time extends beyond 25 time steps for both most central node (Node 59) and least central node (Node 21099) for eigenvector centrality. In both scenarios, it is also observed that the diffusion of memes also extends to hops at a distance of 7 with considerable nodes getting affected in the process.

B. Spread of memes within and across communities

In the context of infecting a randomly selected node in the largest 5 communities, a prominent feature observed is the direct proportionality between the rate of infection and the proportion of nodes that a community holds, as illustrated in Figure 5(a). To provide a benchmark for network behavior, we conducted a simulation of the same scenario using a randomly generated graph. As shown in Figure 5(b), the top communities

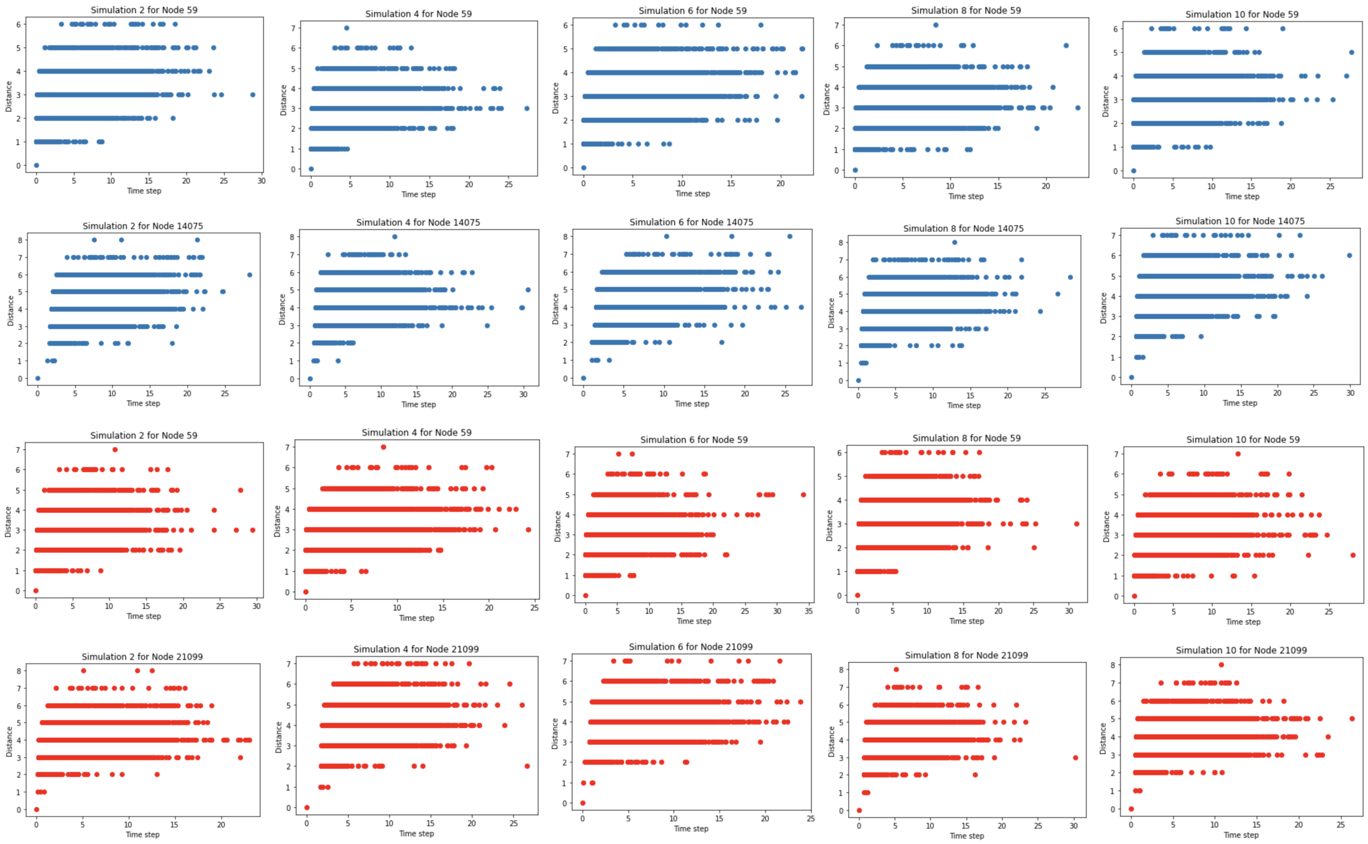

Fig. 4. Evolution of memes spreading through SI epidemic modelling. Row 1 showcases simulations for a node having the highest betweeness centrality. Row 2 showcases simulations for a node having the least betweeness centrality. Similarly, Row 3 showcases simulations for a node having the highest eigenvector centrality and Row 4 showcases simulations for a node having the least eigenvector centrality.

in the simulation had a similar rate of infection in the initial stages, which then accelerated based on the node population in each community at a later stage. This behavior is notably different from the top communities in the RH network, where the largest community exhibits the highest and fastest rate of infection from an early stage, followed by the second-largest community, and so on.

C. Participation of central nodes in meme spreading

The diffusion of memes in a network is a complex phenomenon that is influenced by the role of nodes. To gain a better understanding of this process, we conducted simulations to investigate the impact of removing nodes from the network. Three scenarios were examined, namely: removal of random nodes, removal of nodes with highest betweenness centrality, and removal of nodes with highest eigenvector centrality, as depicted in Figure 5(c)-(e).

Our findings indicate that while the removal of random nodes resulted in a slight delay in the infection rate, the communities were still able to disseminate information. However, the removal of both betweenness and eigenvector central nodes had a more severe impact, with a significant reduction in the spread of meme from and within the community at 5% and a complete stop at higher percentages. We observed that the largest community was unable to spread the meme to other communities when 5% eigenvector central nodes were removed, but it was still possible to spread the meme to other communities when betweenness central nodes were removed.

This finding highlights the critical difference between top betweenness central nodes and eigenvector central nodes, emphasizing the importance of identifying and focusing on the top percentile of eigenvector central nodes as "super-spreaders" within a network, as they play a pivotal role in the dissemination of memes. Targeting these key nodes could significantly enhance the network's ability to spread information quickly and efficiently, with important implications for a wide range of domains, including social media, marketing, and public health.

Upon examining Figure 5(b), we observed an unusual scenario in the 1st and 4th graphs, where memes were unable to spread to other communities. Further investigation revealed that the randomly selected starting node had a low eigenvector centrality measure, indicating that it was not a key node in the network. This observation was consistent with the results of Figure 5(c) 1st graph, which also showed no spread of memes due to the low eigenvector centrality measure of the starting node.

These findings emphasise the critical role of both centrality

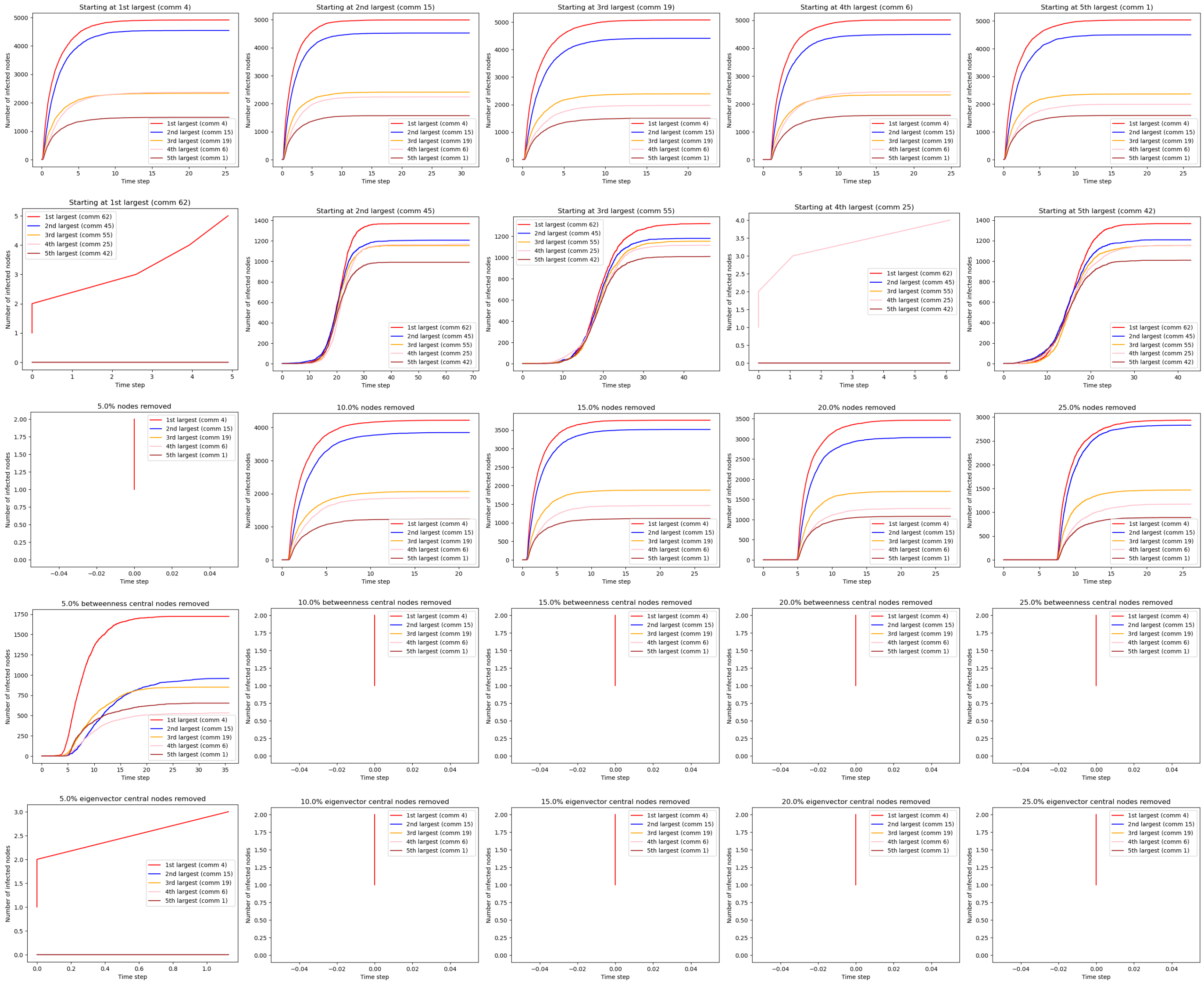

Fig. 5. Evolution of memes spreading. Row 1(a): Infection starting from a randomly selected node in the top 5 communities of Reddit dataset; Row 2(b): In a randomly generated graph, infection starting from a random node among the top 5 sommunities; Row 3(c): Deleting p% of nodes selected randomly from the Reddit dataset; Row 4(d): Deleting top p% of betweenness central nodes from the Reddit dataset; Row 4(e): Deleting top p% of eigenvector central nodes from the Reddit dataset

measures and the selection of the starting node in predicting the spread of memes in a community. By carefully selecting high eigenvector centrality nodes as the starting point for the diffusion process, it may be possible to significantly enhance the spread of information and ideas within a network.

VI. Conclusion and Perspectives

Based on the analysis presented in this paper, it is evident that eigenvector central nodes are instrumental in the spread of information within a network. The centrality measure accounts for the influence of a node's neighbors, prioritizes influence over connectivity and is less susceptible to manipulation. By serving as key players in the formation of large communities and bridging connections between communities, eigenvector central nodes can facilitate the efficient transmission and dissemination of information across the network. In essence, the presence and influence of eigenvector central nodes can greatly impact the network's ability to effectively and rapidly spread information throughout its various components.

One potential avenue for extension could be to explore the effects of different network interventions or policies on information spread. For example, researchers could simulate the impact of removing certain nodes or communities from the network, or examine how the addition of new nodes or the creation of new connections affects information flow. This could provide valuable insights into how network structure can be manipulated to optimize information spread or achieve specific outcomes.

References

[1] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, “The Structural Virality of Online Diffusion,” Management Science, vol. 62, no. 1, p. 150722112809007, Jul. 2015, doi: https://doi.org/10.1287/mnsc.2015.2158.

[2] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, “Community Interaction and Conflict on the Web,” Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18, 2018, doi: https://doi.org/10.1145/3178876.3186141.

[3] A. Mochalova and A. Nanopoulos, “On The Role Of Centrality In Information Diffusion In Social Networks,” ECIS 2013 Completed Research, Jul. 2013, Accessed: Apr. 16, 2023. [Online]. Available: https://aisel.aisnet.org/ecis2013_cr/101/

[4] Z.-W. Wei and B.-H. Wang, “Susceptible-infected-susceptible model on networks with eigenvector localization,” Physical Review E, vol. 101, no. 4, Apr. 2020, doi: https://doi.org/10.1103/physreve.101.042310.

[5] W. Duan, Z. Fan, P. Zhang, G. Guo, and X. Qiu, “Mathematical and computational approaches to epidemic modeling: a comprehensive review,” Frontiers of Computer Science, vol. 9, no. 5, pp. 806–826, Oct. 2015, doi: https://doi.org/10.1007/s11704-014-3369-2.

[6] “SNAP: Social network: Reddit Hyperlink Network,” snap.stanford.edu. https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[7] A. Lancichinetti and S. Fortunato, “Community detection algorithms: A comparative analysis,” Physical Review E, vol. 80, no. 5, Nov. 2009, doi: https://doi.org/10.1103/physreve.80.056117.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: https://doi.org/10.1088/1742-5468/2008/10/p10008.

[9] R. Lambiotte, J.-C. . Delvenne, and M. Barahona, “Laplacian Dynamics and Multiscale Modular Structure in Networks,” IEEE Transactions on Network Science and Engineering, vol. 1, no. 2, pp. 76–90, Jul. 2014, doi: https://doi.org/10.1109/TNSE.2015.2391998.

[10] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and Analysis of Online Social Networks,” 2007. Available: https://conferences.sigcomm.org/imc/2007/papers/imc170.pdf

[11] ”EoN.fast_SIR — Epidemics on Networks 1.2rc1 documentation,” https://epidemicsonnetworks.readthedocs.io/en/latest/functions/EoNḟast_SIR.html