

A centrality measure for quantifying spread on weighted, directed networks

Christian G. Fink (✉ finkt@gonzaga.edu)

Gonzaga University

Kelly Fullin

Ashland University

Guillermo Gutierrez

Georgia Institute of Technology

Nathan Omodt

Gonzaga University

Sydney Zinnecker

Gonzaga University

Gina Sprint

Gonzaga University

Sean McCulloch

Ohio Wesleyan University

Research Article

Keywords:

Posted Date: March 27th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2706014/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A centrality measure for quantifying spread on weighted, directed networks

Christian G. Fink,^{1,*} Kelly Fullin,² Guillermo Gutierrez,³ Nathan Omodt,⁴ Sydney Zinnecker,¹ Gina Sprint,⁵ and Sean McCulloch⁶

¹*Dept. of Physics, Gonzaga University*

²*Dept. of Computer Science, Ashland University*

³*Dept. of Industrial and Systems Engineering, Georgia Institute of Technology*

⁴*Dept. of Mechanical Engineering, Gonzaga University*

⁵*Dept. of Computer Science, Gonzaga University*

⁶*Dept. of Computer Science, Ohio Wesleyan University*

(Dated: March 17, 2023)

While many centrality measures for complex networks have been proposed, relatively few have been developed specifically for weighted, directed (WD) networks. Here we propose a centrality measure for spread (of information, pathogens, etc.) through WD networks based on the independent cascade model (ICM). While deriving exact results for the ICM requires Monte Carlo simulations, we show that our centrality measure (Viral Centrality) provides excellent approximation to ICM results for networks in which the weighted strength of cycles is not too large. We show this can be quantified with the leading eigenvalue of the weighted adjacency matrix, and we show that Viral Centrality outperforms other common centrality measures in both simulated and empirical WD networks.

I. INTRODUCTION

The literature on network centrality measures is vast. Unlike identifying a network's shortest paths [1] or determining the stability of its synchronization [2], there is no single method for identifying the most influential nodes in a network. This is to be expected, since nodes and connections represent vastly different physical/social entities from one network to the next, and the characteristics which confer centrality often depend on the phenomena of interest [3].

Many centrality measures aim to identify nodes that are influential with respect to spreading phenomena, such as the spread of infectious disease [4], rumors [5], or information [6]. The performance of such centrality measures is often determined by comparison with simulations of spread that are highly simplified yet still computationally expensive. The independent cascade and linear threshold models are perhaps the two most popular models for simulating information diffusion through social networks [7].

In this paper we propose a centrality measure based on the independent cascade model (ICM), a generalization of the susceptible-infected-recovered (SIR) epidemiological model originally proposed by Goldenberg et al. [8, 9]. Like the SIR model, the ICM is used to simulate spread through a network in which nodes may be in one of three states: susceptible, active (or infected), and inactive (or recovered). The ICM runs in discrete time steps, with an active node having just one opportunity to activate its neighbors before becoming inactive. The ICM generalizes the SIR model by specifying network structure using a weighted, directed (WD) adjacency matrix \mathcal{P} , with \mathcal{P}_{ij} giving the probability of node j activating node

i , given that node j is itself active.

A tremendous amount of work has gone into identifying the optimal set of seed nodes of a given size to maximize the total number of nodes subsequently activated in the ICM [10–16]. Kempe et al. showed this to be an NP-Hard problem [6], necessitating heuristic approaches. Such heuristic approaches to this combinatorial optimization problem are not the focus of this study. Instead, we address the simpler problem of ranking individual nodes according to the expected number of others they will infect, given a single activated node as the starting point.

Various centrality measures have been proposed to identify such influential spreaders in the SIR model with homogeneous transmission probability, including degree, k -core, betweenness centrality, eigenvector centrality, and PageRank, among others [3]. Such classic approaches suffer from two major drawbacks, however: they do not generally take into account specific details of the spreading model, and they do not apply to WD networks. In cases where a classic centrality measure may be generalized to WD networks, there are often multiple ways of doing so, with no consensus as to which way is best [17–19]. In general, there are relatively few centrality measures that have been developed specifically for WD networks.

In this study, we propose a novel centrality measure for WD networks specifically designed to identify influential spreaders in the ICM. We call our measure the Viral Centrality (VC), and we show that it gives results nearly to identical to the ICM but much faster, so long as the strength of cycles within a network is small. We show that the strength of cycles within a network may be quantified using the leading eigenvalue of the weighted adjacency matrix, and we show that VC's performance compares favorably to several other WD centrality measures in synthetic and empirical networks.

Algorithm 1 Viral Centrality

```

1: for each node as SeedNode do
2:   prob_susceptible = ones(TotalNodes) # initializing probabilities a given node is still susceptible
3:   prev_activated = zeros(TotalNodes) # initializing probabilities a given node was activated on previous time step
4:   cur_activated = zeros(TotalNodes) # initializing probabilities a given node is activated on current time step

5:   prev_activated[SeedNode] = 1 # set seed node to definitely being activated on previous time step...
6:   cur_activated[SeedNode] = 0 # ... and therefore definitely NOT activated on current time step

7:   t=1 #initialize time step
8:   while (termination condition) do

9:     perform breadth-first search to determine all nodes within t edges of SeedNode

10:    for each Node within reach of SeedNode do
11:      prob_uninfected = 1 # initialize probability that a given node is not infected on this time step
12:      for each Neighbor sending a connection to Node do
13:        #  $\mathcal{P}[i, j]$  is probability of node j activating node i, given that node j was activated on previous time step
14:        prob_uninfected = prob_uninfected * (1-prev_activated[Neighbor]* $\mathcal{P}$ [Node, Neighbor])
15:      end for
16:      cur_activated[Node]=(1-prob_uninfected)*prob_susceptible[Node]
17:    end for

18:    for each Node in TotalNodes do # clean-up in preparation for next time step
19:      prev_activated[Node] = cur_activated[Node]
20:      prob_susceptible[Node] = prob_susceptible[Node] - cur_activated[Node]
21:    end for

22:    t = t + 1
23:  end while
24:  viral_centrality[SeedNode] = sum(1-prob_susceptible) - 1 #“-1” discounts initial activation of seed node
25: end for

```

II. RESULTS**Viral Centrality algorithm**

The VC algorithm aims to calculate the expected total number of nodes activated by a given set of initial seed nodes in the ICM. (In this study, we explore VC as a centrality measure, so we focus on the special case of just one initially activated seed node.) As shown in Algorithm 1, for each seed node VC loops over several discrete time steps (Lines 8-22), similar to a Monte Carlo simulation. Rather than prescribe precisely defined states to each node (susceptible, active, or inactive), however, VC continually updates the probability that each nodes remains susceptible (in the array `prob_susceptible`). At the end of the algorithm (Line 23), taking one minus this value gives the probability a given node was activated at some previous point in time, and adding all these probabilities gives the expected number of nodes ultimately activated by the seed node.

Note how Line 8 allows one to specify a condition for terminating the algorithm. This could be simply a fixed number of time steps, or a condition based on the convergence of `prob_susceptible` to stable values. We later explore the consequences of both approaches in applying VC to empirical networks.

The VC algorithm gives exact results for acyclic networks, but strictly overestimates the expected number of nodes infected for networks with cycles. This can be seen most easily with a simple example: suppose node i and node j are reciprocally connected, $i \longleftrightarrow j$, and further suppose node j has no connections with any other nodes. If node i activates node j , the ICM does not allow j to subsequently activate node i (because node i is activated for only one time step, and inactivated thereafter). Yet the VC algorithm allows node j to increase node i 's probability of being activated, resulting in an overestimate of node i 's expected activation probability. While this simple example is for a 2-cycle, it applies to arbitrary n -cycles. It has been shown that properly accounting for all such cycles results in an algorithm that scales exponentially with the number of nodes, making it impractical for large networks [20].

Condition for accuracy

Fortunately, it is possible to determine the credibility of VC's results for a given network by considering the leading eigenvalue of the network's weighted adjacency matrix, \mathcal{P} . To see this, we first devise a measure to quantify the “strength of cycles” within a WD network. First,

the number of paths of length ℓ from node j to node i in an unweighted network with adjacency matrix \mathbf{A} is equal to $(\mathbf{A}^\ell)_{ij}$, and therefore the number of cycles of length ℓ starting and ending on node i is $(\mathbf{A}^\ell)_{ii}$. For a WD network, we may replace the unweighted adjacency matrix \mathbf{A} with the weighted adjacency matrix \mathcal{P} , understanding that each cycle is then weighted by the product of its links.

A reasonable measure of the strength of cycles in a network, SC , is then found by summing over all nodes and all possible cycle lengths:

$$SC = \sum_{\ell=2}^{\infty} \text{Tr}(\mathcal{P}^\ell) \quad (1)$$

We may then use two properties from linear algebra to further simplify this result. First, the trace of a matrix is equal to the sum of its eigenvalues. Second, if λ is an eigenvalue of a matrix \mathbf{M} , then λ^n is the corresponding eigenvalue of \mathbf{M}^n .

For an n -node network, Eq. 1 then becomes

$$SC = \sum_{\ell=2}^{\infty} \lambda_1^\ell + \lambda_2^\ell + \dots + \lambda_n^\ell. \quad (2)$$

This measure will diverge when the magnitude of the weighted adjacency matrix's leading eigenvalue is greater than or equal to 1. In this case, the error (relative to ground truth ICM results) in VC's predicted number of nodes activated will be hopelessly large. On the other hand, VC gives results exactly equal to ICM results in acyclic networks, for which $SC = 0$. We expect a smooth transition in VC's performance between these two extremes, and predict low error between VC and ICM results when $|\lambda|_{\max} \ll 1$.

Results for Erdős-Rényi networks

We explored VC's accuracy as a function of $|\lambda|_{\max}$ in a simple Erdős-Rényi model. We synthesized an assortment of WD 1000-node networks by specifying two different probability values: p_{con} was the probability of a potential directed connection actually being instantiated, and p_{trans} was the transmission probability (fixed to the same constant value for all connections within any particular network). We independently swept over a range of 0.005 to 0.038 for both probabilities (this range was chosen in order to give values for $|\lambda|_{\max}$ ranging from roughly 0 to 1). For each combination of p_{con} and p_{trans} , we generated a corresponding network, then ran ICM Monte Carlo simulations and compared these "ground truth" results to those predicted by VC. We then computed the mean relative error between the ICM and VC results, defined as

$$\epsilon = \frac{\sum_{i=1}^N \frac{|a_i^{\text{ICM}} - a_i^{\text{VC}}|}{a_i^{\text{ICM}}}}{N}, \quad (3)$$

where a_i^{ICM} is the mean number of nodes activated by node i in the ICM (averaged across 10,000 Monte Carlo trials) and a_i^{VC} is the expected number of nodes activated by node i according to Viral Centrality. For Erdős-Rényi networks, the VC algorithm was terminated when either of the following conditions was first met: 1) the maximum relative error (from one time step to the next for a particular node, with the maximum taken across all nodes) first dipped below 10^{-5} , or 2) twenty time steps were completed.

Fig. 1A shows how the mean relative error depended on the magnitude of the leading eigenvalue of the weighted adjacency matrix. As expected, VC performed very well for $|\lambda|_{\max} \ll 1$, and the error steadily increased as $|\lambda|_{\max}$ approached and exceeded 1. (Note that the error did not go to zero as $|\lambda|_{\max}$ approached zero due to imprecision in the Monte Carlo results, resulting from a finite number of ICM trials for each network.) Panels B, C, and D compare Monte Carlo and VC results for 100 randomly selected nodes (out of 1000) in three representative networks. (We do not show results for all 1000 seed nodes in order to aid visualization.) Note the excellent match in B (small $|\lambda|_{\max}$) and the worsening performance as $|\lambda|_{\max}$ increases (C and D). Also note how VC tends to overestimate the number of nodes activated.

Congressional Twitter network

We applied VC to an empirical network by using Twitter's API to construct the Twitter interaction network for the 117th United States Congress House of Representatives. We quantified empirical transmission probabilities according to the fraction of times one member retweeted, quote tweeted, replied to, or mentioned another member's tweet. (See Methods section for more details.) Requiring a minimum of 100 Tweets over a four-month span gave a network of 475 Congressional members, with transmission probabilities ranging between 0.00053 and 0.13, and a leading eigenvalue for the adjacency matrix of 0.208.

Fig. 2A shows excellent agreement between VC and Monte Carlo ICM results. (For this figure, the VC computation terminated with a tolerance of 0.001, meaning every seed node's expected number of nodes activated had to change by less than 0.1% from the previous iteration.) Fig. 2B shows the distribution of transmission probabilities (i.e., connection weights) was well approximated by a lognormal distribution (see Methods section), which we used later in formulating the Higgs Twitter dataset (discussed in the next section).

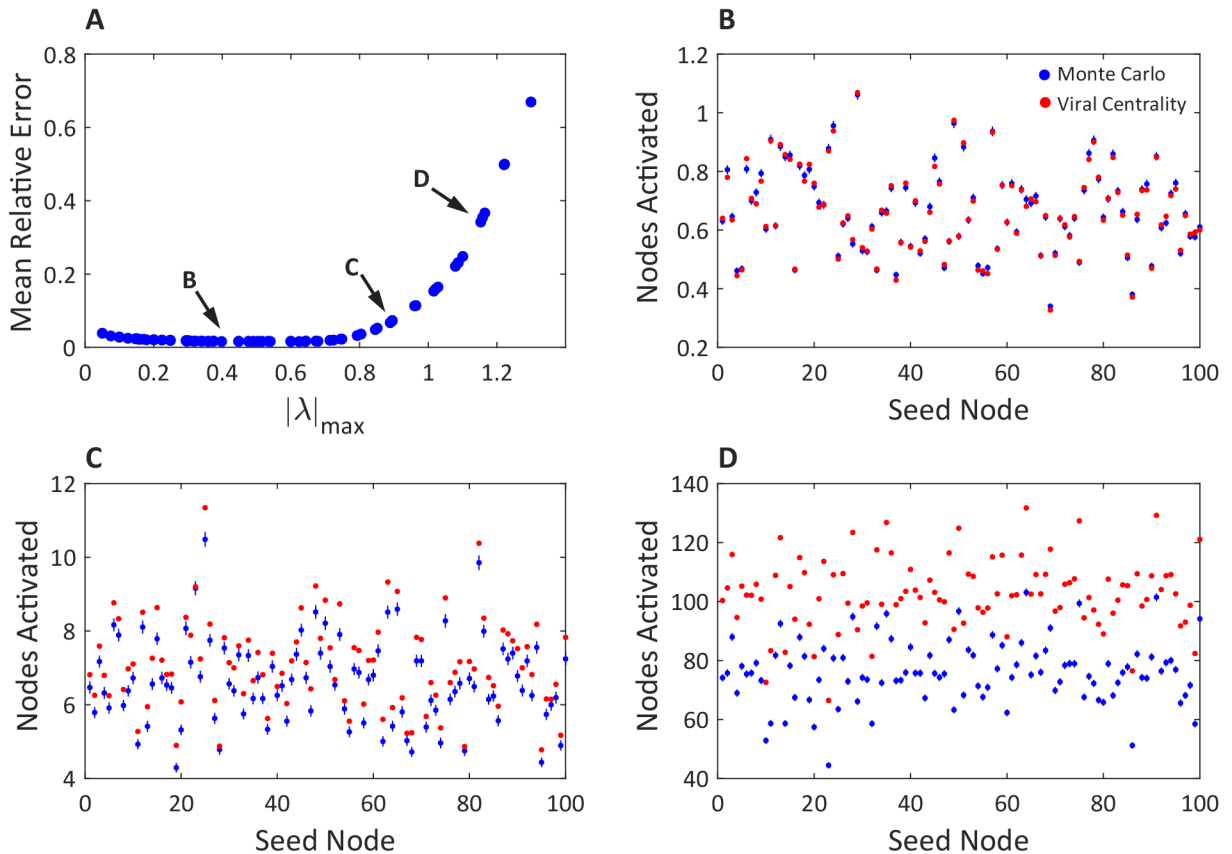


FIG. 1. Performance of Viral Centrality in Erdős-Rényi networks. Various combinations of connection probability (p_{con}) and transmission probability (p_{trans}) were used to synthesize 1000-node networks. 10,000 Monte Carlo simulations of the Independent Cascade Model were conducted for each network, and the average number of nodes activated for each seed node was compared to that predicted by Viral Centrality. A: Each point represents a single network, with different points implying different combinations of p_{con} and p_{trans} . The mean relative error between ICM and VC is plotted as a function of the magnitude of the leading eigenvector of each network’s adjacency matrix. B: Results for a network with $p_{\text{con}} = 0.02$ and $p_{\text{trans}} = 0.02$. The figure shows the mean (ICM Monte Carlo) or expected (VC) number of nodes activated by 100 randomly selected seed nodes. Error bars for Monte Carlo results represent standard error of the mean over 10,000 trials. C: Same as B, but for a network with $p_{\text{con}} = 0.03$ and $p_{\text{trans}} = 0.03$. D: Same as B and C, but for a network with $p_{\text{con}} = 0.034$ and $p_{\text{trans}} = 0.034$.

In Table I, we compare the performance of VC to other common centrality measures in ranking the influence of members of Congress on Twitter. We use expected number of nodes activated in Monte Carlo ICM simulations for the ground truth ranking, then compute Kendall’s tau correlation [21] with the rankings obtained by WD versions of degree, k-core [17–19], and PageRank [22]. We also compare to diffusion degree, which is a generalization of weighted degree that takes second-order neighbors into account [23–25].

The table shows how VC outperforms all these centrality measures, albeit at the cost of taking much longer to compute. Diffusion degree comes closest to matching the performance of VC, giving a Kendall’s tau value nearly identical to VC when ranking all nodes, but clearly inferior when ranking just the top 10% of nodes. It is perhaps not surprising that diffusion degree’s performance is

closest to VC, since it is similar (but not identical) to running VC for two time steps. Diffusion degree starts with the weighted out-degree of a node (first-order neighbors), then adds the products of the two connection weights required to reach each second-order neighbor. This is similar to two-step VC, except that diffusion degree’s simplistic approach does not attempt to make the second-order contributions equal to probabilities of activation of second-order neighbors (although for small transmission probabilities, it gives a good approximation). We therefore expect that if we run VC for two time steps (rather than to a convergence tolerance), it will give results roughly on par with those of diffusion degree, and that running for more time steps will further improve performance. Table I confirms this expectation, demonstrating that running VC for a small, fixed number of time steps can expedite the VC algorithm at the cost of

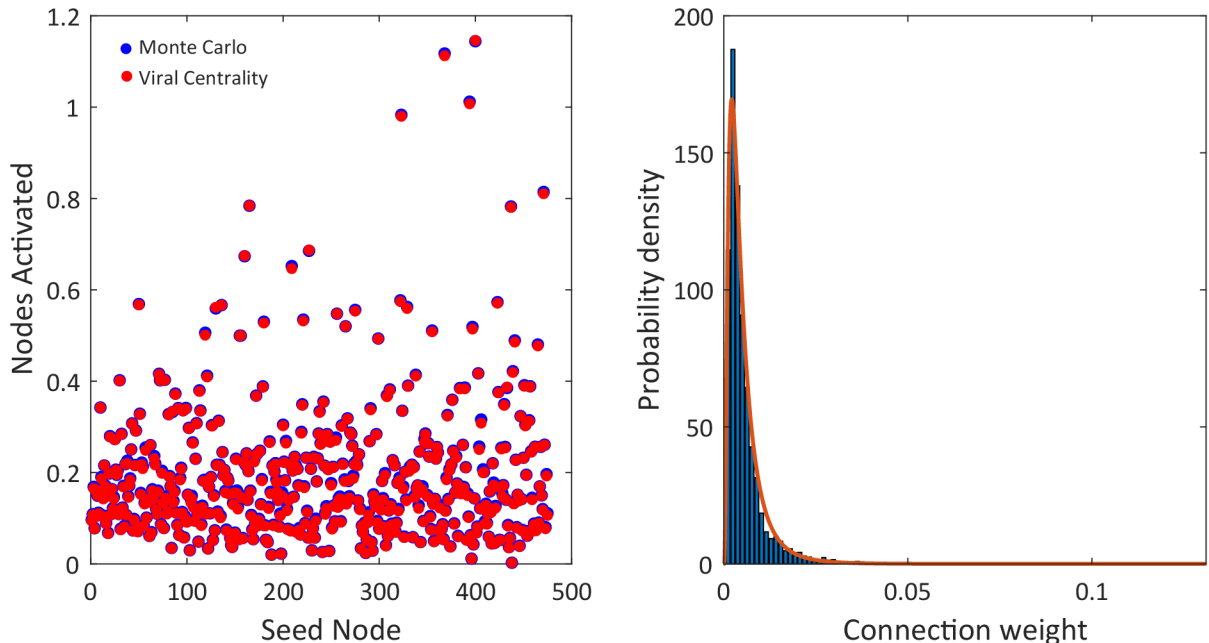


FIG. 2. Viral Centrality and Independent Cascade Model results for Congressional Twitter network. A: Mean (ICM Monte Carlo) or expected (VC) number of nodes activated by each of the 475 seed nodes. Monte Carlo results represent averages over 10^6 trials, with error bars displayed but too small to be seen. B: Distribution of transmission probabilities for this network. Red line represents fit by a lognormal distribution.

diminished accuracy.

Higgs networks

We applied VC to the Higgs Twitter dataset, a much larger network of 456,626 nodes derived from Tweets related to the discovery of the Higgs boson in 2012 [26] (see Methods for more details). Fig. 3 overlays the average number of nodes activated according to Monte Carlo ICM and the expected number according to VC (run with a convergence tolerance of 0.001). Table II shows the Kendall's tau between these two results was 0.967 for all nodes, and 0.927 for the top 10% of nodes. Similar to the Congressional network, the diffusion degree matched this performance for all nodes but was inferior for the top 10%. VC run with two time steps gave virtually identical results to diffusion degree, whereas running it with three time steps gave even better performance than with the convergence tolerance (suggesting that most seed nodes required only two time steps to converge).

While VC is obviously much more computationally expensive than more traditional centrality measures (requiring about 400 compute-hours when run with convergence tolerance), it still provided roughly an order of magnitude speed-up over Monte Carlo simulations, which required approximately 5500 compute-hours to complete 10,000 trials per seed node.

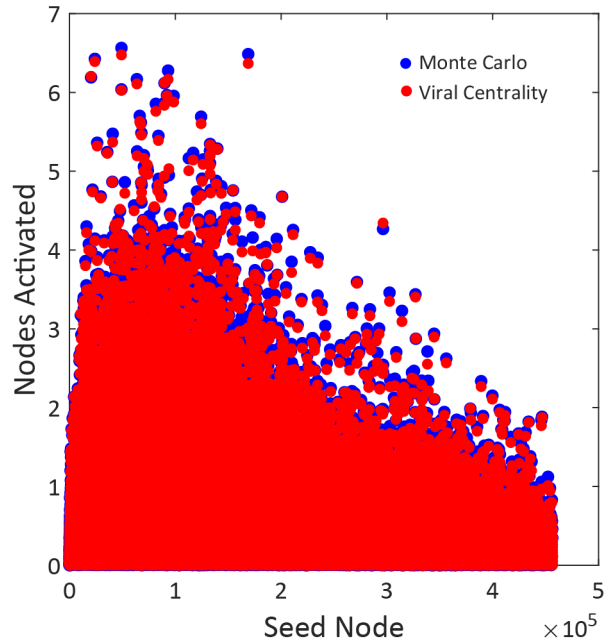


FIG. 3. Viral Centrality and Independent Cascade Model results for Higgs Twitter network. Plot displays mean (ICM Monte Carlo) or expected (VC) number of nodes activated by each of the 456,626 seed nodes. Monte Carlo results represent averages over 10^4 trials, with error bars displayed but too small to be seen.

Measure	Kendall's Tau (top 10%)	Kendall's Tau (all nodes)	Mean Rel. Error	Compute time (s)
WD Out-degree	0.882	0.949	N/A	$< 10^{-7}$
WD k-core	0.223	0.672	N/A	0.031
WD PageRank	0.415	0.515	N/A	0.047
Diffusion degree	0.882	0.989	0.025	0.125
Viral Centrality (tol=0.001)	0.994	0.991	0.017	5.03
Viral Centrality (t=2)	0.965	0.990	0.029	2.89
Viral Centrality (t=3)	0.994	0.997	0.0047	7.58

TABLE I. Comparison of VC to other centrality measures for the Congressional Twitter network. Kendall's tau correlation between Monte Carlo (10^6 trials per seed node) and centrality rankings was computed for all nodes in the network, as well as for just the top 10% of nodes in the network. Mean relative error was also computed between average number of nodes activated in Monte Carlo simulations and the number predicted by diffusion degree and VC. VC outperformed all other measures, though at the cost of substantially increased runtime. Monte Carlo simulations required approximately 6 compute-hours. All compute times were determined using the Neuroscience Gateway computing cluster.

Measure	Kendall's Tau (top 10%)	Kendall's Tau all nodes)	Mean Rel. Error	Compute time (hr)
WD Out-degree	0.872	0.702	N/A	5.0×10^{-5}
WD k-core	0.355	-0.017	N/A	2.40
WD PageRank	0.316	0.144	N/A	26.1
Diffusion Degree	0.877	0.967	0.113	0.08
Viral Centrality (tol=0.001)	0.927	0.967	0.093	398.9
Viral Centrality (t=2)	0.877	0.967	0.113	202.5
Viral Centrality (t=3)	0.938	0.977	0.058	789.0

TABLE II. Comparison of VC to other centrality measures for the Higgs Twitter network. Kendall's tau correlation between Monte Carlo (10,000 trials per seed node) and centrality rankings was computed for all nodes in the network, as well as for just the top 10% of nodes in the network. Mean relative error was also computed between average number of nodes activated in Monte Carlo simulations and the number predicted by diffusion degree and VC. As in the Congressional network, VC generally outperformed all other measures, though at the cost of substantially increased runtime. Monte Carlo simulations required approximately 5500 compute-hours.

III. DISCUSSION

In this study we have proposed a novel centrality measure, Viral Centrality (VC), for weighted, directed (WD) networks. VC is designed to identify influential spreaders within networks for which probability of transmission between nodes can be quantified. We have shown that VC well approximates the expected number of nodes activated in the Independent Cascade Model (ICM), so long as the leading eigenvalue of the WD adjacency matrix is much less than 1. Because VC is modeled after the ICM, it is more accurate than other traditional centrality measures in ranking nodes with respect to the number activated in Monte Carlo ICM simulations.

The trade-off for this improved accuracy is slower run time. VC is much slower than degree-based centrality measures, and is too slow to be used as a heuristic to

find an optimal set of multiple seed nodes in the ICM [10, 27]. For such a combinatorial optimization problem, faster centrality measures such as diffusion degree [24] are required for large-scale networks. VC is best applied to networks composed of up to tens of thousands of nodes.

Nevertheless, we have shown that VC is generally more accurate than diffusion degree, making it an ideal choice for situations where accuracy is paramount, yet speed much faster than Monte Carlo simulations is still required. We have also shown that this superior accuracy can be expected when the leading eigenvalue of a network's WD adjacency matrix is much less than 1 (Fig. 1A suggests $|\lambda|_{\max} = 0.5$ is a conservative and safe cut-off). Although the calculation of the leading eigenvalue of a matrix is generally expensive, we note that Gershgorin's Circle Theorem may be used to place an upper bound on the leading eigenvalue of an adjacency matrix based purely on its row sums. This provides an inex-

pensive, sufficient condition for trusting the accuracy of VC’s results. Our results for the Congressional Twitter network suggest that transmission probabilities in empirical networks may often be small enough for VC to well approximate ICM results.

It should be noted that there are many generalizations and extensions of the ICM that our centrality measure does not address [6, 28–30]. Further research is necessary to develop efficient algorithms for identifying influential spreaders in these more complex models.

IV. METHODS

Comparison centrality measures

WD PageRank

We used an extension of PageRank to WD networks proposed by Zhang et. al. [22]. The original PageRank (PR) algorithm [31] applied to unweighted networks, and defined PR iteratively according to

$$PR(i) = \alpha \sum_j \frac{a_{ij}}{d_j^{\text{out}}} PR(j) + \frac{1 - \alpha}{n}, \quad (4)$$

where a_{ij} is a binary value (0 or 1) specifying whether or not a connection exists from j to i . The extension to WD networks is achieved by introducing a second parameter, $\theta \in [0, 1]$, that determines the degree to which the centrality score is determined by the weighted adjacency matrix (w_{ij}) versus the unweighted one (a_{ij}). Calling the WD PR centrality score ϕ , it may be computed iteratively according to:

$$\phi(i) = \alpha \sum_j \left(\theta \frac{w_{ij}}{s_j^{\text{out}}} + (1 - \theta) \frac{a_{ij}}{d_j^{\text{out}}} \right) \phi(j) + \frac{1 - \alpha}{n}, \quad (5)$$

where s_j^{out} is the weighted out-degree of node j . Note that in this approach, nodes “score points” according to the *incoming* connections they receive. However, in this study influence was based on outgoing connections, so we simply ran the WD PR algorithm on the transpose of our adjacency matrices. Also, in this study we set $\theta = 0.5$.

WD k -core

Inspired by [17–19], we implemented a WD version of k -core decomposition using the following approach:

- Calculate the weighted out-degree, s^{out} , of each node in the network.
- Determine the smallest weighted out-degree, s_{\min}^{out} .

- Remove all nodes for which $s^{\text{out}} = s_{\min}^{\text{out}}$.
- Update the values of s^{out} for the nodes that remain in the network.

This process terminates when all nodes in the network have been removed. Nodes are ranked according to the order in which they are removed from the network, with those surviving the longest ranked the highest.

Diffusion degree

The diffusion degree, DD, generalizes weighted degree to include second-order neighbors in addition to first-order neighbors. Given a weighted adjacency matrix w_{ij} , it is defined mathematically by

$$DD(i) = \sum_j w_{ji} \times \left(1 + \sum_k w_{kj} \right). \quad (6)$$

Datasets

Congressional Twitter network

The Twitter interaction network for the 117th United States Congress House of Representatives was constructed by first obtaining members’ official Twitter handles from <https://pressgallery.house.gov/member-data/members-official-twitter-handles>. The Twitter API was then used to obtain all Tweets by members of Congress between February 9, 2022, and June 9, 2022. (The latter date corresponds to when the data were requested, and the earlier date was the farthest point backward in time at which one user’s number of Tweets exceeded the 3200-Tweet limit set by the API.) Only the 475 representatives (out of 535) who issued 100 or more tweets during this time frame were included in the network.

A weighted, directed network was constructed such that the connection from user j to user i was given by

$$\mathcal{P}_{ij} = \frac{n_{ij}^{\text{retweet}} + n_{ij}^{\text{quote}} + n_{ij}^{\text{reply}} + n_{ij}^{\text{mention}}}{N_j^{\text{tweets}}}, \quad (7)$$

where n_{ij}^{retweet} is the number of times user i retweeted user j , n_{ij}^{quote} is the number of times user i quote tweeted user j , n_{ij}^{reply} is the number of times user i replied to user j , n_{ij}^{mention} is the number of times user i mentioned user j , and N_j^{tweets} is the total number of Tweets issued by user j . The weighted connection, \mathcal{P}_{ij} , can therefore be interpreted as the probability that user i is observably influenced by user j .

The distribution of non-zero elements w of \mathcal{P}_{ij} was approximated by a lognormal distribution with pdf

$$p(w) = \frac{1}{sw\sqrt{2\pi}} \exp\left(-\frac{\log^2(w)}{2s^2}\right). \quad (8)$$

The parameter $s = 0.7893$ was found to best fit to data (see Fig. 2B).

Higgs Twitter dataset

The Higgs Twitter dataset is derived from all tweets from July 1 to July 7, 2012, that contained the following keywords or hashtags: `lhc`, `cern`, `boson`, `higgs` [26]. The dataset is available for download on the Stanford Network Analysis Project website (<https://snap.stanford.edu/data/higgs-twitter.html>). While the original dataset includes information about retweets, replies, and mentions, we chose not to use this information in constructing our network because most users issued only a few tweets in total, resulting in poor statistics for inferring transmission probabilities.

Instead, we started with the follower network, which consisted of 456,626 nodes and 14,855,842 directed connections. (If user j followed user i , this constituted a connection from node i to node j , consistent with the presumed direction of influence.) We then randomly assigned weights (i.e., transmission probabilities) to each connection, using the lognormal distribution empirically determined from the Congressional Twitter network (see Fig. 2B). The magnitude of the largest eigenvalue of the resulting adjacency matrix was 0.074.

Acknowledgements

This work was supported by the Gonzaga Science Research Program, the Ohio Wesleyan Summer Science Research Program, and the National Science Foundation (1658998). Simulations were conducted on the Neuroscience Gateway computing cluster.

* finkt@gonzaga.edu

- [1] E. W. Dijkstra *et al.*, *Numerische Mathematik* **1**, 269 (1959).
- [2] L. M. Pecora and T. L. Carroll, *Physical Review Letters* **80**, 2109 (1998).
- [3] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, *Physics Reports* **650**, 1 (2016).
- [4] M. E. Newman, *Physical Review E* **66**, 016128 (2002).
- [5] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, *Physica A: Statistical Mechanics and its Applications* **374**, 457 (2007).
- [6] D. Kempe, J. Kleinberg, and É. Tardos, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003) pp. 137–146.
- [7] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo, in *Diffusion in Social Networks* (Springer, 2015) pp. 35–48.
- [8] J. Goldenberg, B. Libai, and E. Muller, *Marketing Letters* **12**, 211 (2001).
- [9] J. Goldenberg, B. Libai, and E. Muller, *Academy of Marketing Science Review* **9**, 1 (2001).
- [10] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2007) pp. 420–429.
- [11] W. Chen, Y. Wang, and S. Yang, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2009) pp. 199–208.
- [12] W. Chen, C. Wang, and Y. Wang, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010) pp. 1029–1038.
- [13] Y. Wang, G. Cong, G. Song, and K. Xie, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010) pp. 1039–1048.
- [14] B. Min, *The European Physical Journal B* **91**, 1 (2018).
- [15] F. Altarelli, A. Braunstein, L. Dall’Asta, J. R. Wakeling, and R. Zecchina, *Physical Review X* **4**, 021024 (2014).
- [16] F. Morone and H. A. Makse, *Nature* **524**, 65 (2015).
- [17] A. Garas, F. Schweitzer, and S. Havlin, *New Journal of Physics* **14**, 083030 (2012).
- [18] B. Wei, J. Liu, D. Wei, C. Gao, and Y. Deng, *Physica A: Statistical Mechanics and its Applications* **420**, 277 (2015).
- [19] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, *Physica A: Statistical Mechanics and its Applications* **468**, 278 (2017).
- [20] W. Yang, L. Brenner, and A. Giua, *IEEE Access* **7**, 13745 (2019).
- [21] M. G. Kendall, *Biometrika* **30**, 81 (1938).
- [22] P. Zhang, T. Wang, and J. Yan, *Physica A: Statistical Mechanics and its Applications* **586**, 126438 (2022).
- [23] S. Kundu, C. Murthy, and S. K. Pal, in *International Conference on Pattern Recognition and Machine Intelligence* (Springer, 2011) pp. 242–247.
- [24] S. K. Pal, S. Kundu, and C. Murthy, *Fundamenta Informaticae* **130**, 317 (2014).
- [25] I. Gaye, G. Mendy, S. Ouya, and D. Seck, in *2015 3rd International Conference on Future Internet of Things and Cloud* (IEEE, 2015) pp. 675–680.
- [26] M. De Domenico, A. Lima, P. Mougél, and M. Musolesi, *Scientific Reports* **3**, 1 (2013).
- [27] N. Sumith, B. Annappa, and S. Bhattacharya, *Future Generation Computer Systems* **89**, 777 (2018).
- [28] K. Jung, W. Heo, and W. Chen, in *2012 IEEE 12th International Conference on Data Mining* (IEEE, 2012) pp. 918–923.
- [29] Y. Qin, J. Ma, and S. Gao, *Soft Computing* **21**, 827 (2017).
- [30] B. Liu, G. Cong, D. Xu, and Y. Zeng, in *2012 IEEE 12th International Conference on Data Mining* (IEEE, 2012) pp. 439–448.
- [31] S. Brin and L. Page, *Computer Networks and ISDN Systems* **30**, 107 (1998).