Inferring Network Structure from Cascades

Sushrut Ghonge[1, 2, *] and Dervis Can Vural[2, †]

[1] *Department of Physics, Indian Institute of Technology Delhi, India*
[2] *Department of Physics, University of Notre Dame, USA*

(Dated: July 7, 2017)

Many physical, biological and social phenomena can be described by cascades taking place on a network. Often, the activity can be empirically observed, but not the underlying network of interactions. In this paper we offer three topological methods to infer the structure of any directed network given a set of cascade arrival times. Our formulas hold for a very general class of models where the activation probability of a node is a generic function of its degree and the number of its active neighbors. We report high success rates for synthetic and real networks, for several different cascade models.

I. INTRODUCTION

Neural networks, ecosystems, epidemics, range expansions, gene-protein interactions, diffusion in evolutionary landscapes and many other interesting biological and social phenomena are naturally encoded by cascades on complex networks. Often we can observe when people adopt certain ideas, but we cannot see what social exchanges lead to it. We can observe when species go extinct, but do not know why [1]. We see when the same content appears in several websites and blogs over time, but since we do not know who copied from whom, we cannot tell who follows who [2]. A sequence of neural firings can be observed by flourescent imaging, but it is not trivial to infer neural connectivity [3].

Establishing network structure empirically is tedious. Ideally, to determine the presence of an edge between a node pair, one must perturb one while measuring the response of the other, making sure the rest is unchanged. Proxies such as correlation coefficients can be used, but these yield unreliable results (e.g. [4]). Furthermore, proxies depend on specific models, e.g. the presence of an edge could just as well imply a lack of correlation.

The problem of topological inference has been previously addressed as a convex optimization problem, and only specific cases have been solved [1, 5, 6]. Others have considered inferring topology when each cascade affects only few nodes, and only when few several such cascades take place simultaneously [7].

This work concerns with a very general class of cascade models where the probability that a node activates depends on the degree of the node and the states of the neighboring nodes. In this class of models the activation of every node is permanent till the cascade ends. We present three very generally applicable methods to determine network structure from time-of-activation data. We then evaluate our success for 3 real networks, synthetic random networks and for 5 different kinds of cascade models. For one of the models we evaluate success for a full range of model parameters.

We cite an incomplete list of the systems and models for which our methods are applicable [1, 2, 5–19]. In some of these models, nodes activate when a critical fraction of their providers (in-neighbors) activate [8, 19] . In others, nodes do not deterministically activate when the number of active providers meet a threshold; instead their probability of activation jumps to a different value [20]. In several other models, every active node linearly adds to the activation probability of their common neighbor. In general, a node can respond to its neighbors arbitrarily. The problem of inferring network structure finds applications in many diverse areas such as biochemistry and bioinformatics [12–14, 16], political science [9], social networks, blogs [2, 5, 17], sociology [1, 8] and modelling aging [19]

II. DIFFUSION MODEL

We consider the general model where the probability that a node activates is an arbitrary function $f(m/k)$ of the ratio of the number of active providers m and its indegree k.

We denote the fraction of nodes that activate at time t by $D(t)$ and the fraction of nodes active at t by $Q(t)$, so that $Q(t) = \sum_{\tau=1}^{t} D(\tau)$.

For both the forward solution and the topological inversion, the probability that m out of k providers of a node have activated after a time t, can be approximated as $B(m, k, Q(t)) = \binom{k}{m}(Q(t))^m(1-Q(t))^{k-m}$, where $Q(t)$ is the fraction of nodes active at t. $Q(t)$ is to be determined recursively.

For a node with k providers, the probability $D(t)$ of activating is the sum over all possible number of activated providers of the product of the probability of that number of providers being active and the value of f at that number. Since $\Gamma(k)$ is the fraction of nodes with indegree k, for a random node with unknown indegree,

$$D(t) = \sum_k \Gamma(k) \sum_{m=0}^{k} B(m, k, Q(t-1)) f(m/k) \qquad (1)$$

Since all nodes are inactive at $t = 0$, $D(1) = f(0)$. For $t \geq 2$, $D(t)$ is obtained in terms of $Q(t-1) = \sum_{\tau=1}^{t-1} D(i)$

[*] sushrutghonge@gmail.com
[†] dvural@nd.edu

from equation (1) which is easily iterated. This recursive equation was studied in detail in [20].

Interestingly, knowing the forward dynamics gives little hint about the inverse problem of obtaining network topology, given node activation times. This is an ill posed problem: generally speaking, two different networks (even those with different Γ) can have similar mean field behavior. Thus, the methods we develop will be probabilistic, i.e. we will output the network structure which is *most likely* according to the method used.

III. TOPOLOGICAL INVERSION

We assume that an unknown network undergoes cascades numerous times and that we are given the times when each node activates in each cascade. Throughout, we will short-handedly denote a directional connection from i to j, and that lack of, as $\overrightarrow{ij}$ and $\overset{\not\to}{ij}$ respectively.

Bayes theorem is frequently used in inverse problems related to networks. It has been successfully applied in several problems where network properties need to be inferred [12–16]. Bayesian methods have also been used to infer Bayesian networks [21].

Let N and E be the network size and edge number. In the absence of any information, the probability that $\overrightarrow{ij}$ for randomly chosen nodes i and j is the fractional edge density (ratio of edges to number of possible edges).

$$P(\overrightarrow{ij} \mid \Omega) = \frac{E}{N(N-1)} \equiv \omega$$

were Ω denotes absence of information.

The Bayes theorem is used to update our probabilities when new information arrives. For events A and B, it states that $P(A|B) = P(A|\Omega)P(B|A)/P(B)$. In the present problem, when we get the data from the first experiment giving us the time when i and j activate (let E_1 denote this event), the theorem gives us

$$\begin{aligned} P_{1;i\to j} &= \omega P(E_1|\overrightarrow{ij})/P(E_1) \qquad (2)\\ P(E_1) &= \omega P(E_1|\overrightarrow{ij}) + (1-\omega)P(E_1|\overset{\not\to}{ij}) \end{aligned}$$

We update our probabilities iteratively. As more cascades happen we get more pairs of times for the activation of i and j.

$$P_{n;i\to j} = P_{n-1;i\to j}P(E_n|\overrightarrow{ij})/P(E_n) \qquad (3)$$

where,

$$\begin{aligned} &P(E_n) = P_{n-1;i\to j}P(E_n|\overrightarrow{ij}) + P_{n-1;i\not\to j}P(E_n|\overset{\not\to}{ij}),\\ &P_{n;i\to j} = P(\overrightarrow{ij}|t_{i;1}, t_{j;1}, t_{i;2}, t_{j;2}, \cdots t_{i;n}, t_{j;n}),\\ &P_{n;i\not\to j} = 1 - P_{n;i\to j} \end{aligned}$$

After all experiments are completed, we will get a probability corresponding to each ordered pair of nodes $P(\overrightarrow{ij} \mid \text{all data})$, and choose E edges with the highest probabilities and infer that they must be true edges.

We must now find how the probability that one node activates at t_1 and the other at t_2 is affected by the presence of a directed edge between the two nodes. To do so, we offer two methods: **(M1)** obtaining it theoretically and **(M2)** obtaining it semiempirically from a surrogate network with similar statistical properties. We can also infer networks heuristically without using Bayes Theorem **(M3)**. The latter method has the advantage that it does not require the degree distribution of the network, but has less overall success and requires more experiments. We find that it is possible to use **(M3)** to obtain the degree distribution when its success is above 80% and then use this as an input for **(M1)** or **(M2)** which give far superior outcomes.

We evaluate the success of our three methods in Fig. 2 in detail for a particular forward model. We evaluate our success in Table 1 for other forward models. In all cases more number of experiments give higher overall accuracy. We supplement this letter with the working code that implements these methods. Further details of our three methods are outlined below.

(1) Theoretical Method. Here we theoretically derive an approximation for $P(t_i, t_j \mid \overrightarrow{ij})$ and $P(t_i, t_j \mid \overset{\not\to}{ij})$. Let $\overrightarrow{ij}$, and j have an indegree k. At a time step when i is inactive, j has a total of $k-1$ providers which could possibly have activated. We assume all of them to be equivalent (i.e. equally likely to have activated). The probability that m of those providers have activated at the given time will be a binomial distribution.

After i has activated, there are still $k-1$ providers to choose from but there is an extra node which has activated. So the probability that j is active is given by

$$Q_j(t) = \sum_k \Gamma(k) \sum_{m=0}^{k-1} B(m, k-1, Q(t-1))h(t) \qquad (4)$$

Where,

$$h(t) = \begin{cases} f(m/k) & t \le t_i \\ f((m+1)/k) & t > t_i \end{cases} \qquad (5)$$

To find $P(t_i, t_j \mid \overrightarrow{ij})$, we need the probability that j activates exactly at t_j. This is equal to the difference of the probabilities that j is not active at $t_j - 1$ and the probability that it is active at t_j. We multiply this by the probability that i activates at t_i.

$$P(t_i, t_j \mid \overrightarrow{ij}) = P(t_j \mid \overrightarrow{ij} \cap t_i)P(t_i)$$

$$P(t_i, t_j \mid \overrightarrow{ij}) = D(t_i)\big[Q_j(t_j) - Q_j(t_j - 1)\big]$$

Note that since the activation time of each node in each cascade is known, $D(t)$ and $Q(t)$ can be obtained by simply counting the number of activations at that time.

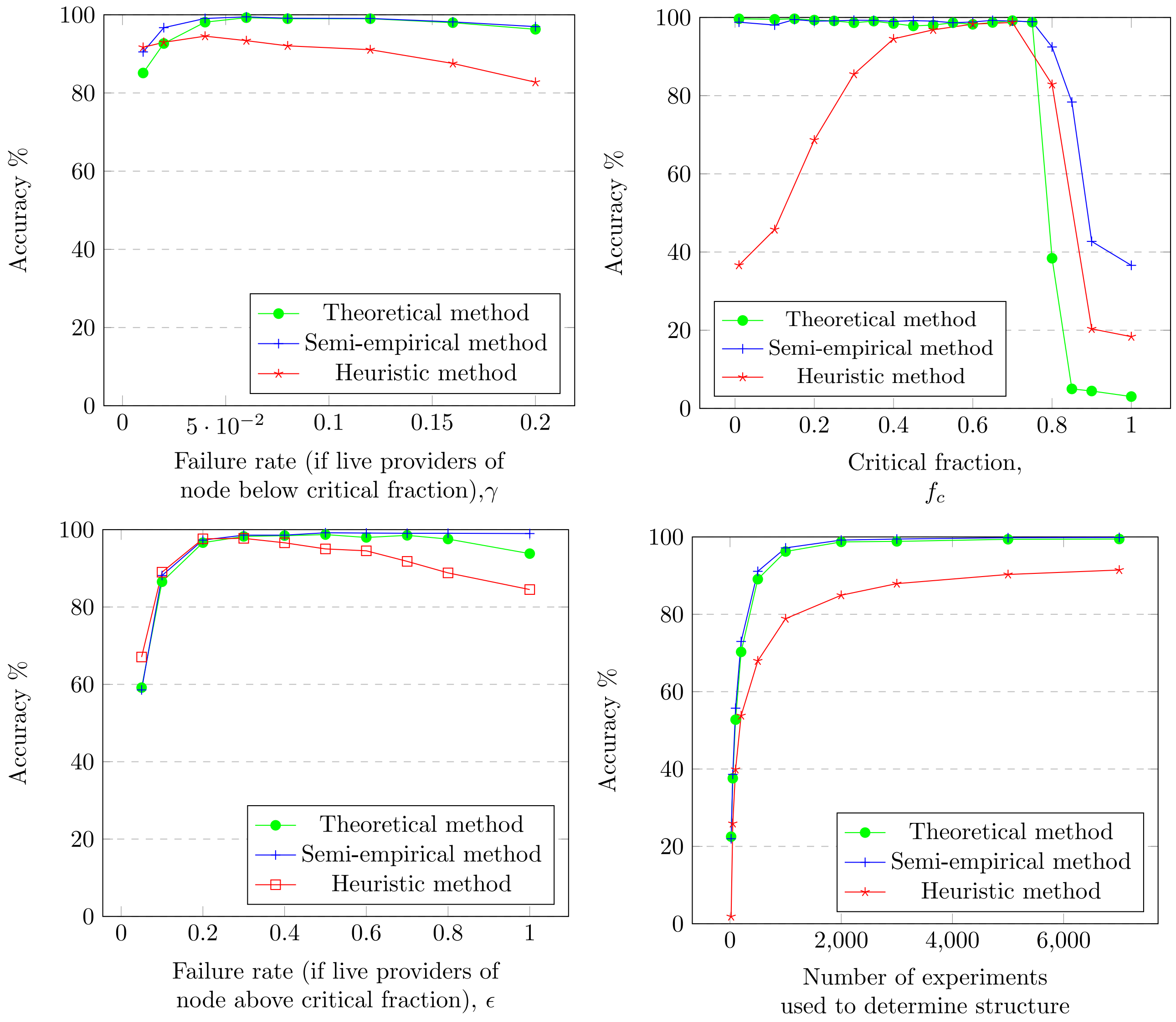

FIG. 1. **Accuracy versus Model Parameters, for a specific model.** We evaluate our success rate using the three reported methods here, for a threshold model $f(m/k)$ that is equal to γ for $m/k < f_c$ and ϵ when $m/k > f_c$. We sweep the parameter space and plot success rate as a function γ, ϵ, f_c, (while keeping the other two constant at $\gamma = 0.04$, $\epsilon = 0.6$ $f_c = 0.4$). For all runs the edges and network size are $E = 1484$ and $N = 200$. Number of experiments(cascades) is 2000 for Semiempirical and Theoretical methods except in the bottom right plot where we plot accuracy vs. number of experiments. For Heuristic method, the number of experiments is appropriately chosen according to the bottom right plot to give high accuracy. We evaluate other models $f(m/k)$ in Table 1

In a large network, the activation of two arbitrarily chosen nodes at two different times are approximately independent. So $P(t_1, t_2) \approx D(t_1)D(t_2)$ for two times t_1 and t_2. This observation is used to obtain $P(t_i, t_j \mid \overset{\not\rightarrow}{ij}$ which is required in the Bayes' theorem (3), as follows-

$$D(t_i)D(t_j) = \omega P(t_i, t_j|\overrightarrow{ij}) + (1 - \omega)P(t_i, t_j|\overset{\not\rightarrow}{ij})$$

(2) Semiempirical Method This is a simple method in which we construct another (surrogate) network with similar statistical properties. Now we can do as many experiments on this network to "measure" $P(t_i, t_j \mid \overrightarrow{ij})$ for all times. Then we use the values of this function in (1) and (2) to get a probability for every entry in the connectivity matrix to be a true edge. A network with the same indegree distribution as that of the unknown network can be easily constructed by starting with an empty network and adding random edges to every node one by one until the exact degree distribution is reached.

(3) Heuristic Method. When the degree distribution is not known, the edges can be considered to be pairs of random variables. Some methods have been developed to infer network structures by finding joint information or correlation between these variables [10]. Here we exploit

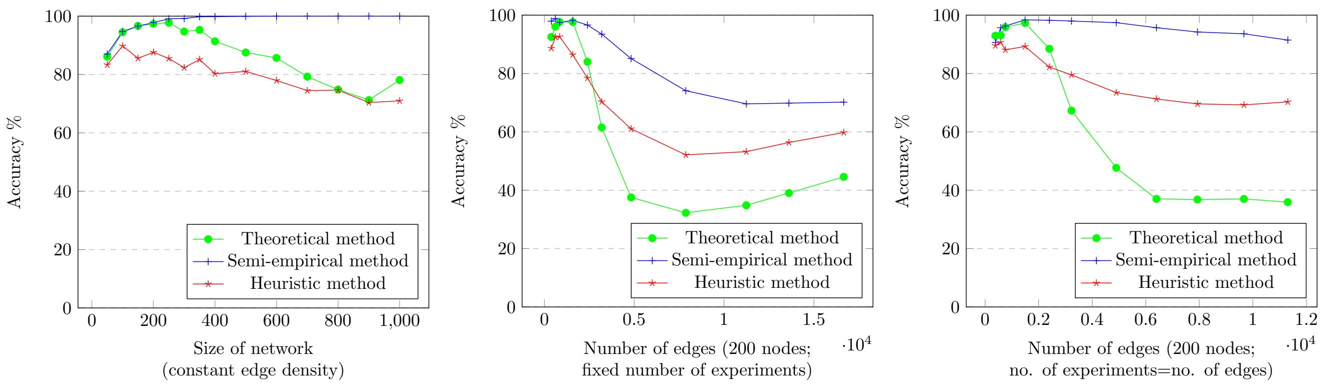

FIG. 2. **Scalability of our methods, for a specific model.** We evaluate our success rate for networks of various sizes and densities. The threshold model with $f_c = 0.4$, $\gamma = 0.04$ and $\epsilon = 0.6$ is used for cascades. Number of experiments(cascades) is 1600 in the first plot, is equal to the number of edges in the second plot and is equal to 0.4 times the number of possible edges (i.e. $0.4N(N-1)$) in the last plot. Edge density if fixed at 4% in the last plot.

$g(m/k)$	Theoretical	Semiempirical	Heuristic
m/k	98.78	99.12	95.05
$(m/k)^2$	95.89	99.12	95.96
$1-(1-m/k)^2$	99.66	99.80	90.36
$1-exp(-m/k)$	97.71	97.57	88.54
$1-exp(-3m/k)$	99.80	99.73	81.67

TABLE I. Accuracy(%) of inversion for some general models of the kind $f(m/k) = 0.04+0.96g(m/k)$, for N=200, E=1563, obtained from 2000 cascades on a random network.

	Threshold Model			$g(m/k) = m/k$		
Expt.	Theo.	SE	Heur.	Theo.	SE	Heur.
20	59.34	64.84	0	48.90	63.74	0
50	78.02	80.22	0	70.33	79.67	45.05
100	84.62	84.62	73.63	78.57	87.36	68.68
200	86.81	85.16	76.92	89.01	91.76	84.62
500	88.46	88.46	85.71	90.11	91.21	86.26

TABLE II. Accuracy(%) of inference for Gagnon and Macrae prison network: N=67, E=182, for when **left:** f(m/k) is a step function (threshold model) with lower value $\gamma = 0.04$, higher value $\epsilon = 0.6$ and threshold point $f_c = 0.4$ and **right:** $f(m/k) = 0.04 + 0.96m/k$

.

the observation that if a node activates at some time, it is quite likely that one or more of its providers activated just before it. We find how often one node activates right after another, and choose the edges between nodes with highest number of such consecutive activations.

IV. EVALUATION

To test the accuracy of our methods we simulated various models on known synthetic and real networks and used the activation time of nodes from the simulations as if experimental data. We then compared our inferred networks to the actual ones.

	Advice,E=480		Discussion,E=565	
Experiments	Theoretical	SE	Theoretical	SE
25	56.67	62.08	55.40	59.29
50	73.12	78.33	72.92	74.69
100	82.92	83.12	82.83	83.36
200	87.92	88.12	89.03	87.61
500	92.71	71.45	92.92	83.62

TABLE III. Accuracy(%) of inference for physician networks (N=246): Threshold model with $\epsilon = 0.6$, $f_c = 0.4$ γ=4%

Experiments	Theoretical	SE	Heuristic
100	64.74	79.49	63.78
200	65.38	84.61	69.23
500	69.87	85.90	73.72
1000	72.75	85.90	76.28

TABLE IV. Accuracy(%) of inference for Zachary's Karate club network: N=34, E=156, γ=4%, $\epsilon = 0.6$, $f_c = 0.4$

.

As a first example, we inverted a generalized version of [8] such that $f(m/k) = \gamma$ if $m/k < f_c$ and ϵ if $m/k \geq f_c$. In other words, a node changes its activation probability if more than a critical number of providers activate. We varied all model parameters for this example and plotted our accuracy in Fig.2.

In all plots and tables, we do not report accuracy as defined by the fraction of correctly identified connectivity matrix elements, but fraction of correctly identified edges. For example, in a network of 100 nodes and 100 edges we must decide whether $\sim 10^4$ entries of the connectivity matrix is a 0 or 1. If we identify 10 false edges (and hence, also not identify 10 true edges), we report our accuracy rate as $90/100 = 90\%$ instead of $9980/10^4 = 99.8\%$. In addition to the threshold model we also evaluate others models (without varying all possible parameters of these models). Our success rates are reported in Table 1.

We tried to infer friendships between inmates of the

Gagnon and Macrae prison using synthetic data. The network consists of 67 prisoners(nodes) which have 182 friendships (edges) [22]. Success rates are reported in Table 2. We have also used our methods on some undirected graphs such as the Zachary's Karate club network [23]. It has 34 members of a karate club (nodes) and have 156 friendships (edges) between members. The experiments simulated resemble studying the spread of an opinions and practices among friends. γ is included to represent opinion formation due factors other than friends, success rates are reported in Table 3.

Several physicians were surveyed in [24] and [25] to study how information about a new medicine spreads among physicians that do friendly discussions or take professional advice. This was later modeled as a network problem in [26], and effects of marketing were studied in [27]. In our simulations, γ simulates the effect of marketing and the jump at f_c assumes that a physician starts prescribing a medicine with an increased probability ϵ if their colleagues prescribe it. The results of inferring physicians' relationships with their colleagues using synthetic data of cascades (i.e. medicine prescriptions) is given in Table IV.

See Supplemental Material at [28] for computer programs of all of our inversion methods and instructions for using them.

V. LIMITATIONS

We conclude our study by discussing our limitations. Our methods do not produce accurate results when the critical fraction is so high that most nodes activate not due to interactions, but randomly. Since in this case, the structure of the network plays little role in the cascade dynamics, it becomes difficult to extract the structure. We also observe that the theoretical method does not work well for very dense networks (Fig. 2). This is because our (approximate) formulas depend only on the indegree distribution $\Gamma(k)$. However, in dense networks, higher order, conditional indegree distributions (such as the probability that a degree k node has a degree m connection $\Gamma(k,m)$) plays an important role. The semiempirical method works best for random networks and its success is slightly lower for other kinds of networks. This is because we match only the indegree distribution of the surrogate network and the outdegree distribution may not be well matched for other kinds of networks. This method is essentially a binary classification of individual edges, but we can also calculate conditional probabilities of trees in the network using Bayes theorem. Relying only on binary classification leads to poor accuracies at higher link densities where higher order structures like trees and cycles play a major role in cascade propagation. Another limitation can be seen in Table IV, where as the number of experiments increases, the accuracy may decrease. This is a common and well-known issue with naive Bayesian classifiers [29]. Lastly, the binomial approximation in (1) works less successfully in networks for which the providers have different likelihood of activating. Nevertheless, our success with heterogeneous networks (cf. Tables II-IV) show that this inaccuracy is not very crucial.

[1] M. G. Rodriguez, J. Leskovec, and A. Krause, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2010) pp. 1019–1028.

[2] T. M. Snowsill, N. Fyson, T. D. Bie, and N. Cristianini, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2011) pp. 466–474.

[3] V. Ntziachristos, Nature methods , 603 (2010).

[4] D. Berry and S. Widder, Frontiers in microbiology **5**, 219 (2014).

[5] M. G. Rodriguez, J. Leskovec, and B. Schlkopf, in *Proceedings of the sixth ACM international conference on Web search and data mining* (ACM, 2013) pp. 23–32.

[6] M. Gomez Rodriguez, B. Schölkopf, L. J. Pineau, *et al.*, in *29th International Conference on Machine Learning (ICML 2012)* (International Machine Learning Society, 2012) pp. 1–8.

[7] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. Schlkopf, Network Science **2**, 26 (2014).

[8] D. J. Watts and P. S. Dodds, Journal of consumer research **34**, 441 (2007).

[9] B. A. Desmarais, J. J. Harden, and F. J. Boehmke, American Political Science Review **109**, 392 (2015).

[10] S. Hempel, A. Koseska, J. Kurths, and Z. Nikoloski, Physical Review Letters **107**, 054101 (2011).

[11] S. Myers and J. Leskovec, in *Advances in Neural Information Processing Systems* (2010) pp. 1741–1749.

[12] M. Komorowski, B. Finkenstdt, C. V. Harper, and D. A. Rand, BMC bioinformatics **10**, 1 (2009).

[13] D. J. Wilkinson, Briefings in bioinformatics **8**, 109 (2007), pmid:17430978.

[14] A. Golightly and D. J. Wilkinson, Biometrics **61**, 781 (2005).

[15] R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood, Statistics and Computing **18**, 125 (2008).

[16] K. Faust and J. Raes, Nature Reviews Microbiology **10**, 538 (2012).

[17] E. Adar and L. A. Adamic, in *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence* (IEEE Computer Society, 2005) pp. 207–214.

[18] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, in *Proceedings of the 13th international conference on World Wide Web* (ACM, 2004) pp. 491–501.

[19] D. C. Vural, G. Morrison, and L. Mahadevan, Physical Review E **89**, 022811 (2014).

[20] J. P. Gleeson, Physical Review E **77**, 046117 (2008).

[21] N. Friedman and D. Koller, Machine learning **50**, 95 (2003).

[22] D. MacRae, Sociometry **23**, 360 (1960).

[23] W. W. Zachary, Journal of anthropological research , 452 (1977).

[24] R. S. Burt, American journal of Sociology , 1287 (1987).

[25] J. Coleman, E. Katz, and H. Menzel, Sociometry **20**, 253 (1957).

[26] T. W. Valente, *Network models of the diffusion of innovations*, 303.484 V3 (1995).

[27] C. Van den Bulte and G. L. Lilien, American Journal of Sociology **106**, 1409 (2001).

[28] .

[29] I. Rish, in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3 (IBM New York, 2001) pp. 41–46.