The Impact of Graph Structure on Small-World Shortest Paths

Pegah Hozhabrierdi and Reza Zafarani

Data Lab, Syracuse University
Syracuse, NY 13244, USA
{pegah, reza}@data.syr.edu

Abstract. Numerous studies have been devoted to modeling and estimating shortest-paths in complex networks. To maintain generality, these studies have neglected a common property of complex social networks; small-world phenomenon (colloquially stated as *six degrees of separation*). Based on the intuition behind the flow of information in small-worlds, we propose a small-world representation for social networks. In this new representation, we study the influence of different network measures on the shortest-paths. We perform a comprehensive analysis on a large set of local and global network measures and report our findings for various social networks. The results of our analyses show that: (1) shortest path lengths in small worlds are strongly correlated to the maximum degree centrality and the diameter. In fact, using these two features one can predict the average path length more accurately than using any other feature alone; (2) when nodes are ranked according to their average shortest-path lengths, we can approximate this ranking by a shifted standard normal distribution with minimum information loss. The shift can be estimated by the rank of the node with maximum local clustering coefficient, which can be computed in linear or constant time [11].

Keywords: small-world · graph structure · shortest path · community.

1 Introduction

Network connectivity and shortest paths have been widely used to study information diffusion and rumor propagation [23]. Shortest paths provide the fastest and, usually, the strongest interaction between actors (nodes) in a network [16]. Theoretical studies on shortest paths in social networks often neglect one of the most well-known properties of these networks: small-world phenomenon [13,15,20]. Small-world phenomenon, first popularized by Milgram in the 60's [19], indicates that individuals in a social network are connected via short paths of friendships. Later studies found a similar pattern in online social networks and further extended the theory behind this phenomenon [12,17,21]. In this study, we analyze the relationship between measures that define a network structure and the small-world shortest paths. We introduce the *small-world representation* of a social network and discuss how it can realistically model the flow of information.

Using the small-world representation of 10 real-world social networks, we investigate the relationship between a large set of local and global network properties and the shortest paths. In this process, we make the following contributions:

- **Small-World Representation**. We offer a simplified representation for social networks which models a realistic flow of information (Section 2);
- **Average Shortest-Path Length Analysis**. We consider local and global network measures and show the correlation and importance of each measure with regards to the average shortest path lengths. Using this information, we build a linear model to predict the average shortest path lengths and present the challenges (Section 3.1); and
- **Shortest-Path Distribution and Local Information**. Finally, we focus on local information with low-cost computation (Section 3.2). If nodes in the graph are ranked based on their average shortest-path length, we show that a shifted standard normal distribution is capable of approximating this ranking with minimum information loss. The required shift is estimated by the rank of the node with maximum local clustering coefficient.

2 Data Description

The focus of this study is to investigate the real-world networks of individuals, be it online or physical. The ideal dataset for this study is a real-world network with (1) nodes representing human users, clients or members and (2) edges representing relationships between those nodes that are (3) large enough to provide statistically significant observations. The fundamental concept behind small-world is *reachability* of the users. Hence, a relationship between users that do not enable them to contact or reach each other is not of interest. For example, the connections between individuals who are recipients of the same email do not imply that these users can necessarily reach each other. Reachability through edges in a network can be inferred via the network's *small-world representation*.

Definition 1. *(Small-world representation). The small-world representation of network $G = (V, E)$ is the undirected network $G' = (V, E')$ such that for all $(v_i, v_j) \in E$, there exists exactly one edge $(v_i, v_j) \in E'$ that represents the flow of information from v_i to v_j and vice versa.*

The intuition behind this definition can be better understood from Milgram's broker experiment [19]. In his experiment, Milgram chose a set of individuals at random and asked each of them to send a letter to a specific broker through their connections. Each individual had to choose a person among their acquaintances to pass the letter on. Intuitively, the chosen candidate should have the highest possibility to reach the broker through his/her connections. The flow of information from A to B (i.e. passing the letter from A to B) was entirely dependent on the 'acquaintanceship' of A and B. A way to extend this experiment to *virtual societies* (e.g. social media platforms) is through asking someone to pass a message, rumor, or news to a target individual using only their acquaintances.

Dataset	Edge Type	|V|	|E|	Avg. Deg.	Max Deg.	Clustering Coeff. (%)	Diameter	ASP
zachary	tie	34	78	4.59	17	57.06	5	2.41
train_bombing	contact	64	243	7.59	29	62.23	6	2.69
residence_hall	friendship	217	1,839	16.95	56	36.28	4	2.39
haggle	contact	274	2,124	15.5	101	63.27	4	2.42
infectious	contact	410	2,765	13.49	50	45.58	9	3.63
hamster	friendship	2,000	16,098	16.1	273	54.01	10	3.59
adolescent_health	friendship	2,539	10,455	8.24	27	14.67	10	4.56
ego_facebook	friendship	2,888	2,981	2.06	769	2.72	9	3.87
advogato	trust	5,042	39,227	15.56	803	25.27	9	3.27
pretty_good_privacy	interaction	10,680	24,316	4.55	205	26.59	24	7.49

Table 1: The network characteristics of 10 real-world datasets. This information belongs to the small-world representation of each network. Abbreviations: Avg.: average, Deg.: degree, Coeff.: coefficient, ASP: average shortest path.

The acquaintanceship in online networks cannot be defined as straightforward as in physical societies. For example, in an online network like Twitter, one might claim that the *Following* relationship makes a one-sided flow of information from the followee to the follower but not vice versa. We argue that, in terms of the information flow in Milgram's small-world experiment, the flow of information can go from the follower to followee as well. Consider A, the subject of our experiment, to follow B and B to follow C. If A is asked to pass a message to C through his/her acquaintances, B will be the optimal receiving end of the message despite the fact that B does not follow A. In general, in networks like Twitter, posting content to be seen by one's followers is not the only way of transferring information. Another way is to receive content from the people whom one is followed by in different ways such as *tagging* a person. As a result, we find the small-world representation of a directed social network, such as Twitter, a more reasonable graph model to study paths that deal with information flow. The small-world representation of our example, Twitter network, is its undirected counterpart. Following this strategy, we have selected ten real-world networks with a type of connection among individuals that has a small-world representation. In the following, we introduce each network's type of connection and how they can be modeled as small-world graphs.

1. **Zachary's Karate Club [22]:** an undirected network of ties among members of a Karate club after the club splits into two groups.
2. **Train Bombing [9]:** an undirected network of contacts among the suspected terrorists in Madrid's train bombing incident in 2004. The original network contains edge weights to show the strength of the connections. However, these weights do not change the *reachability* of the nodes; i.e. an edge between terrorists i and j implies that i can contact j and vice versa regardless of the strength of their relationship (there are no edges with weight zero). So, the small-world representation of the network is the unweighted counterpart of this graph.
3. **Residence Hall [8]:** a directed network of friendships among residents of a residence hall on Australian National University campus. A directed edge from i to j shows that i considers j to be a friend. This also implies that

i and j know each other whether j considers i as a friend or not. So, the small-world representation will be the undirected counterpart of this graph.

4. **Haggle [4]:** an undirected network of contacts between individuals, obtained through carried wireless devices.
5. **Infectious [6]:** a multi-edge undirected network of face-to-face contacts among exhibition visitors at Dublin's Science Gallery in 2009. The contacts have been active for at least 20 seconds and multiple contacts could have occurred between two individuals. The small-world representation of this network is the single-edge undirected counterpart.
6. **Hamster [5]:** an undirected network of friendships between users in Hamster online social network.
7. **Adolescent Health [1]:** a weighted directed network of friendships among students created from a survey in 1994/1995. Each student was asked for the name of his/her top five friends and the edge weights show the frequency of interaction between them. The small-world representation of this network is the unweighted undirected counterpart (same as *residence hall* dataset).
8. **Ego Facebook [3]:** an undirected friendship network of Facebook users.
9. **Advogato [2]:** a directed network of trust among developers in Advogato platform. The edges have positive weights (the amount of trust between two users) and the nodes can contain self-loops (one can trust himself). The trust between user i and j can imply the prior acquaintanceship between i and j that makes the flow of information possible in both directions. Hence, the small-world representation of this network is the undirected network with no weights or self-loops.
10. **Pretty Good Privacy [7]:** the undirected interaction network between users through Pretty Good Privacy (PGP) software.

In all networks, unless otherwise stated, the small-world representation is the same as the original network. The detailed information of these small-world representations can be found in Table 1.

3 Small-World and Network Structure

A network structure can be described through various measures, such as degree distribution, average local clustering coefficient, number of triads, diameter and centrality measures. We investigate a comprehensive set of network measures and group them into three categories:

- **General Network Measures.** number of nodes and number of edges
- **Local Network Measures.** node degrees (degree distribution, average degree, and maximum degree), clustering coefficient (local clustering coefficient distribution and average clustering coefficient), node centrality (degree centrality distribution, eigenvector centrality distribution, average degree centrality, and average eigenvector centrality).
- **Global Network Measures.** diameter, communities (number of communities, community coverage, and community size), number of bridges, degree

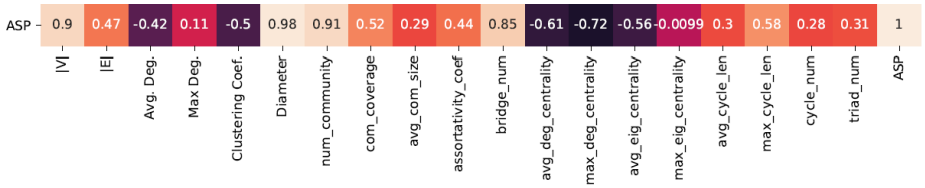


Fig. 1: Correlation between Network Measures and the Average Path Length.
From left to right: average shortest path length (ASP), number of nodes, number of edges, average degree, maximum degree, clustering coefficient, diameter, number of communities, community coverage, average community size, assortativity coefficient, number of bridges, average degree centrality, maximum degree centrality, average eigenvector centrality, maximum eigenvector centrality, average cycle length, maximum cycle length, number of cycles, and number of triads.

assortativity coefficient, special subgraphs (number of triads, number of cycles, the average cycle length, and maximum cycle length).

Local measures can be obtained for any subset of nodes in the graph while *global measures* require all the nodes/edges in the graph.

Community Detection. Finding meaningful communities in a network relies, mainly, on the type of network and the relationships presented in the graph [18]. Intuitively, we expect the shortest paths in a graph to be heavily influenced by the density of the connected components. Having components with higher internal edge-density (intra edges) and comparatively less edge-density when connecting to other components (inter edges) can trap the paths inside the components. Following this intuition, among many existing community detection algorithms, we choose modularity maximization that relies on the network topological structure rather than node/edge similarities. To speed up the process, we choose Louvain algorithm [10], which optimizes modularity gain. We consider the number and size of the communities, and the average community coverage as network measures. The community coverage is defined as

$$C = \frac{\sum_i e_i}{2|E|}, \qquad i = 1, ..., k \tag{1}$$

where we have k communities and e_i is the number of intra-community edges.

3.1 Average Shortest Path Length Analysis

We start our analysis by considering the network measures that have an impact on the average shortest path length of the network.

Correlation. Figure 1 shows how different network measures correlate with the average shortest path (ASP in the Figure) length of the network. A significance test on these correlations reveals significant correlations (p-value less than 0.05) between average path length and the (1) diameter (0.98), (2) number of

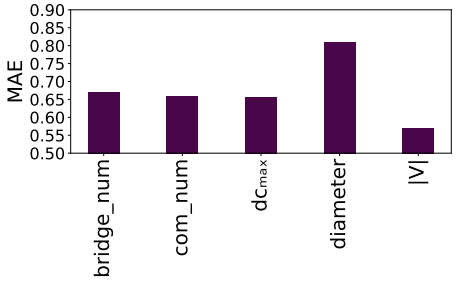


Fig. 2: Mean Absolute prediction Error (MAE) in absence of each feature. Higher MAE indicates higher importance of the feature in predicting ASP. From left to right: number of bridges, number of communities, maximum degree centrality, and number of nodes.

communities (0.91), (3) number of bridges (0.85), (4) maximum degree centrality[1] (-0.72), and (5) number of nodes (0.90). The strong correlation between diameter and number of bridges, and average path length was to be expected; the former is derived from shortest paths and the latter limits the flow of information between connected components. Interestingly, the number of communities has the strongest positive correlation with average path length after diameter while the size and coverage of these communities seem to be completely uncorrelated.

Regression. To test the significance of these correlations, we fit a linear regression model on the data and predict the average path length for each network using cross-validation. Figure 2 shows the importance of each feature in training the regression model based on the Mean Absolute Error (MAE) of the predictions using the leave-one-out approach. The high MAE for the absence of *diameter* indicates the importance of this feature, which was expected. However, our further experiments with other subsets of these five features showed that the combination of *diameter* and *maximum degree centrality* gives the lowest prediction error[2] (0.27). Figure 3 depicts the actual and estimated average shortest path length of the networks using the regression model built on diameter and maximum degree centrality. The fitted model is $0.24 \cdot Diameter - 0.98 \cdot \frac{d_{\max}}{|V|-1} + 1.69$, where the weights are rounded to the nearest second decimal.

In practice, global measures (e.g. diameter) are computationally expensive. Previous studies, such as that conducted by Chechik et al. [14], offer relatively time efficient algorithms to estimate the diameter. These algorithms, however, take considerably more time than the computation of local measures (e.g. max-

[1] Note the difference between the ASP correlation with *maximum degree* (d_{max}) and *maximum degree centrality*. Maximum degree centrality is measured as $\frac{d_{max}}{|V|-1}$ and $|V|$ is different for each network. As each maximum degree is normalized by a different value, the difference in correlation is to be expected.

[2] The mean absolute error of predicting ASP using only maximum degree centrality and diameter.

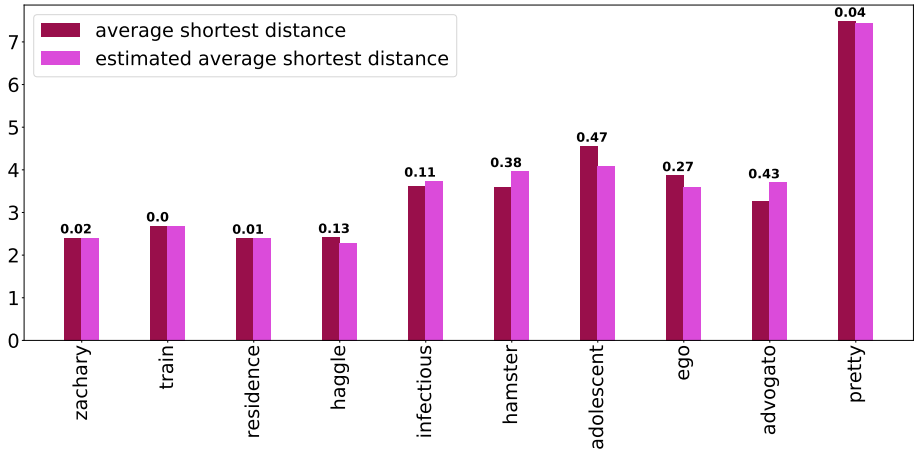


Fig. 3: Estimated sverage Shortest path lengths vs. the actual average shortest path lengths using linear regression on the diameter and maximum degree centrality. The numbers on top of each pair of bars indicates the absolute error of prediction.

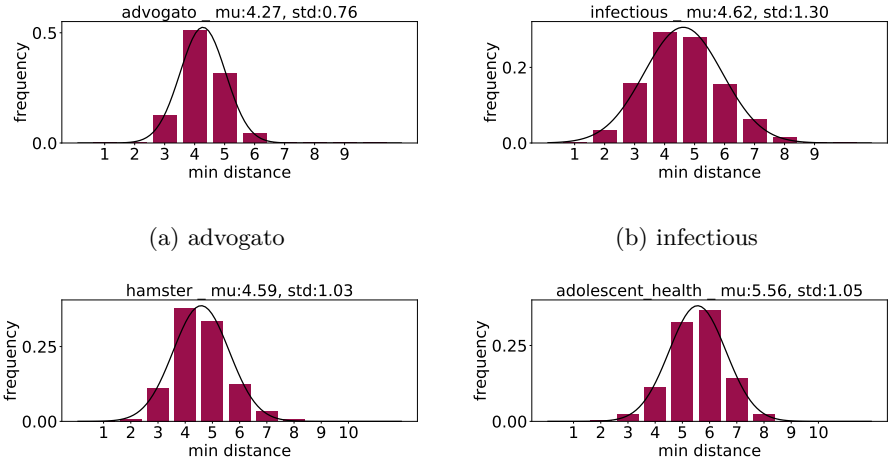
imum degree centrality). In this respect, it is more practical to study the relationship between local network measures and average path lengths. Next section will discuss this relationship in more depth.

3.2 Shortest-Path Distribution and Local Information

In this section, we focus on distributions of local information, i.e. degree distribution, degree centrality distribution, eigenvector centrality distribution, and local clustering coefficient distribution. To infer a meaningful comparison between these distributions and that of shortest paths, we consider the shortest-path distribution of each node. In this distribution, a Shortest-Path Number (SPN) is assigned to each node which is defined in equation 3. The sum of SPN index for all nodes in the graph is $|V|$ times the average path length of the graph. For each node i in graph G, the shortest-path number of i is defined as

$$\mathrm{SPN}(i) = \frac{\sum_{j \neq i} d_{\min}(i,j)}{|V| - 1} \tag{2}$$

We also used the SPN defined as the median of the shortest paths from i which gave the same results as the average. The distribution of SPNs for all nodes in the graph (SPN distribution) is of our interest. We use Kullback-Leibler (KL) divergence to measure the difference between SPN distribution and that of local distributions. We also test the SPN distribution against a modified version of three standard distributions:



(a) advogato (b) infectious

(c) hamster (d) adolescent

Fig. 4: The Shortest-path distribution can be modeled by normal distribution. We chose four datasets randomly for brevity but the same bell-shape has been observed in all of our datasets.

1. **Naïve Uniform**: This distribution models the *random guess* for predicting the SPN of a node. We use this model to test the significance of KL divergence. Any KL divergence value above the corresponding value in naïve uniform model is insignificant. In this model, it is assumed the SPN of each node is drawn from a uniform distribution between the minimum and maximum possible SPN in a graph. Nodes with degree $|V|-1$ give the minimum possible SPN and maximum SPN occurs if the nodes form a chain such as in $a \to b \to c$. In this case, the maximum SPN from equation 2 will be

$$\text{SPN}_{\text{max}} = \frac{1+2+...+|V|-1}{|V|-1} = \frac{|V|}{2} \tag{3}$$

   So, the naïve uniform will be defined as $Unif(1, \frac{|V|}{2})$.
2. **Small-World Uniform**: This model is an improvement of random guess. Naive uniform models the true random guess for the SPN of each node with no prior knowledge about the network. However, from small-world phenomenon, we know that the average shortest path from each node is most probably a number less than 10. We use this prior knowledge to make more educated guesses with uniform distribution. We estimate the SPN_{max} as

$$\text{SPN}_{\text{max}} \approx \text{SPN}(n_{mcc}) \tag{4}$$

   in which n_{mcc} represents the node with the highest local clustering coefficient (LCC) in the graph. This choice has been made due to (1) the fast calculation

Dataset	deg. dist.	deg. cent.	eig. cent.	local clust. coeff.	naïve uniform	uniform	normal
zachary	0.32	0.32	0.24	0.29	0.14	0.04	**0.03**
train_bombing	0.44	0.44	0.88	0.57	0.17	0.09	**0.06**
residence_hall	0.13	0.13	0.26	0.09	0.21	0.08	**0.04**
haggle	1.29	1.29	1.13	0.65	0.23	0.07	**0.04**
infectious	0.28	0.28	1.66	0.19	0.24	0.1	**0.02**
hamster	0.67	0.67	1.53	0.19	0.25	0.1	**0.02**
adolescent_health	0.19	0.19	1.43	0.35	0.24	0.11	**0.01**
ego_facebook	1.57	1.57	1.78	1.82	0.27	0.07	**0.04**
advogato	1.06	1.06	1.24	0.32	0.25	0.09	**0.03**
pretty_good_privacy	0.78	0.78	0.38	0.19	0.26	0.15	**0.02**

Table 2: KL divergence between SPN distribution and local information distributions. Abbreviations from left to right: degree distribution, degree centrality, eigenvector centrality,and local clustering coefficient.

of LCC, and (2) the relatively small KL divergence between LCC distribution and SPN (see Table 2).

3. **Estimated Normal:** The intuition behind choosing this distribution is the bell shape of the shortest-path length distribution appearing in all of our datasets (Figure 4). We use a standard normal distribution which is shifted by $\text{SPN}_{\max}$ as defined in 4.

The KL divergence between all distributions and SPN can be found in Table 2. From the table, it is evident that standard normal distribution shifted by the approximation in 4 models the SPN distribution with the least information loss. This result confirms our previous intuition in Figure 4. In majority of our datasets, local information perform worse than a random guess (naïve uniform). Using small-world phenomenon for more educated random guesses improves the similarity measure significantly. This can be considered as another proof of the existence of small-world in all social networks used in this study.

4 Conclusion

In this paper, we studied the relationship between different network measures in real-world social networks and their average path lengths. We found that a regression model built on the combination of maximum degree centrality ($\frac{d_{\max}}{|V|-1}$) and network diameter can predict the average path length with high accuracy. This shows the surprising importance of maximum degree centrality in predicting the average shortest path length. Our experiments on local network measures based on the defined ranking measure, SPN, shows that these measures do not exhibit similar strong prediction abilities. However, a standard normal shifted by the SPN of the node with the maximum local clustering coefficient can predict the shortest-path distribution with the least information loss.

References

1. Adolescent health network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/moreno_health
2. Advogato network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/advogato
3. Facebook (nips) network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/ego-facebook
4. Haggle network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/contact
5. Hamsterster full network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/petster-hamster
6. Infectious network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/sociopatterns-infectious
7. Pretty good privacy network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/arenas-pgp
8. Residence hall network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/moreno_oz
9. Train bombing network dataset – KONECT (Apr 2017), http://konect.uni-koblenz.de/networks/moreno_train
10. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
11. Boot, C.: Algorithms for Determining the Clustering Coefficient in Large Graphs. B.S. thesis (2016)
12. Centola, D.: The spread of behavior in an online social network experiment. science **329**(5996), 1194–1197 (2010)
13. Chai, Y., Song, C., Nie, P., Yuan, X., Ge, Y.: Community structure based shortest path finding for social networks. In: International Conference on Database and Expert Systems Applications. pp. 303–319. Springer (2018)
14. Chechik, S., Larkin, D.H., Roditty, L., Schoenebeck, G., Tarjan, R.E., Williams, V.V.: Better approximation algorithms for the graph diameter. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. pp. 1041–1052. Society for Industrial and Applied Mathematics (2014)
15. Gubichev, A., Bedathur, S., Seufert, S., Weikum, G.: Fast and accurate estimation of shortest paths in large graphs. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 499–508. ACM (2010)
16. Katzav, E., Biham, O., Hartmann, A.K.: Distribution of shortest path lengths in subcritical erdős-rényi networks. Physical Review E **98**(1), 012301 (2018)
17. Kleinberg, J.: The small-world phenomenon: An algorithmic perspective. Tech. rep., Cornell University (1999)
18. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Physical review E **80**(5), 056117 (2009)
19. Milgram, S.: The small world problem. Psychology today **2**(1), 60–67 (1967)
20. Potamias, M., Bonchi, F., Castillo, C., Gionis, A.: Fast shortest path distance estimation in large networks. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 867–876. ACM (2009)
21. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. nature **393**(6684), 440 (1998)
22. Zachary, W.: An information flow model for conflict and fission in small groups. J. of Anthropological Research **33**, 452–473 (1977)
23. Zhang, R., Li, D.: Rumor propagation on networks with community structure. Physica A: Statistical Mechanics and its Applications **483**, 375–385 (2017)