Optimal network clustering for information diffusion

Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn

School of Informatics and Computing, Indiana University, Bloomington, IN 47408 USA
(Dated: January 8, 2014)

We investigate the impact of community structure on information spreading with the linear threshold model. Contrary to the common belief that communities hinder information diffusion, we show that strong communities can facilitate global cascades by enhancing local, intra-community spreading. Using both analytical approaches and numerical simulations, we demonstrate the existence of optimal clustering, where global cascades require the minimal number of early adopters.

PACS numbers: 89.75.-k Complex systems; 89.65.-s Social and economic systems; 89.20.Ff Computer science and technology

The study of information diffusion — fads, diffusion of innovations, collective action, viral memes, etc. — is relevant to a number of disciplines, including mathematical, physical and social sciences, communication, marketing and economics [1–6]. The most common approach is focusing on the resemblances between information contagion and infectious diseases [7, 8]; a piece of information can travel from one individual to another through social contacts and the ‘infected’ individuals can, in turn, propagate the information to others, possibly generating a large-scale cascade, like an epidemic outbreak [9, 10]. In addition to classical epidemic models, two main types of information diffusion models have been proposed: the independent cascade model, which was initially adopted to study the dynamics of viral marketing [11–15], considers information contagion as a simple branching process; the threshold model, which was originally proposed to study collective social behaviors [2, 16–18], incorporates the idea of ‘social reinforcement’, assuming that each adoption requires a certain number of exposures.

Since information spreads through social contacts, the structure of the underlying social network is a crucial ingredient in studying information diffusion. In particular, the role of hubs and degree distribution have been studied extensively, mainly because of their critical role in epidemic spreading [19–21]. However, it is not yet fully understood how information diffusion differs from epidemic spreading and how the impact of network structure is changed by the differences. One notable difference is the social reinforcement: unlike epidemic spreading where each exposure acts independently, each additional exposure to the information seems to markedly increase the probability of adoption [22–24]. As a consequence, it has been suggested that clustering plays a more important role in information spreading than in epidemics [6, 22–24]. These results suggest that clustering is a crucial element in understanding information diffusion, and that social reinforcement is a key ingredient in modeling information diffusion.

Some studies have investigated the role of communities in information diffusion [25–30], but most of them either overlook the effect of social reinforcement or do not fully explore the implications of community structure. The most common notion about the role of communities is that they simply hinder the spreading of information, keeping it confined in the community of origin, as one would expect for epidemics [25, 31]. However, recent empirical work has suggested that clustering may, counterintuitively, facilitate information diffusion [23] and therefore may provide an important clue to understanding fundamental differences between epidemics and information contagion [6].

In this letter we use the linear threshold model — which incorporates the simplest form of social reinforcement — to study how the global information cascades are affected by community structures. We expose two roles played by community structure: enhancing local spreading and hindering global spreading. Strong communities facilitate social reinforcement and thereby enhance local spreading [6, 24]; weak community structure makes global spreading easier, because it provides more bridges between communities. The more cohesive communities are, the better information can spread locally and the worse it can spread between communities. We show that there exists an optimal balance between these two effects, where community structure counterintuitively enhances — rather than hinders — global cascades of information. This is analogous to the small-world phenomenon where a small number of shortcuts greatly reduces the average path length of the network while maintaining high clustering [32]; a small number of bridges between communities allow inter-community spreading while maintaining intra-community spreading.

We adopt the linear threshold model to account for recent observations and experiments that have demonstrated the relevance of social reinforcement in the diffusion of information [6, 22–24]. Let us formally define the linear threshold model first. Consider a set of \( N \) nodes (agents) connected by \( M \) undirected edges. The state of an agent \( i \) at time \( t \) is described by a binary variable \( s_i(t) = \{0, 1\} \), where 1 represents ‘active’ state and 0 indicates ‘inactive’. At time \( t = 0 \) a fraction \( \rho_0 \) of randomly selected agents, ‘seeds’, is initialized in the active state. At each time step, agents’ states are updated...
synchronously according to the following threshold rule:

\[
s_i(t+1) = \begin{cases} 
1 & \text{if } \theta k_i < \sum_{j \in \mathcal{N}(i)} s_j(t), \\
0 & \text{otherwise}, 
\end{cases}
\]

where \(\theta\) is the threshold parameter, \(k_i\) is the degree of node \(i\), and \(\mathcal{N}(i)\) the set of \(i\)'s neighbors. This rule implies that (i) the dynamics is deterministic; (ii) once a node becomes active, it will remain so forever; and (iii) if \(s_i(t+1) = s_i(t)\) for all nodes, then the system is in a steady state. The linear threshold model exhibits various critical behaviors. For instance there is a critical threshold parameter at which a single active node can trigger a global cascade [17]: there also exists a sharp transition, at a constant threshold parameter, from an inactive state with no cascade to an active state with global cascades, triggered at a critical fraction of initially active nodes [33]. We focus on the latter transition based on the number of seeds and let \(\theta\) constant throughout this letter.

To systematically investigate the impact of community structure, we prepare an ensemble of networks with two communities with varying degree of strength, using the block-model approach [34–36]. First, half of the nodes are randomly selected and assigned to community \(A\), and the other half are assigned to community \(B\). Then, \((1 - \mu)M\) links are randomly distributed among node pairs in the same community and \(\mu M\) is randomly distributed among node pairs that belong to different communities (see Fig. 1). \(\mu\) controls the strength of the community structure; a large value of \(\mu\) means more links between the two communities and thus weak community structure. Finally, we plant the seeds in \(A\), assuming that the contagion originates from the community \(A\).

Let us introduce two analytical approaches — mean-field (MF) and tree-like (TL) approximations — to understand the behaviors of our system. We first assume that the underlying network has a given degree distribution \(p(k)\) but is otherwise random. We aim to compute the final density of active nodes (\(\rho_\infty\)) given the initial density of seeds (\(\rho_0\)). When there is no community structure, using the mean-field approximation, \(\rho_\infty\) can be computed as the smallest stable solution of the equation:

\[
\rho_\infty = \rho_0 + (1 - \rho_0) \sum_{k=1}^{\infty} p(k) \sum_{m=\lceil \theta k \rceil}^{\infty} \binom{k}{m} \rho_\infty^m (1 - \rho_\infty)^{k-m}.
\]

(1)

The probability that a node of degree \(k\) is in the active state at stationarity is the sum of two contributions: (i) the probability that the node is active at \(t = 0\) (\(\rho_0\)), and (ii) the probability that the node is not active at \(t = 0\) \((1 - \rho_0)\) but has at least \(\theta k\) active neighbors at \(t = \infty\) (the second summation). The sum over \(k\) accounts for the different degrees a node may have. The equation can be solved iteratively.

Now let us extend Eq. 1 to deal with networks with communities. While is easy to generalize for arbitrary configurations of communities, here we focus on the case with two communities. When there are two communities the equations for the fraction of active nodes \(\rho^A\) (resp., \(\rho^B\)) in the community \(A\) (resp., \(B\)) can be written as:

\[
\rho_\infty^A = \rho_0^A + (1 - \rho_0^A) \sum_{k=1}^{\infty} p(k) \sum_{m=\lceil \theta k \rceil}^{\infty} \binom{k}{m} q^A_m (1 - q^A_m)^{k-m}.
\]

(2)

where \(\rho_0^A\) is the density of seeds in the community \(A\), and \(q^A = (1 - \mu)\rho_\infty^A + \mu \rho_\infty^B\) is the probability that a neighbor of a node is active, which is the sum of: (i) the probability that the neighbor is in the same community \((1 - \mu)\) and is active \((\rho_\infty^A)\), and (ii) the probability that it is in the other community \((\mu)\) and is active \((\rho_\infty^B)\). Finally, \(\rho_\infty = (\rho_\infty^A + \rho_\infty^B)/2\).

A more sophisticated framework is the tree-like (TL) approximation [26, 37]. This approximation considers trees of infinite depth and assumes that the nodes at level \(n\) is only affected by the ones at level \(n - 1\). The fraction of active nodes in community \(A(B)\) is computed using an auxiliary variable \(y^A_\infty\) obtained by the following iteration over all the levels in the tree:

\[
\bar{y}^A_{n+1} = \rho_0^A + (1 - \rho_0^A) \sum_{k} \frac{k}{z} p(k)
\]

\[
\times \sum_{m=\lceil \theta k \rceil}^{k-1} \binom{k-1}{m} (\bar{y}^A_m)^m (1 - \bar{y}^A_m)^{k-1-m},
\]

(3)

where \(z\) is the average degree and \(\bar{y}^A_\infty = (1 - \mu) y^A_\infty + \mu y^B_\infty\). The fraction of active nodes is given by:

\[
\rho_\infty^A = \rho_0^A + (1 - \rho_0^A) \sum_{k=0}^{\infty} p(k) \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} (\bar{y}^A_\infty)^m (1 - \bar{y}^A_\infty)^{k-m}.
\]

(4)
FIG. 2. The tradeoff between intra- and inter-community spreading. Stronger communities (small \( \mu \)) facilitate spreading within the originating community (local) while weak communities (large \( \mu \)) provide bridges that allow spreading between communities (global). There is a range of \( \mu \) values that allow both (optimal). The blue squares represents \( \rho_A^\infty \), the final density of active nodes in the community \( A \), and the red circles represents \( \rho_B^\infty \). The parameters for the simulation are: \( \rho_0 = 0.17, \theta = 0.4, N = 131056, \) and \( z = 20 \).

Now we aim to answer the question: how do communities affect the information cascade? Before jumping into the details, let us highlight the tradeoff caused by the strength of communities. As \( \mu \) decreases, the nodes in \( A \) have more neighbors in \( A \). Thus the number of seed nodes to which nodes in \( A \) are exposed also increases because the seeds exist only in \( A \) (\( \rho_0^A = 2\rho_0 \) and \( \rho_0^B = 0 \)). In other words, strong communities enhance local spreading. By contrast, the spreading in the community \( B \) is triggered entirely by the nodes in \( A \), as \( \rho_0^B = 0 \). Therefore, larger \( \mu \) (weak clustering) helps the spreading of the contagion to the community \( B \). The fact that strong clustering (smaller \( \mu \)) facilitates the spreading in the originating community, but weak clustering (larger \( \mu \)) helps inter-community spreading, raises the following question: is there an optimal strength of clustering that facilitates both intra- and inter-community spreading?

Figure 2 demonstrates that there is indeed a range of \( \mu \) values that enables both. In the range in blue (‘local’) strong clustering allows intra-community spreading in \( A \); in the range in red (‘global’) weak clustering allows inter-community spreading from \( A \) to \( B \). The range that blue and red overlap (purple, ‘optimal’) is where there is the right amount of clustering and where the global cascade happens. Here the clustering is strong enough to initiate the local spreading and weak enough to induce inter-community spreading. If \( \mu \) is too small, although \( A \) is fully saturated, the contagion does not propagate into \( B \) because there are not enough inter-community bridges. If \( \mu \) is too large, although there are enough bridges, \( \rho_\infty^B \approx 0 \) because the clustering is too weak to initiate the intra-community spreading in \( A \).

Let us analyze the issue in more detail. Figure 3 summarizes our results, showing the analytical results from MF and TL theories, along with the numerical simulations. In our numerical simulation, we calculate the mean of \( \rho_\infty \) across 1,000 runs of the model, where each run assumes a different realization of the network and of the seed nodes. We fix the threshold (\( \theta = 0.4 \)) throughout all simulations.

Figure 3 (a) shows the phase diagram with three phases: no cascade (white), local cascades that saturate the community \( A \) (blue), and global cascades (red). The dotted and dashed lines indicate the values of \( \rho_0 \) shown in (b) and (c). (b) the cross-sections of the phase diagram (dotted lines in (a)). TL (solid lines) shows excellent agreements with the simulation while MF (dotted lines) overestimate the possibility of global cascades. (c) the cross-sections represented in dashed lines in (a).
of \( \mu \) is more interesting, exhibiting qualitatively different patterns depending on \( \rho_0 \).

Figure 3 (b,c) illustrates a set of possible scenarios, using both numerical simulations and analytic calculations. For small values of \( \rho_0 \) (black, \( \rho_0 = 0.10 \)), nodes are hardly activated even in the originating community, and activation essentially fails to propagate, regardless of \( \mu \). By increasing \( \rho_0 \) (blue, \( \rho_0 = 0.13 \)) one reaches a threshold where the contagion can spread to the whole originating community if \( \mu \) is small. However, after a critical value of \( \mu \), the internal connectivity becomes insufficient to allow the spreading of the contagion to the whole originating community. As the originating community is not saturated, the cascade does not spread to the other community as well. This corresponds to a situation where there is no overlap between the blue and red area in Fig. 2.

A larger value of \( \rho_0 \) (red, \( \rho_0 = 0.17 \)) finally allows the global cascade. The range of \( \mu \) that allow full activation in the originating community is even further extended (fewer internal links are needed), until a sufficient number of links can be spared to induce full activation in the second one. If the number of intra-community links becomes too small (large \( \mu \)), though, activation fails to spread in the originating community and therefore it cannot be transmitted over the entire network despite the increased number of cross-community links, creating a range of community strength that allows global spreading.

Even larger values of \( \rho_0 \) (red and magenta) simply extends the range of \( \mu \) for which the activation of the entire network is achieved. When \( \rho_0 \) becomes larger than the critical value for the transition in networks without communities, increasing \( \mu \) never blocks the local spreading, and thus the global cascade always happens as long as the network has enough bridges. Notice that \( \rho_\infty \) is always larger for intermediate values of \( \mu \) with respect to the no-community case \( (\mu = 1/2) \) and indeed full activation can be obtained in an ample set of values of \( \rho_0 \) if \( \mu \) is properly chosen. The smallest value of \( \mu \) that allows full activation of the second community is essentially independent of \( \rho_0 \), for fairly large values of \( \rho_0 \); once the first community is fully active it is only a matter of providing sufficient external links, therefore the precise value of \( \rho_0 \) does not matter. Specifically, using the TL formulation, we obtained that \( \mu_c \approx 0.2175 \) requires the minimal amount of seeds to allow global cascades, given the other parameters. The value of \( \mu \) at which the decay of \( \rho_\infty \) sets in, instead, results from not having sufficient internal links given the initial seed to achieve full activation of the originating community, and it depends therefore on \( \rho_0 \).

In summary, our analysis shows that there exists optimal strength of community structure for facilitating global cascades. We demonstrate that the presence of the right amount of community structure may, counterintuitively, enhance the diffusion of information rather than hinder it. A tight community, with its high level of internal connectivity, can act as an incubator for the localized information contagion and help to achieve a critical mass. Information can then spread outside the community effectively as long as sufficient external connectivity is guaranteed. Our results enrich the growing body of literature that stresses the influence of the community structure in a large number of processes, including epidemics, viral marketing, opinion formation, and information diffusion. Our findings offer insights to understand recent empirical observations, such as the counterintuitive behavior of information contagion in clustered networks 24 or the strong link between viral memes and the community structures in Twitter 6. Further work is needed to understand how our results generalize if different mechanism of transmission are considered, or a richer and more complex organizations of communities is assumed.

We thank James P. Bagrow and Filippo Menczer for helpful discussions and suggestions. AF acknowledges NSF (grant CCF-1101743) and the McDonnell Foundation. AF and EF acknowledge DARPA (grant W911NF-12-1-0037).

* Corresponding author: yyahn@indiana.edu

[1] B. Ryan and N. Gross, Rural Sociol. 8, 15 (1943).
[2] M. Granovetter, Am. J. Sociol. 1, 1420 (1978).
[3] E. M. Rogers, Diffusion Of Innovations (Free Press, 2003).
[4] 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.
[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, in Proceedings of the 21st International Conference on World Wide Web (ACM, 2012) pp. 519–528.
[6] L. Weng, F. Menczer, and Y.-Y. Ahn, Sci. Rep. 3, 2522 (2013).
[7] W. Goffman and V. A. Newill, Nature 204, 225 (1964).
[8] D. J. Daley and D. G. Kendall, Nature 204, 1118 (1964).
[9] N. Bailey, The Mathematical Theory of Infectious Diseases and Its Applications, 2nd ed. (Griffin, London, 1975).
[10] R. M. Anderson, R. M. May, and B. Anderson, Infectious Diseases of Humans: Dynamics and Control (Oxford University Press, USA, 1992).
[11] J. Goldenberg, B. Libai, and E. Muller, Market. Lett. 12, 211 (2001).
[12] K. Saito, R. Nakano, and M. Kimura, in Knowledge-Based Intelligent Information And Engineering Systems (Springer, 2008) pp. 67–75.
[13] P. Domingos and M. Richardson, in Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (ACM, 2001) pp. 57–66.
[14] M. Richardson and P. Domingos, in Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (ACM, 2002) pp. 61–70.
[15] J. Leskovec, L. A. Adamic, and B. A. Huberman, ACM Transactions on The Web (TWEB) 1, 5 (2007).
[16] T. C. Schelling, J. Math. Sociol. 1, 143 (1971).
[17] D. Watts, Proc. Nat. Acad. Sci. 99, 5766 (2002).
[18] P. L. Krapivsky, S. Redner, and D. Volovik, J. Stat. Mech., P12003 (2011).
[19] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. 86, 3200 (2001).
[20] R. Albert and A.-L. Barabási, Rev. Mod. Phys. 74, 47 (2002).
[21] M. E. J. Newman, Networks: An Introduction (Oxford University Press, USA, 2010).
[22] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, in Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (ACM, 2006) pp. 44–54.
[23] D. M. Romero, B. Meeder, and J. Kleinberg, in Proceedings of the 20th International Conference on World Wide Web (ACM, 2011) pp. 695–704.
[24] D. Centola, Science 329, 1194 (2010).
[25] J.-P. Omela, J. Saramäki, J. Hyvönem, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Nat. Acad. Sci. 104, 7332 (2007).
[26] J. P. Gleeson, Phys. Rev. E 77, 046117 (2008).
[27] R. Lambiotte and P. Panzarasa, J. Informetr. 3, 180 (2009).
[28] Y. Ikeda, T. Hasegawa, and K. Nemoto, in Journal of Physics: Conference Series, Vol. 221 (IOP Publishing, 2010) p. 012005.
[29] A. Hackett, S. Melnik, and J. P. Gleeson, Phys. Rev. E 83, 056107 (2011).
[30] K. Chung, Y. Baek, D. Kim, M. Ha, and H. Jeong, (2013), arXiv:1312.0573 [physics.soc-ph].
[31] X. Wu and Z. Liu, Physica A 387, 623 (2008).
[32] D. J. Watts and S. H. Strogatz, Nature 393, 440 (1998).
[33] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, Sci. Rep. 3, 2330 (2013).
[34] M. Girvan and M. Newman, Proc. Nat. Acad. Sci. 99, 7821 (2002).
[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E 78, 046110 (2008).
[36] B. Karrer and M. E. Newman, Phys. Rev. E 83, 016107 (2011).
[37] J. P. Gleeson and D. J. Cahalane, Phys. Rev. E 75, 056103 (2007).