Optimal Network Modularity for Information Diffusion

Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn*
School of Informatics and Computing, Indiana University, Bloomington, Indiana 47408, USA
(Received 8 January 2014; revised manuscript received 9 July 2014; published 18 August 2014)

We investigate the impact of community structure on information diffusion with the linear threshold model. Our results demonstrate that modular structure may have counterintuitive effects on information diffusion when social reinforcement is present. We show that strong communities can facilitate global diffusion by enhancing local, intracommunity spreading. Using both analytic approaches and numerical simulations, we demonstrate the existence of an optimal network modularity, where global diffusion requires the minimal number of early adopters.

DOI: 10.1103/PhysRevLett.113.088701 PACS numbers: 89.75.Hc, 89.65.Ef, 89.75.Fb

The study of information diffusion—fads, innovations, collective actions, viral memes—is relevant to a number of disciplines, including mathematical, physical, and social sciences, communication, marketing, and economics [1–6]. The most common approach is to focus on the affinities between information diffusion and infectious diseases spreading [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 diffusion event similar to 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], describes information diffusion as a branching process; the threshold model, originally proposed to study collective social behavior [2,16–18], incorporates the idea of "social reinforcement" by assuming that each adoption requires a certain number of exposures. Although it is not yet fully understood how the microscopic mechanisms underlying information diffusion differ from those in epidemic spreading, it has been pointed out that social reinforcement could be a crucial one: unlike epidemic spreading, where each exposure acts independently, social reinforcement provisions that each additional exposure to a piece of information sensibly increases the probability of its adoption [19–21].

Since information spreads through social contacts, the structure of the underlying social network is a crucial ingredient in modeling information diffusion. The role of hubs and degree distribution have been studied extensively due to their critical role in epidemic spreading [22–24]. Another obvious network feature that has implications on information diffusion is the presence of a modular structure. Several studies investigated the role of communities in information diffusion [25–30], mostly ignoring the effect of social reinforcement.

Epidemic spreading is hindered by the presence of communities or modular structure, since this helps confining the epidemics in the community of origin [25,31]. This may naturally lead to the expectation that the same is true for information diffusion, given the similar approaches used in modeling epidemic and information diffusion. However, recent empirical work suggested that modular structure may, counterintuitively, facilitate information diffusion [21]. Other studies also proposed that network modularity plays a more important role in information diffusion than in epidemic spreading [6,19,32]. These findings reinforce the need to systematically explore how mechanisms like social reinforcement interact with the ubiquitous presence of modular structure in the underlying network.

In this Letter, we use the linear threshold model—which incorporates the simplest form of social reinforcement—to systematically study how community structure affects global information diffusion. It is worth stressing that both cooperative interactions (as those provisioned by social reinforcement) and modular structure are common in a variety of phenomena. The results described here could be, therefore, directly relevant in several different areas. Examples include neural networks [33], systems with Ising-like dynamics evolving on a nonhomogeneous substrate [34], and more in general, in the study of phenomena that can be interpreted in terms of spreading.

Here, we expose two roles played by modular structure: enhancing local spreading and hindering global spreading. Strong communities facilitate social reinforcement and, thereby, enhance local spreading [6,21]; weak community structure makes global spreading easier, because it provides more bridges among communities. We show that there exists an optimal balance between these two effects, where community structure counterintuitively enhances—rather than hinders—global diffusion of information. This draws a parallel with the "small world" phenomenon, where the presence of a small number of shortcuts greatly reduces the average path length of the network while maintaining high clustering [35]. In information diffusion, a small number of bridges between communities allows intercommunity diffusion while maintaining intracommunity diffusion.

We adopt the linear threshold model to account for recent observations and experiments that demonstrated the impact of social reinforcement in information diffusion [6,19–21]. 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 the "active" state and 0 the "inactive" one. At time $t = 0$, a fraction ρ_0 of randomly selected agents, or "seeds," is initialized in the active state. At each time step, every agent's state is updated synchronously according to the following threshold rule:

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

where θ 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 macroscopic cascade [17]; there also exists a sharp transition, at a constant threshold parameter, from an inactive state where no diffusion occurs, to an active state with global diffusion, triggered at a critical fraction of initially active nodes [36]. In the following, we focus on the latter transition based on the number of seeds and let θ remain constant.

To systematically investigate the impact of community structure, we prepare an ensemble of networks with two communities with a varying degree of strength, using the block-model approach [37–39]. 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 μM are randomly distributed among node pairs that belong to different communities (see Fig. 1). The parameter μ controls the strength of the community structure: a large value of μ yields more links between the two communities and, thus, a weak community structure. Finally, we plant the seeds in A, assuming that the diffusion originates from the community A.

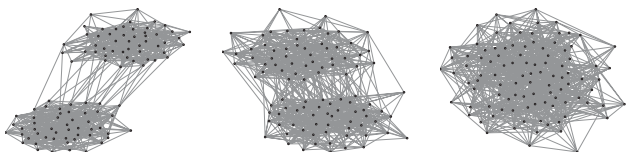

FIG. 1. Example of networks with different degrees of clustering μ (0.03, 0.12, and 0.3 from left to right, respectively). Parameters values are set to $N = 100$, $M = 750$, and $n = 2$.

Let us introduce two analytic approaches—mean-field (MF) and tree-like (TL) approximations—to understand the behavior 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 (ρ_∞) given the initial density of seeds (ρ_0). When there is no community structure, using the mean-field approximation, ρ_∞ can be computed as the smallest stable solution of the equation

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

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 $t = 0$ (ρ_0), and (ii) the probability that the node is not active at $t = 0$ $(1-\rho_0)$ but has at least θk 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 it is easy to generalize it for arbitrary configurations of communities, here, we focus on the case with two communities. In such a case, the equations for the fraction of active nodes ρ^A (respectively, ρ^B) in the community A (respectively, B) can be written as

$$\rho_\infty^{A(B)} = \rho_0^{A(B)} + (1-\rho_0^{A(B)}) \sum_{k=1}^{\infty} p(k) \times \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} (q^{A(B)})^m (1-q^{A(B)})^{(k-m)}, \tag{2}$$

where $\rho_0^{A(B)}$ is the density of seeds in the community $A(B)$, and $q^{A(B)} = (1-\mu)\rho_\infty^{A(B)} + \mu\rho_\infty^{B(A)}$ 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 ($\rho_\infty^{A(B)}$), and (ii) the probability that it is in the other $B(A)$ community (μ) and is active ($\rho_\infty^{B(A)}$). Finally, $\rho_\infty = (\rho_\infty^A + \rho_\infty^B)/2$.

A more sophisticated framework adopts the TL approximation [26,40]. It approximates the underlying network with a tree of infinite depth and assumes that the nodes at level n are only affected by those at level $n-1$. The fraction of active nodes in community $A(B)$ is computed using an auxiliary variable $y_\infty^{A(B)}$ obtained by the following iteration over all the levels in the tree

$$y_{n+1}^{A(B)} = \rho_0^{A(B)} + (1-\rho_0^{A(B)}) \sum_k \frac{k}{z} p(k) \times \sum_{m=\lceil \theta k \rceil}^{k-1} \binom{k-1}{m} (\bar{y}_n^{A(B)})^m (1-\bar{y}_n^{A(B)})^{k-1-m}, \tag{3}$$

where z is the average degree and $\bar{y}_n^{A(B)} = (1-\mu)y_n^{A(B)} + \mu y_n^{B(A)}$. The fraction of active nodes is given by

$$\rho_\infty^{A(B)} = \rho_0^{A(B)} + (1-\rho_0^{A(B)}) \sum_{k=0}^{\infty} p(k) \times \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} (y_\infty^{A(B)})^m (1-y_\infty^{A(B)})^{k-m}. \quad (4)$$

Now, we address the issue of how communities affect information diffusion by first highlighting the trade off due to the strength of communities. As μ decreases, nodes in A have increasingly 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 community B is triggered entirely by the nodes in A, as $\rho_0^B = 0$. Therefore, larger μ (smaller modularity) helps the spreading of the contagion to community B. The fact that large modularity (smaller μ) facilitates the spreading in the originating community, but small modularity (larger μ) helps intercommunity spreading, raises the following question: is there an optimal modularity that facilitates both intracommunity and intercommunity spreading?

Figure 2 demonstrates that there is, indeed, a range of values of μ that enables both. In the blue range ("local"), strong cohesion allows intracommunity spreading in the originating community A; in the red range ("global"), weak modular structure allows intercommunity spreading from A to B. The interval where blue and red overlap (purple, "optimal") provides the right amount of modularity to enable global diffusion. Here, the modularity is large enough to initiate the local spreading and small enough to induce intercommunity spreading. If μ is too small, the contagion cannot propagate into B, even if A is fully saturated, because there are not enough intercommunity bridges. If μ is too large, although there are enough bridges, $\rho_\infty^B \simeq 0$ because the modularity is too small to initiate intracommunity spreading from A.

Let us analyze the issue in more detail. Figure 3 summarizes our results, derived analytically by MF and TL approximations, and by numerical simulations. In our numerical simulations, we compute the mean of ρ_∞ across

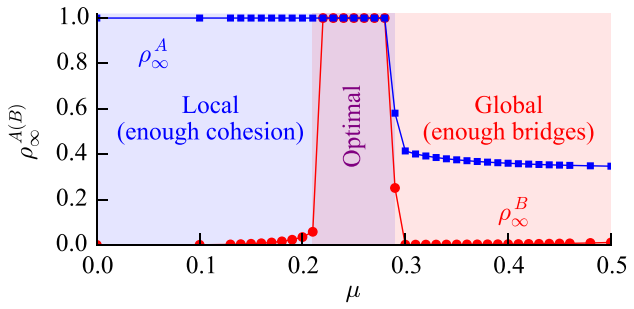

FIG. 2 (color online). The tradeoff between intracommunity and intercommunity spreading. Stronger communities (small μ) facilitate spreading within the originating community (local), while weak communities (large μ) provide bridges that allow spreading between communities (global). There is a range of μ values that allow both (optimal). The blue squares represent ρ_∞^A, the final density of active nodes in the community A, and the red circles represent ρ_∞^B. The parameters for the simulation are $\rho_0 = 0.17$, $\theta = 0.4$, $N = 131056$, and $z = 20$.

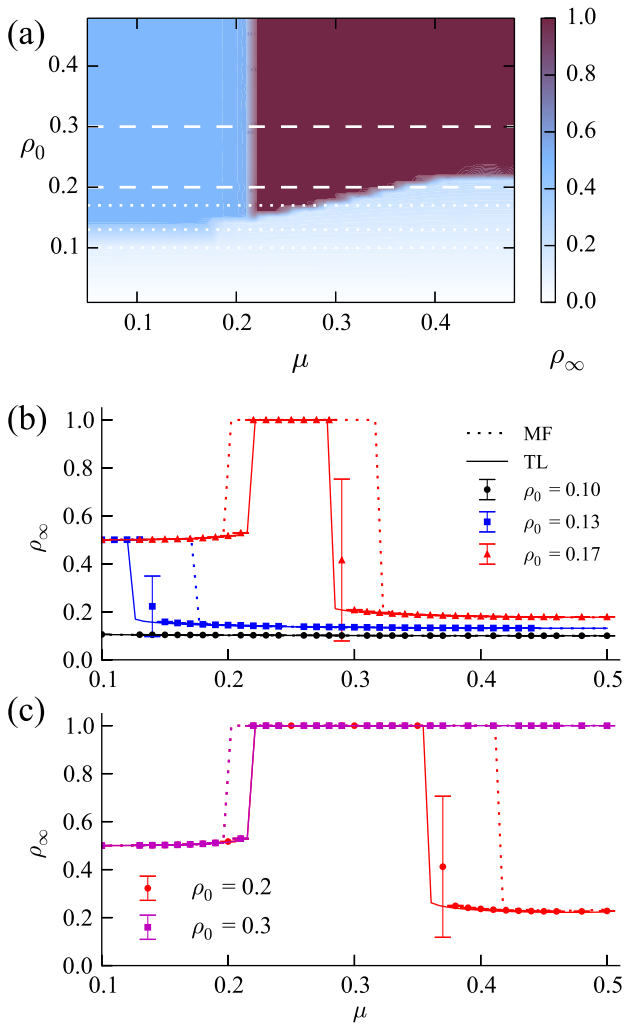

FIG. 3 (color online). (a) Phase diagram of the threshold model in the presence of community structures with $N = 131\,056$, $z = 20$, and $\theta = 0.4$. There are three phases: no diffusion (white), local diffusion that saturates the community A (blue, light-gray), and global diffusion (red, dark-gray). The dotted and dashed lines indicate the values of ρ_0 shown in (b) and (c). (b) Cross sections of the phase diagram [dotted lines in (a)]. TL approximations (solid lines) show excellent agreement with the simulation while MF approximations (dotted lines) overestimate the possibility of global diffusion. (c) Cross sections represented by dashed lines in (a).

1000 runs of the model, each assuming a different realization of the network and of the seed nodes. We fix the threshold ($\theta = 0.4$) throughout all simulations. We discuss the effect of threshold and other parameters, including number of communities and more general degree distributions in the Supplemental Material [41].

Figure 3(a) shows the phase diagram with three phases: no diffusion (white), local diffusion (blue, light-gray), and global diffusion (red, dark-gray). As expected, a cross section for $\mu = \text{const}$ shows that ρ_∞ is an increasing function of ρ_0. The system undergoes a sharp transition for a broad range of values of μ, including the case in which communities are absent ($\mu = 1/2$) [36]. The behavior of ρ_∞ as a function of μ is more interesting, in that it exhibits qualitatively different patterns depending on ρ_0.

Figures 3(b) and 3(c) illustrate a set of possible scenarios, using both numerical simulations and analytic calculations. For small values of ρ_0 (black, $\rho_0 = 0.10$), nodes are hardly activated even in the originating community; the activation essentially fails to propagate, regardless of μ. By increasing ρ_0 (blue, $\rho_0 = 0.13$), one reaches a threshold where the contagion can spread to the whole originating community if μ is sufficiently small. However, when a critical value of μ is exceeded, the internal connectivity becomes insufficient to spread the contagion to the whole originating community. As the originating community is not saturated, the diffusion does not spread to the other community as well. In this situation, there is no overlap between the blue and red area in Fig. 2.

A larger value of ρ_0 (red, $\rho_0 = 0.17$) finally allows the global diffusion. The range of values of μ that allows 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, however, the number of intracommunity links becomes too small (large μ), the 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. The above reflects in a finite, intermediate range of community strength that allows global spreading.

Even larger values of ρ_0 (red and magenta) simply extend the range of μ for which the activation of the entire network is achieved. When ρ_0 becomes larger than the critical value for the transition in networks without communities, increasing μ never blocks the local spreading, and thus, the global diffusion always happens as long as the network has enough bridges. Notice that ρ_∞ is always larger for intermediate values of μ with respect to the no-community case ($\mu = 1/2$), and indeed, full activation can be obtained in an ample set of values of ρ_0 if μ is properly chosen. The smallest value of μ that allows full activation of the second community is essentially independent of ρ_0, if ρ_0 is sufficiently large: once the first community is fully active, it is only a matter of providing sufficient external links; therefore, the precise value of ρ_0 does not matter. Specifically, using the TL formulation and the present value of θ, we obtained that $\mu_c \simeq 0.2175$ requires the minimal amount of seeds compatible with global diffusion. The value of μ for which the decay of ρ_∞ sets in, instead, results from not having sufficient internal links to achieve full activation of the originating community given the initial seed. The value of μ depends, therefore, on ρ_0.

Although, here, we present results only for the case of random networks with two communities and a specific value of θ, our results are more general. In the Supplemental Material [41], we provide evidence that our results are robust under changes in the number of communities and assuming degree distributions more general than that induced by the random arrangement of links described above. Our results include experiments run on Lancichinetti-Fortunato-Radicchi benchmark graphs [38] that provision for a power law degree distribution both for the degree and the size of multiple communities. It is also worth stressing that both the MF and TL methods are flexible enough to handle arbitrary (and community-specific) degree distributions, and arbitrary intercommunity connectivity patterns. To adapt MF to this general case, one would need to replace [e.g., in the equation for ρ_∞^A in Eq. (2)] $p(k)$ with the specific degree distribution of community A and q^A with $\sum_{J\in\mathcal{C}} p_{AJ}\rho_\infty^J$, where $\mathcal{C}$ is the set of communities and p_{AJ} is the probability that a link departing from a node in A ends in J. In the Supplemental Material [41], we also provide evidence that our results are qualitatively unchanged by varying the system size N, the average degree z, and other parameters. Finally, our results are also robust for changes in the threshold θ for a pretty wide range of values (see Supplemental Material [41]).

In summary, our analysis shows that there exists an optimal strength of community structure that facilitates global diffusion. 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 diffusion 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 can be generalized, and offer insights for understanding recent empirical observations, such as the counterintuitive behavior of information diffusion in clustered networks [21], or the strong link between viral memes and the community structures in Twitter [6,32]. Further work is needed to understand how our observations hold if different mechanisms of transmission are considered, or a

richer and more complex organization of communities is assumed.

We thank James P. Bagrow and Filippo Menczer for helpful discussions and suggestions. A. F. acknowledges support from the NSF (Grant No. CCF-1101743) and the McDonnell Foundation. A. F. and E. F. acknowledge support from DARPA (Grant No. W911NF-12-1-0037).

*Corresponding author.
yyahn@indiana.edu

[1] B. Ryan and N. C. Gross, Rural Sociol. **8**, 15 (1943).
[2] M. Granovetter, Am. J. Sociology **83**, 1420 (1978).
[3] E. M. Rogers, *Diffusion Of Innovations* (Free Press, New York, 2003).
[4] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, in *Proceedings of the 13th International Conference on World Wide Web* (ACM Press, New York, 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 Press, New York, 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 (London) **204**, 225 (1964).
[8] D. J. Daley and D. G. Kendall, Nature (London) **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, New York, 1992).
[11] J. Goldenberg, B. Libai, and E. Muller, Marketing letters **12**, 211 (2001).
[12] K. Saito, R. Nakano, and M. Kimura, in *Knowledge-Based Intelligent Information And Engineering Systems* (Springer, New York, 2008), pp. 67–75.
[13] P. Domingos and M. Richardson, in *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2001), pp. 57–66.
[14] M. Richardson and P. Domingos, in *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2002), pp. 61–70.
[15] J. Leskovec, L. A. Adamic, and B. A. Huberman, ACM Trans. Web **1**, 5 (2007).
[16] T. C. Schelling, J. Math. Sociol. **1**, 143 (1971).
[17] D. Watts, Proc. Natl. Acad. Sci. U.S.A. **99**, 5766 (2002).
[18] P. L. Krapivsky, S. Redner, and D. Volovik, J. Stat. Mech. (2011) P12003.
[19] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, in *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2006), pp. 44–54.
[20] D. M. Romero, B. Meeder, and J. Kleinberg, in *Proceedings of the 20th International Conference on World Wide Web* (ACM Press, New York, 2011), pp. 695–704.
[21] D. Centola, Science **329**, 1194 (2010).
[22] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[23] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[24] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
[25] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Natl. Acad. Sci. U.S.A. **104**, 7332 (2007).
[26] J. P. Gleeson, Phys. Rev. E **77**, 046117 (2008).
[27] R. Lambiotte and P. Panzarasa, J. Informet. **3**, 180 (2009).
[28] Y. Ikeda, T. Hasegawa, and K. Nemoto, J. Phys. Conf. Ser. **221**, 012005 (2010).
[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, Phys. Rev. E **89**, 052811 (2014).
[31] X. Wu and Z. Liu, Physica (Amsterdam) **387A**, 623 (2008).
[32] L. Weng, F. Menczer, and Y.-Y. Ahn, in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM'14)* (AAAI, Ann Arbor, MI, 2014).
[33] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge University Press, Cambridge, England, 2002).
[34] H. Chen and Z. Hou, Phys. Rev. E **83**, 046124 (2011).
[35] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[36] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, Sci. Rep. **3**, 2330 (2013).
[37] M. Girvan and M. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[38] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[39] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[40] J. P. Gleeson and D. J. Cahalane, Phys. Rev. E **75**, 056103 (2007).
[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.113.088701 for the robustness analyses under the variation of parameters such as threshold and number of communities.