# Network modularity controls the speed of information diffusion

Hao Peng,[1] Azadeh Nematzadeh,[2] Daniel M. Romero,[1] and Emilio Ferrara[3,*]
[1]*School of Information, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2]*S&P Global, New York, New York 10004, USA*
[3]*Information Sciences Institute, University of Southern California, Los Angeles, California 90292, USA*

(Received 12 August 2020; accepted 8 November 2020; published 30 November 2020)

The rapid diffusion of information and the adoption of social behaviors are of critical importance in situations as diverse as collective actions, pandemic prevention, or advertising and marketing. Although the dynamics of large cascades have been extensively studied in various contexts, few have systematically examined the impact of network topology on the efficiency of information diffusion. Here, by employing the linear threshold model on networks with communities, we demonstrate that a prominent network feature—the modular structure—strongly affects the speed of information diffusion in complex contagion. Our simulations show that there always exists an optimal network modularity for the most efficient spreading process. Beyond this critical value, either a stronger or a weaker modular structure actually hinders the diffusion speed. These results are confirmed by an analytical approximation. We further demonstrate that the optimal modularity varies with both the seed size and the target cascade size and is ultimately dependent on the network under investigation. We underscore the importance of our findings in applications from marketing to epidemiology, from neuroscience to engineering, where the understanding of the structural design of complex systems focuses on the efficiency of information propagation.

DOI: 10.1103/PhysRevE.102.052316

## I. INTRODUCTION

The spread of information in complex networks controls or modulates fundamental processes that can have local effects on individual actors and groups thereof, and macroscopic effects on the whole system (e.g., global information cascades). Information diffusion has been studied by drawing analogies with epidemics. Many social behaviors, for example, act like infectious diseases: once triggered, they can spread to the entire population in a very short amount of time, generating a contagion process similar to an epidemic outbreak. Examples include collective actions such as voting and participation in social movements, the adoption of innovations such as vaccination and emerging technologies, the diffusion of viral memes in social media, and the spread of norms and cultural fads. The dynamics of these intriguing and complex phenomena have attracted research interest from a number of disciplines [1–3].

There are two major models for the study of information diffusion: the *independent cascade model* and the *linear threshold model*. The former assumes that, similar to disease transmission, each exposure is independent from each other and a person has only one chance to "infect" their neighbors [4,5]. The latter postulates that social reinforcement, or exposure to multiple sources, is needed in the contagion process and each person has a threshold to be met for successful adoption [6,7]. The independent cascade model suits well with the *simple contagion* scenario, where the goal is to inform people rather than to convince them to take actions [1,4]. It thus has been adopted in the study of word-of-mouth spreading and viral marketing [4,8]. However, some studies revealed that the threshold model is more applicable to the spread of risky or contentious social behaviors for which each additional exposure increases the likelihood of adoption [7,9–12]. We thus examine the efficiency of information diffusion in the latter case, which is sometimes referred to as *complex contagion*.

Social behaviors spread through social contacts, thus the structure of the underlying social network plays an important role in the process of information diffusion [2,7,13,14]. Recent studies have examined the effects of different network properties on the dynamics of information diffusion [4,7,15].

One prominent network feature is *modular structure*—the separation of a network into several subsets of nodes within which connections are dense, but between which connections are sparser [16,17]. Networks with many "bridges" connecting nodes in different communities tend to have low modularity [18,19]. Note that we distinguish modularity from another related concept, *clustering*, which refers to the network transitivity and is quantified by the clustering coefficient [18,20].

The *strength of weak ties* theory suggests that networks with weak modular structure will promote both the scale and the speed of diffusion since enough shortcuts, which tend to be weak ties, link relatively separated groups and diffuse information across communities [1,20]. In contrast, the *weakness of long ties* theory predicts that, in the case of complex contagion where the adoption requires multiple exposures, networks with strong modular structure, and thus an abundance of strong ties, can enhance the spread of certain social behaviors [10,21]. The two competing hypotheses based on prior theoretical work manifest the interplay between social

*Corresponding author: emiliofe@usc.edu

 

reinforcement and network modularity in most real social networks. Yet empirical studies seem to reveal inconsistent results regarding the role played by community structure in complex contagion [21–23]. Recent findings reveal that network modularity plays two different roles in information diffusion: (1) enhancing intracommunity spreading and (2) hindering intercommunity spreading [24], providing an in-principle unifying explanation to the competing empirical evidence.

Overall, prior work on the relationship between network modularity and large cascades has mainly focused on one aspect of information diffusion—the size of information cascades, i.e., the total number of "infected" individuals in the steady state. Another important cascade feature—the efficiency of information diffusion, i.e., the total time it takes to reach the steady state—has been underexplored [25–27]. A better understanding of information diffusion speed can have many practical applications, such as informing the design of communication and social networks where the efficiency of information flow needs to be prioritized. For instance, a get-out-the-vote campaign on election day may need to be optimized for adoption speed since the operation will be useless after the election is over.

The extant literature has also demonstrated how insights about the interplay between network modularity and information spread can provide a principled understanding of various complex system dynamics, from characterizing neuronal communication in human connectomics [28], to optimizing immunization strategies for public health and animal welfare [29,30].

## II. MODELS

### A. Diffusion model

Here we systematically examine the effects of network modularity on the *speed* of information diffusion in complex contagion by utilizing the linear threshold model [6,7]. We define diffusion speed as the average rate of a spreading process, measured as the eventual growth of the cascade divided by the time it takes to reach equilibrium. We show that, in complex networks, there always exists an optimal amount of modularity for the most efficient information diffusion process.

In the linear threshold model, a node can be in two states: either active or inactive. Each node $a$ is assigned a threshold $\theta_a$ uniformly at random from the interval $[0, 1]$. Initially all nodes are inactive. At time step $t = 0$, a fraction $\rho_0$ of $N$ nodes (the seeds) are switched into active state. In the subsequent time steps, a node can become active if its fraction of active neighbors exceeds the threshold, and it stays active forever once being activated. Following these rules, we update a fraction $f$ of all nodes (selected randomly) at each step. In the synchronous updating scenario, where $f = 1$, the contagion process unfolds in a deterministic manner until the network reaches the steady state [5,7,24]. This model can be adapted to the case of asynchronous updating by setting $f < 1$. We assume that all nodes have the same threshold $\theta$ [24,31]. We measure the time steps $t_s$ it takes to reach steady state and the total fraction $\rho_{t_s}$ of active nodes across the network at $t_s$. The average speed of diffusion is $\bar{v} = (\rho_{t_s} - \rho_0)/t_s$.

### B. Network model

We adopt the stochastic block model (SBM) to generate networks with community structure [32]. The underlying network consists of $N$ nodes partitioned into $d$ communities $\{C_1, C_2, \ldots, C_d\}$. Let $|C_i|$ be the size of $C_i$, and $\rho_t^{(i)}$ be the fraction of active nodes in $C_i$ at time $t$. Each community $C_i$ has a specified degree distribution $p_k^{(i)}$ and a mean degree $z^{(i)} = \sum k p_k^{(i)}$. The edges in the network are randomly distributed according to a $d \times d$ mixing matrix $\mathbf{e}$, with $e_{ij}$ defined as the fraction of edges that connect nodes in $C_i$ to nodes in $C_j$. Although studies have indicated that tie strength is an important factor in modeling information diffusion [4,13], here we consider edges to be unweighted, due to the unclear relationship between tie strength and network topology—some studies argue that strong ties mostly reside within tightly knit clusters and weak ties tend to link together distant communities [1,4,10,13], while other empirical work reveals the opposite conclusion in social and scientific collaboration networks [33–35].

### C. Numerical simulation

We use numerical simulations to compare the speed of diffusion across an ensemble of networks with different strength of network modularity. For simplicity, here we consider the case of two equally sized communities: let $d = 2$, $|C_1| = |C_2| = N/2$, and the seed nodes are randomly selected from $C_1$, thus $\rho_0^{(1)} = 2\rho_0$, $\rho_0^{(2)} = 0$. We construct the network such that $p_k^{(1)}$ and $p_k^{(2)}$ both follow a Poisson distribution, with $z^{(1)} = z^{(2)} = z$. The expected total number of edges is $M = zN/2$. Let $\mu M$ edges be randomly distributed between $C_1$ and $C_2$, and the remaining $(1 - \mu)M$ edges be randomly placed between node pairs in the same community, thus $\mathbf{e} = \frac{1}{2}\begin{bmatrix} 1-\mu, & \mu \\ \mu, & 1-\mu \end{bmatrix}$. Note that, to generate the network with other degree distributions, the configuration model should be used [16]. Here $\mu$ controls the strength of network modularity, which turns out to be $Q = 1/2 - \mu$, based on the current partition. A larger $\mu$ gives a network with weaker network modularity since there are more edges running between two communities. For each $\mu$, we run 100 simulations, with each assuming a different realization of the network and the seeds.

### D. Analytical approximation

We also study the dynamics in our system analytically. The cascade size $\rho_t$ is equal to the probability that a randomly chosen node is active at time $t$. The topology of such a large network can be approximated by a tree structure with infinite depth and a single node at the top, a.k.a. a *tree-like approximation* [36]. The top node is connected to $k_a$ neighbors at the next lower level, while any other node $a$ at level $n$ is connected to $k_a - 1$ neighbors at level $n - 1$, where $k_a$ is the degree of node $a$. At any level, the probability that a node in $C_i$ is among the seeds is $\rho_0^{(i)}$. In synchronous updating, the tree level $n$ can be directly mapped to the time step $t$ used in simulations [37], which means that $\rho_t$ can be approximated as the probability $\rho_n$ that the top node is active, assuming that it resides at level $n = t$, since the top node can be infected only by nodes at most $n$ levels below. We can calculate its probability of being

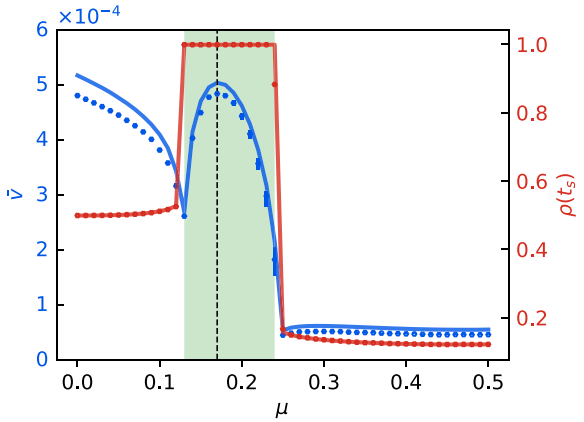


FIG. 1. Simulation results (dots) and analytical predictions (lines) of Eq. (5) (cf. Materials and Methods). Blue axis: the average speed of information diffusion, $\bar{v}$. Red axis: the size of information cascade, $\rho_{t_s}$. The $x$ axis represents the strength of network modularity controlled by $\mu$. Green area: the range of $\mu$ that can enable global cascades ($\rho_{t_s} \geqslant 0.99$). The dashed vertical line corresponds to $\mu = 0.17$ that yields the highest $\bar{v}$ by prediction. The simulation results are averaged over 100 realizations of the network for each $\mu$, with $N = 1 \times 10^5$, $z = 10$, $\rho_0 = 0.1$, $\theta = 0.35$, $f = 0.01$. The error bars indicate the interquartile ranges.

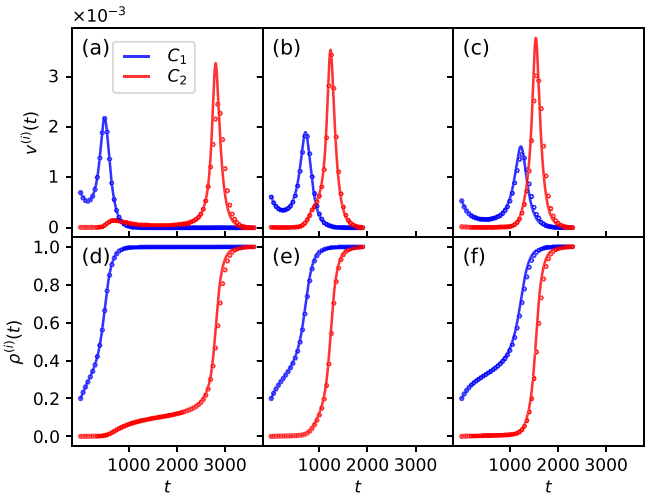


FIG. 2. Cross sections of three different $\mu$ values in Fig. 1 that enable global cascades. (a, d) $\mu = 0.13$; (b, e) $\mu = 0.17$; (c, f) $\mu = 0.21$. (a–c) The diffusion speed $v_t^{(i)}$ in $C_1$ and $C_2$ as a function of time step $t$. (d–f) Same as (a)–(c), but for the cumulative cascade size $\rho_t^{(i)}$. The theoretical predictions of Eq. (4) (lines) show excellent agreement with the numerical simulations (dots), averaged over 100 runs. The optimal $\mu = 0.17$ achieves the shortest total diffusion time, thus the highest average diffusion speed.

active from nodes at the bottom level ($n = 0$) to the top node ($n = t$), one level at a time, according to the linear threshold model. See the derivation of $\rho_n$ in Materials and Methods.

## III. RESULTS

### A. Optimal modularity for the speed of global cascades

Figure 1 displays an interval of network modularity that can trigger global cascades, which concurs with the findings in Ref. [24]. Intuitively, one would imagine that a stronger modularity (smaller $\mu$) increases diffusion speed in $C_1$ since nodes in $C_1$ are exposed to more seeds, while a weaker modularity (larger $\mu$) increases diffusion speed in $C_2$ because more bridges connect nodes in $C_2$ to the seeds. This observation raises the following question: is there an ideal network modularity at which the global cascade reaches the highest average diffusion speed?

Let us first analyze the behavior of our system when only local diffusion is possible. Figure 1 indicates that, when the network modularity is too strong (very small $\mu$), information spreads only among nodes in $C_1$ due to the lack of bridges between two communities, thus decreasing modularity (increasing $\mu$) decreases the average diffusion speed because it takes longer for spreading in $C_1$ and the cascade size stays the same.

When a global cascade is achieved, however, there is a quadratic relationship between the average diffusion speed and network modularity: decreasing modularity first increases the average diffusion speed, but only up to a critical point, after which a further reduction in modularity slows down the overall diffusion dynamics. The global cascade thus reaches its highest average speed at the optimal network modularity ($\mu = 0.17$). The analytical predictions show excellent agreement with the simulations (Fig. 1).

Next, we analyze the cascade dynamics in more detail to understand this phenomenon. Figure 2 shows the diffusion speed per time step in each community, for three different levels of network modularity. The time lags of spreading in two communities can help us to explain the influence of network modularity on the average diffusion speed of global cascades.

At $\mu = 0.13$, we reach the lower bound of the window for global cascades. However, the time difference between $C_1$ and $C_2$ is the longest: the spreading in $C_2$ merely gets started after $C_1$ reaches steady state [Fig. 2(a)]. Thus the relatively long diffusion time in $C_2$ is the bottleneck for the average diffusion speed at global scale.

One may, therefore, predict that the highest average diffusion speed can be achieved when the time lag between the two communities is reduced as much as possible. For instance, since the time difference to finish spreading at $\mu = 0.21$ [Fig. 2(c)] is shorter than that at $\mu = 0.17$ [Fig. 2(b)], the average diffusion speed would be predicted to be faster in the former case [Fig. 2(c)]. However, such an inference is incorrect, as the diffusion at $\mu = 0.21$ takes longer than the scenario when $\mu = 0.17$, for which the global cascade finishes in the shortest amount of time.

Comparing the optimal network modularity [Fig. 2(b)] to the first scenario [Fig. 2(a)], it takes slightly more time to finish spreading in $C_1$, due to the decreasing number of edges in $C_1$. But the increasing connections between the two communities reduces the diffusion time in $C_2$. The time lag between $C_1$ and $C_2$ is much shorter, but not close to zero. Figure 2 indicates that, at this optimal network modularity, neither $C_1$ nor $C_2$ achieves its highest diffusion speed, but both are pretty close to it, resulting in the most efficient global cascade.

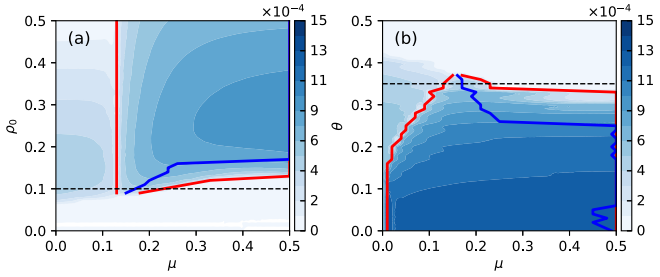


FIG. 3. Phase diagrams of the average diffusion speed $\bar{v}$ as a function of seed size $\rho_0$ (a) and threshold $\theta$ (b) on SBM networks. The two red curves mark the region for global cascades. The blue curve represents the $\mu$ value that yields the highest $\bar{v}$ for a given $\rho_0$ or $\theta$. The results are based on simulations, averaged over 100 runs for each combination of $(\rho_0, \mu)$ or $(\theta, \mu)$. Simulation parameters are $N = 1 \times 10^5$, $z = 10$, $f = 0.01$, with $\theta = 0.35$ (a) and $\rho_0 = 0.1$ (b). The seeds are randomly selected from a single community. The dashed line is a slice at $\rho_0 = 0.1$ ($\theta = 0.35$) in Fig. 1.

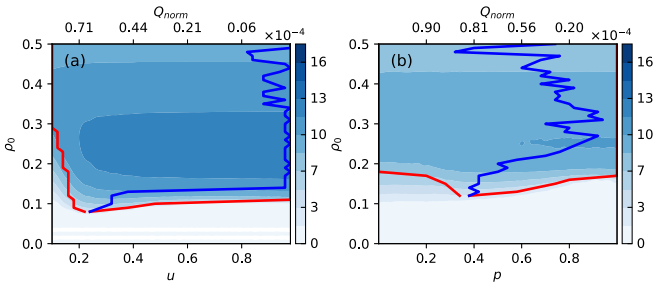


FIG. 4. Phase diagrams of the average diffusion speed $\bar{v}$ on the LFR (a) and Twitter (b) networks. Parameters $\mu$ and $p$ on the $x$ axis control the network modularity. The blue curve indicates the optimal $\mu$ (or $p$) for $\bar{v}$ for a given seed size $\rho_0$. The normalized modularity $Q_{\text{norm}}$ with respect to $\mu$ (or $p$) is shown on the top axis. Network statistics are $N = 25\,000$, $z = 10$, $\gamma = 2.5$, $\beta = 1.5$, $k_{\max} = 30$ (for LFR); $N = 81\,306$, $z = 16$ (for Twitter). Simulations are averaged over 100 runs, with $\theta = 0.3$, $f = 0.01$. The seeds are randomly selected across the whole network.

However, at $\mu = 0.21$, the further reduction of the number of edges in $C_1$ slows down the speed of local spreading in $C_1$, and this becomes the bottleneck of the average speed at global scale. Although, under this condition, $C_1$ and $C_2$ reach the steady state almost concurrently (with a time lag close to zero), it cannot counteract the increase in diffusion time for both communities [Fig. 2(c)].

### B. The effects of seed size and threshold

Figure 3 presents two phase diagrams of the average diffusion speed $\bar{v}$ as a function of the seed size $\rho_0$ and the threshold $\theta$. It indicates that, in the region of global cascades, there always exists an optimal modularity for the most efficient information diffusion, and this critical value of $\mu$ depends on both $\rho_0$ and $\theta$.

Note that a minimal seed size is needed to trigger global cascades, and once above this threshold, when $\rho_0$ is not too large (e.g., $\rho_0 = 0.1$), the average speed of global cascades first increases and then decreases as one reduces the modularity (increasing $\mu$), resulting in an intermediate value of $\mu$ as the optimal modularity [Fig. 3(a)]. However, when $\rho_0$ is sufficiently large (e.g., $\rho_0 = 0.2$), the average speed of global cascades always increases as one increases the number of cross-community links, making the network with no-community structure ($\mu = 0.5$) the ideal case for the most efficient spreading process. This is because, when increasing $\mu$ never blocks local spreading in $C_1$ due to the presence of enough seeds in $C_1$, more external links are always going to make the diffusion faster in $C_2$. Similar patterns emerge for the threshold $\theta$ when the seed size is fixed [Fig. 3(b)].

We obtain consistent results on SBM networks with different network sizes, variable average degrees, different seed arrangements between $C_1$ and $C_2$, and arbitrary number of communities, based on both simulations and analytical predictions [38].

### C. Simulations on non-SBM networks

Although the SBM provides reproducible and well-controlled modular networks for modeling the speed of information diffusion using tractable computational approaches, it is clearly appealing to test the generalizability of our findings on networks without the assumptions of equally sized communities and randomly distributed edges. To this end, we perform simulations on networks with more complex structure, such as heterogeneous communities, high clustering, and power-law degree distribution. We also randomly select the seed nodes across the network, instead of placing them in a single community.

We use the LFR benchmark graph [39] to generate synthetic networks with community structure similar to that observed in real-world networks (see Materials and Methods). We also simulate information diffusion on a Twitter network (see Materials and Methods) and six other real-world networks [38].

The phase diagrams for both types of networks are shown in Fig. 4. An optimal modularity for the most efficient global cascades still emerges as in the case of SBM networks [Fig. 3(a)]. Figure 4 shows that the minimal seed size required to trigger global cascades depends on the network under investigation: a 10% random sample of all nodes is enough to generate global diffusion on LFR networks for a wide range of modularity, while the same fraction of seed nodes generates only small cascades on the Twitter network. For the same reason, the optimal normalized modularity (see Materials and Methods) for a given seed size also changes across networks. Differently from the SBM, when the seed size is large enough, small changes in modularity result in only small changes in the average diffusion speed since the diffusion tends to reach global cascades rapidly due to the fact that seeds are randomly distributed over the whole network. Thus, for large seed sizes, the modularity for the fastest global cascades fluctuates on both LFR and Twitter networks, as opposed to the case of SBM networks, where the optimal values are always the same [Fig. 3(a)]. This also explains why the position of the phase

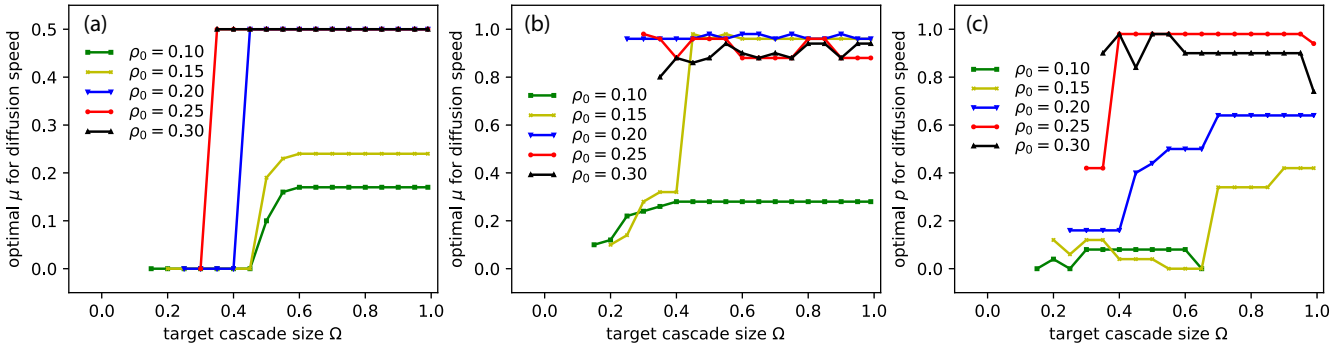


FIG. 5. The optimal network modularity for fast information diffusion changes as a function of the target cascade size on three different types of networks. The modularity is controlled by $\mu$ for SBM (a) and LFR (b) networks [or by $p$ for the Twitter network (c)]. The optimal values of $\mu$ and $p$ are selected from those that can achieve the given target size $\Omega$. All simulation parameters are the same as in Fig. 3(a) and Fig. 4. Note that a small $\rho_0$ may not be able to reach all target cascade sizes, e.g., a seed size of $\rho_0 = 0.1$ on the Twitter network is able to infect only up to about $\Omega = 65\%$ of all nodes. Line plots for different $\rho_0$ start from different $\Omega$ since $\Omega > \rho_0$.

transition at which global cascades emerge moves to the highest modularity for large seed sizes (Fig. 4).

Overall, the optimal modularity is ultimately dependent on the network under investigation. This is because their overall network structure is very different from each other, such as the degree distribution, the clustering coefficient, the community sizes, etc., and the interactions between modularity and these network properties can greatly impact diffusion dynamics. However, the general trend—that the optimal modularity decreases as the seed size increases—is preserved for a variety of complex networks [38].

### D. Optimal modularity for different cascade sizes

The objective of certain diffusion scenarios is not always to reach the global cascades. For instance, in the case where an organization needs to get at least $x$ signatures among its members before a certain date in order to get an initiative on a ballot, the goal is to activate just a fraction of the whole population. This example prompts us to ask: how does the optimal modularity for speed change with different target cascade sizes?

To answer this question, for a fixed $\Omega$ (i.e., the target cascade size) and $\rho_0$ (i.e., the seed size), we determine the optimal value of $\mu$ (or $p$) that minimizes the time it takes for the cascade to reach $\Omega$. Figure 5 indicates that the optimal modularity for the average diffusion speed typically decreases as the target cascade size $\Omega$ increases. For instance, the optimal $\mu$ changes from $\mu = 0$ to $\mu = 0.17$ as $\Omega$ increases from $\Omega = 0.15$ to $\Omega = 0.99$ for $\rho_0 = 0.1$ on SBM networks. The intuition behind this result is that since the originating communities already contain enough nodes to satisfy the small target cascade size, it is better to have strong modularity to facilitate local spreading (Fig. 1). However, when the seed size is large enough (e.g., $\rho_0 = 0.3$), the optimal modularity tends to be small for both large and small (relative to $\rho_0$) target cascades. In this case, originating communities can quickly be saturated, and thus the best strategy is to promote large cascades through intercommunity edges.

This observation provides a more complete picture of our findings: the best network to optimize diffusion speed is not always the same, suggesting that the target cascade size, together with the seed size, should all be taken into consideration when designing the most efficient network.

Beyond a constraint on the cascade size, there are other situations where one needs to optimize the diffusion speed with a time budget (or equivalently to maximize the cascade size in a given time window). We thus further examine the diffusion dynamics by considering having a limit on the diffusion time, which shows that the optimal modularity tends to decrease as the time budget increases [38].

## IV. DISCUSSION

We investigate the effect of community structure, as measured by network modularity, on the speed of information diffusion. Through simulations and analytical approximations, we reveal that there always exists an optimal strength of modularity—under which information or behavior diffuses at the highest average speed. We demonstrate that such an efficient spreading behavior is achieved by making the right compromise between internal connectivity and cross-community bridges for synchronized diffusion in different communities. We also find that the optimal modularity varies with respect to the seed size and the target cascade size. These findings are consistent on both synthetic and real-world networks.

Our findings provide insights for many real-world applications that allow for the optimization of network structure to enable rapid diffusion or adoption. For instance, it may help to design better organizational structure for firms with many different functional departments where the efficiency of diffusion is important (e.g., the adoption of social norms and work habits such as working hard). Drawing on the communication network of employees in a company (e.g., from email or social media), managers could make office assignments as an intervention to help change the interaction patterns such that the network approaches its optimal modularity, and thus making the process of social contagion more efficient.

In preventative health, one intervention used by practitioners to address public health challenges like obesity is to modify the contact network of a community to promote the spread of healthy behaviors, such as by providing role models or "health buddies" to mothers, young children, or users in online health communities [21,40,41]. Our finding suggests that the modularity should be considered in the network modification procedure to maximize the speed of behavioral change.

Online networks can be reshaped to influence information diffusion dynamics: social media platforms, for example, can design their friend recommendation algorithms to change the network modularity to promote (e.g., advertisements) or suppress (e.g., participation in illicit activities) diffusion processes.

Although our study is postulated upon the premise that one can alter network structures to maximize diffusion speed, our findings still have implications for real-world networks with a structure that cannot be modified: one can quantify the degree of efficiency the network is functioning at and determine the optimal seed size for a given network and diffusion process.

This study also has implications for online campaigns. Social media users often receive content relevant to their interests in trending discussions or ephemeral events. Our study suggests that advertisers can target networks with a high level of modular structure to maximize the campaign reach and inform a large audience in a short period of time. For instance, a petition to the White House that needs to gather 100 000 signatures in just 30 days can be promoted within high-modularity social networks to increase the chance of success.

From a methodological standpoint, by incorporating the effect of network modularity on the diffusion speed, machine learning algorithms can utilize modularity to better predict the efficiency of information cascades. Our framework can allow the study of many naturally occurring complex systems in biological networks and enable the understanding of evolutionary dynamics in complex networks exhibiting a certain level of modularity that facilitates or hinders diffusion speed. For example, network modularity has already been used to study spreading dynamics on the human connectome and to explain global communication on brain networks [28], but the communication speed in this context is unexplored.

Future work can focus on the empirical validation of the relationship between network modularity and the efficiency of information diffusion, and on examining its variations by considering other diffusion mechanisms (e.g., the independent cascade model) on networks with even more complex structure such as the hierarchical organization of communities.

## V. MATERIALS AND METHODS

### A. The calculation of diffusion speed

The tree-like approximation deals only with probabilities; it does not represent the actual diffusion process on a particular network, where the spreading always starts from the seeds, not from nodes at the bottom level. Here $\rho_n = \sum_i \rho_n^{(i)}|C_i|/N$, with $i$ in $\rho_n^{(i)}$ indicating that the top node at level $n = t$ belongs to $C_i$ and $|C_i|$ is the size of each community. We can iteratively calculate the cascade size using the following updating equations [38,42]:

$$\bar{q}_n^{(i)} = \frac{\sum_j e_{ij} q_{n-1}^{(j)}}{\sum_j e_{ij}} = \frac{1}{d}\sum_j e_{ij} q_{n-1}^{(j)}, \tag{1}$$

$$q_n^{(i)} = \rho_0^{(i)} + \left(1-\rho_0^{(i)}\right)\sum_k \tilde{p}_k^{(i)} \sum_{m=\lceil *\rceil \theta k}^{k-1} \binom{k-1}{m} \times \left(\bar{q}_n^{(i)}\right)^m \left(1-\bar{q}_n^{(i)}\right)^{k-1-m}, \tag{2}$$

$$\rho_n^{(i)} = \rho_0^{(i)} + \left(1-\rho_0^{(i)}\right)\sum_k p_k^{(i)} \sum_{m=\lceil *\rceil \theta k}^{k} \binom{k}{m} \times \left(\bar{q}_n^{(i)}\right)^m \left(1-\bar{q}_n^{(i)}\right)^{k-m}. \tag{3}$$

In synchronous updating ($f = 1$), the diffusion speed in $C_i$ at time $t$ can be approximated as

$$v_t^{(i)} = d\rho_t^{(i)}/dt = \left[\rho_{t+1}^{(i)} - \rho_t^{(i)}\right]^+, \tag{4}$$

where the notation $[\cdot]^+$ stands for $\max(0,\cdot)$. The overall diffusion speed $v_t$ at time $t$, the total diffusion time $t_s$, and the average diffusion speed $\bar{v}$ are

$$v_t = \sum_i \frac{|C_i|}{N} v_t^{(i)}, \quad t_s = t \mid v_t = 0, \quad \bar{v} = \frac{\rho_{t_s} - \rho_0}{t_s}. \tag{5}$$

Note that this approximation is based on the assumption that the network is locally tree-like, such that the seeds infect nodes one step away at each time step. It becomes the exact solution on large SBM networks ($n \to \infty$) when the average degree remains small. However, it does not work well on non-tree-like networks. We thus perform simulations only on LFR and empirical networks.

### B. LFR network

The node degrees and community sizes in LFR networks both follow a power-law distribution, with exponents $\gamma$ and $\beta$, respectively. The typical values of the exponents are $2 \leqslant \gamma \leqslant 3$, $1 \leqslant \beta \leqslant 2$. Here we let $\gamma = 2.5$, $\beta = 1.5$. Similar to SBM, LFR networks also use a parameter $\mu$ to control for the modularity, which is defined as the fraction of a node's edges to others outside its community. Unlike SBM, the node partition in LFR networks is not fixed for different $\mu$, and $0 \leqslant \mu \leqslant 1$ since the number of communities is typically larger than 2 [38].

### C. Twitter network

The Twitter network data are obtained from Ref. [43]. The largest connected component (LCC) of its undirected network consists of 81 000 nodes and 1.3 million edges, for which we detect 70 communities using Ref. [44].

We rewire edges to change the network modularity. For each rewire, we do the following: (1) with probability $p$, we randomly select a pair of communities and randomly select a within-community edge from each community, and then swap the edge ends if it is possible (no parallel edges are allowed); (2) with probability $(1-p)$, we randomly select a pair of communities and randomly select two cross-community edges running between them, and then swap the edge ends to create

two within-community edges if it is possible. The above process is repeated about 650 000 times so that each edge can be rewired once on average. Parameter $p$ here is similar to $\mu$ used in SBM and LFR networks: a small $p$ increases the modularity, while a large $p$ decreases the modularity.

Note that only rewired Twitter networks are used in Fig. 4. This rewiring process does not change the degree distribution, but would alter other network structure such as clustering besides modularity. However, changes in features other than modularity have relatively small impact on diffusion dynamics [38].

### D. Normalized network modularity

The network modularity $Q$ quantifies the number of intra-community edges minus the expected number if edges are placed at random, for a given node partition. It achieves the maximum value $Q_{\max}$ on a perfectly mixed network where all edges connect nodes in the same community. However, $Q_{\max}$ typically varies from network to network. To compare the strength of modularity across different networks, we therefore use the normalized value of the modularity: $Q_{\text{norm}} = Q/Q_{\max}$ [16]. Note that, in Figs. 1–3, $Q = 1/2 - \mu$, $Q_{\max} = 1/2$, and $Q_{\text{norm}} = 1 - 2\mu$. For LFR (Twitter) networks, the relationship between $Q_{\text{norm}}$ and $\mu$ (or $p$) is nonlinear, as shown in Fig. 4.

All seven empirical networks used in this study is available at Ref. [43]. A public repository with code to reproduce our results is available at Ref. [45].

## ACKNOWLEDGMENTS

We thank Aparna Ananthasubramaniam, Ashok Deb, Ceren Budak, Danaja Maldeniya, Ed Platt, Minje Choi, and Zhuofeng Wu for helpful discussions and suggestions. This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) SocialSim program under contract W911NF-17-C-0094 and by the Air Force Office of Scientific Research under Award No. FA9550-19-1-0029.

H.P., A.N., D.M.R., and E.F. collaboratively conceived and designed the study. H.P. carried out the experiments and performed the analyses. H.P., D.M.R., and E.F. drafted and revised the final manuscript.

[1] M. Granovetter, The strength of weak ties, Am. J. Sociol. **78**, 1360 (1973).

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, The role of social networks in information diffusion, in *Proceedings of the 21st International Conference on World Wide Web* (ACM, New York, 2012), pp. 519–528.

[3] J. Cheng, L. Adamic, A. Dow, J. Kleinberg, and J. Leskovec, Can cascades be predicted? in *Proceedings of the 23rd International Conference on World Wide Web* (ACM, New York, 2014), pp. 925–936.

[4] J. Goldenberg, B. Libai, and E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing Letters **12**, 211 (2001).

[5] D. Kempe, J. Kleinberg, and É. Tardos, Maximizing the spread of influence through a social network, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2003), pp. 137–146.

[6] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. **83**, 1420 (1978).

[7] D. J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. USA **99**, 5766 (2002).

[8] J. Leskovec, L. A. Adamic, and B. A. Huberman, The dynamics of viral marketing, ACM Transactions on the Web **1**, 1 (2007).

[9] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, Group formation in large social networks: Membership, growth, and evolution, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2006), pp. 44–54.

[10] D. Centola and M. Macy, Complex contagions and the weakness of long ties1, Am. J. Sociol. **113**, 702 (2007).

[11] D. Romero, B. Meeder, and J. Kleinberg, Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter, in *Proceedings of the 20th International Conference on World Wide Web* (ACM, New York, 2011), pp. 695–704.

[12] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, Evidence of complex contagion of information in social media: An experiment using Twitter bots, PLoS ONE **12**, e0184148 (2017).

[13] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. USA **104**, 7332 (2007).

[14] M. Smolla and E. Akçay, Cultural selection shapes network structure, Sci. Adv. **5**, eaaw0609 (2019).

[15] A. Galstyan and P. Cohen, Cascading dynamics in modular networks, Phys. Rev. E **75**, 036109 (2007).

[16] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[17] S. Fortunato, Community detection in graphs, Phys. Rep. **486**, 75 (2010).

[18] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[19] M. E. J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[20] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).

[21] D. Centola, The spread of behavior in an online social network experiment, Science **329**, 1194 (2010).

[22] L. Weng, F. Menczer, and Y.-Y. Ahn, Virality prediction and community structure in social networks, Sci. Rep. **3**, 2522 (2013).

[23] D. Romero, C. Tan, and J. Ugander, On the interplay between social and topical structure, in *Proceedings of the 7th International AAAI Conference on Web and Social Media* (AAAI, California, 2013), pp. 516–525.

[24] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn, Optimal network modularity for information diffusion, Phys. Rev. Lett. **113**, 088701 (2014).

[25] J. L. Iribarren and E. Moro, Impact of Human Activity Patterns on the Dynamics of Information Diffusion, Phys. Rev. Lett. **103**, 038702 (2009).

[26] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, Small but slow world: How network topology and burstiness slow down spreading, Phys. Rev. E **83**, 025102(R) (2011).

[27] J.-C. Delvenne, R. Lambiotte, and L. E. C. Rocha, Diffusion on networked systems is a question of time or structure, Nat. Commun. **6**, 7366 (2015).

[28] B. Mišić, R. F. Betzel, A. Nematzadeh, J. Goni, A. Griffa, P. Hagmann, A. Flammini, Y.-Y. Ahn, and O. Sporns, Cooperative and competitive spreading dynamics on the human connectome, Neuron **86**, 1518 (2015).

[29] S. Yan, S. Tang, W. Fang, S. Pei, and Z. Zheng, Global and local targeted immunization in networks with community structure, J. Stat. Mech.: Theory Exp. (2015) P08010.

[30] P. Sah, S. T. Leu, P. C. Cross, P. J. Hudson, and S. Bansal, Unraveling the disease consequences and mechanisms of modular structure in animal social networks, Proc. Natl. Acad. Sci. USA **114**, 4165 (2017).

[31] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, Threshold-limited spreading in social networks with multiple initiators, Sci. Rep. **3**, 2330 (2013).

[32] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E **83**, 016107 (2011).

[33] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, On Facebook, most ties are weak, Commun. ACM **57**, 78 (2014).

[34] Q. Ke and Y.-Y. Ahn, Tie strength distribution in scientific collaboration networks, Phys. Rev. E **90**, 032804 (2014).

[35] A. M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers, Proc. Natl. Acad. Sci. USA **112**, E4671 (2015).

[36] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, The unreasonable effectiveness of tree-based theory for networks with clustering, Phys. Rev. E **83**, 036112 (2011).

[37] J. P. Gleeson, Cascades on correlated and modular random networks, Phys. Rev. E **77**, 046117 (2008).

[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.102.052316 for details on the analytical approximation, results on SBM networks with different parameters, simulations on empirical networks, and different notions of diffusion speed.

[39] A. Lancichinetti, S. Fortunato, and F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E **78**, 046110 (2008).

[40] S.-J. Salvy, K. de la Haye, T. Galama, and M. I. Goran, Home visitation programs: An untapped opportunity for the delivery of early childhood obesity prevention, Obesity Rev. **18**, 149 (2017).

[41] B. Wilder, Han C. Ou, K. de la Haye, and M. Tambe, Optimizing network structure for preventative health, in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (ACM, New York, 2018), pp. 841–849.

[42] M. E. J. Newman, Mixing patterns in networks, Phys. Rev. E **67**, 026126 (2003).

[43] J. Leskovec and A. Krevl, SNAP datasets: Stanford large network dataset collection (2014), https://snap.stanford.edu/data/.

[44] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech.: Theory Exp. (2008)P10008.

[45] https://github.com/haoopeng/diffusion_speed.