Network Modularity Controls the Speed of Information Diffusion (Supplemental Material)

Hao Peng,[1] Azadeh Nematzadeh,[2] Daniel M. Romero,[1] and Emilio Ferrara[3]

[1] *School of Information, University of Michigan*

[2] *S&P Global*

[3] *Information Sciences Institute, University of Southern California*

(Dated: November 6, 2020)

I. INTRODUCTION

The network representation of social relationships between people is a core ingredient in modeling the dynamics of information diffusion since the adoption of ideas or social behaviors are often influenced by one's social neighbors. Therefore the structure of the underlying social network strongly affects the process of information diffusion. In this paper, we study how a salient network property—the modular structure—influences the speed of information diffusion by using the linear threshold diffusion model on networks with varying degree of network modularity. Through both simulations and an analytical approximation, we demonstrate that there exists an optimal network modularity for the most efficient information diffusion at global scale.

In this supplemental material, we present further evidence to support our findings by examining the behavior of our diffusion model under more general conditions with a wide range of parameters. We investigate the average speed of information diffusion on SBM networks with varying (i) network size N, (ii) average degree z, (iii) anti-modular structure, (iv) seed arrangements, and (v) number of communities d. Additionally, we report results based on many real-world networks where the seed nodes are randomly selected across the whole network instead of from a single community. Finally, we present qualitatively similar results by considering a different constraint—fixing the diffusion time instead of fixing the cascade size—in measuring the average diffusion speed.

II. THE TREE-LIKE APPROXIMATION OF DIFFUSION SPEED

As discussed in the main paper, $\rho_n = \sum_i \rho_n^{(i)} |C_i|/N$, with i in $\rho_n^{(i)}$ indicating that the top node belongs to C_i and $|C_i|$ is the size of each community. To calculate $\rho_n^{(i)}$, we introduce two auxiliary variables: $q_n^{(i)}$ and $\bar{q}_n^{(i)}$. Let $q_n^{(i)}$ be the probability that a node in C_i at level n is active, conditioning on its parent being inactive, and $\bar{q}_n^{(i)}$ be the probability of reaching an active child at level $n-1$ by following an edge from an inactive node in C_i at level n. We can update $q_n^{(i)}$ and $\bar{q}_n^{(i)}$ using Eq. 1-2:

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

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

where $\tilde{p}_k^{(i)}$ is the probability that a node in C_i reached by following an edge from its inactive parent has degree k, thus $\tilde{p}_k^{(i)} = k p_k^{(i)} / z^{(i)}$ [1]. Note that $q_0^{(i)} = \rho_0^{(i)}$. Eq. 2 is the sum of two scenarios: (i) the probability that the node is among the seeds ($\rho_0^{(i)}$), and (ii) the probability that the node is not among the seeds ($1-\rho_0^{(i)}$) but is connected to at least $\lceil \theta k \rceil$ active children (the second summation, note that this node connects to $k-1$ children), summed over all possible degrees k of that node (the first summation).

Similar to $q_n^{(i)}$, we calculate $\rho_n^{(i)}$ as (note that the top node connects to k children since it has no parent, and its degree is distributed according to $p_k^{(i)}$ instead of $\tilde{p}_k^{(i)}$)

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

In synchronous updating ($f = 1$), the diffusion speed in C_i at time t can be approximated as: $v_t^{(i)} = d\rho_t^{(i)}/dt = [\rho_{t+1}^{(i)} - \rho_t^{(i)}]^+$, 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

$$v_t = \sum_i \frac{|C_i|}{N} v_t^{(i)}, \qquad t_s = t \mid v_t = 0, \qquad \bar{v} = \frac{\rho_{t_s} - \rho_0}{t_s}. \tag{4}$$

These equations can be adapted for asynchronous updating, provided that the fraction f of nodes updated at each time step is sufficiently small such that they may be considered to be independent of each other [2]. We introduce the following notation: $\bar{q}(t)$, $q(t)$ and $\rho(t)$. The evolution equations for asynchronous updating are

$$\bar{q}^{(i)}(t) = \frac{1}{d} \sum\nolimits_j e_{ij} q^{(j)}(t-1), \tag{5}$$

$$dq^{(i)}(t)/dt = f[g^{(i)}(\bar{q}^{(i)}(t+1)) - q^{(i)}(t)]^+, \tag{6}$$

$$v^{(i)}(t) = d\rho^{(i)}(t)/dt = f[h^{(i)}(\bar{q}^{(i)}(t+1)) - \rho^{(i)}(t)]^+, \tag{7}$$

with $q^{(i)}(0) = \rho^{(i)}(0) = \rho_0^{(i)}$. The speed is calculated as,

$$v(t) = \sum_i \frac{|C_i|}{N} v^{(i)}(t), \qquad \bar{v} = \frac{\rho(t_s) - \rho(0)}{t_s}. \tag{8}$$

III. RESULTS

A. Network size

Figure S1 and Figure S2 present results based on SBM networks with different number of nodes, derived through the analytical approach and the numerical simulation, respectively. It shows that the network size does not change our finding of the most efficient spreading behavior with respect to the network modularity.

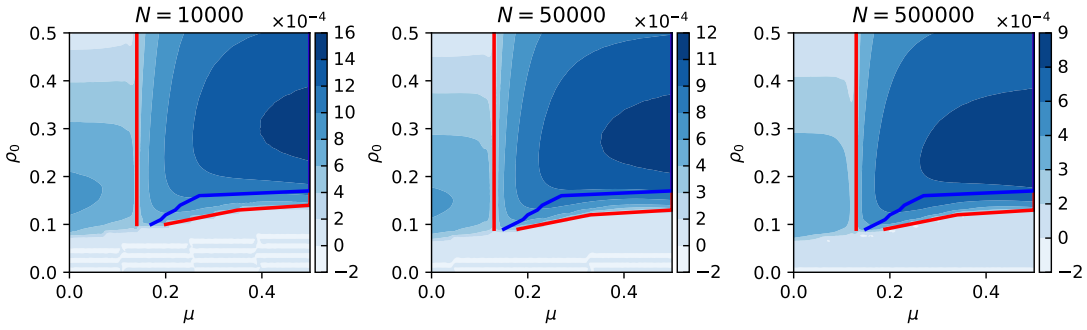

Figure S1. Phase diagrams of the average diffusion speed on SBM networks with different number of nodes N. The results are derived from the analytical approach. Other model parameters are: $z = 10, \theta = 0.35, f = 0.01$.

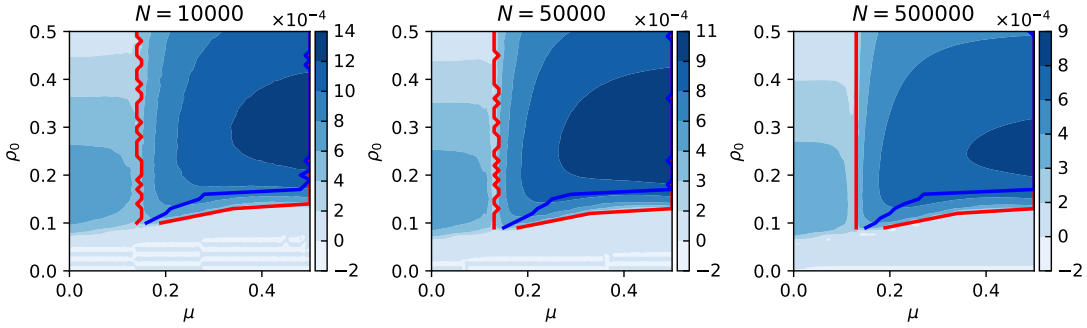

Figure S2. Phase diagrams of the average diffusion speed on SBM networks of different sizes N, derived from numerical simulations (averaged over 100 runs). Other model parameters are: $z = 10, \theta = 0.35, f = 0.01$.

B. Average degree

Figure S3-S4 show the average diffusion speed as a function of the seed size and the network modularity, on SBM networks with different average degrees. The results indicate that, as one increases the average degree, the minimal number of inter-community edges (or the maximum modularity) required to generate global cascades also increases, so does the minimal number of seeds. This is expected because more active neighbors are needed to achieve the same adoption threshold when the nodes' neighbor size increases.

However, the optimal network modularity for the overall fastest information diffusion always exists when global cascades are enabled. And the optimal value depends on the seed size, which agrees with our finding in the main

text. In other words, the average degree does not change the behavior of our system qualitatively.

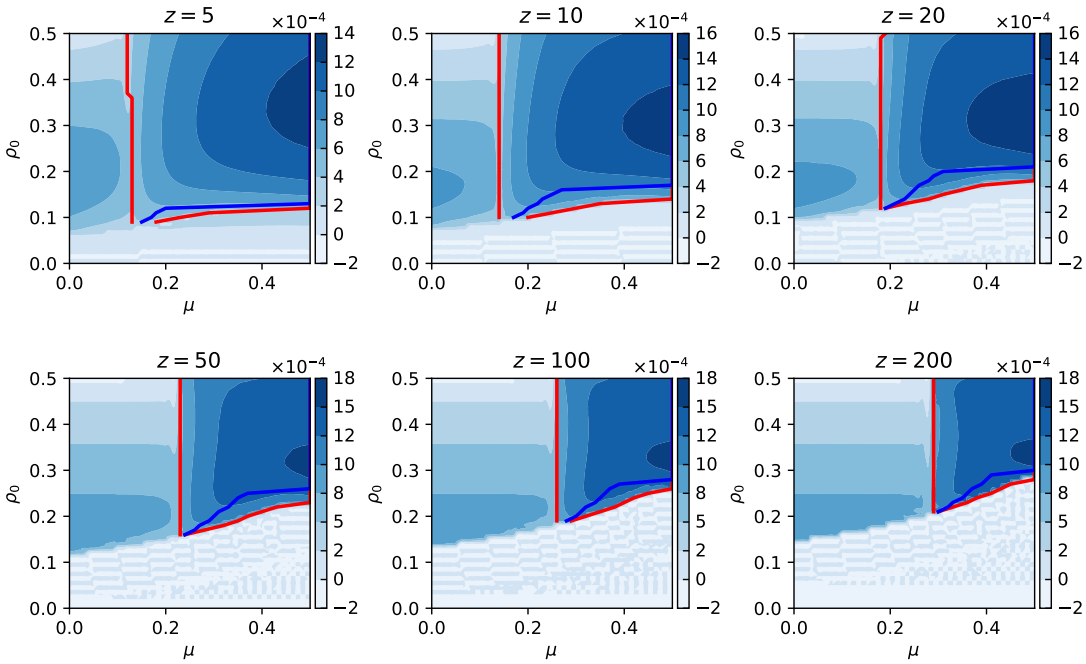

Figure S3. Phase diagrams of the average diffusion speed on SBM networks derived from the analytical approximation. Each subplot corresponds to networks with a specific average degree z. Other model parameters are: $N = 1\times10^4, \theta = 0.35, f = 0.01$.

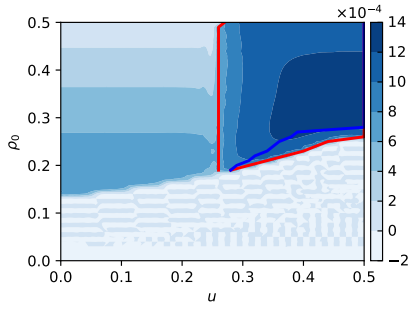

Figure S4. Phase diagrams of the average speed of information diffusion on SBM networks with average degree $z = 100$. The results are obtained from numerical simulations. Other model parameters are: $N = 1\times10^4, \theta = 0.35, f = 0.01$.

C. Anti-modular SBM networks

Although our focus is on networks with community structure, it is intriguing to examine the diffusion dynamics on anti-modular networks. Figure S5 shows the average diffusion speed in the whole range of μ on SBM networks, where the network shifts from exhibiting a modular structure to displaying a bipartite structure. Interesting patterns emerge: different from the dynamics on modular networks, where information spreads from the originating community to the other, the diffusion process on anti-modular networks temporally alternates between the two communities.

In such a scenario, global cascades still require a minimal number of seeds, but unlike modular networks, when ρ_0 is

not too large (e.g., $\rho_0 = 0.2$), strong anti-modular structure (large μ) always promotes the diffusion speed, making the strict bipartite networks the ideal conditions for global cascades. However, when ρ_0 is sufficiently large (e.g., $\rho_0 = 0.4$), the most efficient global cascade happens at an intermediate strength of anti-modular structure (Figure S5).

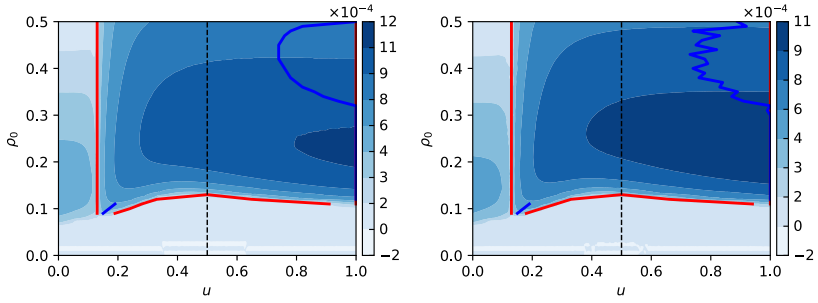

Figure S5. Phase diagrams of the average diffusion speed in the whole range of μ in SBM networks. The results are based on both analytical predictions (left) and numerical simulations (right), averaged over 100 runs. Model parameters are: $N = 1{\times}10^5, z = 10, \theta = 0.35, f = 0.01$. There are three regions: $\mu < 0.5$ (assortative and modular); $\mu = 0.5$ (random); and $\mu > 0.5$ (disassortative and anti-modular). The anti-modular networks behave quite differently from the modular networks.

D. Seed arrangement

We also examine the diffusion dynamics in our system under conditions where the seeds are not entirely placed in a single community. Figure S6 shows that, at any given seed distribution in the network (draw a horizontal slice), when global cascades are possible, there is a window of network modularity for information diffusion at global scale. For example, when all seeds are placed in C_2 (none in C_1), the μ window for global cascades is $[0.13, 0.24]$, and the fastest diffusion process happens at a middle level of modularity ($\mu = 0.17$), which is exactly what we see in Fig. 1 in the main text. The same pattern holds for other seed arrangements in Figure S6. In other words, our finding of an intermediate strength of network modularity being the ideal condition for efficient global cascades can be generalized to all other seed arrangements in the two communities, for the seed size $\rho_0 = 0.1$.

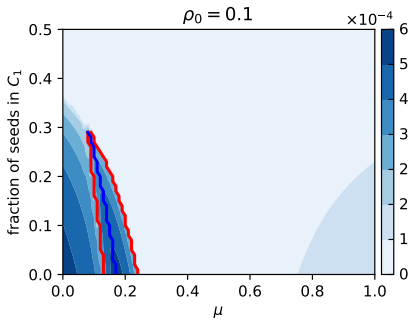

Figure S6. Phase diagrams of the average information diffusion speed in the whole range of μ, as a function of seed arrangements between two communities in SBM networks. The results are based on analytical predictions. The y-axis represents the fraction of seeds placed in C_1. The seed size $\rho_0 = 0.1$. Other model parameters are: $N = 1{\times}10^5, z = 10, \theta = 0.35, f = 0.01$

E. Number of communities

So far, all our experiments on SBM networks are limited to the case of two equally sized communities ($|C_1| = |C_2|$). Here, we examine the diffusion dynamics on SBM networks with different number of communities. As a first step, we assume that all communities still have the same number of nodes and links are randomly placed according to the parameter μ, as is the case in the main text. The mixing matrix is:

$$\mathbf{e} = \frac{1}{d}\begin{bmatrix} 1-\mu & \frac{\mu}{d-1} & \cdots \\ \vdots & \ddots & \\ \frac{\mu}{d-1} & & 1-\mu \end{bmatrix}, \tag{9}$$

where $\mathbf{e}$ is $d \times d$ and d is the number of communities. The diagonal entries of $\mathbf{e}$ are $\frac{1-\mu}{d}$ and the off-diagonal entries are $\frac{\mu}{d(d-1)}$ [3]. The network modularity can be calculated as: $Q = 1-\mu-\frac{1}{d}$, which means that, in order to generate modular networks, μ can be larger than $\frac{1}{2}$ when d is large than 2.

Figure S7 shows the analytical results of the average diffusion speed on SBM networks with different number of communities. Please note that, at any given μ, the number of bridges running between a pair of communities decreases as the number of communities d increases. Thus networks with more communities require smaller adoption threshold θ in order to achieve global cascades. Figure S7 indicates that our finding of the optimal network modularity for the most efficient global diffusion can generalize to networks with multiple communities.

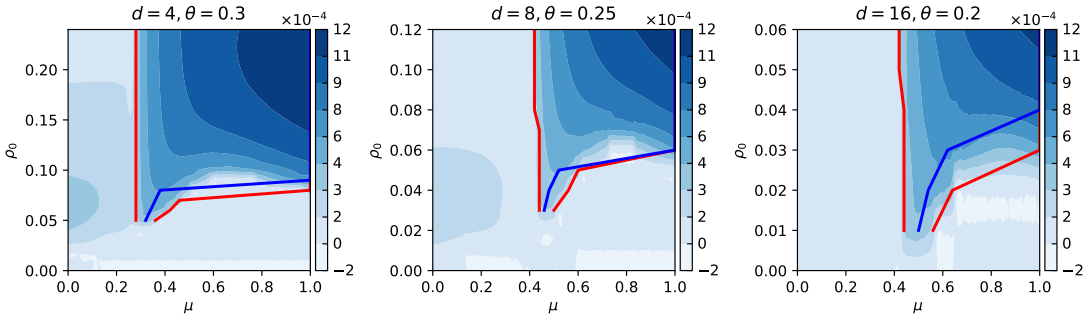

Figure S7. Phase diagrams of the average diffusion speed on SBM networks with different number of equally sized communities d, derived from the analytical approximation. Seeds are randomly selected from a single community. Other model parameters are: $N = 1\times10^5, z = 10, f = 0.01$.

F. Simulations on real-world networks

In the main paper, we showed simulation results on LFR and Twitter networks. Here we extend our experiments to more real-world networks across different domains including social, communication, and collaboration networks from [4]. Like the case of Twitter, we use the largest connected component (LCC) of the undirected version of each network, and use the parameter p to control the network modularity through the edge rewiring process described in the main text. Note that we focused on networks with a LCC that contains at least 50K nodes and excluded those with more than 1M nodes to make the simulations feasible and comparable to SBM and LFR networks. Table S1 summarizes the statistics of networks we test here.

Network Name	Num. of Nodes	Num. of Edges	Avg. Degree	Num. of Communities	$Q_{\mathbf{norm}}$
DBLP	317,080	1,049,866	6.6	239	0.84
Eu Email	224,832	339,925	3	89	0.80
Slashdot	82,168	504,230	12.3	549	0.44
Twitter	81,306	1,342,310	33.0	70	0.86
Epinions	75,877	405,739	10.7	776	0.55
Deezer	54,573	498,202	18.3	24	0.79
FB Pages	50,515	819,090	32.4	34	0.72

Table S1. Statistics of the largest connected component of seven real-world networks we tested. Directed networks are all converted to undirected networks. The communities are detected using the Louvain algorithm [5]. The community sizes are heterogeneous. Note that the Twitter network has been used in the main paper.

Figure S8 shows the phase diagrams for six empirical networks. The pattern looks similar to that on SBM, LFR, and Twitter networks. There exists an optimal modularity for overall fast global cascades, and the optimal value depends on the seed size and the network.

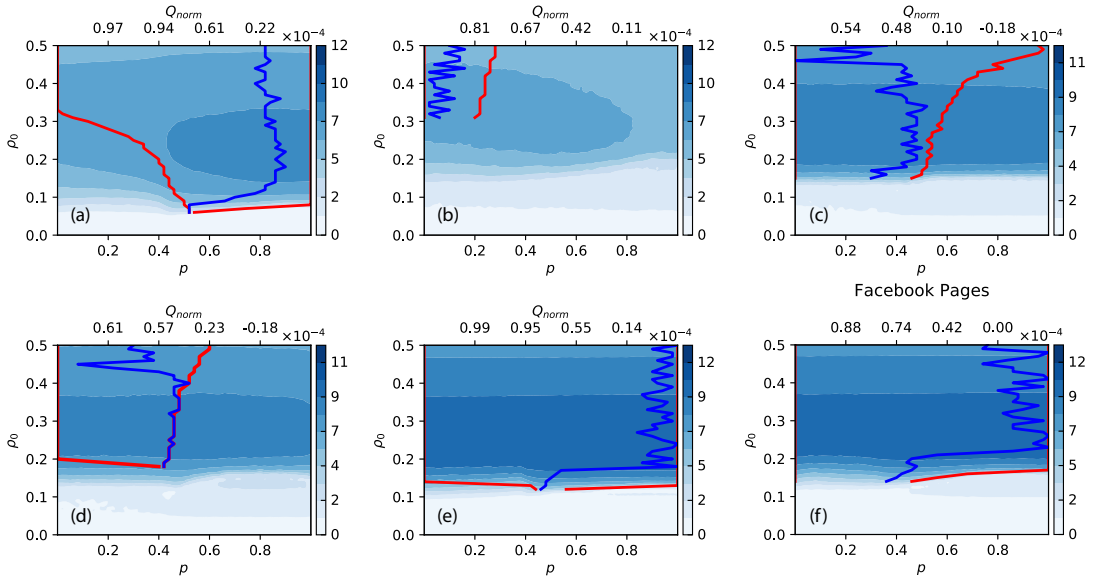

Figure S8. Phase diagrams of the average diffusion speed $\bar{v}$ on six real-world networks. (**a**) DBLP; (**b**) Eu Email; (**c**) Slashdot; (**d**) Epinions; (**e**) Deezer; (**f**) Facebook Pages. Network statistics are shown in Table S1. Network modularity is controlled by parameter p on the x-axis, with the corresponding normalized modularity Q_{norm} shown on the top axis. The blue curve indicates the optimal p for $\bar{v}$ for a given seed size ρ_0 (there is only a single p that maximizes $\bar{v}$ for any given ρ_0). Simulation parameters are: $\theta = 0.3, f = 0.01$. Seed nodes are randomly selected across the whole network.

G. Average diffusion speed with a constraint on time

There are many real world diffusion applications that need to be optimized for the speed with a predefined cascade size. However, there are also cases where one cares about the speed with a time limit (or equivalently the cascade

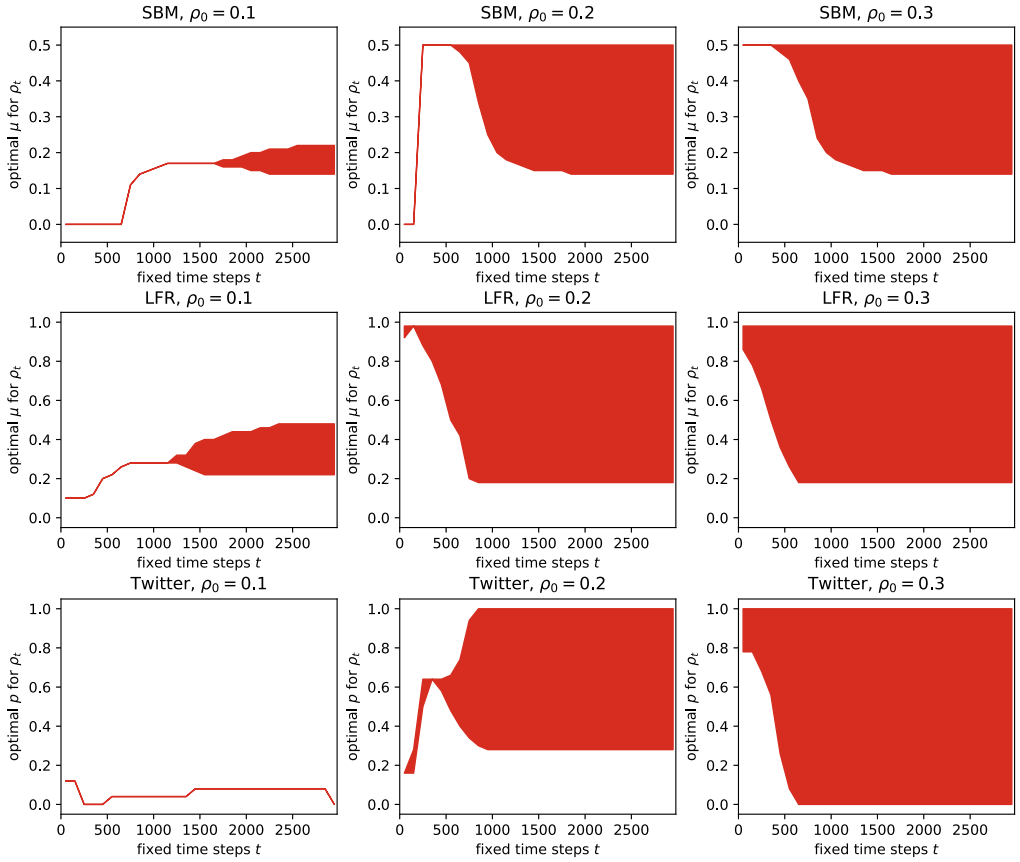

Figure S9. The optimal modularity for the average diffusion speed with a constraint on diffusion time. The results are based on simulations. The most efficient network transitions from having a single optimal modularity to exhibiting a wide range of optimal modularity (controlled by μ or p) as the time increases, especially for small seed sizes. This trend is qualitatively the same for synthetic networks (SBM, LFR) and real-world networks (Twitter).

size for a fixed time window). For example, 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. Additionally, the spread of health behaviors such as wearing masks and social distancing aims to slow down the spread of Coronavirus *before* hospital capacity is surpassed. We thus examine the optimal structure for diffusion speed with a time constraint.

Figure S9 indicates that the best modularity for fast diffusion also tends to decrease as the diffusion time increases. For instance, the optimal μ changes from $\mu = 0$ to $\mu = 0.17$ as t increases from $t = 500$ to $t = 1500$ for $\rho_0 = 0.1$ on SBM networks. In other words, if the goal is to infect as many nodes as possible in a very short period of time, then a higher modularity is better than the optimal for a longer time window when global cascades are achieved. The intuition behind this result is that since the diffusion speed within communities is higher than that for inter-community spreading at the early stage of the diffusion process (see Fig. 2 in the main text), when the time available for diffusion is limited, it is better to have strong modularity to promote local spreading.

Furthermore, unlike conditions with a constraint on cascade size where the optimal modularity is typically a single value, when the constraint is on time, networks can exhibit a wide range of optimal modularity values, especially for a large time budget. Intuitively, if the time is sufficiently long, many modularity values are optimal as long as they are within the window of global cascades. When the seed size is too large (e.g., $\rho_0 = 0.3$), there tends to exist a wide

range of optimal modularity values, regardless of the time budget. The reason is that the diffusion process tends to reach global cascades so quickly that the time budget usually cannot be exhausted.

H. Simulations on the original empirical network vs. rewired networks

Fig. S10 shows that both the average diffusion speed and the cascade size are similar on the original Twitter network or on its rewired network with the same modularity. In other words, although the rewired network can change other network characteristics besides modularity, their changes have relatively little influence on the diffusion dynamics when the network modularity is fixed. This supports our analysis of how modularity affects diffusion speed on empirical networks by systematically rewiring the original network for a continuous change of modularity.

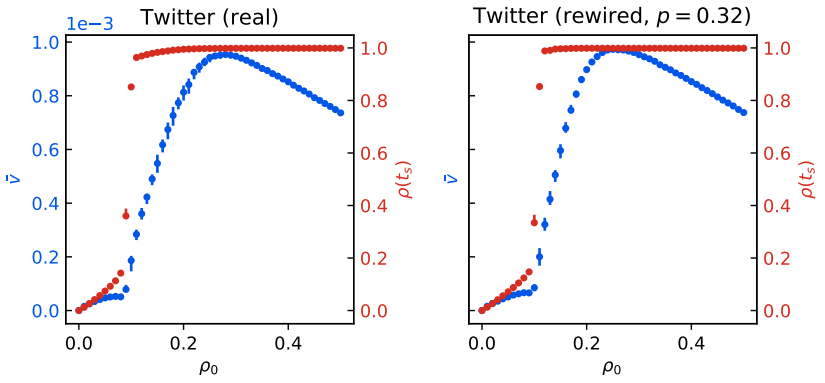

Figure S10. Simulation results of the average diffusion speed ($\bar{v}$, blue axis) and cascade size (ρ_{t_s}, red axis) on the Twitter empirical network (cf. Fig. 4 in the main text), for its original structure (left subplot) and the rewired version with the same normalized modularity (right subplot). The x-axis represents the seed size ρ_0. The simulation results are averaged over 100 runs for each seed size, with $\theta = 0.3, f = 0.01$. The rewired network (right subplot) is constructed with $p = 0.32$ such that its normalized modularity ($Q_{\text{norm}} = 0.86$) is the same as the original structure (cf. Fig. 4 in the main text and Table S1). The error bars indicate the interquartile ranges.

[1] Mark E.J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[2] James P Gleeson. Cascades on correlated and modular random networks. *Physical Review E*, 2008.

[3] Mark E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[4] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection, 2014.

[5] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp*, 2008(10):P10008, 2008.