Social network structure and the spread of complex contagions from a population genetics perspective

Julian Kates-Harbeck
Department of Physics,

Michael M. Desai*
Department of Organismic and Evolutionary Biology,
Harvard University, Cambridge MA 02138, USA
(Dated: August 8, 2022)

Ideas, behaviors, and opinions spread through social networks. If the probability of spreading to a new individual is a non-linear function of the fraction of the individuals' affected neighbors, such a spreading process becomes a "complex contagion". This non-linearity does not typically appear with physically spreading infections, but instead can emerge when the concept that is spreading is subject to game theoretical considerations (e.g. for choices of strategy or behavior) or psychological effects such as social reinforcement and other forms of peer influence (e.g. for ideas, preferences, or opinions). Here we study how the stochastic dynamics of such complex contagions are affected by the underlying network structure. Motivated by simulations of complex epidemics on real social networks, we present a general framework for analyzing the statistics of contagions with arbitrary non-linear adoption probabilities based on the mathematical tools of population genetics. Our framework provides a unified approach that illustrates intuitively several key properties of complex contagions: stronger community structure and network sparsity can significantly enhance the spread, while broad degree distributions dampen the effect of selection. Finally, we show that some structural features can exhibit critical values that demarcate regimes where global epidemics become possible for networks of arbitrary size. Our results draw parallels between the competition of genes in a population and memes in a world of minds and ideas. Our tools provide insight into the spread of information, behaviors, and ideas via social influence, and highlight the role of macroscopic network structure in determining their fate.

Individuals on a social network are subject to influence by their neighbors, affecting their adoption of information [1], ideas [2], and behaviors [3]. The likelihood that a given individual adopts a new idea depends on how many of her neighbors have adopted the idea already. For physically spreading infections, as encountered in traditional epidemiology [4], this dependence is typically linear and leads to a "simple contagion". By contrast, social reinforcement and other forms of peer influence [5, 6], as well as game theoretical considerations of behavior [7], can result in a non-linear dependence of an individual's likelihood of adoption on her neighbors' status [5, 8–16]. A spreading process with such a non-linear likelihood of adoption is a "complex contagion", whose properties can differ significantly from simple contagions [17, 18]. The spread of complex contagions is related intimately to the interplay of network structure and adoption patterns, relying on locally high prevalence and multiple peer influence in order to spread.

The empirical evidence for complex contagions is accumulating [1, 19–23] and several structural features influencing spread have been identified [18, 23–26]. Beyond adoption characteristics and network structure, other factors influencing spread likely include individual heterogeneity, personal characteristics, strategic or reactive adoption, as well as global influences such as mass media [21, 27–29].

Generalized dose response models [30, 31] and threshold models [32] provide a simple and elegant way to capture non-linear adoption. In some cases, the relevant parameters of the model, such as the probability of contagion spreading given one or two exposures, can be empirically measured to calibrate the model [31]. However, these models do not address the temporal dynamics of the contagion or connect its behavior to specific structural properties of the underlying network. In some cases, assuming locally random tree-like networks (i.e. in the absence of significant clustering), general conditions for global spread can be derived [9, 33]. These approaches, however, do not illustrate the dynamics of "small" contagions that never reach macroscopic size, and do not apply to community based or highly clustered networks. They do illustrate a subtle tradeoff between threshold level and degree heterogeneity that we build on in this paper: When an individual's threshold is a function of the fraction (as opposed to the absolute number) of affected neighbors, low degree nodes are easily susceptible, but don't pass on the contagion easily, while high degree nodes are harder to activate but pass it on more widely. It is thus not obvious what effect changing degree heterogeneity may have on the spread of such contagions.

Game theoretic and threshold models have been used successfully to illustrate the key insight - supported by recent empirical work [34, 35] - that clustering and communities can accelerate the spread of a complex contagion by allowing it to quickly reach locally high levels and spread one community at a time [7, 36], whereas simple contagions converge faster for high-dimensional networks dominated by "long ties". Incidentally, similar insights emerge in the context of synergistic coinfections, whose coupled epidemiological dynamics also exhibit non-linearities and thus complex contagion properties [16]. These theoretical studies use approaches focused on deterministic mean field dynamics and convergence times, and

* mdesai@oeb.harvard.edu

are restricted to the regime of strong positive selection (i.e. where convergence is essentially guaranteed) [7].

In this paper, we provide a unified framework for studying complex contagions and formulating the conditions under which they propagate based on the mathematical tools and intuition from population genetics, which offers a straightforward frame for analyzing contagion dynamics at all scales of a network, from the local neighborhood to the community to the global scale. We aim to formulate the general network conditions under which complex contagions propagate. While the influence of clustering, community structure, sparsity, and degree distributions has been illuminated in previous work [17, 18], our approach builds on and supplements prior work by providing a general framework for intuitive derivations for key properties of complex contagions and their dependence on the above network features, accounting for their full stochastic dynamics subject to arbitrary nonlinear adoption patterns and selection regimes, all in the language of population genetics. The mathematical approach highlights parallels between the competitions of genes in a population and the competition of memes in a world of minds and ideas.

In particular, we study here the fate and adoption of a newly arising idea on a network, giving rise to a complex contagion. We model this process in the framework of evolutionary game theory by considering individuals as the nodes of an undirected graph, with edges representing interaction and communication patterns (**?? (a,b)**). We introduce the new idea as a single randomly chosen type B node on a network in which all other nodes are initially of type A. Both types spread by contagion. In particular, we assume that individuals update their type as a continuous stochastic process, where the rate of switching depends on the fraction of neighbors of a given type: a type A node becomes type B at rate $r_1 = y\,[1 + f_1(y)]$ and type B nodes become type A at rate $r_2 = [1 - y]\,[1 + f_2(y)]$, where y is the local fraction of type B neighbors at a given node. Our main aim is to understand how successfully the new idea spreads through the network by calculating how the overall fraction of type B individuals, $y(t)$, changes over time.

In a simple contagion, the rates of switching are proportional to the local density of type B individuals, so f_1 and f_2 are independent of y [4, 37–39]. This is typical in many models of epidemics, where the rate at which a susceptible individual becomes infected is often assumed to be proportional to the number of infected neighbors. However, we consider here complex contagions, in which f_1 and f_2 can be functions of y. This non-linear dependence on y has been observed in the propagation of online contagions [1, 5, 19]. The competing effects of clustering and "long ties" on such complex epidemics have been studied theoretically [6, 7, 13, 14] and empirically [40], and deterministic models studying how heterogeneous adoption thresholds [8, 9] and the form of adoption functions [11] interact with node degree on random networks have also been analyzed. However, this earlier work focuses largely on the deterministic mean-field behavior of large epidemics, or on stochastic analysis characterizing how specific network properties affect the probability of large or global "cascades". In contrast, the effects of general network features on the stochastic dynamics of epidemics of a range of sizes (both the statistical distribution of rare events as well as the probabilities of global cascades) remain poorly characterized. Here we present a unified framework to analyze these stochastic dynamics for arbitrary forms of complex contagions, and apply our model to understand the effects of key network properties such as sparsity, community structure, and degree distributions.

For concreteness we focus primarily on the simple illustrative case where $f_1(y) = \alpha y$ and $f_2(y) = \beta$, with positive α and β. This models "positive frequency dependence" [41], where an idea is unpersuasive while rare but becomes more attractive as it is more widely adopted [6, 7, 13]. This is a natural assumption in many contexts (e.g. political views, preferences, games, or communication habits). However, we note that some ideas may be positively selected at all frequencies (i.e. negative β), in which case they will always tend to spread, and negative frequency dependence (i.e. negative α) may also be relevant in other scenarios (e.g. fashion trends or baby naming). We further assume that $\alpha, \beta \ll 1$, which implies that the strength of selection is relatively weak, such that a preference for one or the other type only emerges on a collective population level (in the opposite case, the idea will tend to very quickly either spread or be eliminated).

To some readers this model may appear reminiscent of SIS or SIR models in epidemiology [42]. Indeed, these models are encompassed by our framework. However, in SIS or SIR models the rate of recovery of a given individual is generally not subject to neighbor influence, while the rate of spread is linear in the neighbors. This leads to simple contagion dynamics for low values of y and a diverging negative frequency dependent selection for large values of y (see **SI**). Therefore, small epidemics are well described with simple contagions, with the additional trivial consequence that large epidemics become exponentially unlikely. We do not study this case here. Instead, our paper is focused on the rich behavior resulting from positive frequency dependence once a sufficient prevalence y is reached. In this case, dynamics for low y are not well described with simple contagion models, considerations of social proof [5, 19] and evolutionary game theory are relevant, and the conclusions and intuitions gained from the model can differ substantially from those implied by epidemic models [7].

In **?? (c-e)**, we explore how the spread of such a complex contagion is influenced by network structure. For this purpose, we consider the Facebook network from the Stanford Large Network Dataset collection [43]. We construct a sequence of networks with variable clustering but unchanged degree sequence by randomly swapping pairs of edges, and study contagions on this set of graphs. We find that the spread of simple contagions is largely insensitive to network structure. However, for complex contagions there is a critical level of clustering required to allow the contagion to spread globally. Below this level, the contagion becomes exponentially unlikely to fix across large networks. We also find that the epidemic fixes one community at a time when clustering is sufficiently high, but for moderate or low clustering values, all communities move through y space more or less in unison.

To analyze these effects, we begin by calculating the global rate at which type A individuals become type B. In our model of epidemic dynamics, this is

$$\text{Rate}_{\text{A}\to\text{B}} = N(1-y)E_A[y(1+f_1(y))] = N(1-y)\left(E_A[y] + \alpha E_A[y^2]\right) \ . \tag{1}$$

Here $N(1-y)$ is the number of type A individuals, and the expectation value gives the mean rate r_1 as averaged over all of these type A nodes. Crucially, this depends on the distribution of y seen by type A individuals, which will depend on the network structure. The rate of the reverse process $\text{Rate}_{\text{B}\to\text{A}}$ has an equivalent form. The relative difference between these rates determines a selection pressure, s, which we define in the standard way from population genetics [44],

$$s \equiv \frac{2\left(\text{Rate}_{\text{A}\to\text{B}} - \text{Rate}_{\text{B}\to\text{A}}\right)}{\text{Rate}_{\text{A}\to\text{B}} + \text{Rate}_{\text{A}\to\text{B}}} \ . \tag{2}$$

This selection pressure determines whether the contagion will on average tend to grow ($s > 0$) or shrink ($s < 0$).

In a well-mixed population, where every node is connected to all other nodes, all individuals see the same global value of y. Thus $E_A[y^2] = y^2$, and hence $s(y) = \alpha y - \beta$, consistent with our model of an idea that is negatively selected when rare but becomes more popular as it increases in frequency. The critical threshold frequency above which the idea becomes positively selected is $y = y_n \equiv \frac{\beta}{\alpha}$. However, standard results from population genetics [44] imply that whenever the number of type B individual is small compared to the inverse of the selection pressure (i.e. when $Ny|s| \ll 1$), the random stochasticity of the process dominates over the effects of selection, and the frequency of the idea is dominated by random "genetic drift." In contrast, when $Ny|s| \gg 1$, selection dominates over random drift, and the idea will tend to deterministically spread or be eliminated from the population.

We define $P_{reach}(y)$ as the probability that the epidemic reaches a given value of at least y before becoming extinct. This function captures the ability of the new idea to invade the population and describes the statistical behavior of the process at both small and large values of y. The selection regimes described above then define various different behaviors of $P_{reach}(y)$. When drift dominates, $P_{reach}(y)$ falls off as $\frac{1}{Ny}$ as in a neutral random walk. In regimes of positive selection, a contagion reaching a given value of y is almost certain to reach continuously higher values of y, so P_{reach} is approximately constant. By contrast, when negative selection dominates, the contagion becomes exponentially less likely to reach ever higher values of y, so P_{reach} falls off exponentially.

In a complex contagion, where s is a function of y, the process can encounter various such regimes of selection, as illustrated in figure **Figure 2 (a-b)**. In our example where $s(y) = \alpha y - \beta$, the epidemic begins with a neutral regime at low y. Depending on the total network size N, the epidemic may then encounter a regime of negative selection before eventually reaching the regime of positive selection above frequency y_n. In our example, the boundaries between these regimes are defined by the points at which $Ny|s| = 1$, while in the more general frequency dependent case for arbitrary $s(y)$ this generalizes to $N|S(y)| = 1$, where $S(y) = \int_0^y s(z)dz$ captures the integrated effect of selection up to y (see **SI**, **Extended Data Figure 1**).

One simple but critical aspect of network structure is that not all nodes are connected. To focus on the effects of this sparsity, we next consider the spread of a contagion on a random regular graph, where each node is connected at random to exactly k other nodes [45]. In such a network, each node will no longer see the "global" value of y, but rather some local value that reflects the fraction of its neighbors that happen to be type B. In principle, determining these local values of y is a complicated problem. However, because the network is random, we expect no strong locality in how type B individuals are distributed, so the neighbors of each type A individual form an approximately random sample of size k of the whole population (the "annealed approximation" [46, 47]). This contrasts with the case of a spatial network (e.g. a square lattice) where locality is fundamental to the network geometry (in this case the epidemic becomes a front propagation problem and must be treated differently [48]). We show the accuracy of our assumption in **Extended Data Figure 2**, and contrast it with the case of spatial networks in **Extended Data Figures 3** and **4**.

In our approximation, $E_A[y] = y$, so a simple contagion is unaffected by network sparsity. However, due to limited connectivity, some type A nodes will have more type B neighbors than others, and hence $E_A[y^2] > E_A[y]^2 = y^2$. Thus for a complex contagion, sparsity increases r_1 and enhances the spread of the epidemic. Specifically, we find (see **SI** for details) that for large networks where $N \gg k$ (and assuming $\alpha, \beta \ll 1$), the selection pressure acting on the contagion is

$$s(y) = \alpha\left(y + \frac{(1-y)}{k}\right) - \beta \ . \tag{3}$$

This reduces to the well-mixed solution $s(y) = \alpha y - \beta$ as k becomes large, but for small k selection is significantly enhanced, as shown in **Figure 2 (c)**. The key point is that for small k, some nodes will happen to have more type B neighbors than others, and because the transition rates increase nonlinearly with y, the increased positive selection on the few individuals that see high values of y outweighs the effect of the reduced value of y seen by individuals with few type B neighbors. It is important to note that this distribution of type B individuals is influenced by the network structure for any contagion process, but it is only for complex contagions that it affects the spread (see "Simple Contagion" in the **SI**).

This example illustrates a general pattern. The structure of the network influences how type B individuals are dis-

tributed during the contagion. This distribution, according to the expectations in Eq. (1), interacts with the specific form of $f_{1/2}(y)$ to produce s. This selection strength then determines regions of selection and the overall behavior of the contagion. Moreover, s defines an effective diffusion process [44] (see "Basic Model" in the **SI**) capturing the behavior of $y(t)$, which we can easily solve using standard methods to obtain $P_{reach}(y)$, the fixation probability P_{fix}, properties of the temporal evolution [14], or any other quantities of interest. Thus we can reduce the problem to calculating the distribution of y in the neighborhoods of type A and type B individuals. In general, s at any point in time will depend on the full configuration of the type B individuals on the network. However, using key assumptions about the dynamics, we can often significantly reduce the degrees of freedom on which s depends. In the above example, by assuming no locality, we reduced the complexity of the process to a single degree of freedom: y. **Figure 3** shows that the resulting theory accurately predicts the results of numerical simulations of the process.

Another key feature of networks is community structure. To analyze this effect, we consider random graphs that consist of randomly connected communities of m individuals each. In particular, we assume every individual has exactly k_i random connections within the community and k_e outside of it, where $k_i + k_e \equiv k$. By tuning k_i/k, we can vary the strength of community structure. As $k_i/k \to 1$, we have very strong and cohesive communities, while $k_i/k \to \frac{m}{N}$ reduces to the case of a random regular graph of degree k. To analyze the contagion on such a graph, we must understand how type B individuals distribute themselves across the network for a given fixed value of y. We make the key assumption that for any given y, the distribution of y as seen within a community reaches a quasi-steady-state before y can change significantly across the whole graph. This approximation assumes that within-community dynamics are fast compared to global changes of y across the whole network; we expect this to hold when communities are small and well-connected compared to the overall network. By finding the steady state of a non-linear dynamical system that tracks the number of communities with a given fraction of type B individuals (see **SI**), we compute this equilibrium distribution of type B individuals, which then allows us to compute the distribution of y as seen by any given node (we show that this approximation accurately predicts the results of simulation in **Extended Data Figure 5**). From this distribution, we can find the effective selection strength acting on the contagion (**Figure 2 (d)**). The agreement between our theoretical predictions and numerical simulations are shown in **Figure 3 (d-f)**.

When communities are cohesive, type B individuals are concentrated in just a few communities. The resulting distribution of y as seen by type A individuals is broad, and the spread of the contagion is enhanced. The concentration arises because for high k_i, the many connections within a community can "conduct" influence between the types and thus cause rapid fluctuations of y within the community, but only slow fluctuations between communities (see **SI** for details). The rate of fluctuations are fastest when there are approximately equally many type B and type A individuals in a community. By contrast, fluctuations are slow when nearly all the nodes within a community have the same type. The values of y within a community (which are subject to random diffusion) will therefore spend most of their time at extreme values of $y \to 0$ or $y \to 1$. This intuition is confirmed in that we observe a critical level of community strength $\frac{k_i}{k}$ above which the equilibrium distribution of y within a community turns from a narrow distribution (concentrated around the global mean y across the whole network) to a U-shaped distribution (same mean, but concentrated at the extreme values), as shown in **Extended Data Figure 5**. The resulting variance in y as seen by individuals is high, and selection is enhanced. Intuitively, it is much easier for the contagion to randomly reach a "critical mass" of popularity within a single community and experience positive selection there, compared to across the whole network. The contagion simply fixes one community at a time, as visualized in **Extended Data Figure 6** as well as **Supplementary Videos 1-3**. These effects explain our observations on the role of clustering and community strength on real social networks in **Figure 1**. It is important to note that the unequal distribution of type B individuals among communities (just like the broader distribution of y in sparse networks) is again a feature purely of the network structure and arises with or without complex contagion. However, it is only in the former case that this distribution has an effect on the spread.

Finally, we consider graphs with variable degree distributions. In this case, it is no longer possible to calculate a selection strength s that depends only on y. Instead, we must compute the fraction of nodes of each degree k' that are type B, $y_{k'}$. This requires an explicit analysis of the fraction of type B individuals for each degree k', which leads to a high-dimensional diffusion process. We can solve this process using the no-locality approximation, in which nodes of degree k see a distribution of y identical to that for a k-regular random graph, with the global frequency y replaced by the effective frequency $z_k = \sum_{k'} P(k'|k) y_{k'}$, where $P(k'|k)$ is the neighbor degree distribution of a node of degree k (see **SI**).

To vary both the mean and variance of the degree distribution continuously, we consider graphs where the degree of each node is drawn from a Gamma distribution with mean k and variance σ_k. We compare our predicted local distribution of y to observations in **Extended Data Figure 7**. Our theoretical predictions show excellent agreement with full numerical simulations on networks, as demonstrated in **Figure 3 (g-i)**. We find two competing effects. On the one hand, high degree type B nodes are able to convert many other nodes. However, it is easier to convert low degree nodes to type B for the same reason that low k increases selection for the random regular graph. Overall, we find that broader degree distributions dampen the effects of selection (whether positive or negative) on the epidemic, both for simple and for complex contagions. Another effect is the consistent suppression of the epidemic for very low y (see **Figure 3 (h)**), which is enhanced for distributions with significant degree correlations (see **Extended Data Figure 8**). We give intuition and a derivation for this effect in the **SI**.

Whenever it is possible to compute an $s(y)$, our framework implies a simple condition under which the contagion can

spread globally with finite probability even in arbitrarily large networks (i.e. global cascades are possible, see "phase transitions" in the **SI**): the width of a region of negative $s(y)$ around $y = 0$ must scale as $N^{-\gamma}$, with $\gamma \geq 1$. That is, the contagion must need to tunnel through at most a finite number of individuals to reach a frequency above which it is positively selected. Otherwise, the process encounters negative selection and is exponentially unlikely to spread globally for large N. Using Eq. (3), this leads to the critical sparsity $k_{crit} = \frac{1}{y_n} = \frac{\alpha}{\beta}$ below which global contagion is possible (**Figure 4 (a)**). For community-based networks, we find that the effective selection strength $s(y = 0) = -\beta$, but jumps higher as $y \to \frac{m}{N}$ (see **Figure 2** (d)). Thus global contagion is possible provided that $s(\frac{m}{N}) \geq 0$. We find this implies a critical community strength k_i/k above which complex contagions are able to spread globally by appearing popular and reaching critical mass in one community at a time, even though they do not have critical mass on the global network (**Figure 4 (b)**), as seen in our initial simulations of contagions on real social networks **Figure 1 (c-e)**.

Our results demonstrate the broad importance of interactions between non-linear adoption probabilities and network structure on the dynamics and outcomes of complex contagions. Our general framework for analyzing these dynamics makes it possible to calculate how network structure interacts with arbitrary non-linear adoption probabilities to modulate the effects of both selection and stochasticity. This determines the statistical properties of both large and small contagions, as well as the probability of global cascades. These results help explain why the spread of even initially unpopular ideas and opinions can be enhanced both by overall sparsity and by cliques and other forms of community structure. They also show that in contrast to simple contagions (where the existence of highly-connected individuals always enhances spread), broad degree distributions dampen both positive and negative selection for complex contagions and hence have more subtle effects.

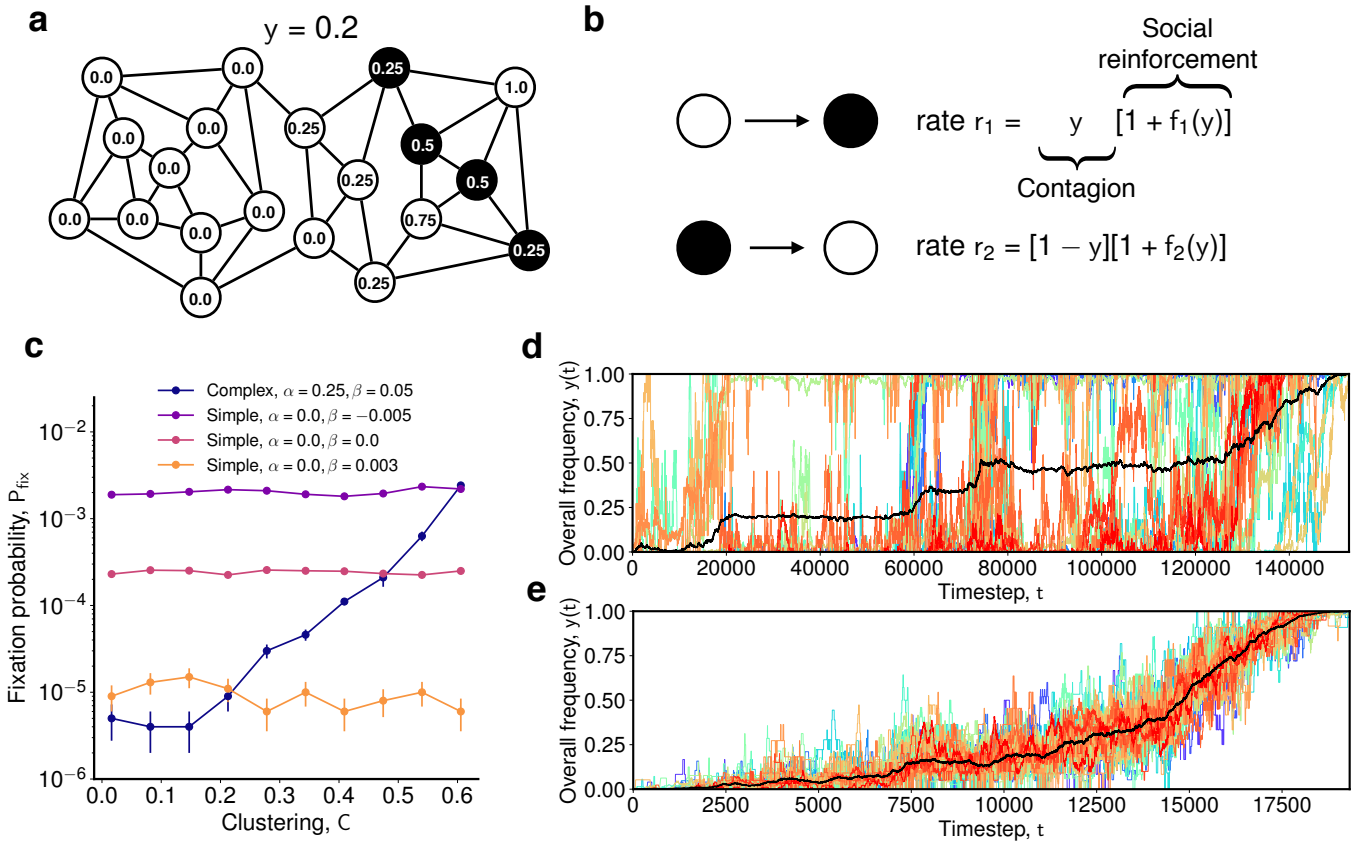


FIG. 1: Model and simulations on real social networks. a, We model a complex contagion on a network where each individual can be type B or type A. We denote the global frequency of type B individuals as y, but each node sees a local fraction of type B neighbors (node labels). **b,** Transition rates between type B and type A individuals occur at rates r_1 and r_2; the form of $f_{1/2}(y)$ determines the non-linear adoption probabilities in complex contagions. **c,** Simulations on networks of variable clustering derived by swapping pairs of edges in a Facebook network[43] ($N = 4039$, $k = 43$) show that the spread of complex (but not simple) contagions are highly sensitive to clustering. **d,e,** Example frequency trajectories for contagions that fixed in our simulations. Each colored line shows the frequency within a given community as detected by a standard community detection algorithm [49], while the black line shows overall frequency. If the community structure is strong, the contagion fixes one community at a time, rapidly gaining and maintaining local popularity which helps the spread (**d**, Clustering $C = 0.6$). If the community structure is weaker (but still detectable [49]), the contagion instead spreads uniformly across the entire network (**e**, $C = 0.2$). This is much less likely, so P_{fix} is lower in this case. Simulations assume $f_1(y) = \alpha y$, $f_2(y) = \beta$, $\alpha = 0.25$, and $\beta = 0.05$.

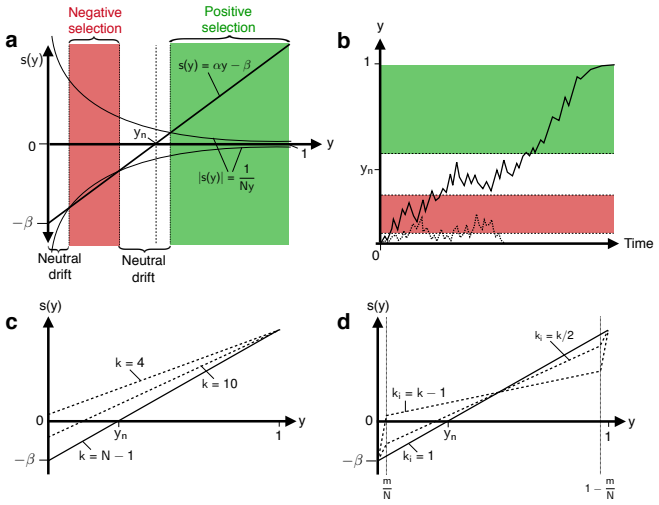


FIG. 2: Selection and genetic drift in complex contagions. **a,** The condition $N|s(y)|y = 1$ distinguishes regimes where contagion dynamics are dominated by genetic drift, negative selection, or positive selection (this is an approximation to the exact condition $N|S(y)| = 1$, see **SI**). **b,** A contagion can spread globally if it reaches high enough frequency to be positively selected; this may require "tunneling" through a regime of negative selection at lower frequencies. **c,d,** Sparsity (**c**) and community structure (**d**) can change the shape of $s(y)$ and hence alter the contagion dynamics.

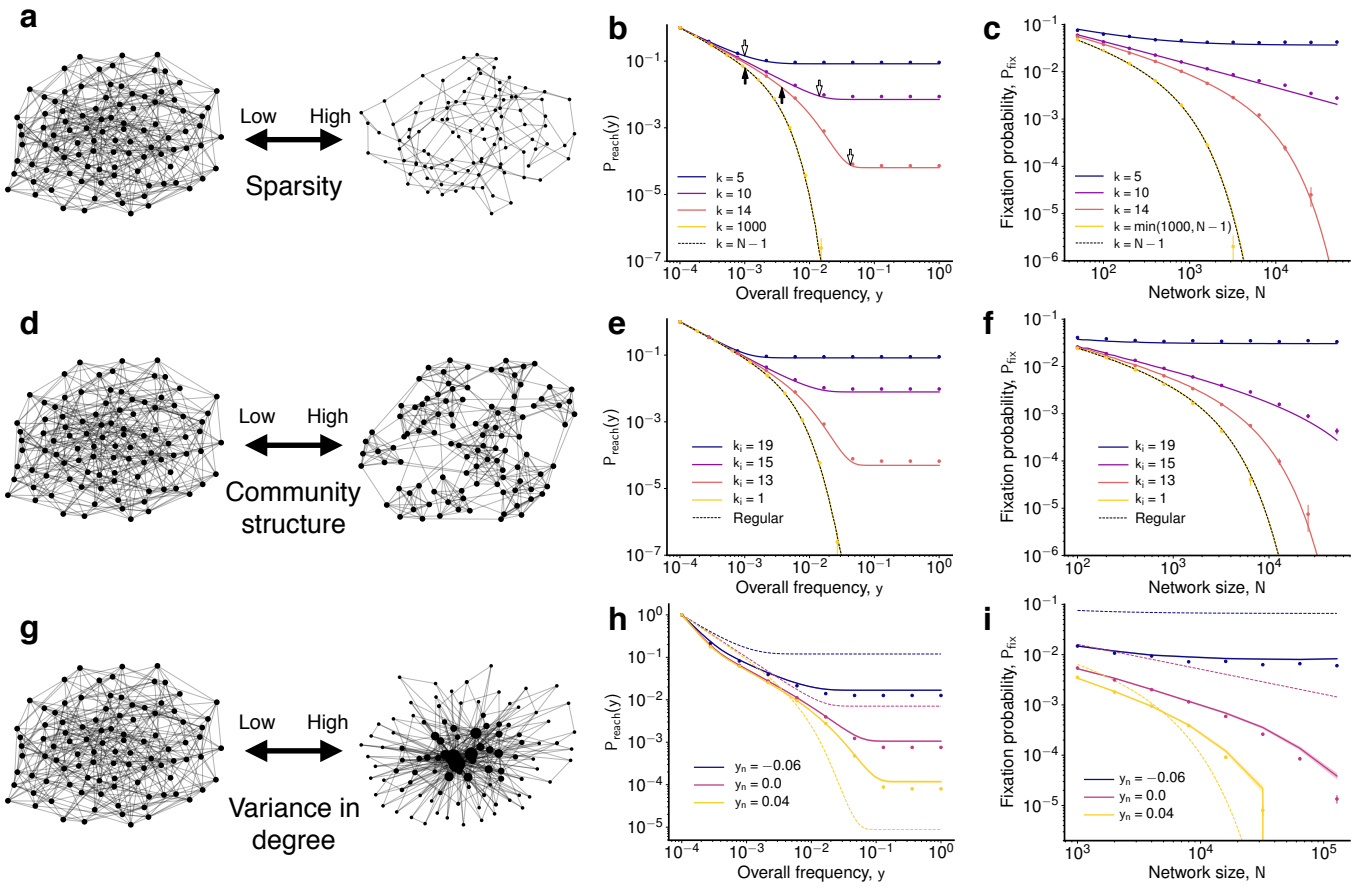


FIG. 3: Network structure and the dynamics of complex contagions. a, Illustration of networks that are more (right) or less (left) sparse. **b,** Theoretical predictions (solid lines) and simulated results (for $N = 10,000$; dots) for $P_{reach}(y)$ for networks of different sparsity. Theoretical predictions for the transition to the regime of negative and positive selection are shown as black or white arrows respectively. **c,** Theoretical predictions (solid lines) and simulated results (dots) for the fixation probability $P_{fix}(y)$ as a function of network size N. We show results for five values of k, corresponding to sparsity above (blue), approximately on (purple), and below (red) the phase transitions allowing for global spread, as well as for a large value (yellow) that approximates a fully-connected graph (dotted line). **d,** Illustration of networks with more (right) or less (left) community structure. **e,f,** Theoretical predictions (solid lines) and simulated results (dots) for $P_{reach}(y)$ **(e)** and P_{fix} **(f)** for networks with different strengths of community structure. **g,** Illustration of networks with high (right) or low (left) variance in degree distribution. **h,i,** Theoretical predictions (solid lines) and simulated results (dots) for $P_{reach}(y)$ **(h)** and P_{fix} **(i)** for networks with variance in degree distribution $\sigma_k = 30$ for contagions with $\beta = 0.1$ **(h)**, $\beta = 0.05$ **(i)**, and three different values of α corresponding to the values of y_n shown. For comparison, dotted lines show theoretical predictions for regular graphs (i.e. with $\sigma_k = 0$). Parameters: $\alpha = 1.0$, $\beta = 0.1$ **(b)**, $\alpha = 0.4$, $\beta = 0.04$ **(c)**, $\alpha = 0.88$, $\beta = 0.1$ **(e)**, $\alpha = 0.2$, $\beta = 0.025$ **(f)**.

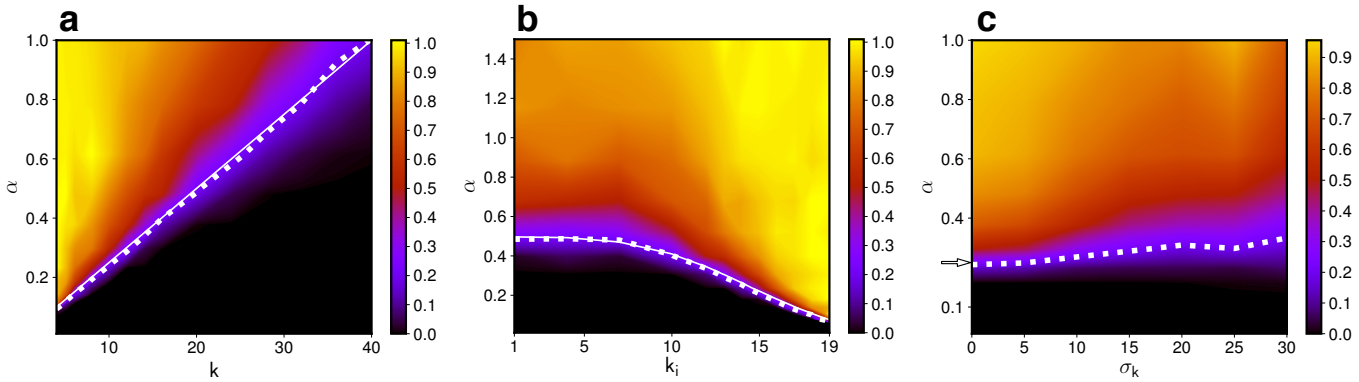


FIG. 4: Phase transitions for complex contagions. a,b,c, Ratio of P_{fix} on a network of size $N_1 = 50000$ to P_{fix} on a network of size $N_2 = 2000$ for contagions with $\beta = 0.025$, different values of α and varying sparsity **(a)**, community structure **(b)**, or degree distributions **(c)**. Values close to one correspond to cases where P_{fix} does not scale strongly with N, so global cascades are possible even in large networks. Solid white lines in **(a,b)** denote the theoretically predicted phase transition, and the thick dashed white line indicates an observed ratio of $1/5 = \sqrt{N_1/N_2}$ (the empirical location of the phase transition; see "Phase Transitions" in the **SI** for details). In **(c)**, the empirical phase transition correctly approaches the theoretical prediction (regular graph limit, white arrow) as $\sigma_k \to 0$. Since wider degree distributions weaken the effect of selection, the "transition regime" becomes noticeably wider for large σ_k.

. Methods

1. Numerical Methods

a. Analytical predictions of contagion statistics For the sparse network we can obtain an analytical solution for the function $s(y)$, while this function is the solution of a numerical procedure for the community based networks (see **SI**). In both cases, we then numerically evaluate the appropriate standard integrals from diffusion theory [44] to obtain solutions for $P_{reach}(y)$.

For the degree distribution network, we explicitly run a simulation of the multi-dimensional diffusion process (see the section "Multi dimensional diffusion" in the **SI**). For each simulation run, we sample the degree sequence from the full degree distribution. We also choose the initial degree of the first type B individual at random from the degree distribution. From then on, for every time step, we calculate y_k — the frequency of type B individuals among nodes of degree k — for all k. Given y_k, we calculate the rates of switching type A and type B individuals of a given degree. We sample the number of switching events from a binomial distribution, where the number of individuals constitutes the number of trials, and the switching rate multiplied by a small time step (chosen such that the success probability is < 0.1) constitutes the probability of success. We have verified that the results of the simulation do not change noticeably with a smaller time step.

b. Network simulations In order to simulate the process on real networks, we generate random graphs according to the structural features in question. We then perform a large number of simulations to obtain statistics on $P_{reach}(y)$ and P_{fix}. In particular, each simulation begins with a network of all type A individuals, with a single randomly placed type B individual. We then update the types of all nodes according to a Gillespie algorithm, where the rates are given by the rates r_1 and r_2 from the main text. The algorithm is terminated when the absorbing states of $y = 0$ (extinction) or $y = 1$ (fixation) are reached.

c. Local distributions of y In Extended Data Figures 2-5, we compare the predicted distributions of y as seen by individual nodes to those observed in simulations, for a given global value of y. Since the global value of y varies over time in simulations, we run simulations as usual starting with $y = \frac{1}{N}$, and use data from all nodes during all time steps where the global value of y is equal to the desired value. Data is collected for 20 separate simulation runs.

2. Generating random networks

a. Real social network with variable clustering For **Figure 1 (c-e)**, we construct a sequence of networks based on the Facebook network ($N = 4039$, $k = 43$) from the Stanford Large Network Dataset collection [43]. By considering random $A-B, C-D \to A-C, B-D$ swaps, we reduce clustering until a desired value is reached, while keeping the degree sequence intact. To test the impact of this reduction in clustering on the community structure of the network, we measure "community overlap" between the original and modified networks, by perform a bipartite matching of communities found by a standard community detection algorithm [49]. We find that the communities still overlap to 80% when clustering is reduced from the original value of 0.6 to 0.2. Community overlap finally drops to below 0.2 as clustering is reduced to that of a random network.

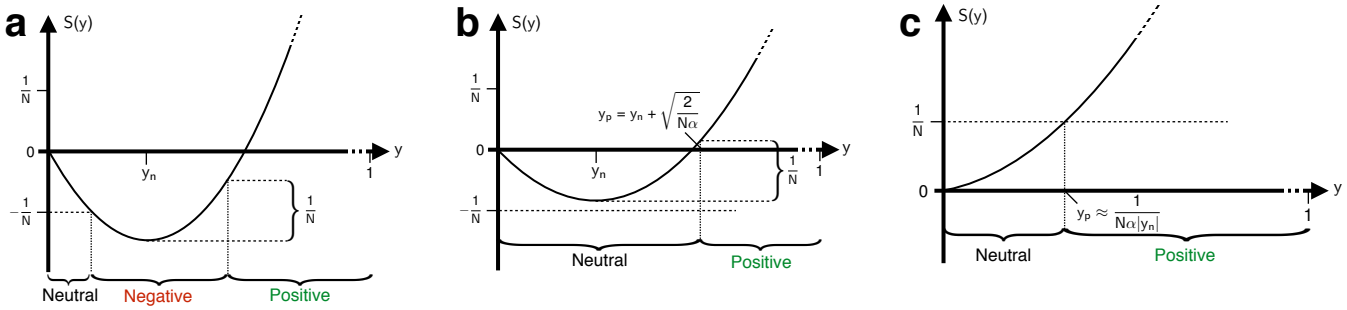
b. Community based network For the community based network model, random graphs are generated such that in the resulting graph every node and community in the network is statistically equivalent. Every community has exactly m individuals, and every individual has exactly k_i random connections to nodes within its community, and k_e random connections to nodes outside the community, where $k_i + k_e \equiv k$. In other words, the network structure is a regular random graph of communities, where each community is itself a regular random graph. First each community is sampled as a regular random graph. Then the connections between communities are sampled on a community basis by sampling the supergraph of community connectivity as a regular random graph (with duplicate edges allowed) with mk_e edges per community. Then the edges incoming to each community are evenly distributed among its member individuals. The resulting inter-community connectivity pattern is then rewired randomly without changing the degree sequence (by considering random $A-B, C-D \to A-C, B-D$ swaps) until all edges are valid (i.e. no duplicate edges).

c. Variable degree distributions Random graphs are generated by sampling a degree sequence from the specified degree distribution. In our simulations, we choose a Gamma distribution with a given mean k and variance σ_k^2. These degrees are then matched up with the stub connect algorithm. The resulting connectivity pattern is then rewired randomly without changing the degree sequence (by considering random $A - B, C - D \to A - C, B - D$ swaps) until all edges are valid (i.e. no duplicate edges).

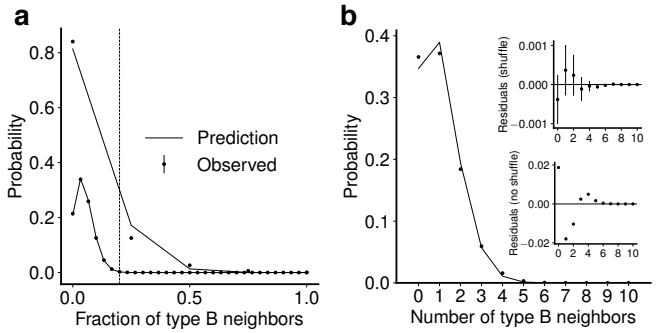
d. Lattice networks The lattice networks in **Extended Data Figures 3** and **4** are generated as linear (1D) and square (2D) lattices with periodic boundary conditions. On the 1D lattice, we connect each node to its k nearest neighbors. On the 2D lattice, we connect each node to its closest neighbors ($k = 4$), or all its second neighbors ($k = 24$).

3. Code and Data availability

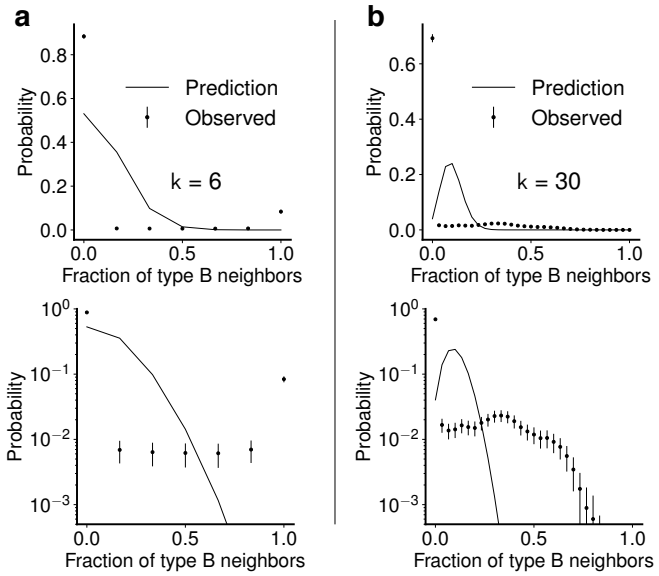
All simulations and numerical calculations were performed with Julia 1.1. Our code is open source and available at `www.github.com/jnkh/epidemics`. The network data used is publicly available [43].



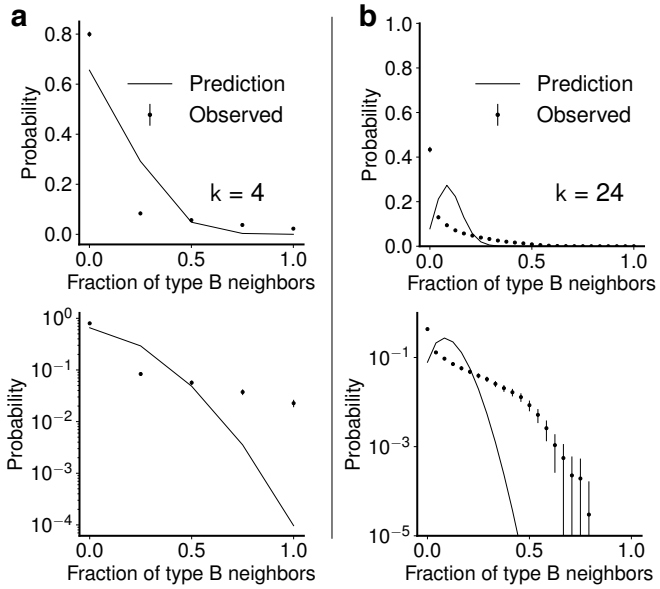
Extended Data Figure 1: Scaling regimes for positively frequency dependent complex contagions. The quantity $NS(y)$ determines the various selection regimes. A linearly increasing selection strength of the form $s(y) = \alpha y - \beta$ leads to a quadratic $S(y) = \frac{\alpha}{2}y^2 - \beta y$ which we can then compare to the relevant scale $\frac{1}{N}$. There are three possibilities: **a,** $S(y)$ dips below the $-\frac{1}{N}$ line and thus enters a regime of negative selection. Once the difference from the lowest point of $S(y)$ to the current position becomes $\frac{1}{N}$, the selection becomes positive (see **SI** for details). **b,c,** $S(y)$ never dips below the $-\frac{1}{N}$ line, so the process experiences neutral drift until $S(y)$ grows by $\frac{1}{N}$ from its minimum value. This happens at $y = y_p$, at which point positive selection takes over. It follows that $P_{fix} = \frac{1}{Ny_p}$. If $S(y)$ dips below 0 initially (**b**), the fixation probability scales like N^{-1}. Otherwise (**c**), the fixation probability does not scale with N, and global cascades are possible for arbitrarily large networks.



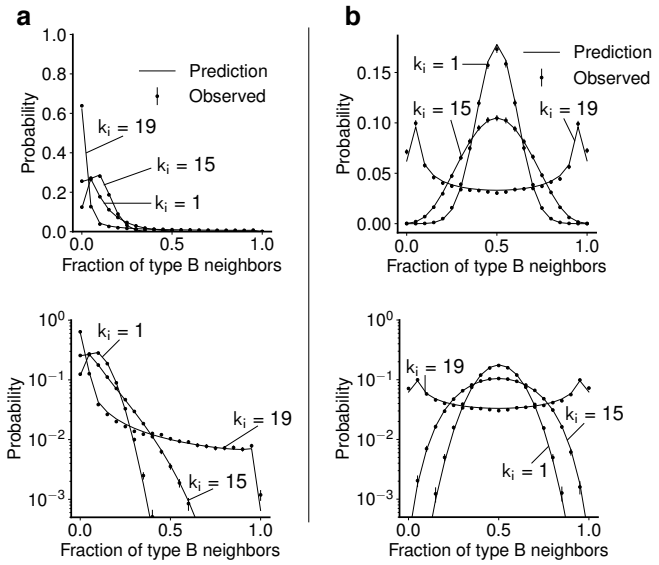
Extended Data Figure 2: Local distribution of y for regular random graphs. The figure shows the expected distribution according to the assumption of "no locality" (see main text) as the black line, and compares with the observed distribution from simulations (dots) on random regular networks of size $N = 1000$. The error bars denote standard error. **a,** Comparison of the distributions for $k = 4$ and $k = 30$, for a situation where $y_n = 0.2$, and the global value of $y = 0.05$. Since $y < y_n$, almost no individuals experience positive selection when k is large and observed values of y are tightly concentrated around the global value. However, for small k, a significant number of nodes do experience positive selection "just by chance". **b,** Predicted and observed values on a network with $k = 10$ and $y = 0.1$. The lower inset shows the residual mismatch between the prediction and the observed values, while the upper inset shows the residuals in simulations where the location of type B individuals is shuffled at every time step (making the no locality assumption exactly true by definition). Shuffling causes the mismatch to disappear.



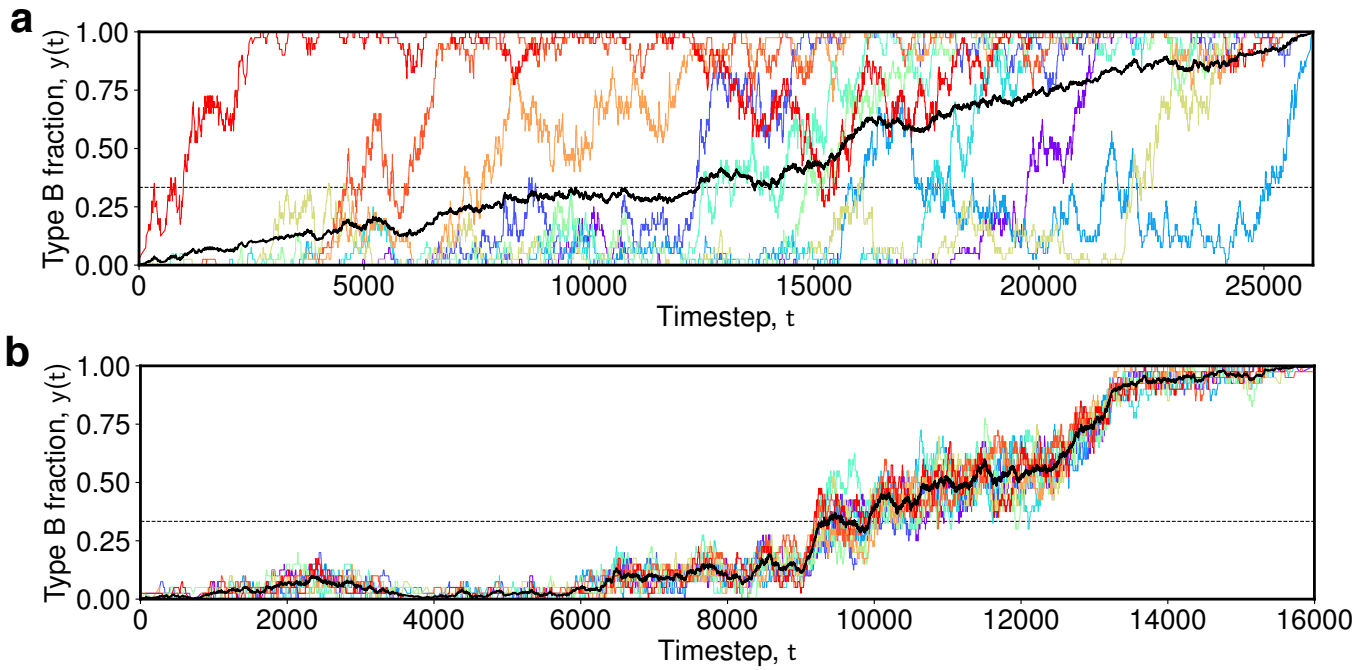
Extended Data Figure 3: Local distribution of y for 1D lattice networks. The figure shows the expected distribution according to the "no locality" assumption for random regular graphs (black line, see main text), and the observed distribution (dots) for a contagion on a 1D lattice network with $y = 0.1$, $N = 1000$ and $k = 6$ (**a**) as well as $k = 30$ (**b**). The error bars denote standard error. The bottom plots show the same data as the top, but on a logarithmic scale. It is clear that the "locality" on the lattice causes more extreme y values that depart significantly from the "no locality" prediction. These patterns arise purely due to the network structure, even for simple contagions (here $\alpha = 0$ and $\beta = 0$).



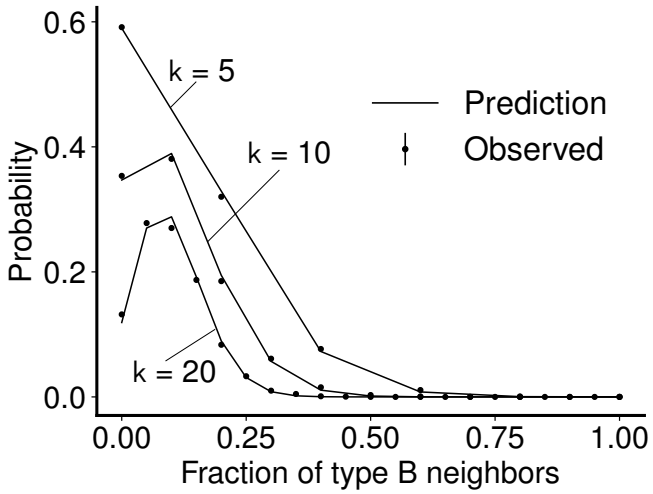
Extended Data Figure 4: Local distribution of y for 2D lattice networks. The figure shows the expected distribution according to the "no locality" assumption for random regular graphs (black line, see main text), and the observed distribution (dots) for a contagion on a 2D lattice network with $y = 0.1$, $N = 1600$ and $k = 4$ (**a**) as well as $k = 24$ (**b**). The error bars denote standard error. The bottom plots show the same data as the top, but on a logarithmic scale. It is clear that the "locality" on the lattice causes more extreme y values that depart significantly from the "no locality" prediction. These patterns arise purely due to the network structure, even for simple contagions (here $\alpha = 0$ and $\beta = 0$).



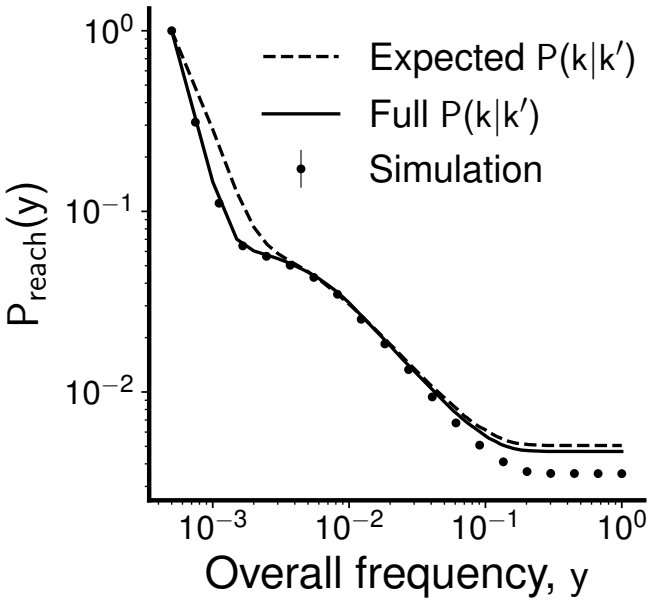
Extended Data Figure 5: Local distribution of y for community based networks. The figure shows the expected distribution according to the equilibrium assumption (black line, see main text), and the observed distribution (dots) for a contagion on a network with $N = 1000$ and $k = 20$. We show distributions for $y = 0.1$ (**a**) and $y = 0.5$ (**b**). The error bars denote standard error. The bottom plots show the same data as the top, but on a logarithmic scale. The theoretical prediction matches well. When clustering and community strength reaches a critical value, the distribution shifts from a tightly concentrated one to a broad distribution with significant probability mass at the extremes. These patterns arise purely due to the network structure, even for simple contagions (here $\alpha = 0$ and $\beta = 0$).



Extended Data Figure 6: Contagion over time for community based networks. The figure shows the temporal evolution of a contagion on a community based network. In the case of high community strength (**a**), as in the high clustering case of the real network in **Figure 1 d**, each community fixes one at a time. In the opposite case (**b**), the y values in each community are tightly coupled to the global value, as in **Figure 1 e**. Parameters: $N = 400$, $y_n = 0.2$ (dashed horizontal line), size of communities $m = 40$, $k_i = 39$ (**a**) and $k_i = 1$ (**b**).



Extended Data Figure 7: Local distribution of y for networks with degree distributions. The figure shows the expected distribution of y as seen by nodes of degree 5, 10 and 20 according to the "no locality" assumption (black line, see main text), as well as the observed distribution (dots), for a contagion on a network with $N = 1000$, mean degree $\bar{k} = 20$, and degree standard deviation $\sigma_k = 10$. Error bars denote standard error. The theoretical prediction matches well.



Extended Data Figure 8: Effect of the neighbor degree distribution. The figure shows $P_{reach}(y)$ as computed according to the "no locality" assumption for a contagion on a network with $N = 10000$, mean degree $\bar{k} = 10$, degree standard deviation $\sigma_k = 30$, and strong positive degree correlations (see **SI** for details). The lines show the prediction assuming the "expected" neighbor degree distribution $P(k|k') = \frac{kP(k)}{\bar{k}}$ (dashed), as well as the actual neighbor degree distribution $P(k|k')$ taking degree correlations into account (solid). The error bars denote standard error. Once the full neighbor degree distribution is taken into account, the results agree well with simulations (dots). We have verified that the residual mismatch at high y disappears when type B nodes are shuffled at every time step (keeping their degree constant), showing the mismatch is due to the "no locality" assumption not being exactly true. Parameters: $\alpha = 1.0$, $\beta = 0.1$.

[1] C. Zhou, Q. Zhao, and W. Lu, Impact of repeated exposures on information spreading in social networks, PloS one **10**, e0140556 (2015).

[2] A. Pentland, *Social physics: How good ideas spread-the lessons from a new science* (Penguin, 2014).

[3] N. A. Christakis and J. H. Fowler, Social contagion theory: examining dynamic social networks and human behavior, Statistics in medicine **32**, 556 (2013).

[4] M. J. Keeling, The effects of local spatial structure on epidemiological invasions, Proceedings of the Royal Society of London B: Biological Sciences **266**, 859 (1999).

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

[6] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn, Optimal network modularity for information diffusion, Physical review letters **113**, 088701 (2014).

[7] A. Montanari and A. Saberi, The spread of innovations in social networks, Proceedings of the National Academy of Sciences **107**, 20196 (2010).

[8] M. Granovetter, Threshold models of collective behavior, American journal of sociology **83**, 1420 (1978).

[9] D. J. Watts, A simple model of global cascades on random networks, Proceedings of the National Academy of Sciences **99**, 5766 (2002).

[10] J. Borge-Holthoefer, R. A. Baños, S. González-Bailón, and Y. Moreno, Cascading behaviour in complex socio-technical networks, Journal of Complex Networks **1**, 3 (2013).

[11] D. López-Pintado, Diffusion in complex social networks, Games and Economic Behavior **62**, 573 (2008).

[12] C. T. Bauch and A. P. Galvani, Social factors in epidemiology, Science **342**, 47 (2013).

[13] D. Centola and M. Macy, Complex contagions and the weakness of long ties, American journal of Sociology **113**, 702 (2007).

[14] D. Eckles, E. Mossel, M. A. Rahimian, and S. Sen, Long ties accelerate noisy threshold-based contagions, Available at SSRN 3262749 **0** (2018).

[15] F. L. Pinheiro, V. V. Vasconcelos, and S. A. Levin, Consensus and polarization in competing complex contagion processes, arXiv preprint arXiv:1811.08525 **0** (2018).

[16] L. Hébert-Dufresne and B. M. Althouse, Complex dynamics of synergistic coinfections on realistically clustered networks, Proceedings of the National Academy of Sciences **112**, 10551 (2015).

[17] D. Guilbeault, J. Becker, and D. Centola, Complex contagions: A decade in review, Complex spreading phenomena in social systems , 3 (2018).

[18] D. Centola, *How behavior spreads* (Princeton University Press, 2018).

[19] 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).

[20] D. M. 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* (2011) pp. 695–704.

[21] B. State and L. Adamic, The diffusion of support in an online social movement: Evidence from the adoption of equal-sign profile pictures, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015) pp. 1741–1750.

[22] D. A. Sprague and T. House, Evidence for complex contagion models of social contagion from observational data, PloS one **12**, e0180802 (2017).

[23] Z. C. Steinert-Threlkeld, Spontaneous collective action: Peripheral mobilization during the arab spring, American Political Science Review **111**, 379 (2017).

[24] D. Centola, The social origins of networks and diffusion, American Journal of Sociology **120**, 1295 (2015).

[25] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, Structural diversity in social contagion, Proceedings of the National Academy of Sciences **109**, 5962 (2012).

[26] D. Guilbeault and D. Centola, Topological measures for identifying and predicting the spread of complex contagions, Nature communications **12**, 1 (2021).

[27] E. Bakshy, S. Messing, and L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science **348**, 1130 (2015).

[28] J. L. Toole, M. Cha, and M. C. González, Modeling the adoption of innovations in the presence of geographic and media influences, PloS one **7**, e29528 (2012).

[29] V. A. Traag, Complex contagion of campaign donations, PloS one **11**, e0153539 (2016).

[30] P. S. Dodds and D. J. Watts, Universal behavior in a generalized model of contagion, Physical review letters **92**, 218701 (2004).

[31] P. S. Dodds and D. J. Watts, A generalized model of social and biological contagion, Journal of theoretical biology **232**, 587 (2005).

[32] P. L. Krapivsky, S. Redner, and D. Volovik, Reinforcement-driven spread of innovations and fads, Journal of Statistical Mechanics: Theory and Experiment **2011**, P12003 (2011).

[33] P. S. Dodds, K. D. Harris, and J. L. Payne, Direct, physically motivated derivation of the contagion condition for spreading processes on generalized random networks, Physical Review E **83**, 056122 (2011).

[34] S. Lehmann and Y.-Y. Ahn, *Complex spreading phenomena in social systems* (Springer, 2018).

[35] P.-M. Hui, L. Weng, A. S. Shirazi, Y.-Y. Ahn, and F. Menczer, Scalable detection of viral memes from diffusion patterns, in *Complex Spreading Phenomena in Social Systems* (Springer, 2018) pp. 197–211.

[36] J. P. Gleeson, Cascades on correlated and modular random networks, Physical Review E **77**, 046117 (2008).

[37] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks, Physical review letters **86**, 3200 (2001).

[38] T. House and M. J. Keeling, Insights from unifying modern approximations to infections on networks, Journal of The Royal Society Interface **8**, 67 (2011).

[39] G. E. Leventhal, A. L. Hill, M. A. Nowak, and S. Bonhoeffer, Evolution and emergence of infectious diseases in theoretical and real-world networks, Nature communications **6**, 6101 (2015).

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

[41] M. Pagel, M. Beaumont, A. Meade, A. Verkerk, and A. Calude, Dominant words rise to the top by positive frequency-dependent selection, Proceedings of the National Academy of Sciences **0**, 201816994 (2019).

[42] L. J. Allen, An introduction to stochastic epidemic models, in *Mathematical epidemiology* (Springer, 2008) pp. 81–130.

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

[44] W. J. Ewens, *Mathematical Population Genetics 1: Theoretical Introduction*, Vol. 27 (Springer Science & Business Media, 2012).

[45] N. C. Wormald *et al.*, Models of random regular graphs, London Mathematical Society Lecture Note Series **0**, 239 (1999).

[46] B. Derrida and Y. Pomeau, Random networks of automata: a simple annealed approximation, EPL (Europhysics Letters) **1**, 45 (1986).

[47] A. Galstyan and P. Cohen, Cascading dynamics in modular networks, Physical Review E **75**, 036109 (2007).

[48] H. Tanaka, H. A. Stone, and D. R. Nelson, Spatial gene drives and pushed genetic waves, Proceedings of the National Academy of Sciences **114**, 8452 (2017).

[49] U. N. Raghavan, R. Albert, and S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Physical review E **76**, 036106 (2007).