SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPLEX NETWORKS
APPLIED MATHEMATICS
EPIDEMIOLOGY
PHASE TRANSITIONS AND CRITICAL PHENOMENA

Received
1 May 2013

Accepted
18 June 2013

Published
10 July 2013

Correspondence and requests for materials should be addressed to L.J.D. (Louis.Dube@phy.ulaval.ca)

Global efficiency of local immunization on complex networks

Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young & Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec (Québec), Canada G1V 0A6.

Epidemics occur in all shapes and forms: infections propagating in our sparse sexual networks, rumours and diseases spreading through our much denser social interactions, or viruses circulating on the Internet. With the advent of large databases and efficient analysis algorithms, these processes can be better predicted and controlled. In this study, we use different characteristics of network organization to identify the influential spreaders in 17 empirical networks of diverse nature using 2 epidemic models. We find that a judicious choice of local measures, based either on the network's connectivity at a microscopic scale or on its community structure at a mesoscopic scale, compares favorably to global measures, such as betweenness centrality, in terms of efficiency, practicality and robustness. We also develop an analytical framework that highlights a transition in the characteristic scale of different epidemic regimes. This allows to decide which local measure should govern immunization in a given scenario.

Epidemics never occur randomly. Instead, they follow the structured pathways formed by the interactions and connections of the host population[1,2]. The spreading processes relevant to our everyday life take place on networks of all sorts: social (e.g. epidemics[3,4]), technological (e.g. computer viruses[5,6]) or ecological (cascading extinctions in food webs[7]). With a network representation, these completely different processes can be modelled as the propagation of a given agent on a set of nodes (the population) and links (the interactions). Different systems imply networks with different organizations, just as different agents require different epidemic models.

There has long been significant interest in identifying the *influential spreaders* in networks. Which nodes should be the target of immunization efforts in order to optimally protect the network against epidemics? Unfortunately, most studies feature two significant shortcomings. Firstly, the proposed methods are often based on optimization or heuristic algorithms requiring nearly perfect information on a static system[8,9]; this is rarely the case. Secondly, methods are usually tested on small numbers of real systems using a particular epidemic scenario[10,11]; this limits the scope of possible outcomes.

We first present a numerical study, perhaps the largest of its kind to date, where we argue that, depending on the nature of the network and of the disease, different immunization tactics have to be taken into consideration. In so doing, we formalize the notion of node influence and illustrate how *local knowledge* around a particular node is usually sufficient to estimate its role in an epidemic. We also show how, in certain cases, the influence of a node is not necessarily dictated by its number of connections, but rather by its role in the network's community structure (see Fig. 1). Far from trivial, it follows that an efficient immunization strategy can be obtained solely from local measures, which are easily estimated in practice and robust to noisy or incomplete information. We further develop an analytical formalism ideally suited to test the effects of local immunization on realistic network structures. Combining the insights gathered from the numerical study and this formalism, we finally formulate a readily applicable approach which can easily be implemented in practice.

Results

Models and measures. There exist two standard models emulating diverse types of epidemics: the *susceptible-infectious-recovered* (SIR) and *susceptible-infectious-susceptible* (SIS) dynamics. In both, an infectious node has a given probability of eventually infecting each of its susceptible neighbors during its infectious period, which is terminated by either death/immunity leading to the recovered state (SIR) or by returning to a susceptible state (SIS). In the SIR dynamics, for a given transmission probability T, the quantity of interest is the mean fraction R_f of recovered nodes once a disease, not subject to a stochastic extinction, has finished spreading (i.e. we focus on the giant component[12]). Since each edge can only be followed once, this dynamics investigates how a population is vulnerable to the *invasion* of a new pathogen. In the SIS dynamics, we are interested in the prevalence I^* (fraction

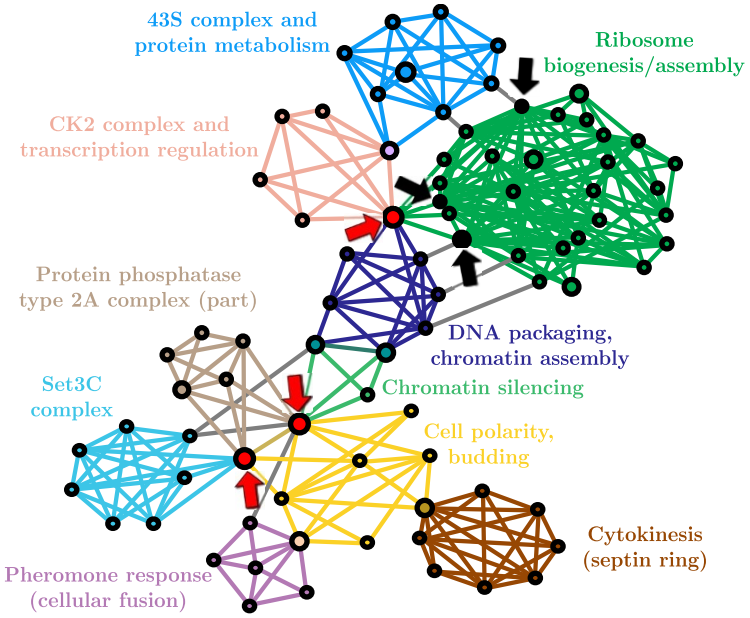


Figure 1 | Protein interactions of *S. cerevisiae* (subset)[22]. The three black nodes correspond to the ones with the highest degree, and the three red ones have the highest membership number. In this particular example, it is readily seen that the latter are structurally more influent.

of infectious nodes) of the disease at equilibrium (equal amounts of infections and recoveries) as a function of the ratio $\lambda = \alpha/\beta$ of infection rate α and recovery rate β. This particular dynamics permits the study of how a given network structure can *sustain* an already established epidemic.

Should a fraction ε of the population be fully immunized, our objective is to identify the nodes whose absence would minimize R_f and I^*. The *epidemic influence* of a node — that is the effect of its removal on R_f and I^* — depends mainly on its role in the organization of the network. Hence to efficiently immunize a population, we must first understand its underlying structure.

Network organization can be characterized on different scales, each of which affect the dynamics of propagation. At the microscopic level, the most significant feature is the *degree* of a node (its number of links, noted k) which in turn defines the degree distribution of the network. The significance of the high-degree nodes (the *hubs*) for network structure in general[13], for network robustness to random failure[14] and for epidemic control[15] has long been recognized.

At the macroscopic level, the role of a node can be described by its *centrality*, which may be defined in various ways. Frequently used in the social sciences is the *betweenness centrality* (b), quantifying the contributions of a given node to the shortest paths between every pair of nodes in the network[16]. Arguably, this method should be among the best estimate of a node's epidemic influence as it directly measures its role in the different pathways between all other individuals[17], yet at a considerable computational cost. A simpler method, the k-core (or k-shell) decomposition[18,19], assigns nodes to different layers (or *coreness* c) effectively defining the core and periphery of a network (high and low c respectively). It has recently been shown that coreness is well suited to identify nodes that are the most at risk of being infected during the course of an epidemic[20]. In light of our results, we will be able to discuss the distinction between a node's vulnerability to infection and its influence on the outcome of an epidemic.

The mesoscopic scale has recently been the subject of considerable attention. At this level of organization, the focus is on the redundancy of connections forming dense clusters referred to as the community structure of the network[21,22]. Nodes can be distinguished by their *membership* number m, i.e., the number of communities to which they belong. We will consider that two links of a given node are part of one community if the neighbours they reach lead to significantly overlapping neighbourhoods[21]. This definition is directly relevant to epidemic dynamics as links within communities do *not* lead to new potential infections. We call *structural hubs* the nodes connecting the largest number of different communities. These nodes act as bridges facilitating the propagation of the disease from one dense cluster to another. Targeting structural hubs to hinder propagation in structured populations has been previously proposed and investigated[10,11], but has yet to be tested extensively.

Note that the microscopic and mesoscopic levels (as defined above) are characterized by *local measures* in the sense that they do not require a complete knowledge of the network, in contrast to *global measures* like the betweenness centrality. Moreover, as we will see, *local measures are less sensitive to incomplete or incorrect information*. Adding, removing or rewiring a link only affects the degree or membership of nodes directly in the neighbourhood of the modification; whereas the same alterations can potentially affect the centrality of nodes anywhere in the network through cascading effects. Furthermore, even if community detection often requires the tuning of a global resolution parameter, we will see that this additional step does not affect the identification of structural hubs, meaning that local information is sufficient to accurately determine a node's memberships.

In our numerical simulations we will have a perfect knowledge of static networks. This will allow us to use global measures as a reference to test the efficiency of local measures best suited in practice. We therefore ask without discrimination: which of the degree k, the coreness c, the betweenness centrality b or the membership number m is the best identifier of the most influential nodes on the outcome of an epidemic? To answer this question, we have simulated SIR and SIS dynamics with Monte Carlo calculations on 17 real-world networks. In each case, a fraction ε of the nodes was removed in decreasing order of the nodes' score for each of the four different measures. By comparing their efficiency to reduce R_f or I^* as a function of ε, we are able to establish which measure is best suited for a given scenario characterized by a network structure, a propagation dynamics and a disease transmissibility (i.e. probability of transmission).

Case study: a data exchange network. We first illustrate our methods using the network of users of the Pretty-Good-Privacy

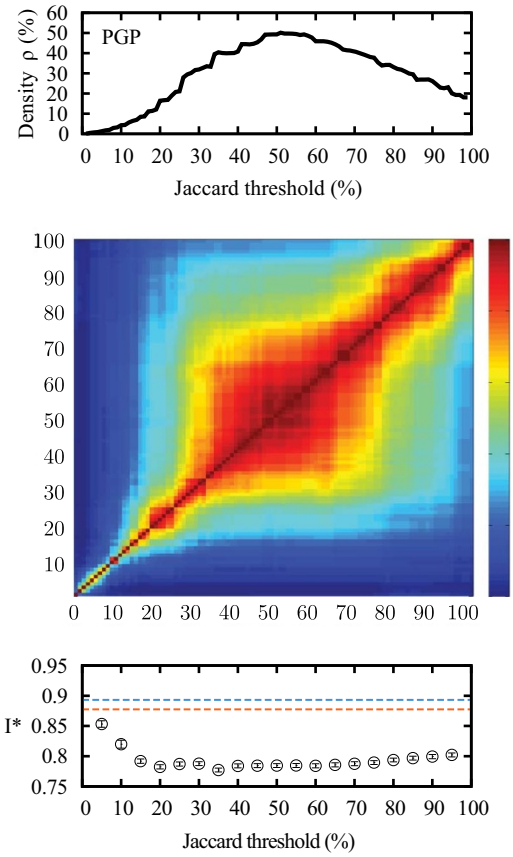


Figure 2 | Robustness of structural hubs in the PGP network. (top) Community density (ρ) obtained through different Jaccard thresholds. (middle) Robustness of the structural hubs identification methods. Element (*i,j*) gives the overlap (normalized) between the structural hubs (top 1%) selected with thresholds *i* and *j*. The highest line and last column of the matrix correspond to the case where the membership number equals the degree. (bottom) Prevalence I^* of SIS epidemics with $\lambda = 5$ when the top 1% of structural hubs are removed (compared with the results without removal in blue or with random targets in orange).

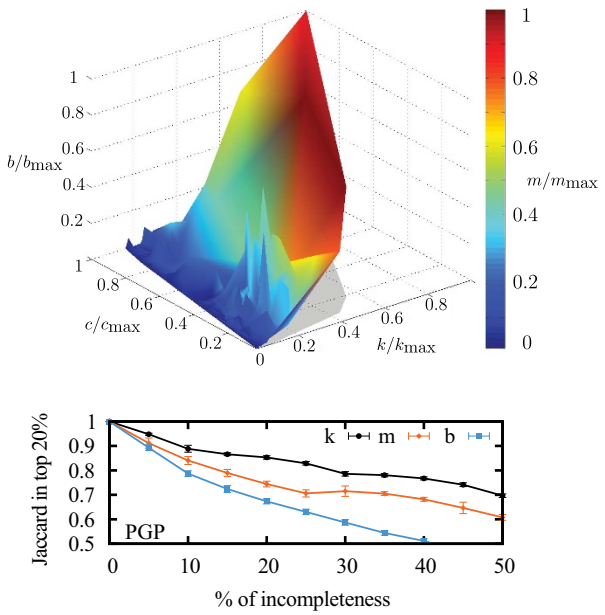


Figure 3 | Difference in immunization targets for the PGP network. (top) We present correlations between the degree (*k*, right axis), the coreness (*c*, left axis), the betweenness centrality (*b*, vertical axis) and the membership number (*m*, color) for each nodes. Each measure is normalized according to the highest value found in the network. Each node is represented in this 4-dimensional space and a simple triangulation procedure then yields a more intelligible appearance. Structural hubs (dark red) can be found even at relatively small degree ($\sim k_{max}/2$), coreness ($\sim c_{max}/5$) and centrality ($\sim b_{max}/3$). (bottom) Jaccard coefficient between the ensemble of nodes identified as part of the top 20% according to a given measure (*k*, *m* or *b*) on two versions of the network: the original complete network and a network ensemble where a certain percentage of links has been randomly removed (horizontal axis). The shorter the range of a measure, the more robust it is to incomplete information.

algorithm for secure information interchange (hereafter, the PGP network)[23], which could be the host of the propagation of computer viruses, rumors or viral marketing campaigns. Results for the 16 other networks are presented and discussed in the next section as well as in the Supporting Information (SI) document.

Communities in the network are extracted with the link community algorithm of Ahn *et al.*[21]. This algorithm groups links — and therefore the nodes they join — into communities based on the overlap of their respective neighbouring nodes. It is this overlap that reduces the number of new potential infections in a community structure, as opposed to a random network. This method thus reflects our understanding of how communities affect disease propagation. While it may not directly detect the social groups or functional modules of a network, it identifies significant clusters of redundant links. This redundancy or overlap is quantified through a Jaccard coefficient, and two links are grouped into the same community when their coefficient exceeds a certain threshold. The threshold value acts as a resolution, enabling to look at different levels of organization. As suggested[21], the value of the threshold is chosen to maximize the average density ρ of the communities (see Methods). As this choice may seem arbitrary, Fig. 2 investigates the similarity between the nodes with the highest membership numbers, for different thresholds. It suggests that the membership number is fairly robust around the threshold. Moreover, Fig. 2 also demonstrates that the effect of the removal of the structural hubs on a SIS epidemics is very robust to the choice of the threshold. Thus, we will henceforth use the membership numbers obtained with the threshold value corresponding to the highest community density.

The differences, if any, between the efficiency of the different methods are due to the immunized nodes not being the same. Figure 3 (top) investigates the correlations between the different properties (*k*, *b*, *c* and *m*) of each node. Perhaps the most important result here is that nodes with a high membership number may have relatively small degree, coreness and betweenness centrality. Hence, we expect the immunizing method based on community structure to have a different influence on the outcome of epidemics. Figure 3 (bottom) shows the consistensy (or lack thereof) of a given measure, depending on the quality of the available data. The robustness of local (micro and meso) measures is of obvious practical advantage. Both robustness and correlations are further investigated in the SI.

To study various epidemic scenarios, we consider both SIS and SIR dynamics (which may behave quite differently) with different values of the transmission probability (λ and T for SIS and SIR, respectively). In fact, each network features an *epidemic threshold*, i.e. critical values λ_c[24] and T_c[25], below which I^* and R_f vanish to zero in an equivalent infinite network ensemble. As we will show, the observed behavior can differ significantly depending whether or not λ and T are close to their critical value.

Figure 4 presents results of different immunization methods against SIS dynamics for different values of λ. On the top figure, where λ is near λ_c, the most successful method of intervention is to target nodes according to their degree. At low transmissibility, the disease follows only a very small fraction of all links. The shortest

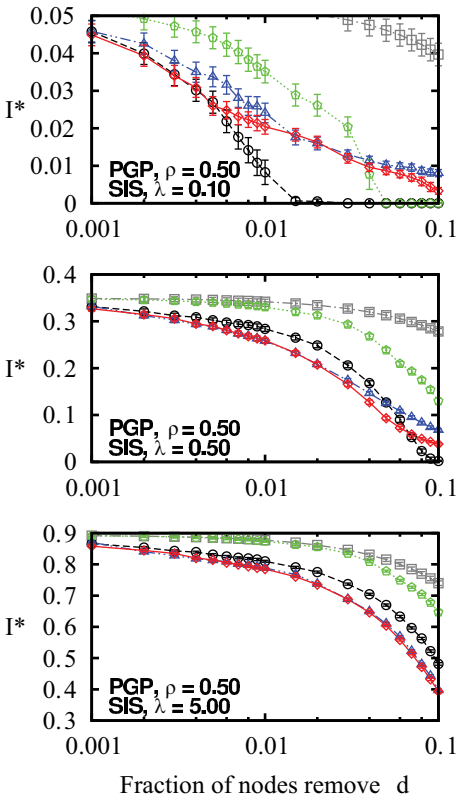


Figure 4 | Efficiency of the immunization methods against an SIS epidemics on the PGP network. Nodes are removed in decreasing order of their score according to each method: coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) and memberships (red diamonds) and the effect of removal is then quantified in terms of the decrease of the prevalence I^*. The prevalence of the epidemics when the removed nodes are chosen at random (grey squares) has been added for comparison. Figures are presented in increasing order of transmissibility (λ) from top to bottom.

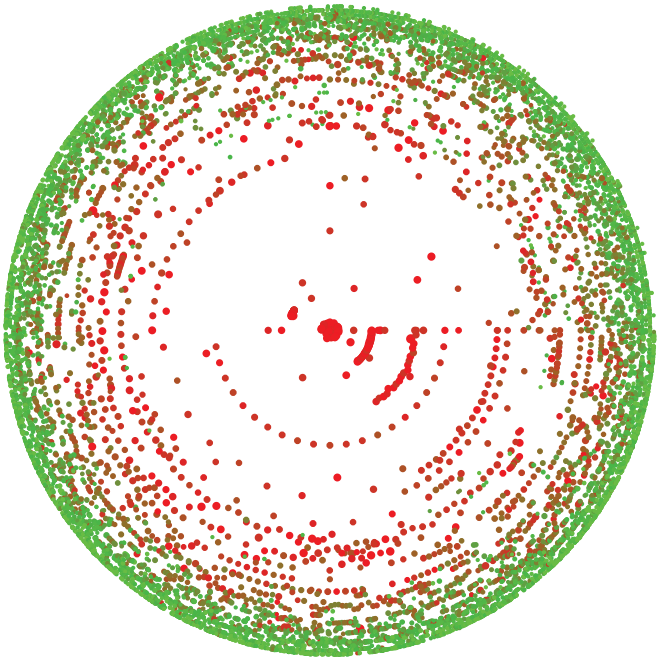


Figure 5 | k-core decomposition of the PGP network. Representation (based on[26]) of the k-shells in the PGP network with nodes colored according to their total infectious period during a given time interval. Red nodes are more likely to be infectious at any given time than green nodes as the color is given by the square of the fraction of time spent in infectious state. Note how the central nodes (the core) of the network are most at risk.

paths are seldom used and the poor performance of betweenness centrality follows. Moreover, the disease will not be affected by the community structure, because even in dense neighbourhoods, most links will not be travelled. We then say that the disease, unaffected by link clustering, follows a tree-like structure (without loops), where community memberships are insignificant. It is therefore better to simply remove as many links as possible.

As λ increases beyond λ_c, we see that immunization based on membership numbers quickly outperforms the other methods. As more links are travelled, the disease is more likely to follow superfluous links in already infected communities. Hubs sharing their many links within few communities are therefore not as efficient in causing secondary infections as one might expect. Similarly, targeting through betweenness centrality also performs better with higher λ, albeit not as well as membership-targeting in this case. For $\lambda \gg \lambda_c$, immunization based on membership numbers (local) and on betweenness centrality (global) converge toward similar efficiency, significantly outperforming degree-based immunization.

Another interesting feature of our results is the poor performance of immunization based on node coreness. A previous study had clearly shown that epidemics mostly flourished within the core of the network (see Fig. 5) because of its density[20]. Ironically, this density also implies redundancy. While the core nodes are highly at risk of being infected, their removal has a limited effect because there exist alternative paths within their neighbourhood: the core offers a perfect environment to the disease and is consequently robust to node removal. It is therefore more effective to stop the disease from reaching, or leaving, the core by removing the nodes bridging other neighbourhoods (i.e. the structural hubs).

Similar conclusions are drawn for the SIR dynamics. As T moves away from T_c, the most significant level of organisation shifts from the degree (microscopic) to communities (mesoscopic) as membership-based immunization progressively outperforms the other strategies.

Results on networks of diverse nature. In this section, we highlight different behaviours observed in social, technological and communication networks using 7 other datasets (full results for the 17 datasets are available in the SI): subset of the World Wide Web (WWW)[13], MathSciNet co-authorship network (MathSci)[27], Western States Power Grid of the United States (Power Grid)[28], Internet Movie Database since 2000 (IMDb)[29], cond-mat arXiv co-authorship network (arXiv)[22], e-mail interchanges between members of the University Rovira i Virgili (Email)[30] and Gnutella peer-to-peer network (Gnutella)[31].

The results for the WWW, MathSci and IMDb networks further support our previous conclusions, with the exception that membership-based immunization performs surprisingly better than the degree-based variant even near the epidemic threshold of the network (see WWW and MathSci). The betweenness-centrality-based immunization was not tested on IMDb because of computational constraints (its computation required over 800 hours with our available ressources and a standard algorithm[32]), which illustrates a significant limit of this measure. Approximations could have been used[33], but the intricate (and mostly unknown) relationship between the efficiency of the measure and the accuracy of the approximation would have only caused additional uncertainties.

The results presented for the Power Grid network illustrate a fundamental difference between the SIS and the SIR dynamics: while we are interested in the fraction of the network sustaining an established epidemic in SIS, it is the fraction of nodes invaded by a new

disease that is relevant in SIR. In fact, the structure of the Power Grid, a chain of small, easily disconnected modules, enhances the qualitative discrepancy between the epidemic influence of nodes subjected to these two dynamics. For the SIS dynamics, the membership-based intervention is the most efficient because it weakens all modules, limiting the prevalence of the disease. In distinction, targeting through betweenness centrality merely separates the modules, so that they indiviually remain infected. For the SIR dynamics, separating the modules is the best approach as it directly stops the infection from spreading; while weakened – but connected – modules still provide pathways. This effect is a direct consequence of the particular structure of the Power Grid and is insignificant on other networks.

Finally, the last set of results, on arXiv, Email and Gnutella, present the effect of the community density ρ on the performance of membership-based immunization. For very small ρ, the paths within communities do not qualitatively differ from the links bridging neighborhoods in their effect on the disease propagation. This targeting method is therefore expected to converge toward degree-based immunization if m and k are strongly correlated. However, as most tested networks had fairly dense communities, $\rho \geq 0.3$, the relevance of memberships should not be understated.

Investigation of the epidemic regimes transition. The results of the previous sections suggest that local information (i.e., degree,

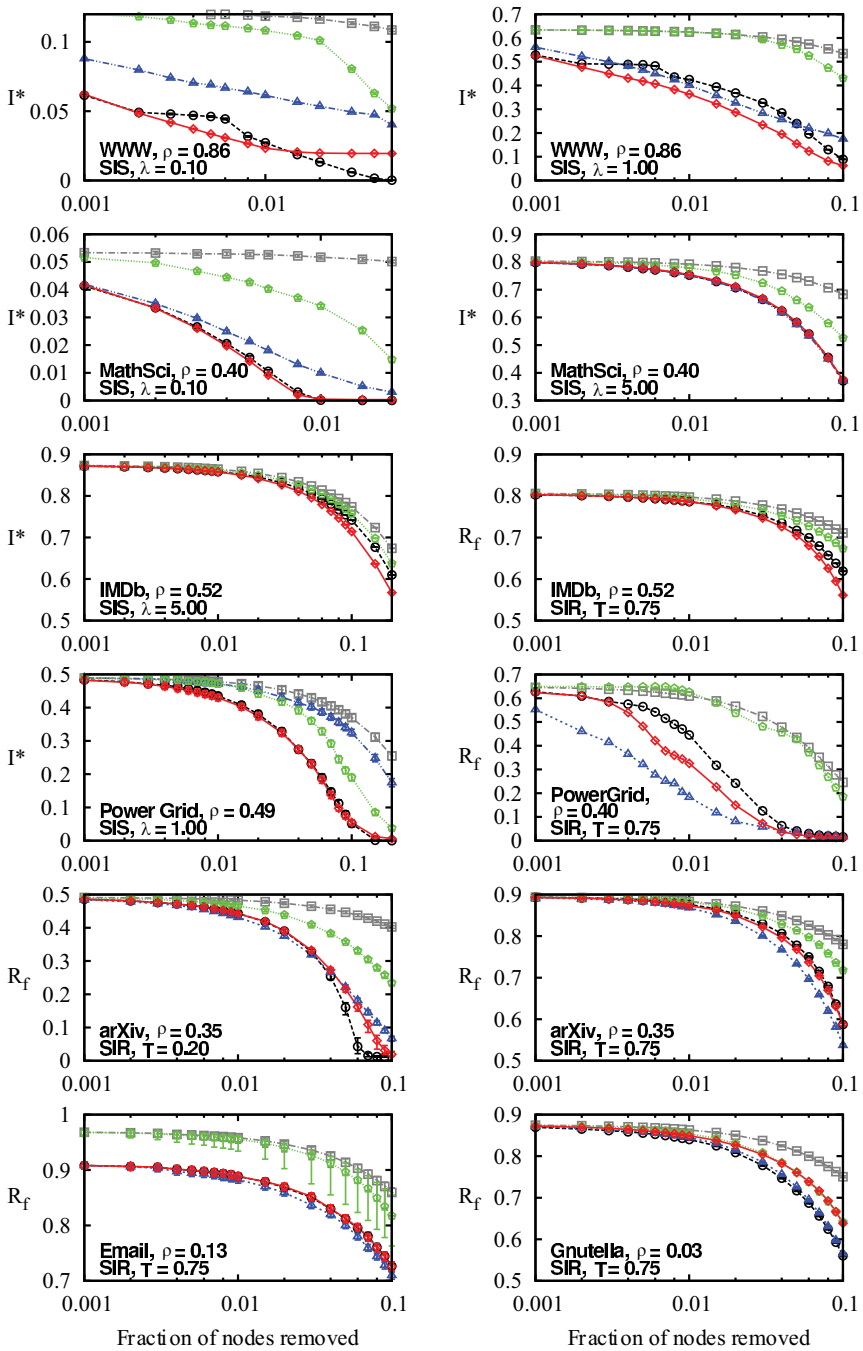


Figure 6 | Efficiency of the immunization methods against SIS and SIR epidemics on several networks. Nodes are removed in decreasing order of their score according to each method: coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) and memberships (red diamonds) to measure efficiency by the decrease of I^* or R_f. The size of the epidemics for random removal of nodes (gray squares) is added for comparison. Error bars have been omitted for clarity of the SIR results on the Power Grid, but are shown in the SI.

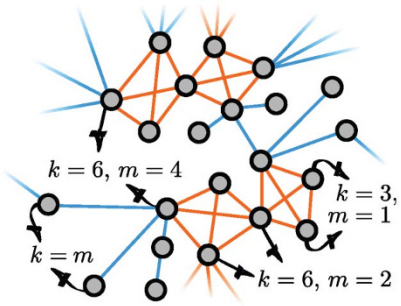


Figure 7 | Synthetic networks with tunable community structure. Orange links belong to motifs of size $M = 4$, and single links are shown in blue. The degree k and membership m of a few selected nodes are indicated. They belong to $i = (k - m)/(M - 2)$ motifs and have $j = [(M - 1)m - k]/(M - 2)$ single links.

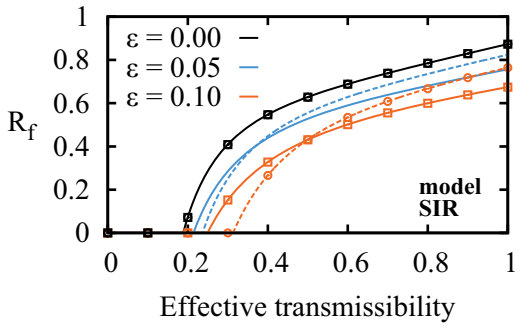


Figure 8 | Results of local immunization methods on synthetic networks. Final sizes of SIR epidemics after immunization of various fractions ε of nodes on synthetic networks with $M = 4$ and an heterogeneous degree distribution (details in SI). Near the epidemic threshold, targeting by degree (dotted curves) is the better choice whereas targeting by memberships (solid curve) should be preferred for higher transmissibility. Monte Carlo simulations were also performed to validate the formalism and indicated on the curves (the case $\varepsilon = 0.05$ is omitted not to clutter the graph) with circles (targeting by degree) and squares (targeting by membership).

membership) is often sufficient for a nearly optimal global immunization. More precisely, we found these two methods to outperform or to be as efficient as the betweenness centrality (the global method used for comparison) in 62 of the 68 studied scenarios (i.e., 17 networks / 2 dynamics / 2 transmissibility regimes). This implies that membership (e.g., on PGP), degree (e.g., Gnutella) or both (e.g. MathSci) lead to an immunization at least as efficient as global methods while having the noteworthy advantage of requiring much less information and of being less sensitive to incomplete information. This section focuses on the conditions guiding the choice between the degree-based or the membership-based immunization strategy. In this respect, Figs. 4 and 6 provide a useful hindsight: the membership-based strategy is more efficient than the degree-based one when transmissibility is high and/or when communities are dense. To further our understanding and test this hypothesis, we introduce a random network model featuring a community structure, and exactly solve its final state (R_f) under SIR dynamics using generating functions.

Our model is a slightly modified version of the configuration model[12,34] where nodes are connected either through single links or through motifs (see Fig. 7 for an example). Motifs are used to simulate the effect of a community structure, that is the redundancy of the neighbourhoods of nodes. Our motifs are composed of M nodes, all connected to each other, and a node belongs to i motifs and has j single links with probability $p(i,j)$. This node therefore has a degree (k) equal to $(M - 1)i + j$ and a membership (m) equal to $i + j$. Networks are generated using a stub pairing scheme: a node belonging to i motifs and having j single links has i "motif stubs" and j "link stubs". Groups and single links are then formed by randomly choosing M motif stubs and 2 link stubs, respectively, and then by linking the corresponding nodes to one another. This last step is repeated until none of the motif and link stubs remains. The distribution $\{p(i,j)\}_{i,j\in\mathbb{N}}$ therefore defines a maximally random network ensemble, and the results obtained are averaged over this ensemble.

Extending previous work[35], we compute the expected value of R_f for the network ensemble just defined where nodes and links are randomly removed to simulate immunization and disease transmission (SIR dynamics), respectively. Full details are given in the SI. Using typical values for $\{p(i,j)\}$, our model illustrates and confirms our hypothesis by clearly showing in Fig. 8a transition of efficiency between the degree-based and the membership-based immunization strategy. Initially less efficient when the transmissibility is low (i.e., higher threshold, lower value of R_f), membership progressively outperforms degree as the transmissibility increases. As mentionned above, for lower values of T, the best option is therefore to immunize the hubs (high k) to shift the degree distribution towards lower degrees. For higher values of T, targeting structural hubs (high m) that act as bridges between "independent" neighbourhoods leads to a more efficient immunization as it reduces the number of paths between different regions of the network. Note that we do not explicitly model the effect of community density. This could have been done by letting links exist independently with a given probability η. This is however identical to letting the disease propagate with probability ηT. Thus, the value of T in Fig. 8 is related to the density of the communities, and our conclusions can therefore be extended to the cases of low/high community densities.

Discussion

One of the main contributions of this work is to offer a formal definition of the epidemic influence of nodes, i.e. the effect of its removal on I^* of R_f, which is open to diverse methods of approximation. Our results confirm that standard measures such as the degree or betweenness centrality are *not always* the best indicators of a node's influence. Moreover, we have highlighted that the coreness, which has recently been proposed as an indicator of nodes' influence[20], offers poor performances. This has brought us to distinguish between individual risk and global influence. We have also illustrated how a universal approach is still wanting, since different networks and different diseases require different methods of intervention.

Consequently, the fact that the numbers of links and/or communities to which a node belongs are excellent measure of its epidemic influence — at times better, at times equivalent, but never much worse than global centrality measures — is a particularly important result. The fact that they both are *local measures* is especially relevant considering that we rarely have access to the exact network structure of a system, either because it is simply too large (WWW), too dynamic (email networks) or because the links themselves are ill-defined (social networks). Not only are local measures computable from a limited subset of a network (which is often the only available information), but a coarse-grained measure like membership is even more interesting as it is easier to estimate than a node's actual degree. For instance, consider how much simpler it is to enumerate your social groups (work, family, etc.) than the totality of your acquaintances.

Finally, the existence of a transition between two epidemic regimes with different characteristic scales may well be the single most important conclusion of this work. In the first regime, for low transmissibility and sparse communities, the microscopic structural

features (i.e. node connectivity or degree) offer the most relevant information; while for higher transmissibility and denser communities, mesoscopic features (i.e node communities or membership) appear more relevant. We expect to see an equivalent transition between any pair of measures which oppose the micro and meso scales (e.g. different range-limited measures of centrality[36]).

Based on our empirical and analytical results, we thus propose a simple procedure on how to judge which local measure can be expected to yield the best results in a given situation. From the available subset of a given network:

1. Obtain the degree distribution to estimate the transmissibility of the disease in relation to the epidemic threshold λ_c[24] or T_c[25].
2. If easily transmissible ($\lambda \gg \lambda_c$ or $T \gg T_c$), evaluate the network's community structure; otherwise, go to 4.
3. If the community density is high ($\rho \gtrsim 0.3$), immunize nodes according to their memberships; otherwise, go to 4.
4. For a transmissibility near the epidemic threshold, or for sparse communities (low ρ), immunize according to the degree of the nodes.

The analytical and numerical frameworks used in this work are expected to guide immunization efforts toward simpler, more precise and efficient strategies. Likewise, the introduction of a node influence classification scheme opens a new avenue for finding better local estimates of a node's role in the global state of its system.

Methods

Betweenness centrality. For all pairs (a,b) of nodes excluding i, list the $n_{a,b}$ shortest paths between a and b. Let $n_{a,b}(i)$ be the number of these paths containing i. The betweenness centrality b_i of node i is then given by:

$$b_i = \sum_{(a,b)} \frac{n_{a,b}(i)}{n_{a,b}}. \quad (1)$$

Coreness. The coreness of node i is the highest integer c_i such that the node is part of the set of all nodes with at least c_i links within the set.

Community detection. Two links, e_{ij} and e_{ik}, from a given node i, are said to belong to the same community if their Jaccard coefficient $J(e_{ij},e_{ik})$ (similarity measure) is above a given threshold J_c:

$$J(e_{ij}, e_{ik}) = \frac{n_+(j) \cap n_+(k)}{n_+(j) \cup n_+(k)} > J_c, \quad (2)$$

where $n_+(u)$ is the set containing the neighbors of u including u.

Community density. The density ρ_i of a community i of $n_i > 2$ nodes and d_i links is the proportion of the possible redundant links that do exist; i.e., the fraction of existing links excluding the minimal $n_i - 1$ links that are needed for this community to be connected:

$$\rho_i = \frac{d_i - (n_i - 1)}{\frac{n_i(n_i-1)}{2} - (n_i - 1)}. \quad (3)$$

The community density ρ is then calculated according to

$$\rho = \frac{1}{D}\sum_i d_i \rho_i \,, \quad (4)$$

where D is the total number of links not belonging to single link communities, for which $\rho_i = 0$[21].

Immunization. To perform the immunization of a fraction ε of the network according to a certain measure Γ, we remove the εN nodes with the highest Γ. When a choice must be made (nodes with equal Γ), all decisions are taken randomly and individually for each simulated epidemics.

Monte Carlo simulations. To investigate the fraction of a network which can *sustain* an epidemics, SIS simulations start with all nodes in an infectious state and are then relaxed until an equilibrium is reached. To investigate the mean fraction of a network which a disease can *invade*, SIR simulations start with a single randomly chosen infectious node and run until there are no more infectious nodes. Results shown in the figures are obtained by averaging over the outcome of several numerical simulations until the minimal possible standard deviation (limited by network structure and finite size) is obtained. For the SIR dynamics, only the simulations leading to largescale epidemics (at least 1% of the nodes) were considered. The complete procedure is given in the SI.

1. Caldarelli, G. & Vespignani, A. *Large Scale Structure and Dynamics of Complex Networks.* World Scientific Publishing Company, Singapore (2007).
2. Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals* Princeton University Press, Princeton (2008).
3. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* Oxford University Press, New York (1991).
4. Keeling, M. J. & Eames, K. T. D. Networks and epidemic models. *J R Soc Interface* **2**, 295–307 (2005).
5. Pastor-Satorras, R. & Vespignani, A. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
6. Gómez-Gardeñes, J., Echenique, P. & Moreno, Y. Immunization of real complex communication networks. *Eur. Phys. J. B* **49**, 259–264 (2006).
7. Dunne, J. A. & Williams, R. J. Cascading extinctions and community collapse in model food webs. *Philos Trans R Soc Lond B Biol Sci* **364**, 1711–1723 (2009).
8. Gallos, L. K., Liljeros, F., Argyrakis, P., Bunde, A. & Havlin, S. Improving immunization strategies. *Phys. Rev. E* **75**, 045104(R) (2007).
9. Chen, Y., Paul, G., Havlin, S., Liljeros, F. & Stanley, H. E. Finding a Better Immunization Strategy. *Phys.Rev. Lett.* **101**, 058701 (2008).
10. Salathé, M. & Jones, J. H. Dynamics and Control of Diseases in Networks with Community Structure. *PLoS comp. biol.* **6**, e1000736 (2010).
11. Masuda, N. Immunization of networks with community structure. *New J. Phys* **11**, 123018 (2009).
12. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
13. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
14. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
15. Pastor-Satorras, R. & Vespignani, A. Immunization of complex networks. *Phys. Rev. E* **65**, 036104 (2002).
16. Freeman L. Centrality in social networks: Conceptual clarification. *Social Networks* **1**, 215–239 (1979).
17. Barthélemy, M. Betweenness centrality in large complex networks. *Eur. Phys. J. B* **38**, 163–168 (2004).
18. Batagelj, V. & Zaveršnik, M. Generalized Cores. *arXiv:cs/0202039v1*.
19. Batagelj, V. & Zaveršnik, M. An O(m) Algorithm for Cores Decomposition of Networks. *arXiv:cs/0310049v1*.
20. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nature Physics* **6**, 888–893 (2010).
21. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
22. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
23. Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A. & Arenas, A. Models of social networks based on social distance attachment. *Phys. Rev. E* **70**, 056122 (2004).
24. Hébert-Dufresne, L., Noël, P.-A., Allard, A., Marceau, V. & Dubé, L. J. Propagation dynamics on networks featuring complex topologies. *Phys. Rev. E* **82**, 036115 (2010).
25. Newman, M. E. J. Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128 (2002).
26. Alvarez-Hamelin, I., Dall'Asta, L., Barrat, A. & Vespignani, A. k-core decomposition: A tool for the visualization of large scale networks. *Advances in Neural Information Processing Systems* **18**, 41–50 (2006).
27. Palla, G., Farkas, I. J., Pollner, P., Derényi, I. & Vicsek, T. Fundamental statistical features and self-similar properties of tagged networks. *New J. Phys.* **10**, 123026 (2008).
28. Watts, D. J. & Strogatz, S. H. Collective dynamics of small-world networks. *Nature* **393**, 440–442 (1998).
29. Hébert-Dufresne, L., Allard, A., Marceau, V., Noël, P.-A. & Dubé, L. J. Structural Preferential Attachment: Network Organization beyond the Link. *Phys. Rev. Lett.* **107**, 158702 (2011).
30. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103(R) (2003).
31. Ripeanu, M., Foster, I. & Iamnitchi, A. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal* **6**, 50–57 (2002).
32. Brandes, U. A Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.* **25**(2), 163–177 (2001).
33. Madduri, K., Ediger, D., Jiang, K., Bader, D. A. & Chavarría-Miranda, D. G. A Faster Parallel Algorithm and Efficient Multithreaded Implementations for Evaluating Betweenness Centrality on Massive Datasets. *Third Workshop MTAAP* (2009).
34. Newman, M. E. J. Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121 (2003).

35. Allard, A., Hébert-Dufresne, L., Noël, P.-A., Marceau, V. & Dubé, L. J. Bond percolation on a class of correlated and clustered random graphs. *J. Phys. A: Math. Theor.* **45**, 405005 (2012).
36. Ercsey-Ravasz, M., Lichtenwalter, R. N., Chawla, N. V. & Toroczkai, Z. Range-limited centrality measures in complex networks. *Phys. Rev. E* **85**, 066103 (2012).

Acknowledgements

The authors wish to thank Louis Roy for the development of a k-core visualization tool; Yong-Yeol Ahn *et al.* for their link community algorithm; all the authors of the cited papers for providing their network data; and Calcul Québec for computing facilities. This research was funded by CIHR, NSERC and FRQ-NT.

Author contributions

L.H.-D. and A.A. designed the study. L.H.-D., A.A. and J.-G.Y. performed the computations. All authors have contributed to the analysis and wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hébert-Dufresne, L., Allard, A., Young, J. & Dubé, L.J. Global efficiency of local immunization on complex networks. *Sci. Rep.* **3**, 2171; DOI:10.1038/srep02171 (2013).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0

Global efficiency of local immunization on complex networks
Supplementary Information

Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young and Louis J. Dubé

http://dynamica.phy.ulaval.ca/

Département de Physique, de Génie Physique et d'Optique
Université Laval, Québec, Québec, Canada G1V0A6

Contents

1 Supplementary discussions on methods **2**

2 Theoretical modelling **4**

3 Introduction to the supplementary results **8**

4 arXiv co-authorship **9**

5 Brightkite online social network **10**

6 University email exchange **11**

7 Enron email exchange **12**

8 Gnutella peer-to-peer network **13**

9 Google weblinks **14**

10 Gowalla social network **15**

11 Internet autonomous systems **16**

12 Internet Movie Database **17**

13 MathSciNet co-authorship **18**

14 Myspace online social network **19**

15 Pretty-Good-Privacy data exchange **20**

16 Power grid **21**

17 Protein interactions network **22**

18 Slashdot online social network **23**

19 Word association network **24**

20 World Wide Web **25**

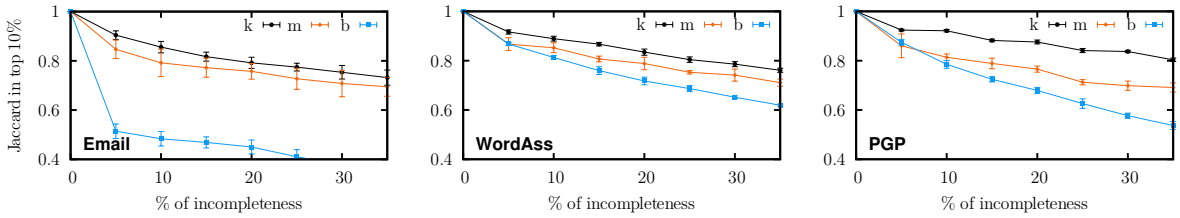
1 Supplementary discussions on methods

Local vs global measures

We differentiate between these two types of measures by the information required to compute them. If this information (per node) is independent of total system size, the measure is considered local; whereas a global measure requires information scaling with system size (often a complete description). For the four properties studied in this paper, we thus consider that:

1. degree is a local measure, as only the number of neighbours of a node is required;
2. membership is a local measure, as the chosen algorithm only requires the neighbourhood of one given node and that of its neighbours to estimate its membership number;
3. coreness is a global measure, as a node's coreness depends on the coreness of its neighbours which in turn depend on the coreness of their neighbours and so on;
4. betweenness centrality is a global measure, since it is calculated by considering the shortest paths between a given node and all of the other nodes in the network.

For obvious reasons, *local measures are less sensitive to incomplete or incorrect information*. Adding, removing or rewiring a link only affects the degree or membership of nodes directly in the neighbourhood of the modification; whereas the same alterations can potentially affect the coreness or betweenness centrality of nodes anywhere in the network through cascading effects. Consequently, measures based on shorter-range information are always more robust to missing information, and sometimes quite significantly, as seen on Suppl. Fig. 1.

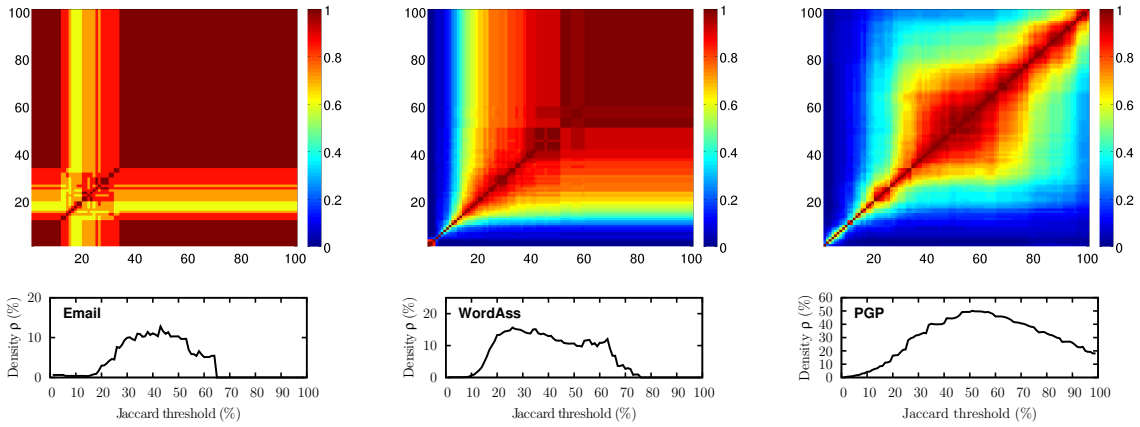


Supplementary Figure 1: Robustness of measured from the micro (degree k), meso (memberships m) and macro (betweenness centrality b) scales when information on the network is removed. The robustness is here measured by comparing (with a Jaccard coefficient) the ensemble of nodes identified as being in the top 10% of nodes when a certain fraction of links are randomly removed (horizontal axis) as opposed to the ensemble obtained by considering the complete data.

Community detection

As mentioned above the *link clustering algorithm* of Ahn *et al.* was chosen in part because it can perform well (and at times even better) using local information instead of the entire network.[1] While we always partitioned the network globally, by setting a resolution threshold, the identification of structural hubs is very robust to this global threshold (see Suppl. Fig. 2). More importantly, this algorithm groups links stemming from a given node in a community based on the similarity of the two neighbourhoods reached through them. Hence, it evaluates the redundancy in second neighbourhoods (how many of my second neighbours are neighbours of more than one of my neighbours?). This redundancy (or overlap) can then serve as an appropriate measure to gauge the major impact of community structure on an epidemic process, namely the loss of potential new infections due to clustering. The link clustering algorithm therefore provides a well-defined method to quantify this loss.

[1] Y.-Y. Ahn, J. P. Bagrow, & S. Lehmann, Link communities reveal multiscale complexity in networks, *Nature*, vol. 466, p.761-764, 2010. See also the corresponding Supplementary Information.



Supplementary Figure 2: Robustness of our ability to identify structural hubs (1% of nodes with the most memberships) as the global resolution varies (the final partition is chosen to maximize the community density shown at the bottom). The color map represents the Jaccard coefficient, i.e. the similarity of ensembles, as measured between the structural hubs identified with two different resolution parameter. These ensembles are always very similar when avoiding extreme resolutions (e.g. low density). Note that in the color map, yellow corresponds to a Jaccard coefficient of 0.6 which implies that 75% of the same structural hubs were found by both partitions.

Supplementary simulation details

SIS. All nodes are initially infectious and we relax the system by iterating a discrete time propagation simulation using time step Δt chosen such that $\alpha\Delta t$ and $\beta\Delta t$ are less than 10^{-3}:

i. at each Δt, every susceptible neighbour S of every infectious individual I is infected with probability $\alpha\Delta t$;
ii. at each Δt every infectious individual I recovers with probability $\beta\Delta t$;

the steady-state is averaged over multiple independent simulations to minimize the standard deviation (due to network structure and finite size).

SIR. A single node is randomly infected and the following stochastic process is iterated until no infectious nodes remain:

i. I nodes infect each of their S neighbours with probability T and then recover.

The final state, considering only epidemics larger than 1% of the system size, is averaged over multiple independent simulations to minimize the standard deviation (due to network structure and finite size).

2 Theoretical modelling

The conclusion drawn in the main text were validated using synthetic networks. In this Section, we present how these synthetic networks are generated. Furthermore, we describe the mathematical framework used to calculate the final outcome of the SIR dynamics on these networks. Finally, we give the parameters used for the main results.

Synthetic networks

The synthetic networks considered are a clustered and multitype generalisation of the Configuration Model.[2] In these networks, nodes are connected either through single links or through motifs (see Fig. 7 in the main text for an example). Motifs are composed of M nodes which are all connected to one another, and a node belongs to i motifs and has j single links with probability $p(i, j)$. This node therefore has a degree (k) equal to $(M-1)i+j$ and a membership (m) equal to $i+j$.

Networks are generated using a stub pairing scheme: a node belonging to i motifs and having j single links has i "motif stubs" and j "link stubs". Groups and single links are then formed by randomly choosing M motif stubs and 2 link stubs, respectively, and then by linking the corresponding nodes to one another. This last step is repeated until none of the motif and link stubs remain. The distribution $\{p(i, j)\}_{i,j\in\mathbb{N}}$ therefore defines a random network ensemble, and the results obtained in this Section are averaged over this ensemble.

Mathematical formalism

As there exists a mapping—under simple assumptions—between the SIR dynamics and bond percolation on networks[3,4]. To calculate the outcome of the SIR dynamics on the networks just described, we use a previously published formalism[2] in which we add the possibility for nodes to exist with a given probability (i.e., site percolation) to simulate immunization strategies. We only give a short outline of this theoretical model as a general and more formal description will be the subject of a subsequent publication.

For each pair (i, j) such that $p(i, j) \neq 0$ we assign a node type denoted by $\{i, j\}$ (the set of such pairs is noted $\mathcal{M}$). As the pair (i, j) is the only information available about the nodes, assigning one node type per pair allows us to simulate very detailed immunization strategies. Indeed, we define $q_{\{i,j\}}$ as the probability for a type-$\{i, j\}$ node to be immunized; the simulated immunization strategy is therefore encoded in the set of probabilities $\{q_{\{i,j\}}\}_{\{i,j\}\in\mathcal{M}}$. Also, as explained in the main text, the infectious agent propagates from an infected node to a susceptible neighbour with probability T. From a percolation point of view, $1 - q_{\{i,j\}}$ is the occupation probability of type-$\{i, j\}$ sites (nodes) and T is the occupation probability of bonds (links).

Solving site/bond percolation in motifs

The mathematical formalism that we have developed relies on probability generating functions (PGFs) and therefore implicitly requires the networks under consideration to have a tree-like structure. As the

[2] A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé (2012). Bond percolation on a class of correlated and clustered random graphs. *J. Phys. A*, 45(40):405005.

[3] M. E. J. Newman (2002). Spread of epidemic disease on networks. Phys. Rev. E, 66(1):016128.

[4] E. Kenah and J. Robins (2007). Second look at the spread of epidemics on networks. Phys. Rev. E, 76(3):1-12.

networks we consider contain motifs, which clearly do not comply with that assumption, we need to solve the bond and site percolation within motifs beforehand.

As previously shown[5], the bond percolation outcome—the distribution of the number of nodes that can be reached by following links from an initial node—can be exactly obtained by iterating a set of simple equations. We denote $\boldsymbol{n}$ the $|\mathcal{M}|$-tuple whose elements[6] $n_{\{i,j\}}$ correspond to the number of nodes of each type (i.e., there are $n_{\{i,j\}}$ type-$\{i, j\}$ nodes). For the remaining, each boldfaced variable will correspond to such $|\mathcal{M}|$-tuple.

Let us define $Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})$ as the probability to find a component of $\boldsymbol{l}$ nodes in a motif of size (and composition) $\boldsymbol{n}$ from an initial type-$\{i, j\}$ node. Although nodes are initially all connected to one another in motifs, we are interested in the number of nodes (and their type) that can be reached from an initial node when links are followed with a probability T (bond percolation). Following previous work[5], $Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})$ is obtained by iterating

$$Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) = Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{l}) \prod_{\{i',j'\}\in\mathcal{M}} \binom{n_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}}{l_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \prod_{\{i'',j''\}\in\mathcal{M}} (1-T)^{n_{\{i',j'\}}(n_{\{i'',j''\}} - l_{\{i'',j''\}})} \tag{2.1a}$$

$$Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{l}) = 1 - \sum_{\boldsymbol{m}<\boldsymbol{l}} Q_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{l}) \tag{2.1b}$$

from the initial condition $Q_{\{i,j\}}(\boldsymbol{\delta_{\{i,j\}}}|\boldsymbol{\delta_{\{i,j\}}})$, where δ_{ij} is the Kronecker delta, and where $\boldsymbol{\delta_{\{i,j\}}}$ is an $|\mathcal{M}|$-tuple whose elements are all equal to 0 except for the $\{i, j\}$-th one that is equal to one. The sum in Eq. (2.1b) is over all $\boldsymbol{m}$ such that $m_{\{i,j\}} \leq l_{\{i,j\}}$ for every node type $\{i, j\}$ but with the additional constraint that $\boldsymbol{m} \neq \boldsymbol{l}$. The initial condition simply states that the probability of finding a component of one type-$\{i, j\}$ node from a type-$\{i, j\}$ node in a motif containing only one type-$\{i, j\}$ node is one (the initial node is always included in the size of the component). Then, Eqs. (2.1) iteratively increase the size of the motif and compute the size distribution along the way until the complete distribution for a motif of the desired size (and composition) is obtained.

Should we be interested in studying bond percolation on motifs solely, we would keep the distribution $\{Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})\}$ for a given size $\boldsymbol{n}$, and discard the distributions for motifs of intermediate size obtained while iterating Eqs. (2.1). These intermediate distributions can however be used to *exactly* predict the distribution of the number of nodes that can be reached by following links from an initial node in motifs where links *and* nodes exist with given probabilities (bond and site percolation). Indeed, as each node exists independently with a given probability, the probability for a motif of original size $\boldsymbol{n}$ to be of size $\boldsymbol{m}$ after the random removal of its nodes is simply

$$W_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{n}) = \prod_{\{i',j'\}\in\mathcal{M}} \binom{n_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}}{m_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \left[1 - q_{\{i',j'\}}\right]^{m_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \left[q_{\{i',j'\}}\right]^{n_{\{i',j'\}} - m_{\{i',j'\}}}, \tag{2.2}$$

where we assume that the initial type-$\{i, j\}$ exists. Then, the probability for a type-$\{i, j\}$ node to lead to a component of size $\boldsymbol{l}$ in a motif of original size $\boldsymbol{n}$ but whose links and nodes have been randomly removed is simply

$$P_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) = \sum_{\boldsymbol{m}=\boldsymbol{\delta_{\{i,j\}}}}^{\boldsymbol{n}} Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{m}) W_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{n}) \,. \tag{2.3}$$

[5] A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé (2012). Exact solution of bond percolation on small arbitrary graphs. EPL, 98(1):16001.

[6] $|\mathcal{M}|$ is the number of elements (i.e., the cardinality) of the set $\mathcal{M}$, hence the number of node types.

The site and bond percolation can therefore be exactly solved for motifs by iterating Eqs. (2.1)–(2.3). As expected, if applied to the simplest motifs, i.e. links, the type-$\{i, j\}$ node at the other end of the link will be reached with probability $T(1-q_{\{i,j\}})$ and will not be reached with probability $1-T(1-q_{\{i,j\}})$.

As motifs and single links are built by randomly matching stubs, the probability for a type-$\{i, j\}$ node to appear in a motif [a link] is proportional to $[ip(i, j)]$ $[jp(i, j)]$. The probability for a motif to be composed by $\boldsymbol{n}$ nodes is given by the multinomial distribution

$$R^{(\mathrm{m})}(\boldsymbol{n}) = \frac{M!}{\left[\sum_{\{i,j\}\in\mathcal{M}} ip(i,j)\right]^M} \prod_{\{i,j\}\in\mathcal{M}} \frac{[ip(i,j)]^{n_{\{i,j\}}}}{n_{\{i,j\}}!} \,. \tag{2.4a}$$

The same applies for the composition of links

$$R^{(\mathrm{l})}(\boldsymbol{n}) = \frac{2!}{\left[\sum_{\{i,j\}\in\mathcal{M}} jp(i,j)\right]^2} \prod_{\{i,j\}\in\mathcal{M}} \frac{[jp(i,j)]^{n_{\{i,j\}}}}{n_{\{i,j\}}!} \,. \tag{2.4b}$$

Finally, for the purpose of calculating the outcome of the bond and site percolation on the synthetic networks, let us define the two following generating functions:

$$\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x}) = \sum_{\boldsymbol{n}} \frac{n_{\{i,j\}}R^{(\mathrm{m})}(\boldsymbol{n})}{\sum_{\boldsymbol{n}'} n'_{\{i,j\}}R^{(\mathrm{m})}(\boldsymbol{n'})} \left[\sum_{\boldsymbol{l}=\boldsymbol{\delta}_{\{i,j\}}}^{\boldsymbol{n}} P^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) \prod_{\{i',j'\}\in\mathcal{M}} [x_{\{i',j'\}}]^{l_{\{i',j'\}}-\delta_{ii'}\delta_{jj'}}\right] \tag{2.5a}$$

$$\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y}) = \sum_{\boldsymbol{n}} \frac{n_{\{i,j\}}R^{(\mathrm{l})}(\boldsymbol{n})}{\sum_{\boldsymbol{n}'} n'_{\{i,j\}}R^{(\mathrm{l})}(\boldsymbol{n'})} \left[\sum_{\boldsymbol{l}=\boldsymbol{\delta}_{\{i,j\}}}^{\boldsymbol{n}} P^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) \prod_{\{i',j'\}\in\mathcal{M}} [y_{\{i',j'\}}]^{l_{\{i',j'\}}-\delta_{ii'}\delta_{jj'}}\right] \tag{2.5b}$$

where the superscript "m" (resp. "l") indicate that the quantities have been solved for motifs (resp. links). In other words, the function $\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})$ generates the probability distribution for the number of nodes of each type that can be reached from a type-$\{i, j\}$ node in a random motif [i.e., whose composition is averaged over $R^{(\mathrm{m})}(\boldsymbol{n})$]. Specifically, the coefficient in front of $x^s_{\{i',j'\}}$ in Eq. (2.5a) is the probability of reaching s type-$\{i', j'\}$ nodes from a type-$\{i, j\}$ node in a random motif. The same applies to $\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})$.

Calculating the average fate of an outbreak

We are now in a position to solve the bond and site percolation on the synthetic networks defined previously. For the purpose of the present study, we are interested in the quantity R_{f}: the relative size of the extensive (giant) component. To highlight the nontrivial effect of immunization[7], R_{f} is expressed in terms of the fraction of the *existing* nodes (i.e., $1-\varepsilon$) that are part of the giant component.

It is convenient to introduce the following function

$$g_{\{i,j\}}(\boldsymbol{x}, \boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^i \left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^j \tag{2.6}$$

generating the distribution of the number of nodes of each type that are in the immediate neighbourhood of a type-$\{i, j\}$ node. The *immediate neighbourhood* refers to the nodes to which the type-$\{i, j\}$

[7] The relative size of the giant component cannot exceed $1-\varepsilon$ on networks for which a fraction ε of the nodes has been removed. This reduction in size obviously occurs during any immunization strategy and, for comparison purposes, must be taken into account.

node is connected either via its single links or via its motifs. Similarly, we define

$$f^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x},\boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^{i-1}\left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^{j} \tag{2.7a}$$

$$f^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{x},\boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^{i}\left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^{j-1} \tag{2.7b}$$

generating the distribution of the number of nodes of each type that are in the immediate neighbourhood of a type-$\{i, j\}$ node that has been reached by either one of its single links or one of the motifs it is a part of (if applicable). In other words, these functions generate the excess degree distribution.

To calculate R_{f}, let us define $a_{\{i,j\}}$ as the probability that a link to a type-$\{i, j\}$ node *does not* lead to the giant component. Similarly, we define $b_{\{i,j\}}$ as the probability that a type-$\{i, j\}$ node reached through a motif does not lead to the giant component. Due to the effective tree-like structure of the networks—recall that the outcome of percolation on the motifs has already been solved—$a_{\{i,j\}}$ and $b_{\{i,j\}}$ must satisfy the following self-consistency relations

$$a_{\{i,j\}} = f^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b}) \tag{2.8a}$$

$$b_{\{i,j\}} = f^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b})\,. \tag{2.8b}$$

Put simply, these equations state that if a type-$\{i, j\}$ node reached from either a link or a motif does not lead to the giant component, then neither should the nodes that can be reached from it. The probability that a type-$\{i, j\}$ node *is* part of the giant component is then $1 - g_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b})$. The probability that a randomly existing node is part of the giant component—which corresponds to its size as well—is therefore

$$R_{\mathrm{f}} = \sum_{\{i,j\}\in\mathcal{M}} \frac{(1-q_{\{i,j\}})p(i,j)\left[1-g_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b})\right]}{\sum_{\{i',j'\}\in\mathcal{M}}(1-q_{\{i',j'\}})p(i',j')}\,. \tag{2.9}$$

The theoretical predictions (lines) on Fig. 8 in the main text were obtained by solving Eqs. (2.1)–(2.9) for various values of T and $\{q_{\{i,j\}}\}_{\{i,j\}\in\mathcal{M}}$. Comparison with results obtained from numerical simulations (symbols) confirms the validity of our theoretical model.

Parameters used for theoretical calculations

Table 1 shows the distribution $\{p(i, j)\}_{i,j\in\mathbb{N}}$ used for the synthetic networks considered in the main text. It also gives the degree ($k_{\{i,j\}}$) and the membership ($m_{\{i,j\}}$) of each node type $\{i, j\} \in \mathcal{M}$. Motifs were composed of $M = 4$ nodes, and numerical simulation results (symbols on Fig. 8) were averaged over 5×10^5 realisations of networks of 2.5×10^5 nodes. For node types with $k_{\{i,j\}} > 2$, we let sets of $M-1$ links to either be part of cliques of M nodes or be single links in order to avoid unintended degree correlations[8]. For a given fraction ε of the nodes to immunize, we have

$$\sum_{\{i',j'\}\in\mathcal{M}} q_{\{i',j'\}}p(i',j') = \varepsilon\,. \tag{2.10}$$

The probabilities $q_{\{i,j\}}$ are chosen to satisfy this condition and in decreasing order of degree or membership.

[8] I.Z. Kiss & D.M. Green (2008), Comment on "Properties of highly clustered networks", *Phys. Rev. E*, 78:048101.

$\{i, j\}$	$p(i, j)$	$k_{\{i,j\}}$	$m_{\{i,j\}}$
$\{0, 1\}$	0.43930	1	1
$\{0, 2\}$	0.13179	2	2
$\{0, 9\}$	0.00712	9	9
$\{1, 3\}$	0.25831	6	4
$\{1, 6\}$	0.04982	9	7
$\{2, 3\}$	0.02325	9	5
$\{3, 0\}$	0.09041	9	3

Supplementary Table 1: Distribution $\{p(i, j)\}_{i,j\in\mathbb{N}}$ used for the synthetic networks discussed in the main text. The degree and the membership of each node type is computed according to $k_{\{i,j\}} = (M - 1)i + j$ and $m_{\{i,j\}} = i + j$, respectively, with $M = 4$.

3 Introduction to the supplementary results

The last sections of this Supplementary Information present a more complete view of the results obtained on empirical networks and are structured as follows. Each section covers one of the 17 datasets used in the study. Firstly, a brief discussion on the nature of each network is given, along with:

- the number of nodes (N), of links (L) and the degree distribution (k links per node);
- the maximal community density ρ and corresponding Jaccard threshold J_ρ.
- the maximal values of degree k, coreness c, betweenness centrality b and memberships m.

Secondly, correlations between degree, betweenness centrality, coreness and memberships are quantified using Spearman's rank correlation coefficient (defined below). We leave to the reader to observe how, given the correlation coefficient between memberships ranking and degree ranking, along with the mean community density, one can somewhat predict if the membership-based immunization will be more or less efficient than the degree-based version. Finally, the results of all immunization methods (random or on the four measures) are presented for SIS and SIR dynamics for a virulence (probability of disease transmission) close and far from the network's epidemic threshold.

Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient quantifies the statistical dependence of two different orderings of the same set of items (nodes) on a scale of -1 (perfectly anti-correlated) to 1 (perfectly correlated).[9]

Consider x_i to be the rank of item i according to measure X, and y_i to be the rank of the same item according to a different measure Y. If for example, 10 items have the same score according to X and would otherwise be ranked from x_j to x_{j+9}, they are all given the rank $\left[\sum_{k=0}^{9} x_{j+k}\right]/10$. The Spearman's rank correlation coefficient $\sigma(X, Y)$ is then given by:

$$\sigma(X, Y) = \left[\sum_i (x_i - \bar{x})\,(y_i - \bar{y})\right] \Big/ \left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2\right]^{1/2},$$

where $\bar{u}$ is the average rank according to measure U (the mean of $\{u_i\}$).

[9] C. Spearman (1904), The proof and measurement of association between two things, *Amer. J. Psychol.*, 15:72101.

4 arXiv co-authorship

The cond-mat arXiv database uses articles published at *http://arxiv.org/archive/cond-mat* between April 1998 and February 2004. In this network, an article written by n co-authors contributes to a link of weight $(n-1)$ between every pair of authors. The unweighted network was obtained by deleting all links with a weight under the selected threshold of 0.1.[10]

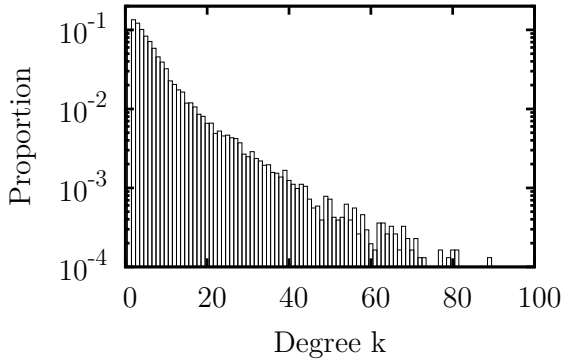
Supplementary Table 2: arXiv statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
30561	125959	191	15	$6.9e+06$	127	0.35

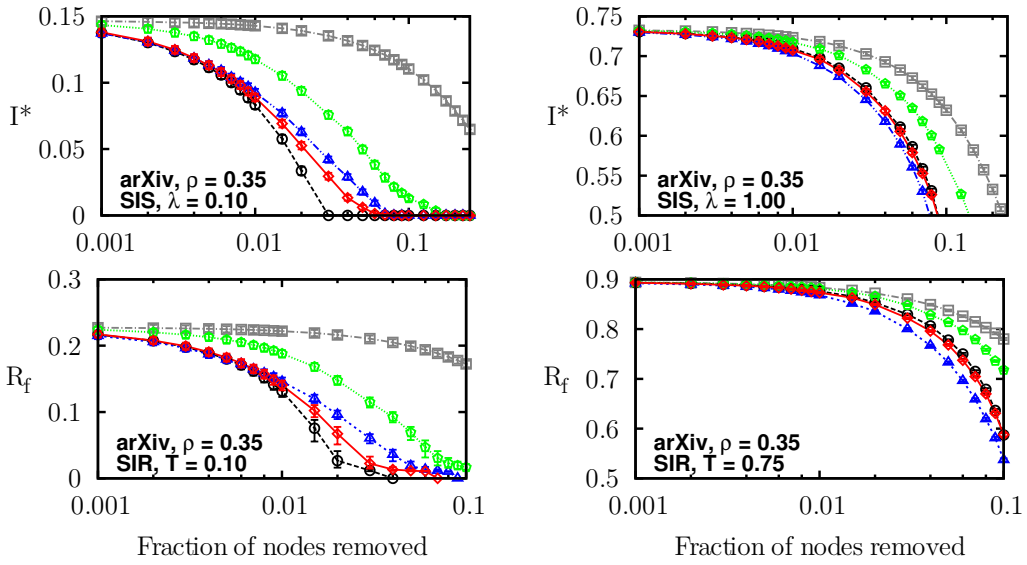
Supplementary Table 3: arXiv correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.7766	0.7717	0.6639	0.7461	0.9411	0.5388

Supplementary Figure 3: arXiv degree distribution



Supplementary Figure 4: Intervention against epidemics on arXiv after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[10]Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814-818

5 Brightkite online social network

Brightkite was a location-based online social network where users could "check in" to the physical places they were visiting to connect with nearby friends. This datasets was obtained from a total of 4,491,143 check-ins over the period of Apr. 2008 - Oct. 2010.[11]

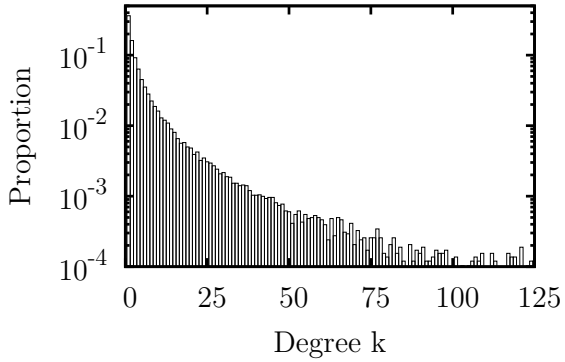
Supplementary Table 4: Brightkite statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
58228	214078	1134	52	$2e+08$	1118	0.55

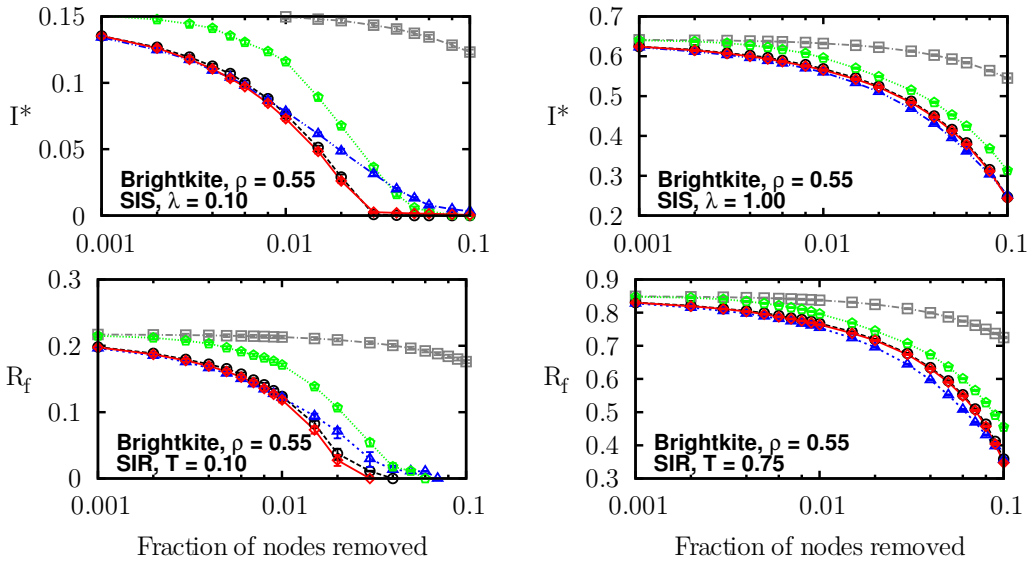
Supplementary Table 5: Brightkite correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9845	0.8919	0.9477	0.8822	0.9659	0.7767

Supplementary Figure 5: Brightkite degree distribution



Supplementary Figure 6: Intervention against epidemics on Brightkite after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[11] Cho, E., Myers, S.A. & Leskovec, J. (2011) Friendship and Mobility: User Movement in Location-Based Social Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

6 University email exchange

Network of email communication between accounts from the University Rovira i Virgili.[12]

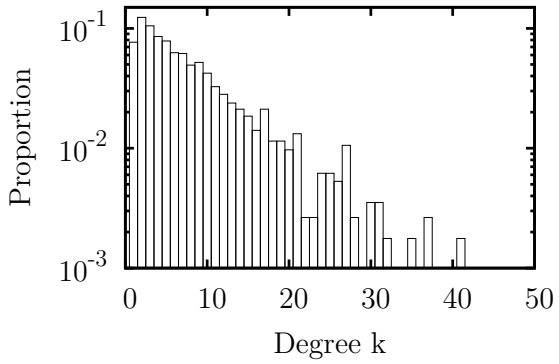
Supplementary Table 6: Email statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
1134	5143	1080	8	$6.1e+05$	929	0.13

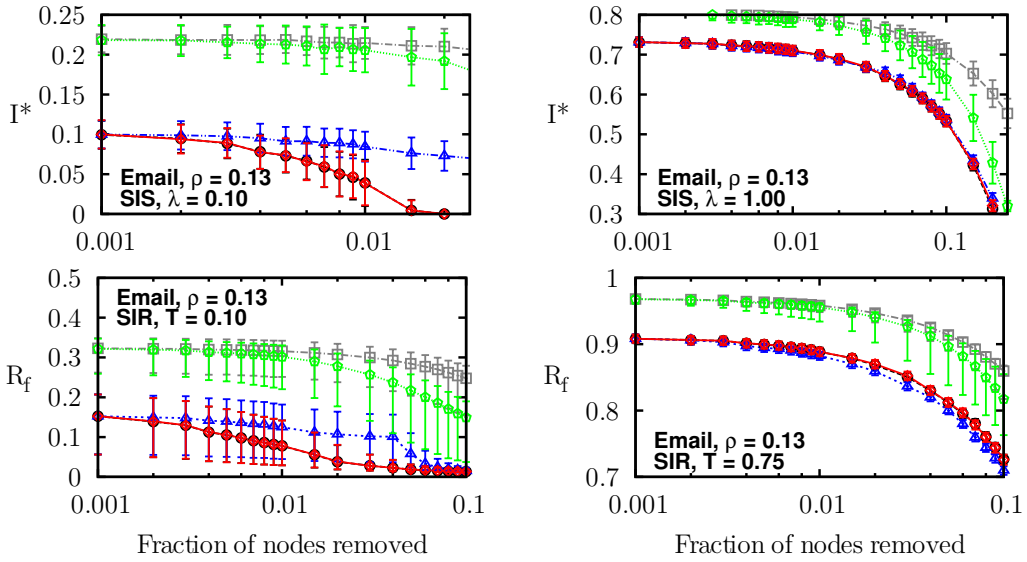
Supplementary Table 7: Email correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9900	0.9474	0.9560	0.9447	0.9613	0.8831

Supplementary Figure 7: Email degree distribution



Supplementary Figure 8: Intervention against epidemics on university email network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[12]Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. (2003) Self-similar community structure in a network of human interactions. *Phys. Rev. E* 68:065103(R)

7 Enron email exchange

Network of email interchanges between all different Enron email addresses built from a dataset of around half million emails (made public by the Federal Energy Regulatory Commission).[13]

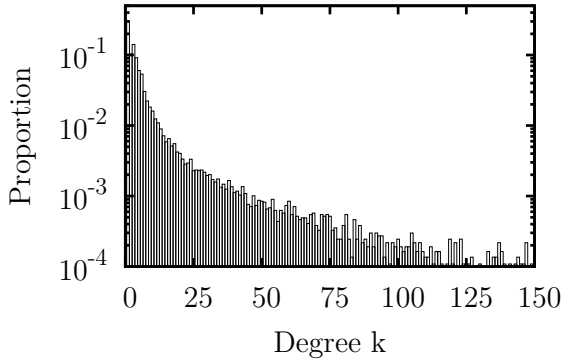
Supplementary Table 8: Enron statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
36692	183831	1383	43	$4.3e+07$	1306	0.61

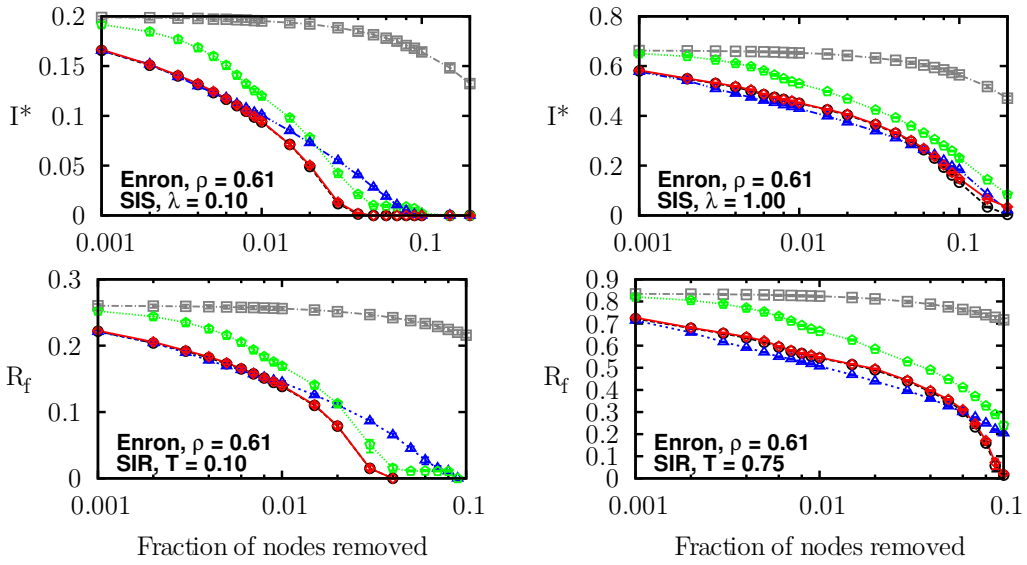
Supplementary Table 9: Enron correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9325	0.7567	0.9173	0.7585	0.9839	0.6862

Supplementary Figure 9: Enron degree distribution



Supplementary Figure 10: Intervention against epidemics on Enron email network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[13] Klimmt, B. & Yang, Y. (2004) Introducing the Enron corpus. *CEAS conference*.

8 Gnutella peer-to-peer network

A snapshot of the Gnutella peer-to-peer network, where nodes are hosts and edges connections, from August 30th 2002. The data is originally directed (files taken *from* one host *to* another), but was made undirected for this work.[14]

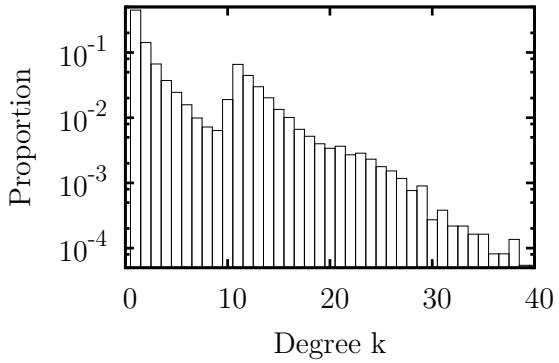
Supplementary Table 10: Gnutella statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
36682	88328	55	7	$5.3e+06$	52	0.03

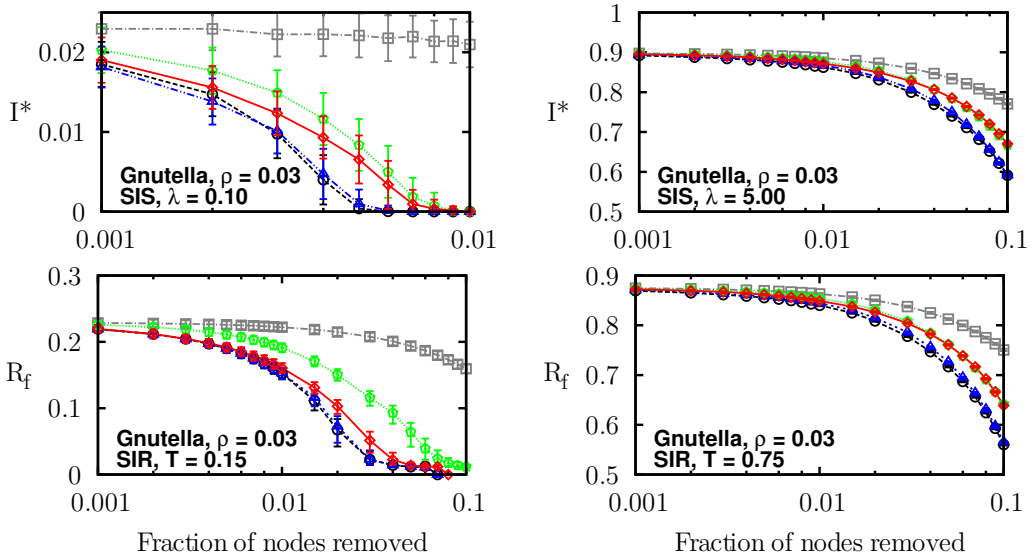
Supplementary Table 11: Gnutella correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9849	0.9848	0.9796	0.9925	0.9823	0.9743

Supplementary Figure 11: Gnutella degree distribution



Supplementary Figure 12: Intervention against epidemics on Gnutella network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[14]Ripeanu, M., Foster, I. & Iamnitchi, A. (2002) Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal* 6:50-57

9 Google weblinks

Directed network of hyperlinks between Google's webpages (considered undirected for this study).[15]

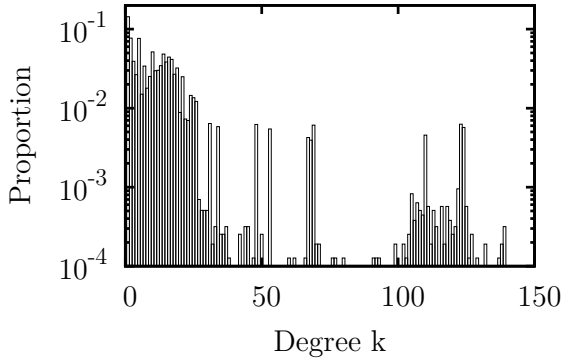
Supplementary Table 12: Google statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
15763	149456	11401	102	$9.0e+07$	2883	0.49

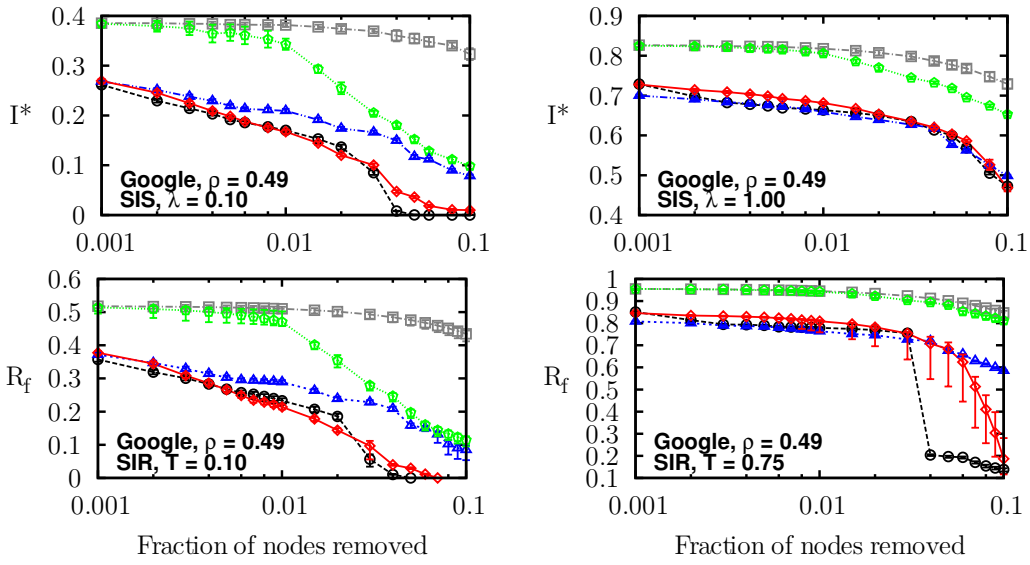
Supplementary Table 13: Google correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8862	0.7941	0.8401	0.7735	0.9723	0.6995

Supplementary Figure 13: Google degree distribution



Supplementary Figure 14: Intervention against epidemics on Google network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[15]Palla, G., Farkas, I.J, Pollner, P., Derényi, I. & Vicsek, T. (2007) Directed network modules. *New. J. Phys.* 9:186

10 Gowalla social network

Gowalla is a location-based social networking website similar to Brightkite. This friendship network is undirected and composed from a total of 6,442,890 check-ins over the period of Feb. 2009 - Oct. 2010.[16]

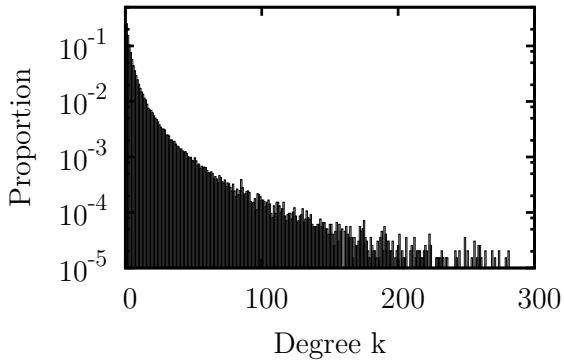
Supplementary Table 14: Gowalla statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
196591	950327	14730	51	$6.3e+09$	14600	0.54

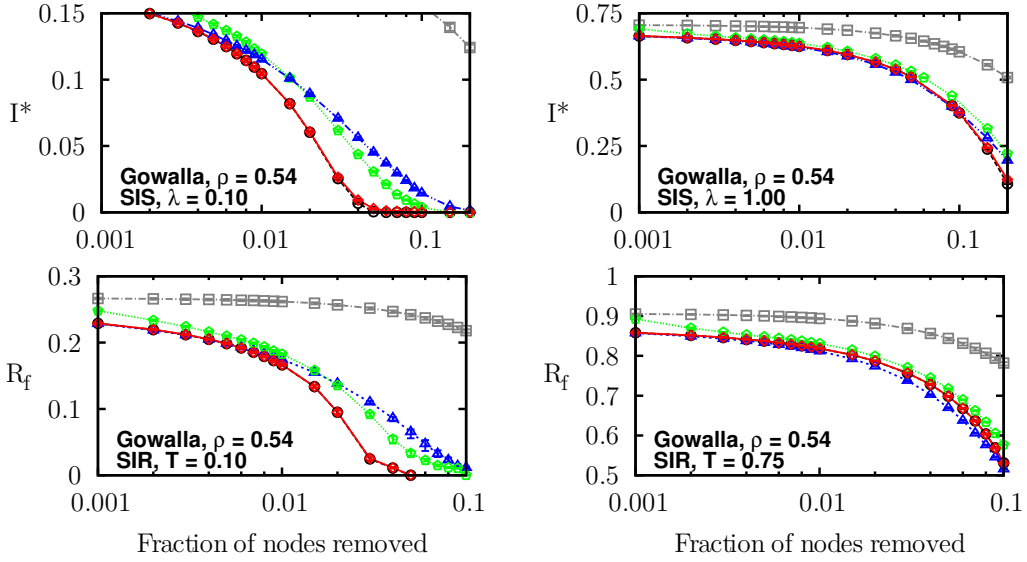
Supplementary Table 15: Gowalla correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9792	0.8363	0.9514	0.8311	0.9724	0.7309

Supplementary Figure 15: Gowalla degree distribution



Supplementary Figure 16: Intervention against epidemics on Gowalla network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[16]Cho, E., Myers, S.A. & Leskovec, J. (2011) Friendship and Mobility: Friendship and Mobility: User Movement in Location-Based Social Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

11 Internet autonomous systems

This dataset is a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted at *archive.routeviews.org*. This snapshot was created by Mark Newman from data for July 22nd 2006.[17]

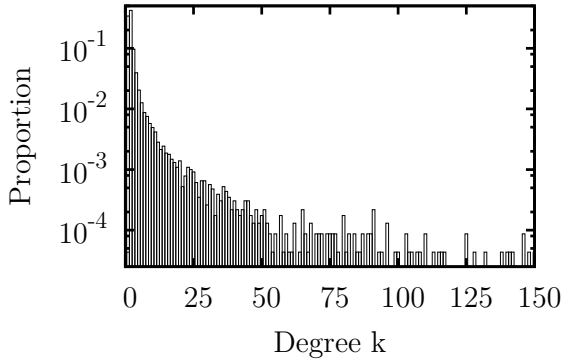
Supplementary Table 16: Internet statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
22963	48436	2390	25	$3.8e+07$	1710	$7e-4$

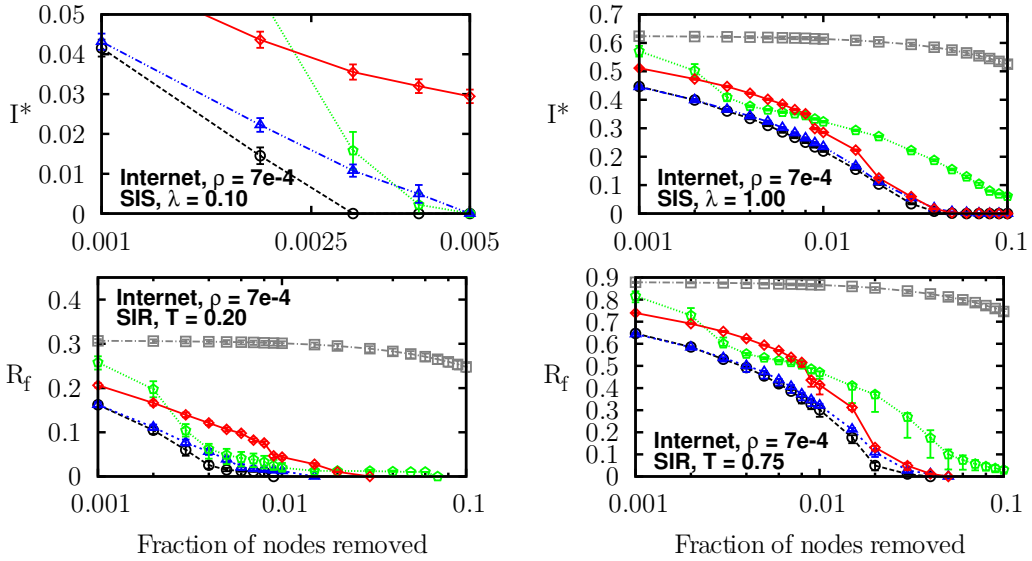
Supplementary Table 17: Internet correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9857	0.7933	0.9469	0.7807	0.9631	0.7079

Supplementary Figure 17: Internet degree distribution



Supplementary Figure 18: Intervention against epidemics on Internet network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[17]Hébert-Dufresne, L., Allard, A., Marceau, V., Noël, P.-A. & Dubé, L.J. (2011) Structural Preferential Attachment: Network Organization beyond the Link. *Phys. Rev. Lett.* 107:158702

12 Internet Movie Database

This dataset details the co-acting network of for movies released after December 31st 1999 as compiled by IMDb.[18][19]

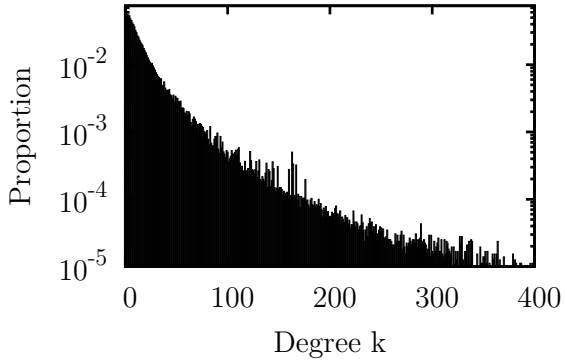
Supplementary Table 18: IMDb statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
716463	7665259	4625	192	N/A	2152	0.52

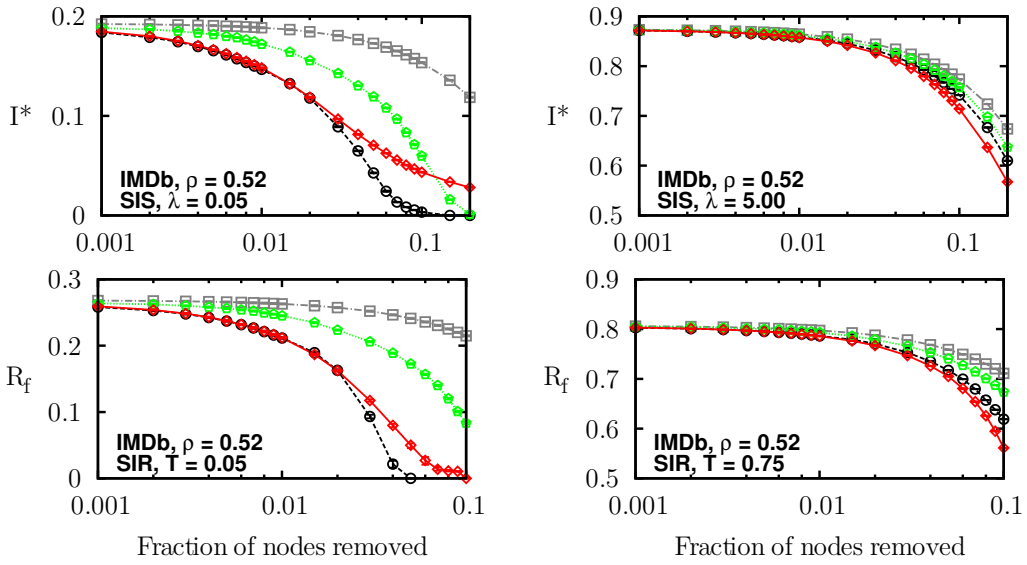
Supplementary Table 19: IMDb correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.6830	N/A	0.6186	N/A	0.9813	N/A

Supplementary Figure 19: IMDb degree distribution



Supplementary Figure 20: Intervention against epidemics on IMDb network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[18] http://www.imdb.com/

[19] Hébert-Dufresne, L., Allard, A., Marceau, V., Noël, P.-A. & Dubé, L.J. (2011) Structural Preferential Attachment: Network Organization beyond the Link. *Phys. Rev. Lett.* 107:158702

13 MathSciNet co-authorship

Co-authorship network of MathSciNet before 2008.[20][21]

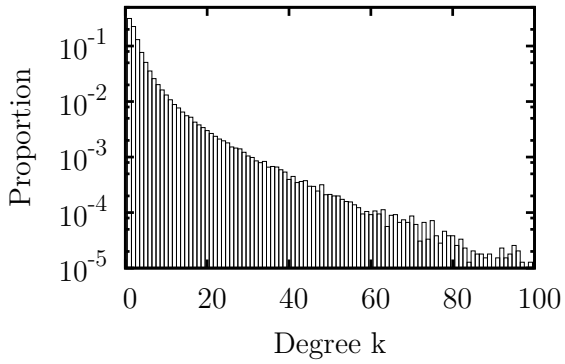
Supplementary Table 20: MathSci statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
391529	873775	496	24	$1.9e+09$	485	0.40

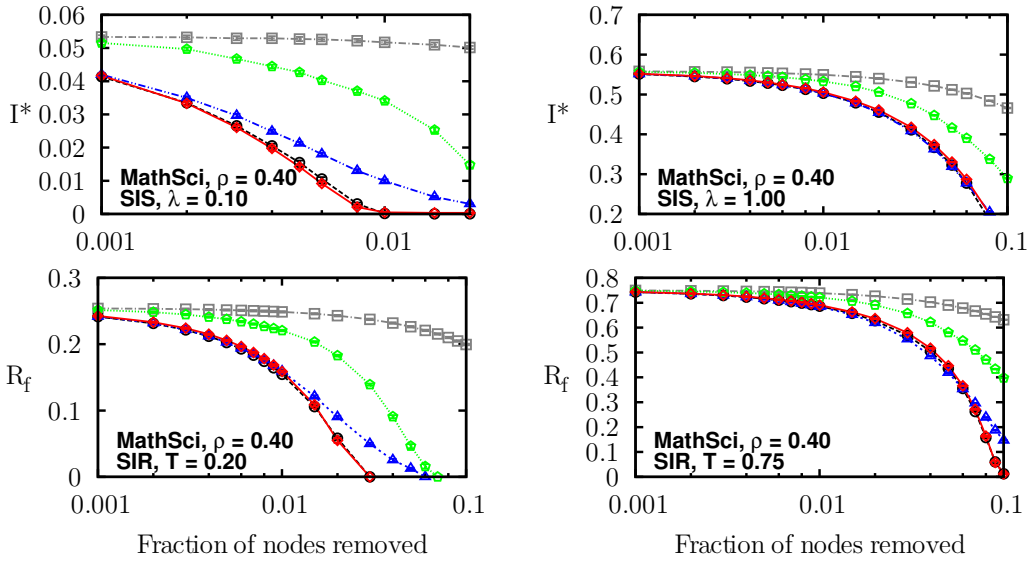
Supplementary Table 21: MathSci correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8645	0.8320	0.7749	0.7835	0.9465	0.6200

Supplementary Figure 21: MathSci degree distribution



Supplementary Figure 22: Intervention against epidemics on MathSci network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[20] http://www.ams.org/mathscinet/

[21] Palla, G., Farkas, I.J., Pollner, P., Derényi, I. & Vicsek, T. (2008) Fundamental statistical features and self-similar properties of tagged networks. *New J. Phys.* 10:123026

14 Myspace online social network

Friendships between the first 100,000 users encountered while crawling Myspace accounts from September to October 2006 (excluding Tom Anderson, the cofounder of MySpace, which is connected to everyone).[22][23]

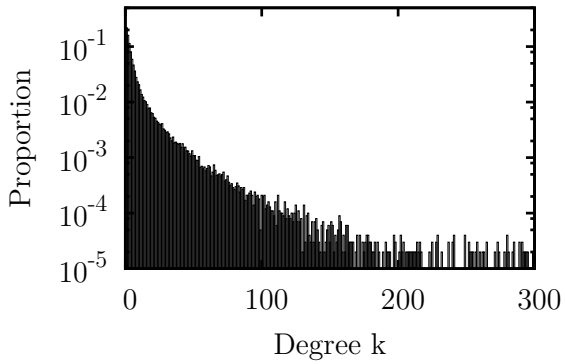
Supplementary Table 22: Myspace statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
100000	841224	59108	78	$2.6e+09$	59102	0.77

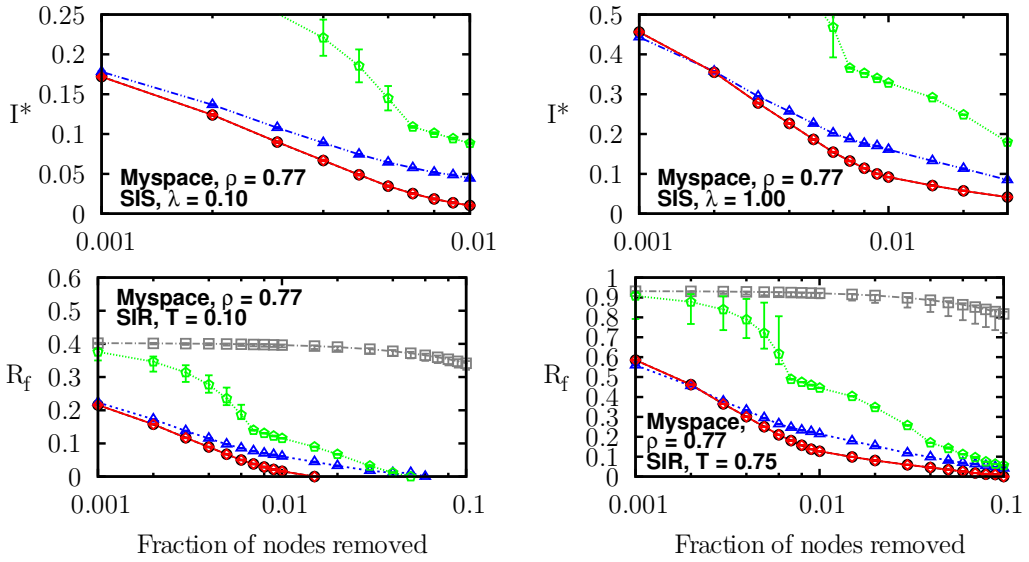
Supplementary Table 23: Myspace correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
1.0000	0.8667	0.9995	0.8667	0.9995	0.8662

Supplementary Figure 23: Myspace degree distribution



Supplementary Figure 24: Intervention against epidemics on Myspace after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[22] http://www.myspace.com/

[23] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S. & Jeong, H. (2007) Analysis of Topological Characteristics of Huge Online Social Networking Services, *Proc. of International World Wide Web Conference*

15 Pretty-Good-Privacy data exchange

Dataset describing the giant component in the network of users of the Pretty-Good-Privacy algorithm for information exchange.[24]

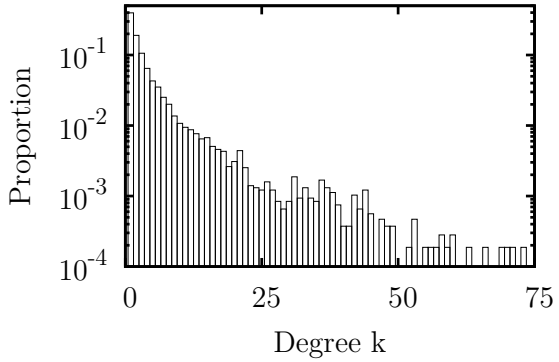
Supplementary Table 24: PGP statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
10680	24316	205	31	$7.5e+06$	110	0.50

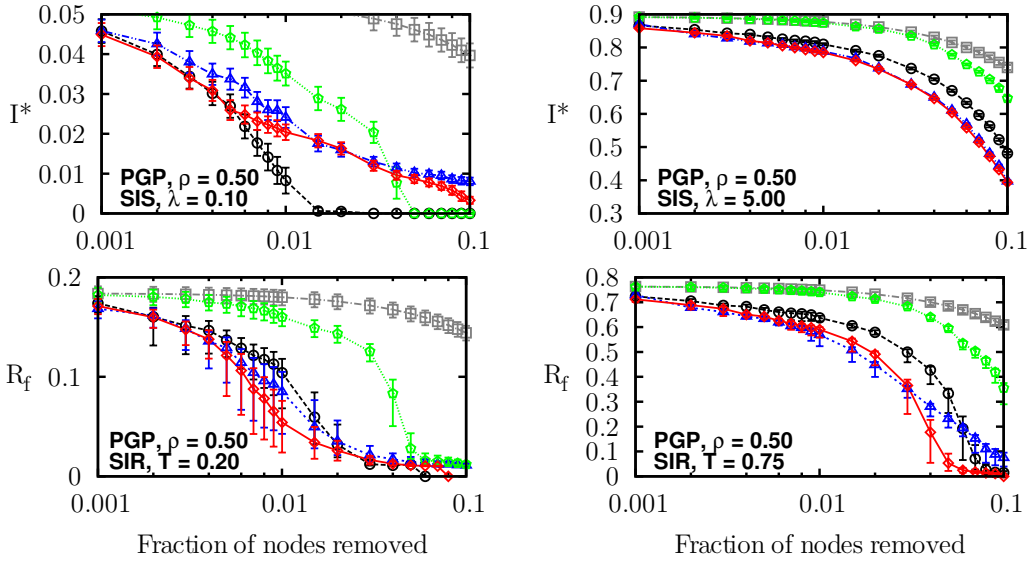
Supplementary Table 25: PGP correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8862	0.8599	0.7256	0.7973	0.8973	0.5464

Supplementary Figure 25: PGP degree distribution



Supplementary Figure 26: Intervention against epidemics on the PGP network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[24]Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A. & Arenas, A. (2004) Models of social networks based on social distance attachment. *Phys. Rev. E* 70:056122

16 Power grid

The topology of the Western States Power Grid of the United States as compiled by Duncan Watts and Steven Strogatz.[25]

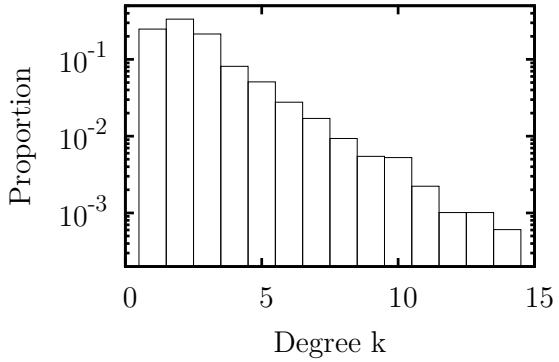
Supplementary Table 26: Power grid statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
4941	6594	19	5	$3.5e+06$	18	0.49

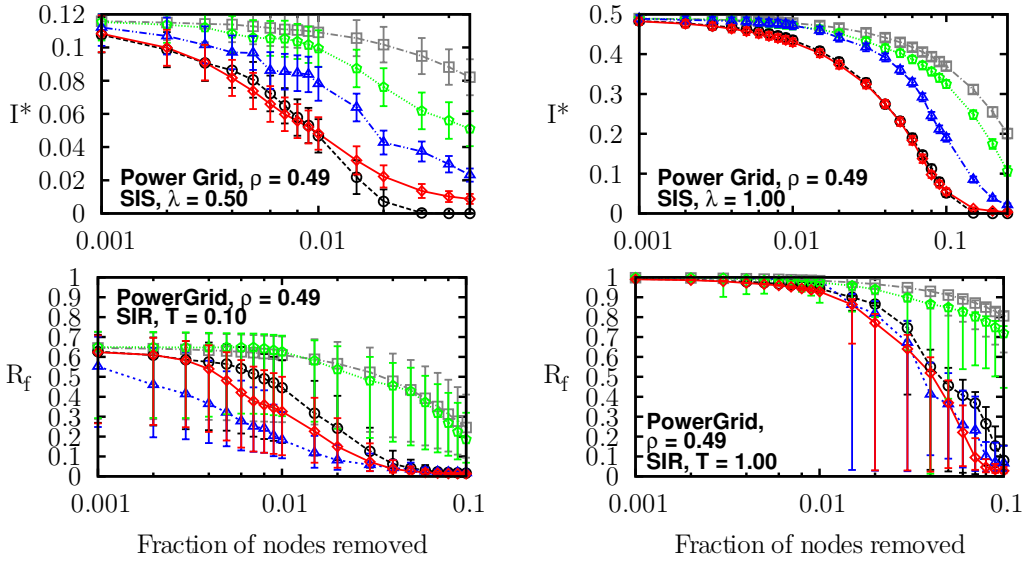
Supplementary Table 27: Power grid correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9192	0.8605	0.6191	0.8042	0.7342	0.5788

Supplementary Figure 27: Power grid degree distribution



Supplementary Figure 28: Intervention against epidemics on the power grid after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[25] Watts, D.J. & Strogatz, S.H. (1998) Collective dynamics of small-world networks. *Nature* 393:440-442

17 Protein interactions network

Protein-protein interactions (ProteinCore) in *S. cerevisiae* as listed by the Database of Interacting Proteins.[26][27]

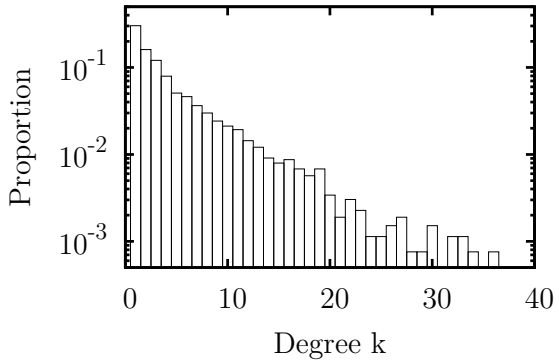
Supplementary Table 28: ProteinCore statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
2640	6600	111	8	$4.0e+05$	71	0.32

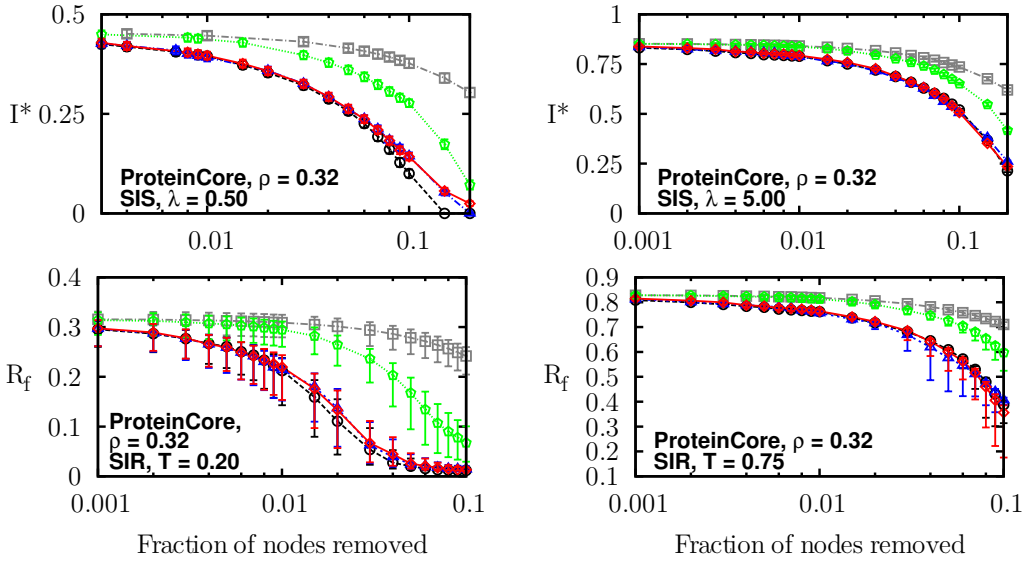
Supplementary Table 29: ProteinCore correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8538	0.9118	0.7702	0.8828	0.9543	0.7712

Supplementary Figure 29: ProteinCore degree distribution



Supplementary Figure 30: Intervention against epidemics on the protein interactions network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[26] http://dip.doe-mbi.ucla.edu/

[27] Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814-818

18 Slashdot online social network

Network of tagged relationships (friends or foes) in the community of the Slashdot news website in November 2008.[28][29]

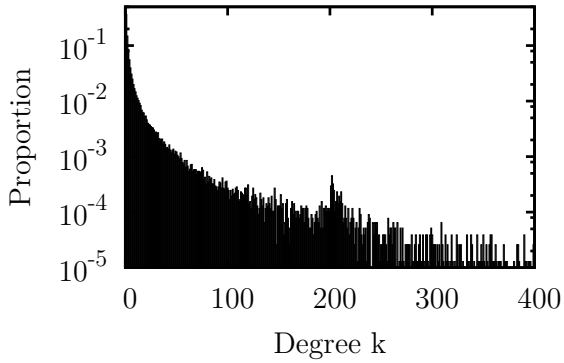
Supplementary Table 30: Slashdot statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
77360	469180	2539	54	$1.2e+08$	2506	0.46

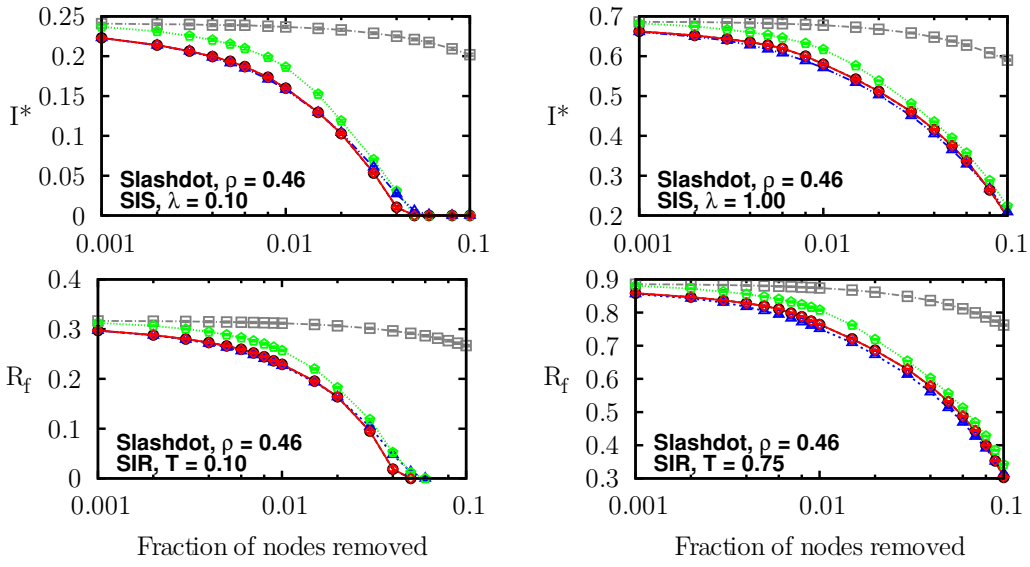
Supplementary Table 31: Slashdot correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9958	0.9373	0.9832	0.9358	0.9870	0.8855

Supplementary Figure 31: Slashdot degree distribution



Supplementary Figure 32: Intervention against epidemics on Slashdot after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[28] http://slashdot.org/

[29] Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. (2009) Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* 6:29-123

19 Word association network

Word association graph (built by survey) obtained from the South Florida Free Association norms.[30][31]

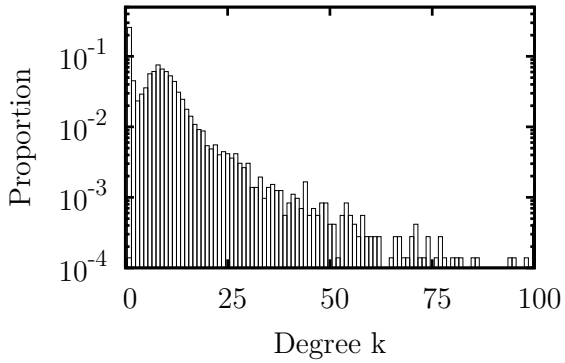
Supplementary Table 32: Word ass. statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
7207	31784	218	7	$1.2e+06$	137	0.16

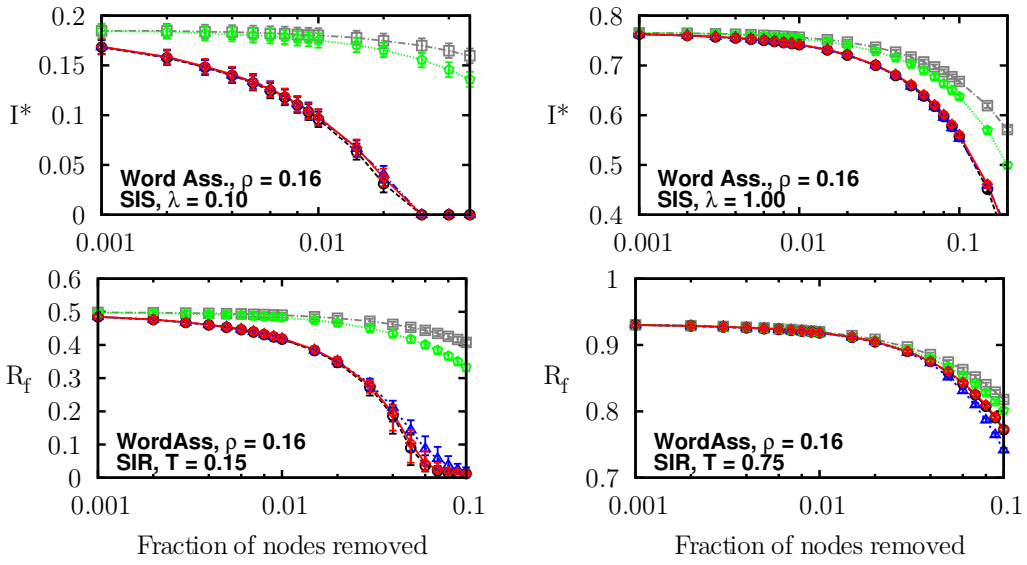
Supplementary Table 33: Word ass. correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9698	0.9230	0.9110	0.9229	0.9281	0.8337

Supplementary Figure 33: Word ass. degree distribution



Supplementary Figure 34: Intervention against epidemics on the word association graph after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[30] http://www.usf.edu/FreeAssociation/

[31] Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814-818

20 World Wide Web

Network of links between the webpages within *nd.edu* domain and considered undirected for this study.[32]

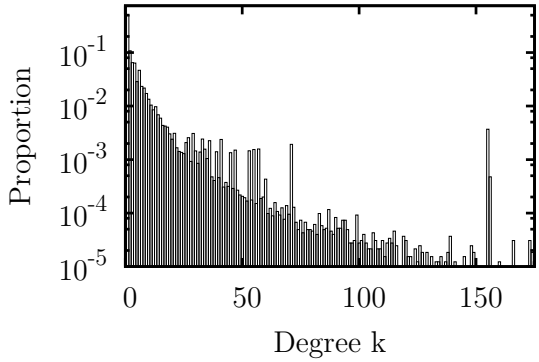
Supplementary Table 34: WWW statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
325729	1090108	10721	155	$2.5e+10$	6993	0.86

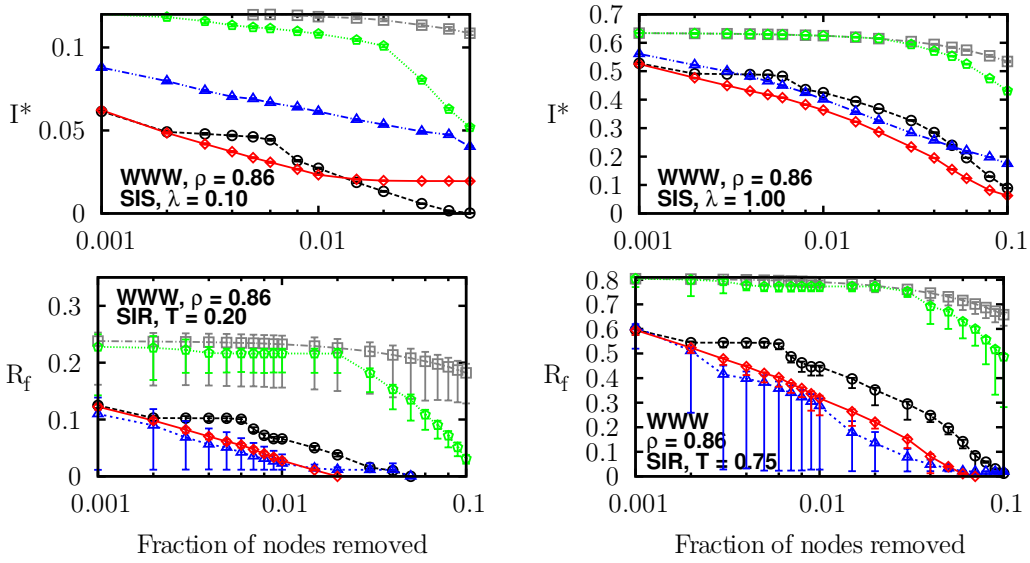
Supplementary Table 35: WWW correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9569	0.8683	0.9020	0.8665	0.9614	0.7905

Supplementary Figure 35: WWW degree distribution



Supplementary Figure 36: Intervention against epidemics on the WWW after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).



[32]Barabási, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science* 286:509-512