nature physics
SUPPLEMENTARY INFORMATION
DOI: 10.1038/NPHYS1746

Identifying influential spreaders in complex networks

SUPPLEMENTARY INFORMATION

I. DATASETS

In this study we have mainly focused on social networks, but our results can be extended to networks from practically any discipline. The datasets that were used in the paper and in this Supplementary Information are the following:

a) *Contact Network of Inpatients.* We use records from Swedish hospitals [16] and establish a link between two inpatients if they have both been hospitalized in the same quarters. We restrict the recording period to one week. All the data have been handled in a de-indentified form. There are 8622 inpatients in the largest component, with an average degree of around 35.1.

b) *IMDB actors in adult films.* We have created a network of connections between actors who have co-starred in films, whose genre has been labeled by the Internet Movie Database [17] as 'adult'. This network is a largely isolated sub-set of the original actor collaboration network. Additionally, all these films have been produced during the last few decades, rendering the network more focused in time. The largest component comprises 47719 actors/actresses in 39397 films. The average degree of the network is 46.0.

c) *Email Contact Network.* The network of email contacts is based on email messages sent and received at the Computer Sciences Department of University College London. The data have been collected in the time window between December 2006 and May 2007. Nodes in the network represent email accounts. We connect two email accounts with an undirected link in the case where at least two emails have been exchanged between the accounts (at least one email in each direction). There are 12701 nodes with an average degree of 3.2.

d) *LiveJournal.com.* The network of friends in the LiveJournal community, as recorded in a 2008 snapshot. We only consider reciprocal links, i.e. when two members are in each other's list of friends. There are 3453394 nodes in the largest component, and the average degree is 12.4.

e) *Cond-mat collaboration network.* This is the network of collaborations between scientists that have posted reprints in the 'cond-mat' e-print archive, between 1995 and 2005. The

DOI: 10.1038/NPHYS1746

Network Name	N	N_E	$< k >$	$< k^2 >$	β_c^{rand}	β	$k_{S_{max}}$
Contact Network of Inpatients	8622	151649	35.1	1633	1.7%	4%	66
Actor Network	47719	1028537	46.0	17483	0.21%	1%	199
Email Contacts	12701	20417	3.2	351.1	0.73%	8%	23
Live Journal	3453394	21378154	12.38	892.45	1.1%	1.5%	100
Cond-mat Collaboration Network	17628	52884	7.0	109.4	5.1%	10%	22
RL Internet	493312	808844	3.3	71.9	4.6%	6%	36
AS Internet	20556	62920	6.1	2111.2	0.23%	n/a	41
Product Space	765	40164	104.8	16931	0.50%	n/a	100

TABLE I: Properties of the real-world networks studied in this work. Here N is the number of nodes, N_E is the number of edges, $< k >$ is the average degree in the network, $< k^2 >$ is the average squared degree in the network, β_c^{rand} is the epidemic threshold for a corresponding random network ($\beta_c^{\text{rand}} \approx \lambda < k > / < k^2 >$), $\lambda = 0.8$ in SIS simulations, β is the value we used in SIR simulations and $k_{S_{max}}$ is the highest k-shell index of the network. We consider only the largest connected cluster of the network if the original network is disconnected.

nodes of the network represent the authors, who are connected if they have co-authored at least one paper. The cond-mat collaboration dataset consists of 17628 authors with average degree 6.0

f) *The Internet at the router level (RL).* The nodes of the RL Internet network are the Internet routers. Two routers are connected if there exists a physical connection between them. Data have been gathered from the DIMES project [13]. The largest connected component of the analyzed dataset contains 493312 routers with an average degree of 3.3.

g) *The Internet at the autonomous system level (AS).* The nodes are autonomous systems which are connected if there exists a physical connection between them. An autonomous system is a collection of connected IP routing prefixes under the control of one or more network operators that presents a common, clearly defined routing policy to the Internet. Data have been gathered by the DIMES project [13]. The largest connected component of the AS Internet consists of 20556 autonomous systems with average degree 6.1.

h) *Product space of economic goods.* This is the network of proximity between products according to Ref. [22]. We use a proximity threshold 0.3, and we recover similar results for

different thresholds, as well.

We outline some of the basic properties for these networks in Table I.

II. THE k-SHELL DECOMPOSITION METHOD

In order to classify the nodes into k-shells we employ the k-shell decomposition algorithm. First, we remove all nodes with degree k=1. After this first stage of pruning there may appear new nodes with k=1. We keep on pruning these nodes, as well, until all nodes with degree k=1 are removed. The removed nodes along with the links connecting them form the $k_S = 1$ k-shell. Next, we repeat the pruning process in a similar way for the nodes of degree k=2 to extract the $k_S = 2$ k-shell and subsequently for higher values of k until all nodes are removed. As a result, the network can be viewed as a set of adjacent k-shells (see Fig. 5).

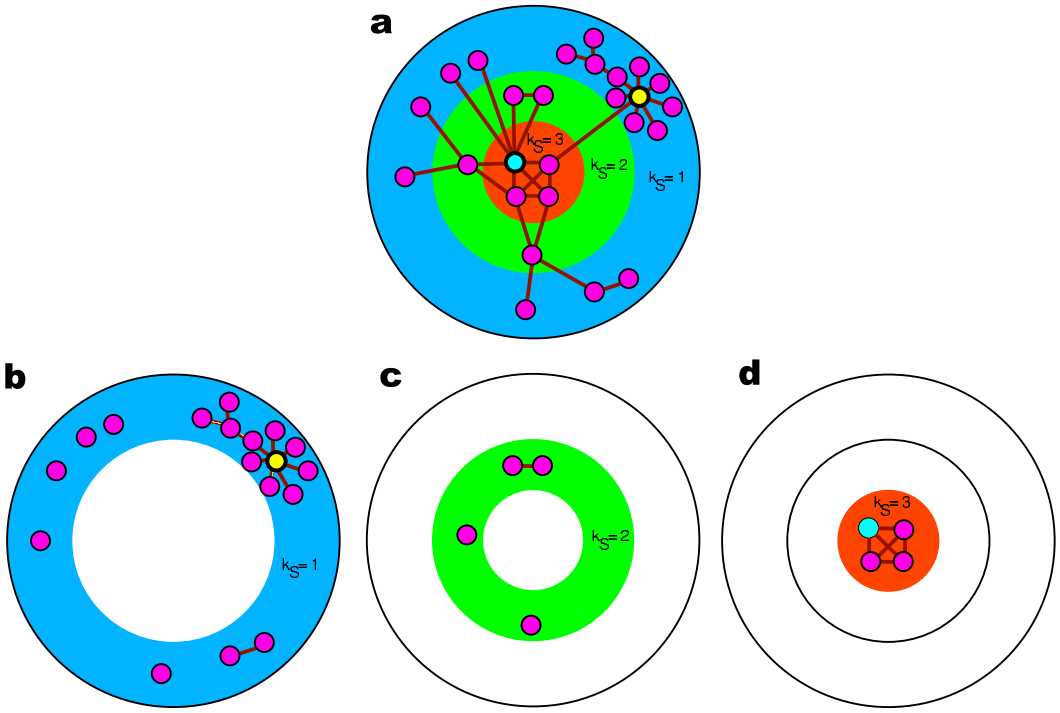

FIG. 5: **The illustration of the k-shell extraction method.** **a,** A schematic network is represented as a set of 3 successively enclosed k-shells labeled accordingly. **b,** Nodes with edges forming $k_S = 1$ shell of the network. **c,** Nodes with edges forming $k_S = 2$ shell of the network. **d,** Nodes with edges forming $k_S = 3$ shell of the network.

The k-shell decomposition method assigns a unique k_S value to each node, that corre-

DOI: 10.1038/NPHYS1746

sponds to the index of the k-shell this node belongs to. The k_S index provides a different type of information on a node than that provided by the degree k. By definition, a given layer with index k_S can be occupied with nodes of degree $k \geq k_S$. In the case of random model networks, such as the configurational model, there is a strong correlation between k and the k_S index of a node and, therefore, both quantities provide the same type of information. Thus, the low-degree nodes are generally in the periphery, and the high-degree nodes are generally in the innermost k-shells. In real networks, however, this relation is often not true. In real networks hubs may have very different k_S values and can be located both in the periphery (yellow node in Fig. 5) or in the core (blue node in Fig. 5) of the network.

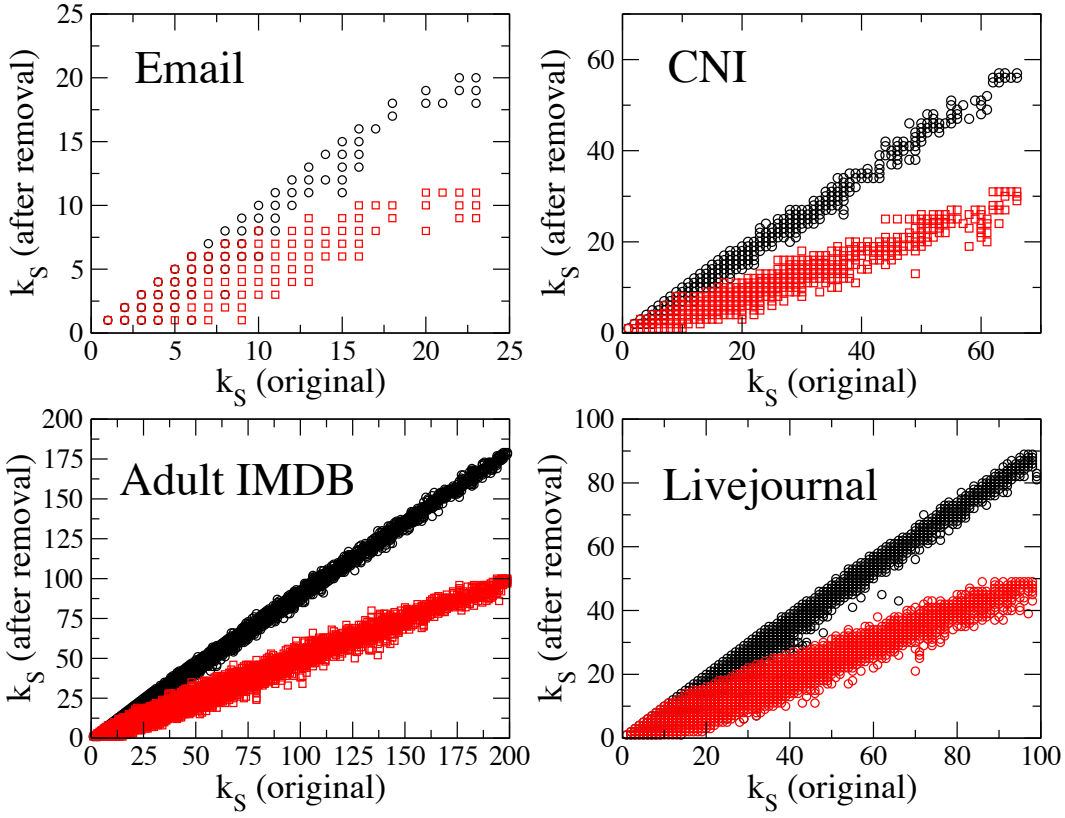

FIG. 6: **Robustness of k_S under incomplete network information**. We randomly remove 10% of the network links and 50% of the network links (results shown in black and red symbols, respectively). The relative ranking of the nodes remains invariant under both removals, for all the networks studied: Email, Hospital, Adult IMDB, and Livejournal.com.

The assignment of a k_S index to a node is also quite robust. We have randomly removed 10% and 50% of the links in the networks that we study, simulating thus incomplete infor-

mation. When we measure the new k_S value for the same nodes in the resulting networks (Fig. 6) we find that their relative ranking remains the same. We recover a practically linear dependence on the k_S values of the original and the incomplete networks, showing that this measure would work equally well for predicting the spreading efficiency of nodes in a network with missing information.

III. PROBABILITY AND TIME OF INFECTION

We have demonstrated that the location of a node, as described through the k_S index, is important for the extent of spreading M_i when this node is the spreading origin. Here, we show that nodes with high k_S are more probable to be infected during an epidemic outbreak and are infected earlier than nodes with low k_S, when spreading starts at a random node. We introduce the quantity E_i, as the probability that a node i is going to be infected during an epidemic outbreak originating at a random location, and T_i, as the average time before node i is infected during the same process.

As shown in Figs. 7a-d all three quantities that characterize the role of a node in an epidemics process, M_i, E_i and T_i are strongly correlated. The nodes that are infected by a given node i form a cluster of size $\overline{M_i}$, and they are statistically the nodes that can reach i when they act as origins themselves. Thus, the probability E_i to reach this node in general is directly proportional to the size M_i, as shown in the plots. The average time T_i to reach a node is inversely proportional to its spreading efficiency M_i, which emphasizes the fact that these nodes are easily reachable from different network locations. In conclusion, the nodes with the largest k_S values consistently a) are infecting larger parts of the network, b) are infected more frequently, and c) are infected earlier, than nodes with smaller k_S values.

IV. THE IMPRECISION FUNCTIONS

We quantify the spreading efficiency of an individual origin i through the infected number of nodes M_i. In order to compare the different methods, we rank all network nodes according to their spreading efficiency, independently of their other properties, and we consider a fraction p of the most efficient spreaders ($p \in [0,1]$). We designate this set by $\Upsilon_{eff}(p)$. Similarly, we define $\Upsilon_{k_S}(p)$ as the set of individuals with highest k-shell values. In order to

DOI: 10.1038/NPHYS1746

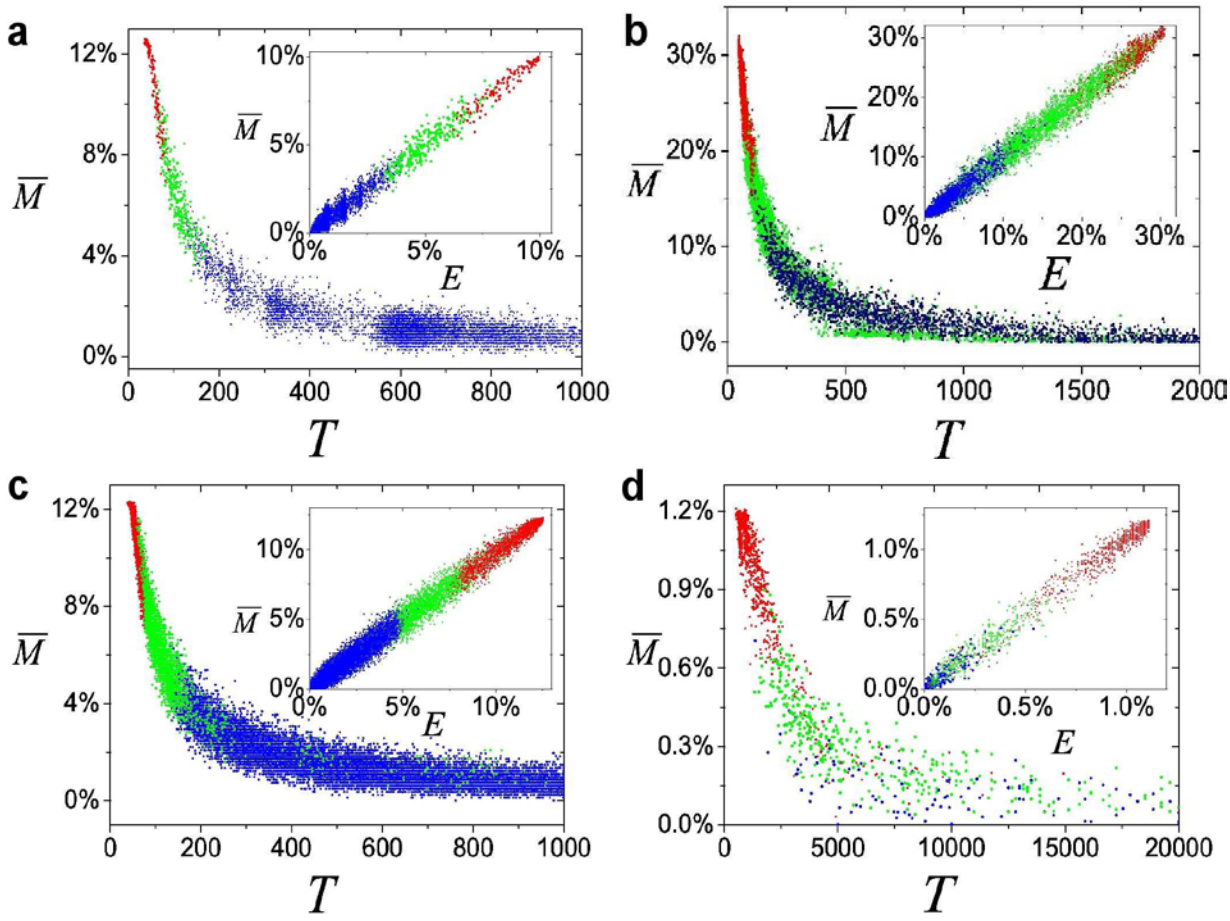

FIG. 7: Cross-plots of M_i as a function of T_i, and M_i as a function of E_i (inset) for a) email, b) hospital inpatients, c) actor network and d) RL Internet. Every point denotes the corresponding quantities for a given node, and the color denotes the k-shell index of this node. The k_S values are aggregated and highlighted with red (large k_S regime), green (intermediate k_S regime) and blue (low k_S values) colors, respectively. A high level of correlation between M_i and E_i indicates that the most efficient spreaders (as measured by M_i) are the most likely to be infected during an epidemic outbreak originating at random inpatient in the network. On the other hand, the anti-correlation between M_i and T_i indicates that the most efficient spreaders are typically infected earlier than other nodes during an epidemic outbreak.

assess the merit of using k-shell decomposition to identify the most efficient SIR spreaders one needs to compare the two sets $\Upsilon_{eff}(p)$ and $\Upsilon_{k_S}(p)$. In order to consider individual M_i values, we calculate the average $M_{eff}(p)$ and $M_{k_S}(p)$ values for the sets $\Upsilon_{eff}(p)$ and $\Upsilon_{k_S}(p)$ respectively: $M_{k_S}(p) \equiv \sum_{i\in\Upsilon_{k_S}(p)} M_i/Np$ and $M_{eff}(p) \equiv \sum_{i\in\Upsilon_{eff}(p)} M_i/Np$, where Np is

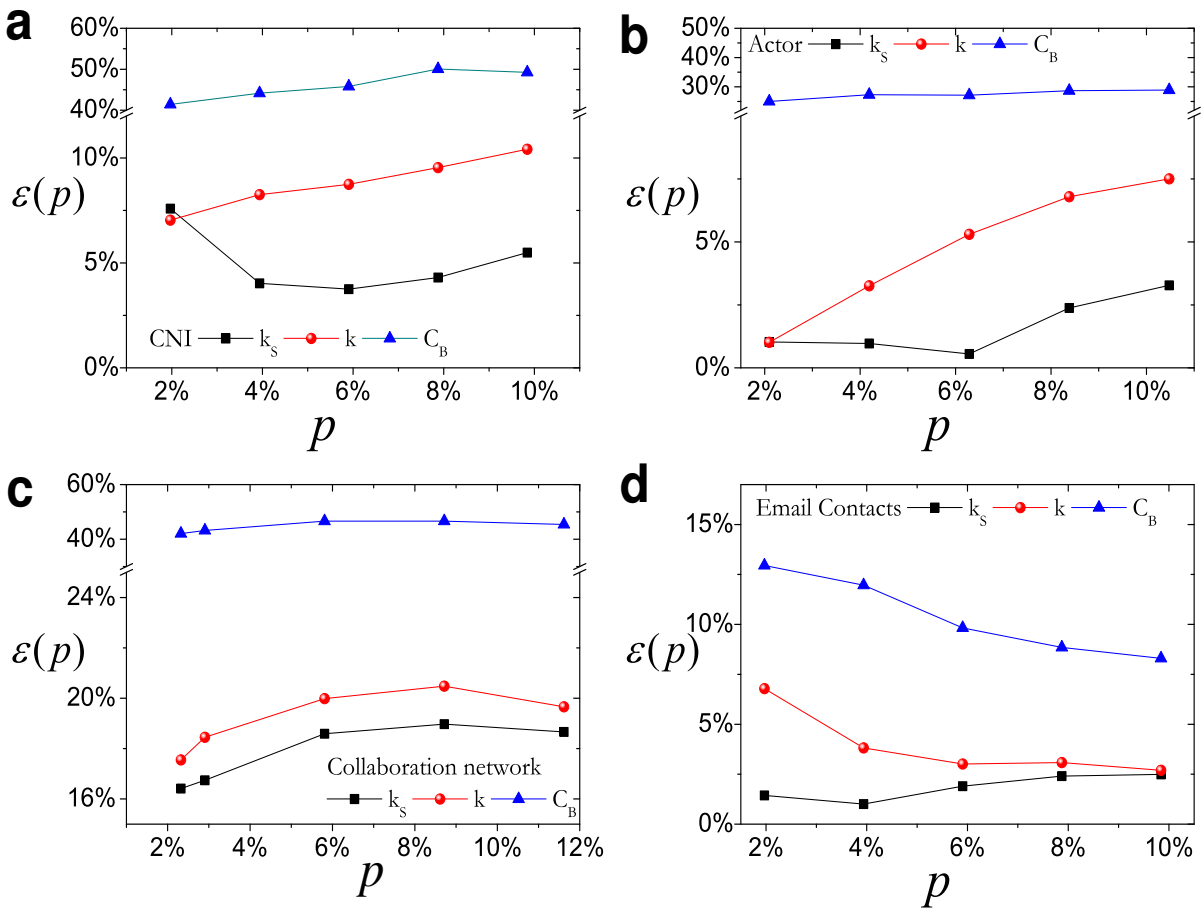

FIG. 8: **The imprecision functions $\epsilon(p)$ test the merit of using k-shell, k and C_B to identify the most efficient spreaders in the CNI, actor, collaboration, and email contact networks.** The k-shell based identification method yields consistently lower imprecision compared to the k and C_B based methods.

the number of nodes that we consider in the comparison. By definition, $M_{eff}(p) \geq M_{k_S}(p)$, and the equality is only reached if $\Upsilon_{eff}(p) = \Upsilon_{k_S}(p)$. We assess the imprecision of k-shell identification by calculating the ratio between $M_{eff}(p)$ and $M_{k_S}(p)$:

$$\epsilon_{k_S}(p) \equiv 1 - \frac{M_{k_S}(p)}{M_{eff}(p)}. \tag{4}$$

Similarly, we can define $\epsilon_k(p)$ and $\epsilon_{C_B}(p)$:

$$\epsilon_k(p) \equiv 1 - \frac{M_k(p)}{M_{eff}(p)}\,,\ \epsilon_{C_B}(p) \equiv 1 - \frac{M_{C_B}(p)}{M_{eff}(p)}. \tag{5}$$

A value for ϵ close to 0 denotes a very efficient process, since the nodes that are chosen are practically those that contribute most to epidemics. In all cases, the k_S method yields a

DOI: 10.1038/NPHYS1746

spreading that is closer to the optimum than either the degree or the betweenness centrality. Additionally, this behavior is independent on the fraction of spreaders p that we consider in each case.

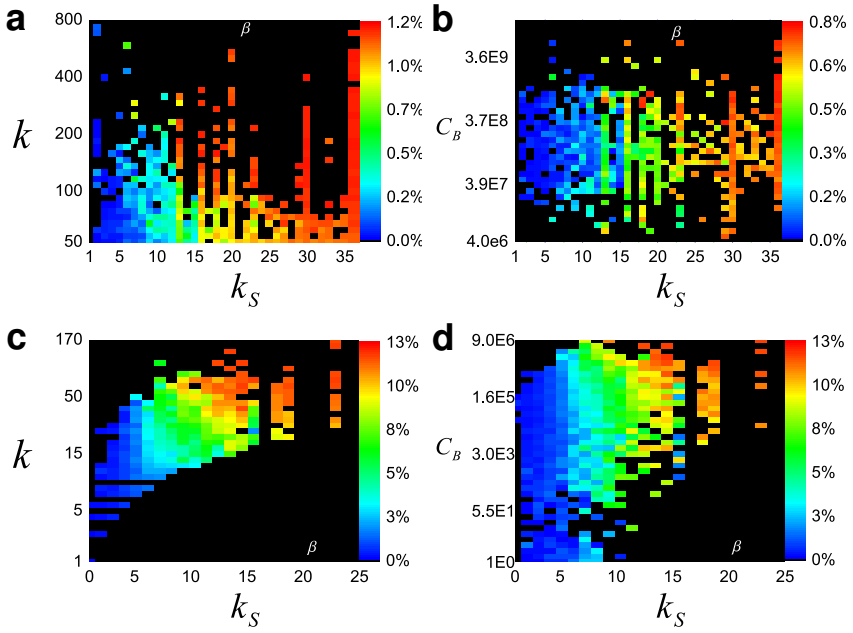

FIG. 9: **The shell index k_S predicts the outcome of spreading more reliably than the degree k or the betweenness centrality C_B.** The networks that were analyzed are: (**a, b**) the RL Internet and (**c, d**) the collaboration network. **a** and **c**, The average infected size $M(k_S, k)$ as a function of (k_S,k) values of the infection origin nodes. **b** and **d**, The average infected size $M(k_S, C_B)$ as a function of (k_S,C_B) values of the infection origin nodes.

V. SIR SPREADING EFFICIENCY

In the main text we present results for $M(k_S, k)$ for the email network, the CNI, the actor network and the Livejournal network. Here, we present additional results of the k-shell analysis of the Internet at the Router Level (RL) and the scientific collaboration network. Figure 9 shows the results for $M(k_S, k)$ and $M(k_S, C_B)$. The conclusion on the spreading importance of high k_S nodes is exactly the same as for the social networks in the main text.

The results on the nodes efficiency are not significantly influenced by the choice of the infected probability value, β. In Fig. 10 we present the infected percentage M for different networks, as an average over nodes that belong in the same k_S range, for different β values.

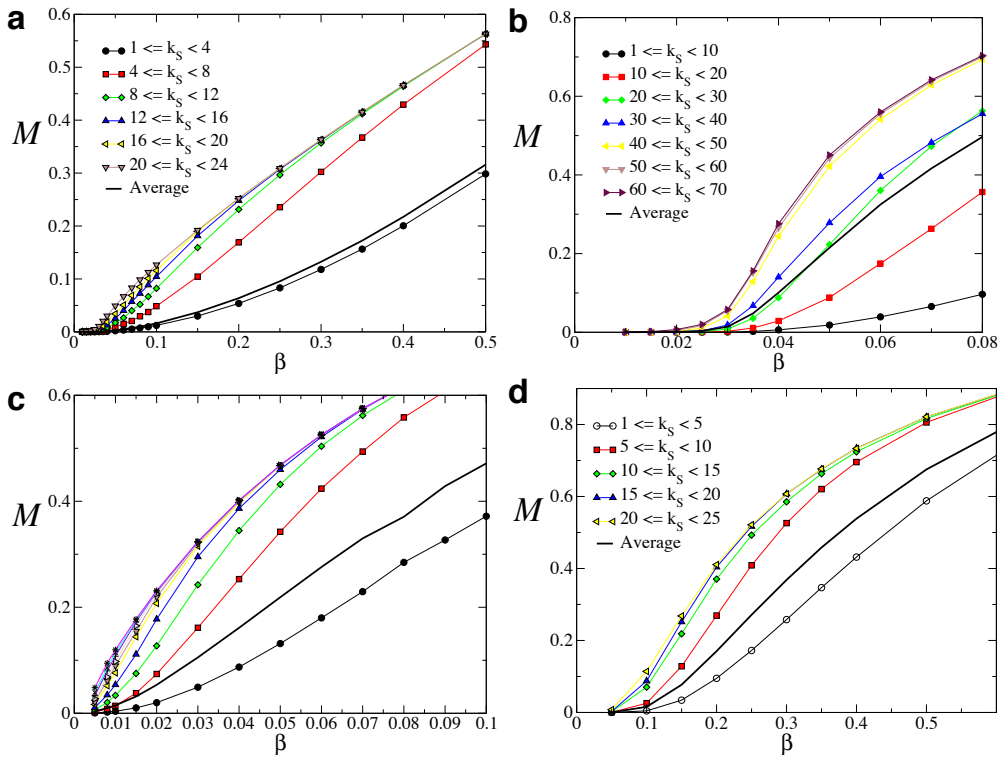

FIG. 10: **The infected percentage is always higher in higher k-shells, independently of the infection probability β.** Nodes are grouped according to their k-shell and we calculate the average infected percentage for each group as a function of β. The solid lines correspond to the grand average over all nodes acting as spreading origins. The networks that were analyzed are: **a,** the email network, **b,** the CNI, **c,** the adult IMDB actors network, and **d,** the cond-mat collaboration network.

The nodes in higher k-shells are consistently reaching a larger fraction of the network. Our main interest is in the β range where we are above the critical point, $\langle M \rangle > 0$, but the average infection reaches a finite but small fraction, in the range of 1-20%. When the average spreading is even larger, nodes of lower k-shells can become efficient too, because in this case there is a high probability to reach the 'core' of the network, and this would enable the spreading to extend over an even larger part of the network.

For β values in this 'intermediate' range, the distribution $P(M)$ of the infected percentage M is composed by two well-defined peaks (Fig. 11). The first is at $M = 0$ and corresponds to those instances where the infection dies within the first few infection steps. The second peak is at a finite fraction M, and it seems to be at the same point for all origins. However, the intensity of each peak strongly differs, depending on the k_S value of the origin. For

DOI: 10.1038/NPHYS1746

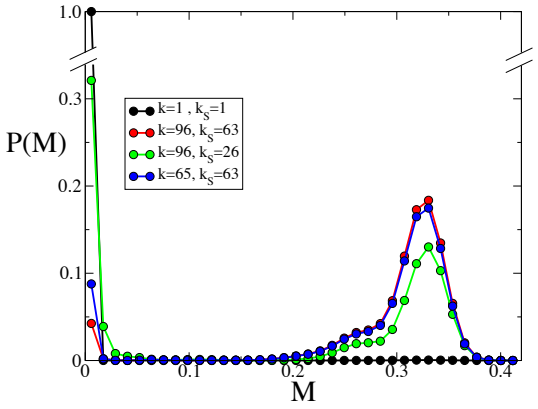

FIG. 11: **Distribution of spreading based on individual origins.** The probability distribution $P(M)$ of the infected percentage for the contact network of inpatients, when the epidemic starts at four nodes of different properties. The infection probability is $\beta = 4\%$, which is above the critical threshold. All distributions exhibit two peaks at similar ranges every time, i.e. around $M = 0$ (epidemics dies very fast) and $M \simeq 33\%$. However, the intensity of each peak differs, and in higher k-shells the majority of the realizations result in large infections, compared to the much higher ratio of zero-spreading realizations for origins of small k_S values.

the higher k_S value in the plot, the stronger peak is at the non-zero value, and very few realizations end up at $M = 0$ even for smaller degrees. On the contrary, an origin with larger degree k, but smaller k_S value results in a stronger peak at $M = 0$. These distributions converge quite well, and we can expect that nodes with small k_S will in general result in a higher peak at $M = 0$. The above means that if an infection can reach a critical mass of nodes then it will eventually cover a significant part of the network. The low k-shell nodes cannot reach this critical mass so that the infection dies at the early stages, resulting to the strong peak at $M = 0$. On the contrary, the neighborhood of high k-shell nodes is favorable for sustaining an infection at early stages, allowing the system to reach this critical mass.

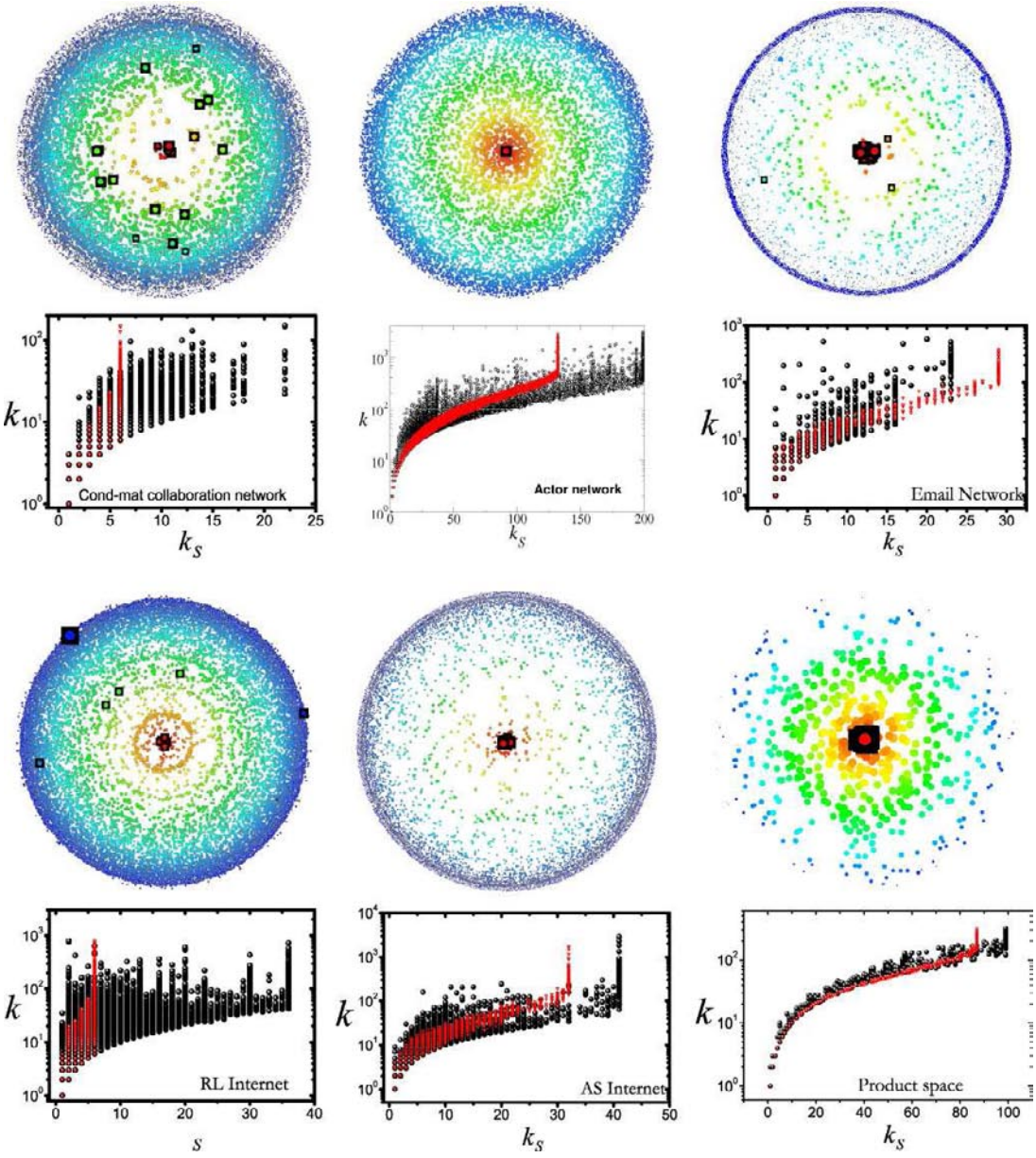

FIG. 12: **k-shell structure of the analyzed networks.** (**Top row**): Visualization of the k-shell structure. We represent networks as sets of concentric circles of nodes, each one corresponding to the particular k-shell, with low k_S values in the periphery and large k_S values towards the center of the network. The size of each visualized node is proportional to the logarithm of its degree value. We highlight the 25 highest degree nodes with black squares. Many of the hubs are found in outer layers. (**Bottom row**): Scatter plots of node degree k as a function of its k-shell index k_S for the original networks (black symbols) and the degree-preserving randomized version of the networks (red symbols). The networks correspond to: the cond-mat collaboration network, the actor network, the email contact network, the RL Internet, the AS Internet, and the Product Space network.

DOI: 10.1038/NPHYS1746

We also highlight the location of the 25 largest hubs in the k-shell structure of the studied networks. Fig. 12 shows the results for the collaboration, actor, email, RL Internet, AS Internet, and Product space networks. High-degree nodes in most of the studied networks are scattered at different k-shells: the high-k nodes appear both in the periphery (starting as low as $k_S = 1$) and in the network center (large k_S value). In certain cases, such as in the actors network, the largest hubs are located in the highest k_S layers. The relation of k_S and k in the AS Internet and the product space is strongly monotonic, and there are very few nodes where k_S is large or small compared to the degree k. This is a typical behavior for random networks, and the structure of these two networks is significantly close to their randomized counterparts. In these cases, choosing a node based on its degree or its k-shell index does not make a difference, since they practically lead to the same nodes.

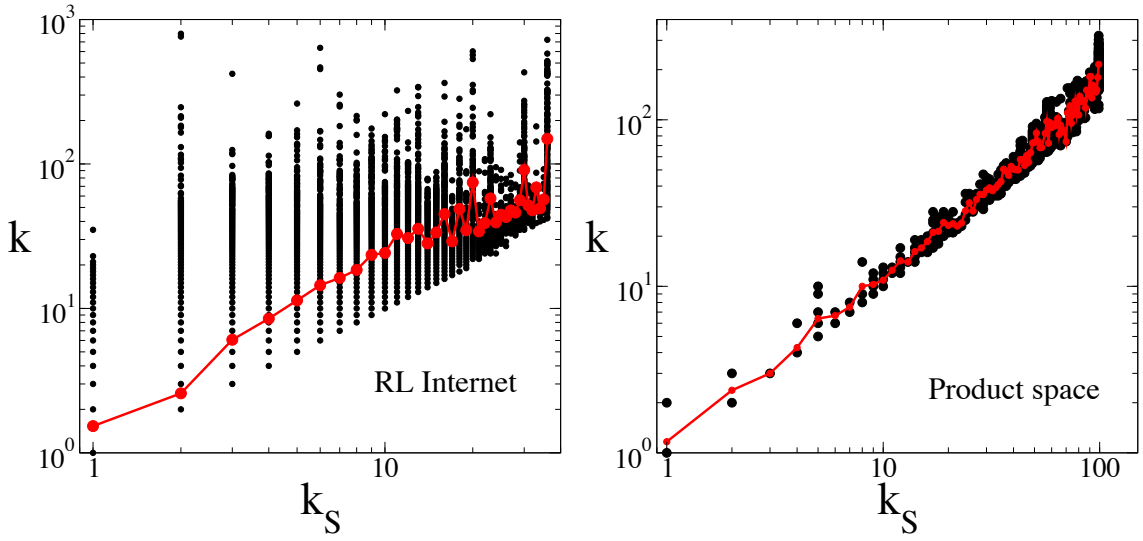

FIG. 13: **Deviations from the average behavior highlight the importance of the k-shell structure.** The average degree (red symbols) for a given k_S index follows roughly a power-law dependence, as a function of k_S. The deviation from this behavior can be significant, e.g. in RL internet, or negligible, as e.g. in the product space network.

It is clear that the assortative behavior in a network can influence the extent to which hubs will appear in the periphery or in the core of a network. In principle, in a highly disassortative network we expect more hubs in the periphery, due to their tendency to connect to low-degree nodes. However, even in assortative networks it is possible that some hubs may still belong to low k-shells, so that the k_S value will appropriately rank even these

exceptions. The average degree of the nodes in a specific shell follows roughly a power law with k_S (Fig. 13). The deviations from this average behavior emphasize the importance of spreaders within the core of the network having high values of k_S and potentially smaller degrees, than those with high k and low k_S values.

The complex organization of the nodes in the k-shells is highlighted when we randomly rewire the links in the networks, yet preserving the nodes degree. This rewiring 'restores' all the hubs to the innermost k-shell of the system and imposes a strict hierarchy of nodes in terms of both k and k_S. The bottom row of plots in Fig. 12 shows the scatter-plots of degree k as a function of k-shell index k_S for every node in the network. In all cases, a monotonic relation of k vs k_S is followed in the 'rewired' networks (red symbols), where now all the hubs appear in the highest k-shell) as opposed to the weak correlation between k and k_S in the original networks (shown in black).

VI. REWIRING HIGHLIGHTS THE IMPORTANCE OF k-SHELL

In Figs. 1a and 1b of the main text we show that the extent of infection can be remarkably different, although we start from two origins with similar degree. The importance of the structure in the dynamics of spreading can be highlighted if we randomly rewire the network. During this process the original degrees of all nodes are preserved, but random neighbors are chosen for each node, destroying thus any correlations and any patterns in the local connectivity. We denote by $P(M|i)$ the probability that a percentage M of the total population will be infected if a disease originates on node i. In Figs. 1a,b of the main text and in Fig.14a we show that two nodes #1 and #2 with similar degree may yield markedly different distributions $P(M|1)$ and $P(M|2)$. After rewiring, these distributions become practically indistinguishable (see Fig. 14b).

VII. VIRUS PERSISTENCE IN SIS

Many infectious diseases, including most sexually transmitted infections, do not confer immunity after infection, so that they cannot be described via the SIR model. These cases are better simulated through the SIS epidemic model [18]. The dynamics of SIS epidemics is different, since the number of infected nodes eventually reaches a dynamic equilibrium

DOI: 10.1038/NPHYS1746

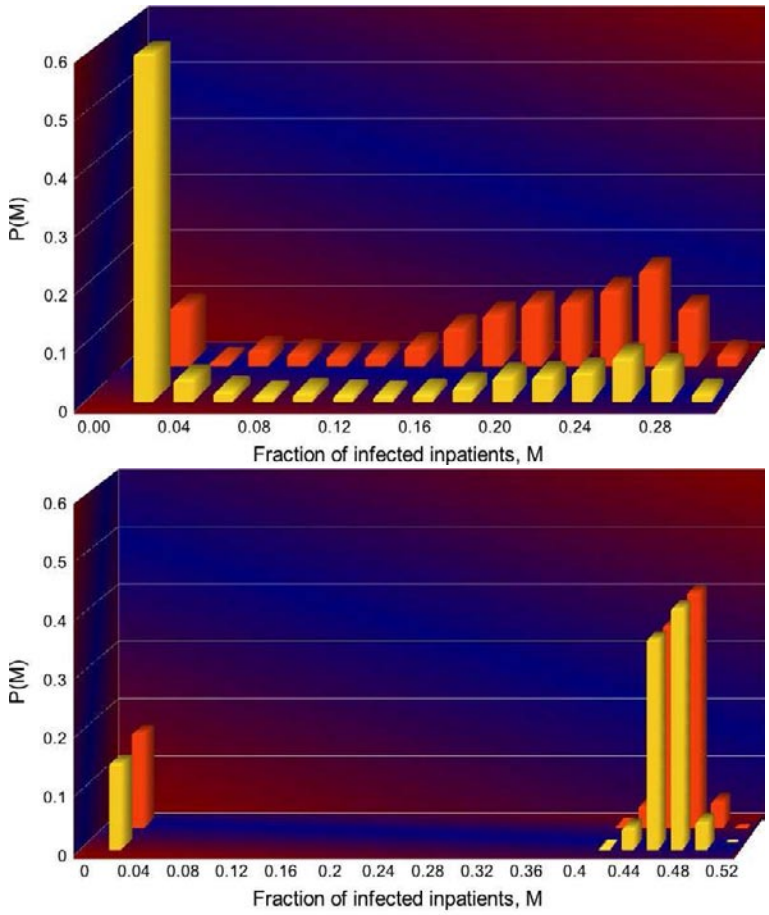

FIG. 14: **Why the hubs may not be good spreaders.** The probability distribution $P(M|i)$ of the infected percentage for the contact network of inpatients, when the epidemic starts at two of the origin hubs in Fig. 1 $i = A, B$ with the same degree ($k = 96$), but different k_S values ($k_S = 63$ and $k_S = 26$, respectively). In each histogram, we use 1000 random realizations of the simulation, starting an SIR epidemic from the same given origin i. Despite the fact that the two origins of the epidemic spreading have the same degree, the two histograms present a radically different character. In one case (red histogram), the hub infects up to 30% of the population, while most of the spreading attempts from the other hub (yellow histogram) practically cannot propagate the infection at all. The importance of the organization of the network is highlighted when we randomly rewire the network (preserving the same degree for all nodes). In this case both distributions $P(M|A)$ and $P(M|B)$ coincide and both hubs contribute equally to spreading. Notice also that spreading in the rewired network extends over a much larger size of the population.

"endemic" state at which exactly as many infectious individuals become susceptible as sus-

ceptible nodes become infected [18]. The quantity characterizing the role of nodes in SIS spreading is the persistence, $\rho_i(t)$, defined as the probability that node i is infected at time t [7]. In an endemic SIS state, which is reached asymptotically, ρ_i becomes independent of t. The persistence ρ has been shown to be higher in hubs which are reinfected frequently due to the large number of their neighbors [7, 24, 25]. To uncover the role of k-shell layers in SIS spreading we use the joint persistence function

$$\rho(k_S, k) \equiv \sum_{i \in \Upsilon(k_S,k)} \frac{\rho_i}{N(k_S, k)}. \quad (6)$$

Here we present results for the virus persistence in the Actor, Collaboration, Email and RL Internet Networks. Similar to Fig. 4, we depict $\rho(k_S, k)$ in both supercritical ($\beta > \beta_c$) and subcritical ($\beta < \beta_c$) regimes, where β_c is the critical threshold. In the supercritical regime, $\rho(k, k_S)$ increases with both k and k_S, with maximum values corresponding to hubs in the innermost layers (see Fig. 15). As depicted in Fig. 15, in the subcritical regime, viruses persist only in the highest k_S layers, while the probability of finding an infected node in low k-shells is negligible.

In order to determine in the above networks the actual epidemic threshold β_c we study the behavior of SIS spreading over a wide range of β values. In order to highlight the role of k-shells in spreading, we organize several groups of nodes based on the k_S layers of each network. Every such group comprises approximately 100 randomly chosen nodes with the corresponding k-shell indices. In order to achieve similar average degree in each of the groups, we pick nodes with uniform probability based on their degree. As shown in Fig. 16, virus persistence is consistently higher in the inner k-shells for all values of β. Moreover, we find substantially lower epidemic thresholds than in the random cases $\beta_c < \beta_c^{\text{rand}}$ in all considered networks except for the Email Contact network.

The results of Figs. 15 and 16 suggest that the observed persistence of a virus is due to the dense sub-network formed by nodes in the innermost k-shell, which helps the virus to consistently survive locally in this area. Indeed, the innermost layers can be regarded as a small subgraph exclusively consisting of hubs. By definition, all nodes in this innermost k-shell will have degrees $k \geq k_{S_{max}}$. Therefore, as a simple approximation, one can regard the innermost core of a network as a regular graph consisting of nodes with the same degree $k = k_{S_{max}}$.

The mean-field solution of the SIS spreading in a regular graph can be found, for instance

DOI: 10.1038/NPHYS1746

in Ref. [24]. We reproduce this solution below for the sake of convenience.

The master equation describing the time evolution at a mean-field level of the average density of infected individuals $\rho(t)$:

$$\frac{d\rho(t)}{dt} = -\rho(t) + \beta k \rho(t)(1-\rho(t)), \tag{7}$$

where k is the degree of all nodes in the regular graph. The first term on the right hand side of Eq. (7) accounts for infected nodes becoming healthy. The second term on the right hand side of Eq. (7) accounts for healthy nodes becoming infected: a randomly chosen node is healthy with probability $1-\rho(t)$, this healthy node can be infected by either of its k neighbor nodes with total probability of $\beta k \rho(t)$. The stationary endemic state is reached when $d\rho(t)/dt = 0$ which leads to

$$\rho = 1 - \frac{1}{\beta k}, \tag{8}$$

indicating the existence of a nonzero epidemic threshold of $\beta = 1/k$. The innermost core of a network consisting only of nodes with degrees $k \geq k_{S_{max}}$ will have epidemic threshold

$$\beta_c \leq 1/k_{S_{max}}. \tag{9}$$

The above inequality holds for all considered networks. Moreover, this inequality becomes an equality for CNI and collaboration networks where nearly all nodes in the innermost cores have degree $k \approx k_{S_{max}}$.

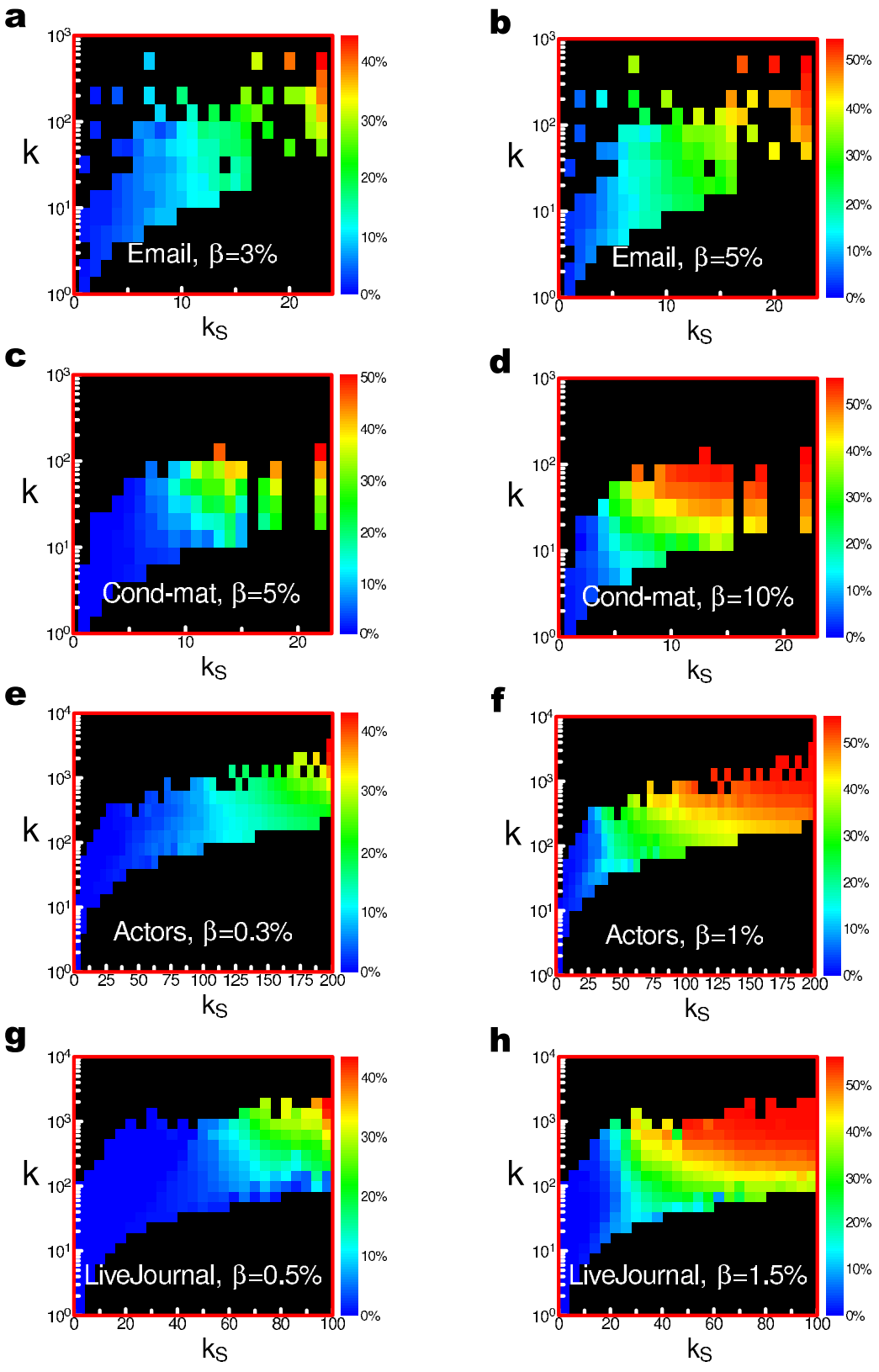

FIG. 15: SIS maps

DOI: 10.1038/NPHYS1746

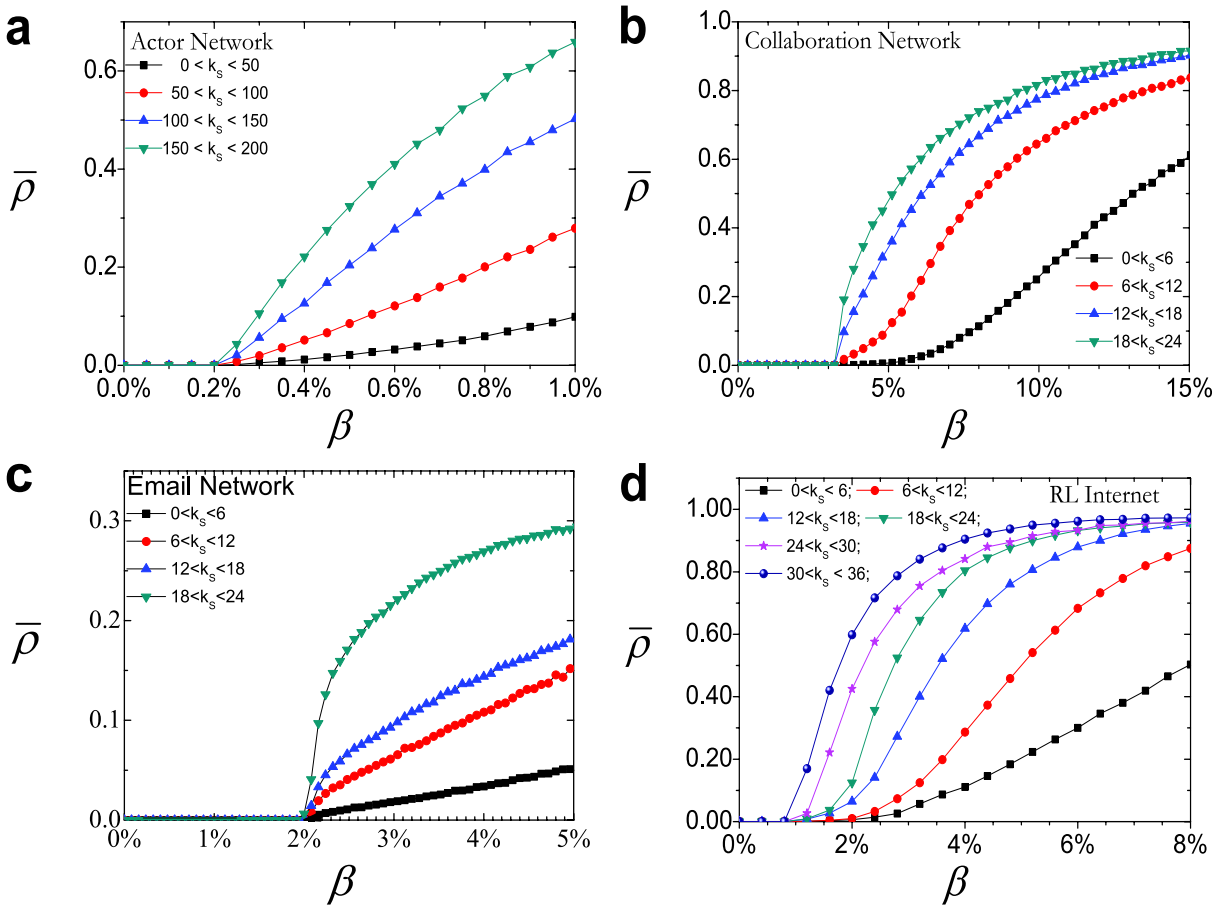

FIG. 16: **How average SIS persistence in different k-shells depends on virus contagiousness.** For every network we randomly sample several groups of nodes based on k-shell index (as described in SI). We plot the average virus persistence $\overline{\rho}$ for every group of nodes as a function of β for the Email, Actor, Collaboration and RL Internet networks. Virus persistence is higher for nodes located in higher k-shells.