SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
STATISTICS
COMPLEX NETWORKS
INFORMATION THEORY AND COMPUTATION

Received
28 September 2014

Accepted
11 March 2015

Published
6 May 2015

Correspondence and requests for materials should be addressed to M.T. (tangminghuang521@hotmail.com)

Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition

Ying Liu[1,2], Ming Tang[1], Tao Zhou[1,3] & Younghae Do[4]

[1]Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, China, [2]School of Computer Science, Southwest Petroleum University, Chengdu 610500, China, [3]Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 610054, China, [4]Department of Mathematics, Kyungpook National University, Daegu 702-701, South Korea.

Identifying the most influential spreaders is an important issue in understanding and controlling spreading processes on complex networks. Recent studies showed that nodes located in the core of a network as identified by the *k*-shell decomposition are the most influential spreaders. However, through a great deal of numerical simulations, we observe that not in all real networks do nodes in high shells are very influential: in some networks the core nodes are the most influential which we call true core, while in others nodes in high shells, even the innermost core, are not good spreaders which we call core-like group. By analyzing the *k*-core structure of the networks, we find that the true core of a network links diversely to the shells of the network, while the core-like group links very locally within the group. For nodes in the core-like group, the *k*-shell index cannot reflect their location importance in the network. We further introduce a measure based on the link diversity of shells to effectively distinguish the true core and core-like group, and identify core-like groups throughout the networks. Our findings help to better understand the structural features of real networks and influential nodes.

The most influential nodes can maximize the speed and scope of information spreading compared with other nodes in a network[1]. Locating these influential nodes is important in improving the use of available resources[2] and controlling the spread of information[3]. A critical issue is how to determine and distinguish the spreading capability of a node. Centrality is usually used to measure the relative importance of nodes within the network, such as degree centrality[4], betweenness centrality[5], closeness centrality[6], eigenvector centrality[7], PageRank centrality[8] and its variance[9]. Nodes with high centrality are considered more influential in the spreading process[10–14]. Among these measures, degree centrality is a simple and effective way, although it is based only on local link information[15,16]. The merit of degree is challenged by a recent study[12], in which the authors pointed out that the most influential spreaders do not correspond to the nodes with largest degree, but are those located in the core of the network as identified by the *k*-shell decomposition[17]. This means the higher coreness of a node, the more influential it is in the spreading dynamics.

The *k*-shell decomposition decomposes a network into hierarchically ordered shells by recursively pruning the nodes with degree less than current shell index (see Methods for details). This procedure assigns each node with an index k_S, representing its coreness in the network. A large k_S value means a core position in the network, while a small k_S value defines the periphery of the network[12]. Because of a low computational complexity of $O(N + E)$[18], where N is the network size and E is the number of edges in the network, this method is extensively used for large-scale network analysis. Generally speaking, it is used to efficiently visualize the structure of large-scale networks[19,20], analyze the core structure of networks[21–25], and capture the essential structural properties of real networks[26]. Since the publication of Ref. 12, the coreness is widely used to quantify the importance of a node in a spreading process[27–29]. For example, in a economic crisis network, nodes with the highest coreness are most likely to spread a crisis globally[30], while in a rumor spreading, nodes with high coreness act as firewalls to prevent the diffusion of a rumor to the whole system[31]. Even nodes with low coreness are considered as bridge elements, which can effectively control the disease in small world networks through an acquaintance-based vaccination strategy[32]. Many works extended the *k*-shell decomposition method, either modify it for a better ranking[33–37] or generalize it to weighted networks[38,39], dynamical networks[40] and multiplex networks[41].

In all these studies, the *k*-shell decomposition is used as a powerful tool to analyze the network structure and identify important nodes. Despite its effectiveness, researchers have noticed that the *k*-shell method has some

defects[28,42]. For example, when there is a lack of the complete network structure, one can not apply the k-shell decomposition to the network. In tree structure and BA model network, the capability of finding influential spreaders is limited due to the low resolution of k_S index. Here, from the perspective of spreading efficiency of cores that have been identified by the k-shell decomposition, we study on whether in different real networks do the core nodes have a higher spreading influence than other nodes. In a common belief, it is. But through intensive computer simulations, we find that it is not the case. In some networks core nodes have the largest spreading efficiency, while in others core nodes have relatively low spreading efficiency. What is the reason for the obvious distinct results? No work has focused on this question to our knowledge. In Refs. 12, 37, the authors pointed out that the performance of centrality measure relates somehow to the infection probability when evaluating the spreading capability of nodes. We find that although the infection probability will cause some fluctuations, the specific structure of real networks is the origin of the distinct performance of coreness in predicting spreading efficiency: in the first case, the core of a network has a link diversity to other shells of the network, while in the latter case the core is linked very locally. We respectively call them true core and core-like group. Then, we propose a measure of information entropy to locate core-like groups in real networks. These findings will help us in understanding the real network structures.

Results

We first calculate the imprecision of coreness and degree in identifying influential spreaders and discover the true core and core-like group in real networks. We then analyze the structural features of the true core and core-like group and uncover their difference. Finally we successfully locate the core-like groups throughout the network by defining a measure of link entropy.

Calculating the imprecision function of coreness and degree in the SIR spreading process. We use a classic Susceptible-Infected-Recovered (SIR) spreading model to simulate the spreading process[43,44], and record the spreading capability (or spreading efficiency) of each node, which is defined as the average size of infected population M for each node as spreading origin (see Methods for details). To evaluate whether the structural centrality of coreness is an effective index to measure the spreading capability of nodes compared with degree, we calculate the imprecision function $\epsilon_{k_s}(p)$ and $\epsilon_k(p)$ proposed in Ref. 12. The imprecision function is defined as

$$\varepsilon_{k_S(k)}(p) = 1 - \frac{M_{k_S(k)}(p)}{M_{eff}(p)}, \tag{1}$$

where p is the fraction of network size $N(p \in [0,1])$, $M_{k_S(k)}(p)$ and $M_{eff}(p)$ are the average spreading efficiencies of pN nodes with highest coreness k_S (degree k) values and largest spreading efficiency, respectively. This function quantifies how close to the optimal spreading is the average spreading of the pN nodes with largest k_S (k) values. The smaller the $\varepsilon_{k_S(k)}$ value, the more accurate the k_S (k) index is a measure to identify the most influential spreaders.

The imprecision functions of nine real networks are shown in Fig. 1. Contrary to common belief, the coreness k_s does not perform

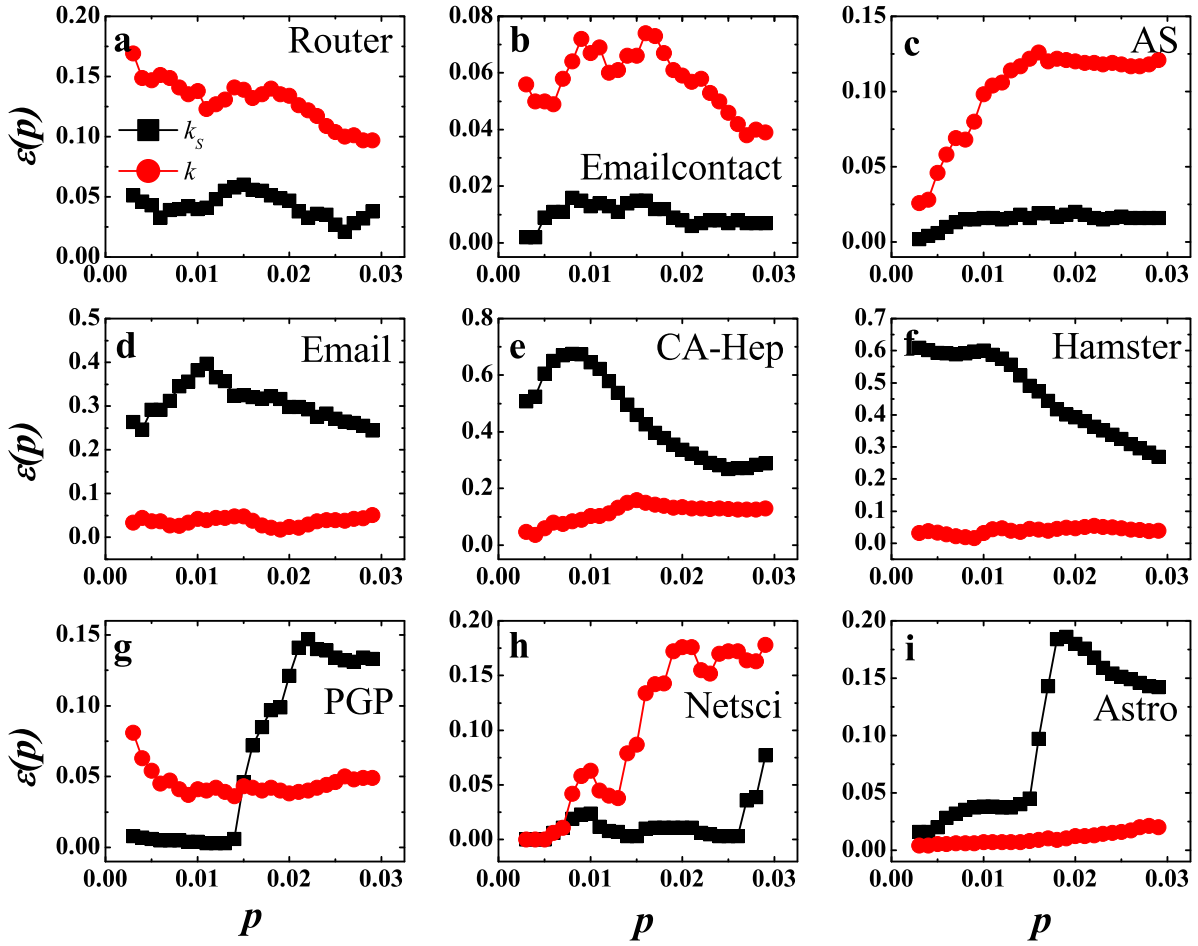

Figure 1 | The imprecision of k_S and k as a function of p for nine real networks. The k_S imprecision (black squares) and k imprecision (red circles) are compared in each network. p is the proportion of nodes calculated, ranging from 0.003 to 0.029. See Fig. S1 for large p plots in SI.

consistently well in all networks. We divide them into three groups. In Router, Emailcontact and AS networks, the k_S imprecision is lower than the k based method. In Fig. 1 (a)–(c), the imprecision $\epsilon_{k_S}(p)$ is very low, under 0.06 in the demonstrated range of $0.003 \le p \le 0.029$, and is much lower than $\epsilon_k(p)$. This means the coreness predicts the outcome of spreading more reliably than degree. However, the imprecision $\epsilon_{k_S}(p)$ for the next three networks (i.e., Email, CA-Hep and Hamster) is much higher than the imprecision $\epsilon_k(p)$. In Fig. 1 (d)–(f), the values of $\epsilon_{k_S}(p)$ is above 0.2 for all the three networks, and is much higher than $\epsilon_k(p)$. This is completely contrary to the case of the first three networks. As for the last three networks of PGP, Netsci and Astro networks, things are more complicated shown in Fig. 1 (g)–(i). In PGP, the k_S method acts better than k when $p < 0.015$. Then there is a sudden rise in $\epsilon_{k_S}(p)$ and it becomes higher than the imprecision of degree. In Netsci, when $p \le 0.026$, the imprecision of k_S is much lower than that of k. There is a fast rise of the $\epsilon_{k_S}(p)$ at $p = 0.027$, and at $p = 0.05$ the k_S imprecision exceeds k imprecision (see Fig. S1 in Supporting Information (SI) for large p plots). In Astro, the sudden rise of k_S imprecision occurs at $p = 0.015$ and the value of $\epsilon_{k_S}(p)$ goes up to around 0.18. This indicates a complex performance of coreness as a measure of spreading efficiency.

Discovering true core and core-like group in real networks. To find out the reason for the distinct performance of coreness in predicting spreading efficiency is the origin of our research interest in this paper. In the following, we first explore the structural characteristics of the first two groups of networks, and then explain the performance of coreness in the last three networks. As we know, the k-shell decomposition tends to assign many nodes with identical k_S value, although their spreading capabilities may be different. When we calculate the imprecision function at a certain p, nodes with the same k_S value are chosen randomly. This will cause some fluctuation in the k_S imprecision curve (fraction of nodes in high shells is shown in the SI table S1). Given this fluctuation, we change to calculate

$$\varepsilon_{k_S}(k_S) = 1 - \frac{M_{k_S}(k_S)}{M_{eff}(k_S)}, \tag{2}$$

where $M_{k_S}(k_S)$ is the average spreading efficiency of the nodes with coreness $k'_S \ge k_S$ (nodes in k_S-core), and $M_{eff}(k_S)$ is the average spreading efficiency of n nodes with highest spreading efficiency, where n equals to the number of nodes with coreness $k'_S \ge k_S$. To compare with k performance, we have

$$\varepsilon_k(k_S) = 1 - \frac{M_k(k_S)}{M_{eff}(k_S)}, \tag{3}$$

where $M_k(k_S)$ is the average spreading efficiency of n nodes with highest degree, and n is as above. The imprecision of k_S is supposed to be low if the nodes in high shells are efficient spreaders. The results are shown in Fig. 2. In the first three networks (a)–(c), the $\varepsilon_{k_S}(k_S)$ is very low and much lower than the imprecision of degree $\varepsilon_k(k_S)$ for large k_S, which means most of nodes in high shells (shells with large k_S value) are efficient spreaders. In the next three networks (d)–(f), the $\varepsilon_{k_S}(k_S)$ is much higher than the $\varepsilon_k(k_S)$ for the innermost core (the shell with the maximum k_S value), and the absolute value is even greater than 0.4, which means many nodes in the innermost core are not influential spreaders. From the perspective of dynamic spreading, we call the innermost core of the first three networks a *true core*, and presumably call that of the other three networks a *false core*, or *core-like group*. This poor k_S performance is obviously different from the fluctuation of imprecision caused by the resolution of k_S index we mentioned above.

Exploring the cause of poor coreness imprecision from structural features. In order to find out the reason for the poor performances of coreness in the spreading process, we first look into the structural properties of the studied real networks. The features of the studied real networks are listed in Table 1. From Table 1, we see that the degree heterogeneity H_k of the first group is sufficiently larger than that of the second group. The degree heterogeneity is defined as $H_k = \langle k^2 \rangle / \langle k \rangle^2$ that evaluates the heterogeneity of degree sequence of a network, where $\langle k^2 \rangle$ and $\langle k \rangle$ are the second moment and first moment of degree respectively. In addition, the degree assortativity r of the first group is negative, which implies that nodes of large

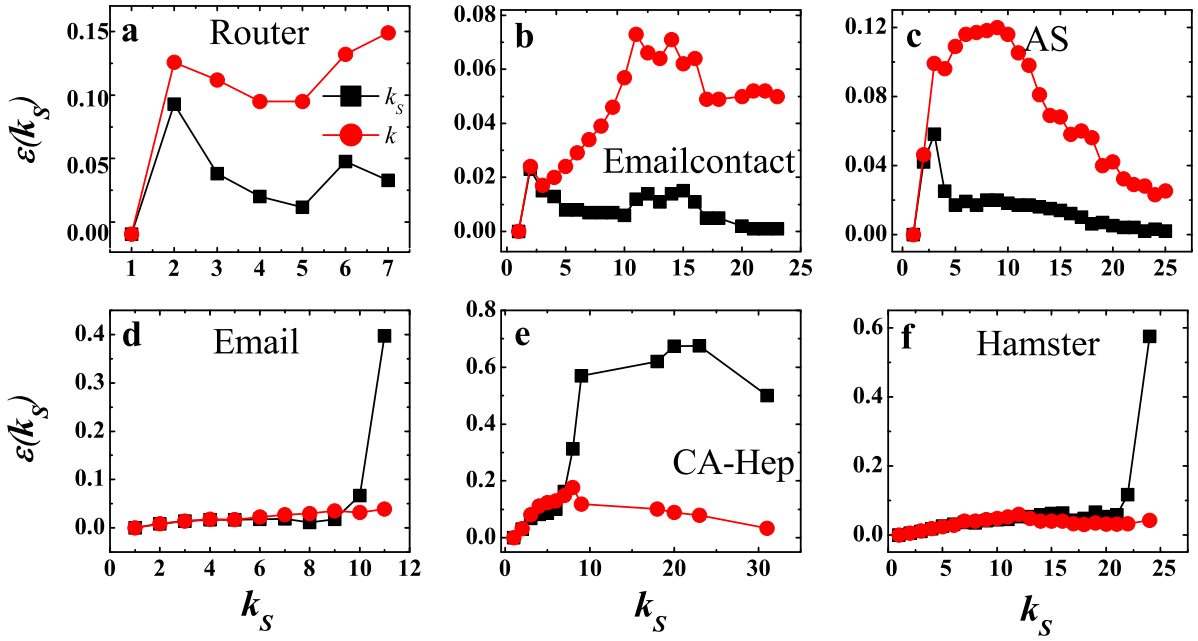

Figure 2 | The imprecision of k_S and k as a function of k_S for six real networks. The k_S imprecision (black squares) and k imprecision (red circles) are compared in each network. Each square represents the k_S imprecision of nodes in k_S-core, and each circle represents the k imprecision of n highest degree nodes, where n equals to the number of nodes in k_S-core. k_S is an integer representing the shell index, ranging from the smallest k_S value to the largest k_S value in the network.

Table 1 | Properties of the real networks studied in this work. Structural properties include number of nodes (N), number of edges (E), average degree ($\langle k\rangle$), maximum degree (k_{max}), degree heterogeneity (H_k), degree assortativity (r), clustering coefficient (C), maximum k_S index ($k_{S_{max}}$), epidemic threshold (λ_c), infection probability in the SIR spreading in the main text (λ)

Network	N	E	$\langle k\rangle$	k_{max}	H_k	r	C	$k_{S_{max}}$	λ_c	λ
Router	5022	6258	2.5	106	5.503	−0.138	0.012	7	0.08	0.27
Emailcontact	12625	20362	3.2	576	34.249	−0.387	0.109	23	0.01	0.10
AS	22963	48436	4.2	2390	61.978	−0.198	0.230	25	0.004	0.13
Email	1133	5451	9.6	71	1.942	0.078	0.220	11	0.06	0.08
CA-Hep	8638	24806	5.7	65	2.261	0.239	0.482	31	0.08	0.12
Hamster	2000	16097	16.1	273	2.719	0.023	0.540	24	0.02	0.04
PGP	10680	24340	4.6	206	4.153	0.240	0.266	31	0.06	0.19
Netsci	379	914	4.8	34	1.663	−0.082	0.741	8	0.14	0.30
Astro	14845	119652	16.1	360	2.820	0.228	0.670	56	0.02	0.05

degrees are inclined to connecting to nodes of small degrees. As nodes in high shells always have large degrees and nodes in low shells (shells with small k_S value) have small degrees, negative assortativity implies a good connection between high shell nodes and low shell nodes. On the contrary, the assortativity of the second group is positive or close to zero, which implies nodes of large degrees are inclined to connect to each other or connect randomly.

To evaluate whether the difference of H_k and r between the two groups of networks results in the distinct performance of k_S imprecision, we randomize the networks using two rewiring schemes (see Methods for details). In the first one, degrees of nodes are preserved after each single rewiring but correlations between the degrees of connected nodes are destroyed[45]. This keeps the H_k unchanged with the original real networks while other structural features destroyed. As is shown in Fig. S2 in SI, the coreness performance is greatly improved: the k_S imprecision is very low and basically lower than or close to the k imprecision in degree-preserving randomized networks. This indicates that the relatively small H_k value of the second group of networks is not the reason of poor k_S imprecision. Next, in the second scheme the rewiring preserves both the degrees of nodes and the joint degree-degree distribution of connected nodes, $P(k, k')$, so that the degree-degree correlations of all nodes are preserved. This keeps both H_k and r unchanged as the original real networks, but as shown in Fig. S3, the k_S imprecision is very low and in general lower than the k imprecision. This implies that the small H_k value and positive r are not the cause of poor k_S imprecision in the second groups of networks. So, what is the real origin of the poor coreness performance?

Analyzing the connectivity between shells. We move to explore the complex connectivity between shells of the studied real networks. Specifically, we consider the link patterns from each shell to its upper shells (shells with greater k_S index), equal shell (the shell with equal k_S index) and lower shells (shells with smaller k_S index). We define the link strength of node i to its upper shells by the proportion function

$$r_i^u = \frac{e_i^u}{k_i}, \tag{4}$$

where e_i^u is the number of links originating from node i to nodes in upper shells, k_i is the total number of links of node i, that is the degree of node i. Large r_i^u indicates more links to the upper shells. Similarly, the link strengths of node i to its equal shells and lower shells are quantified by $r_i^e = e_i^e/k_i$, $r_i^l = e_i^l/k_i$ respectively. The link strengths of k_S-shell to its upper (equal, lower) shells are the average link strength of nodes in that shell, that is

$$R_{k_S}^{u(e,l)} = \frac{\sum_{i\in\Gamma_{k_S}} r_i^{u(e,l)}}{n_{k_S}}, \tag{5}$$

where Γ_{k_S} consists of nodes with coreness k_S, n_{k_S} is the number of nodes in k_S-shell, and $R_{k_S}^u + R_{k_S}^e + R_{k_S}^l = 1$.

From Fig. 3 (a)–(c) for the first group of networks, $R_{k_S}^u$ generally decreases with k_S, this is because the number of nodes in upper shells decreases monotonously with the increase of k_S. $R_{k_S}^e$ remains stable with the increase of k_S. $R_{k_S}^l$ increases with k_S, and in the innermost core, this value goes up to 0.6 and above. For large k_S, $R_{k_S}^l$ is much greater than $R_{k_S}^e$, which means a large proportion of links of high shells point to the their lower shells, obviously higher than the proportion of links within the shell. In Fig. 3 (d)–(f) for the second group, $R_{k_S}^u$ decreases with k_S in Email and CA-Hep, although there is some fluctuation in Hamster. $R_{k_S}^e$ increases with k_S. In the innermost core, $R_{k_S}^e$ is close to 0.7 in Email, close to 1.0 in CA-Hep, and close to 0.8 in Hamster, which is at least 50% larger than that of the first group. $R_{k_S}^l$ increases with k_S at first and falls suddenly at some high shells. For these three networks, $R_{k_S}^l$ are under 0.4 in the innermost core, and is much lower than $R_{k_S}^e$. This indicates that in the second group, the proportion of links from high shells pointing to lower shells is obviously lower than the proportion of links pointing within the shell. This is a sign of densely connected small group within the shell. The average clustering coefficient of nodes in high shells also reflects the overly dense connection in high shells in the second group (See Fig. S4 in SI). We plot the link strength of each shell to its lower shells, equal shell and upper shells in the degree-degree correlation preserving randomized networks in Fig. S5. $R_{k_S}^l$ is promoted above 0.35 and is greater than $R_{k_S}^e$ in most high shells in CA-Hep and Hamster, although in Email there is only a little promotion. The rewiring has changed the dense local link patterns of core-like groups, which is reflect by the increase of $R_{k_S}^l$ and decrease of $R_{k_S}^e$ in high shells. The promoted k_S performance in Fig. S3 is the result of enhanced link diversity.

Next, we focus on the link pattern of the innermost core. The link strength $r_i^{k_S} = e_i^{k_S}/k_i$ defines the ratio of links from a innermost core node i to the shell with index k_S to the degree of node i. $R_{k_{S_{max}}}$ is the average link strength of nodes in the innermost core to the shell with index k_S. Fig. 4 (a) shows the link strength of innermost core to all shells in the first three networks, which is a U-shape curve. In these networks, apart from the link ratio within the core, the largest link ratio points to the shell with most nodes, usually the 1-shell. A U-shape distribution of links from the core is a good feature of core-periphery structure, in which core nodes are well connected to other core nodes and to periphery nodes and periphery nodes are not well connected to each other[46]. In the second group, shown in Fig. 4 (b), the link of innermost core to all shells is different from the first group. Core nodes are very inclined to connecting to core nodes, with a link strength above 0.6. The second largest link ratio points to the adjacent shell of the innermost core, other than the shell with most nodes.

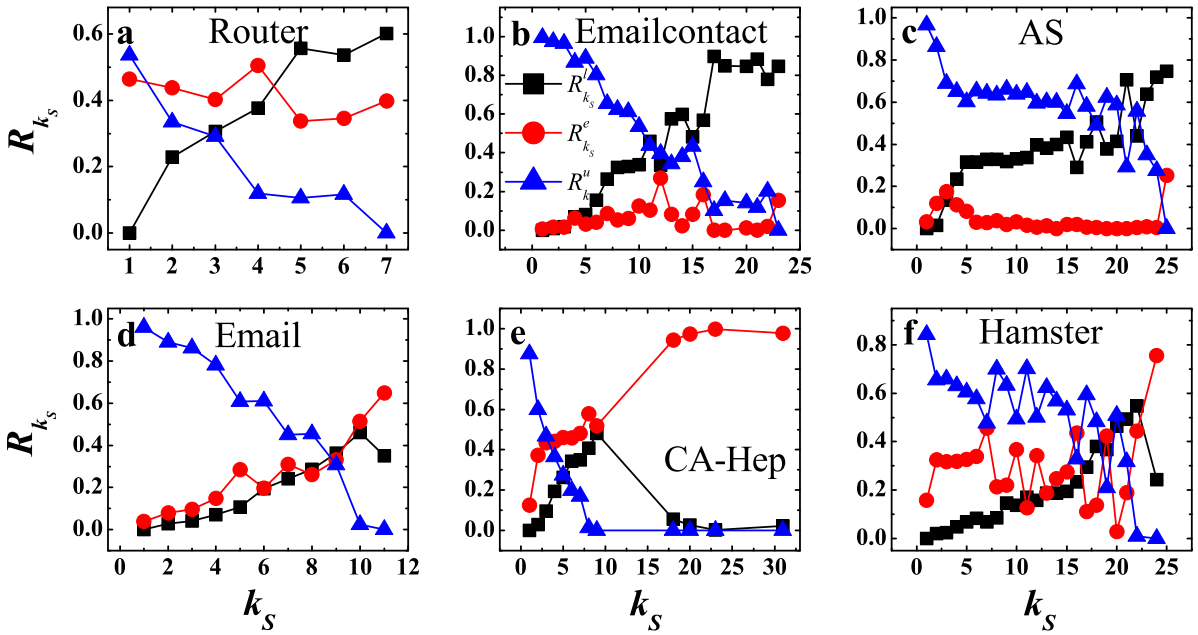

Figure 3 | Link strength of shells for the real networks. The link strength of each shell to its lower shells $R^l_{k_S}$ (black squares), equal shell $R^e_{k_S}$ (red circles) and upper shells $R^u_{k_S}$ (blue triangles) are represented. k_S ranges from the smallest k_S value to the largest k_S value in the network.

When an epidemic spreading origins from core nodes, it is easy to spread throughout the core, but is relatively difficult to spread system wide. This locally connected phenomenon also implies the origin of core-like group (i. e., false core): nodes are densely connected within a small group which contributes much to the k_S index of the nodes, but in the whole network these nodes are not best connected and not located in the most important position for spreading. The link pattern of the second innermost shell is shown in Fig. S6 in SI.

Identifying core-like groups from a structural perspective. The above analysis suggests an obvious structural difference between the two groups of networks: in the first one, the link pattern of innermost core to other shells exhibits a strong diversity, while in the second group, the link of innermost core is very localized within the shell. To quantify the link diversity of a shell with index k_S, we define a link entropy as

$$H_{k_S} = -\frac{1}{lnL}\sum_{k'_S=1}^{k_{S_{max}}} r_{k_S,k'_S} lnr_{k_S,k'_S}, \tag{6}$$

where r_{k_S,k'_S} is the average link strength of k_S-shell to the k'_S-shell and L is the number of shells. The normalized factor lnL measures the entropy when links are uniformly distributed in all shells. This normalization makes the networks with different number of shells comparable. For the innermost core of each network, k_S is set to the maximum k_S value of the network. Entropy of cores of the real network and its degree-preserving randomized version are shown in Fig. 5. In Fig. 5 (a), true cores have a link entropy $H_{k_{S_{max}}}$ higher than 0.6 while false cores have a link entropy lower than 0.5. But in the randomized network, Fig. 5 (b), all the cores have a link entropy $H_{k_{S_{max}}}$ higher than 0.6. Fig. S7 in SI shows the core entropy of degree-degree preserving randomized networks, which is above 0.5

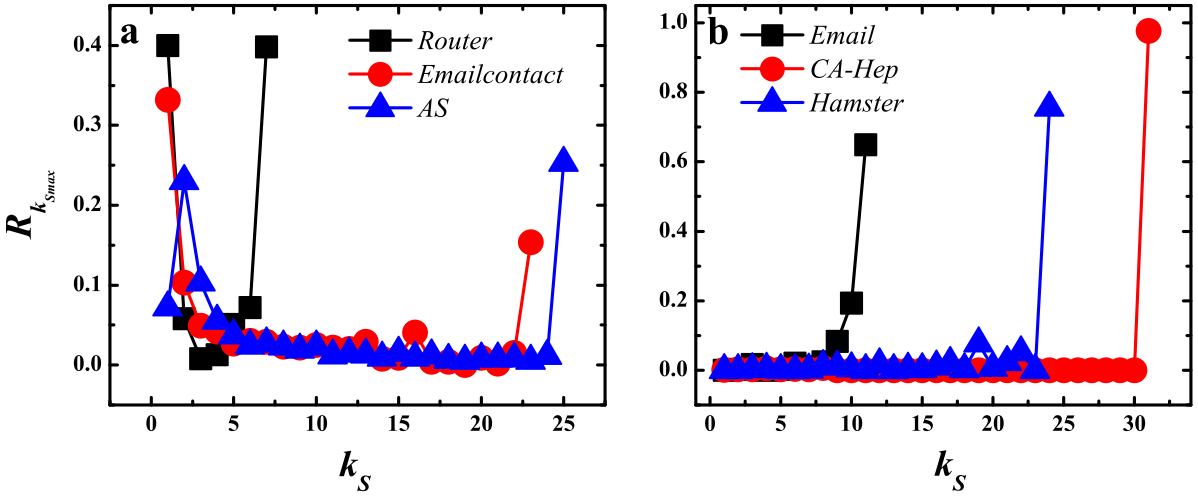

Figure 4 | Link strength of the innermost core to each shell of the network. (a) The link strength of the innermost core to each shell exhibits a U-shape curve in Router (black squares), Emailcontact (red circles) and AS (blue triangles) networks. (b) The link strength of the innermost core to each shell exhibit a slope in Email (black squares), CA-Hep (red circles) and Hamster (blue triangles) networks. k_S ranges from the smallest k_S value to the largest k_S value in the network.

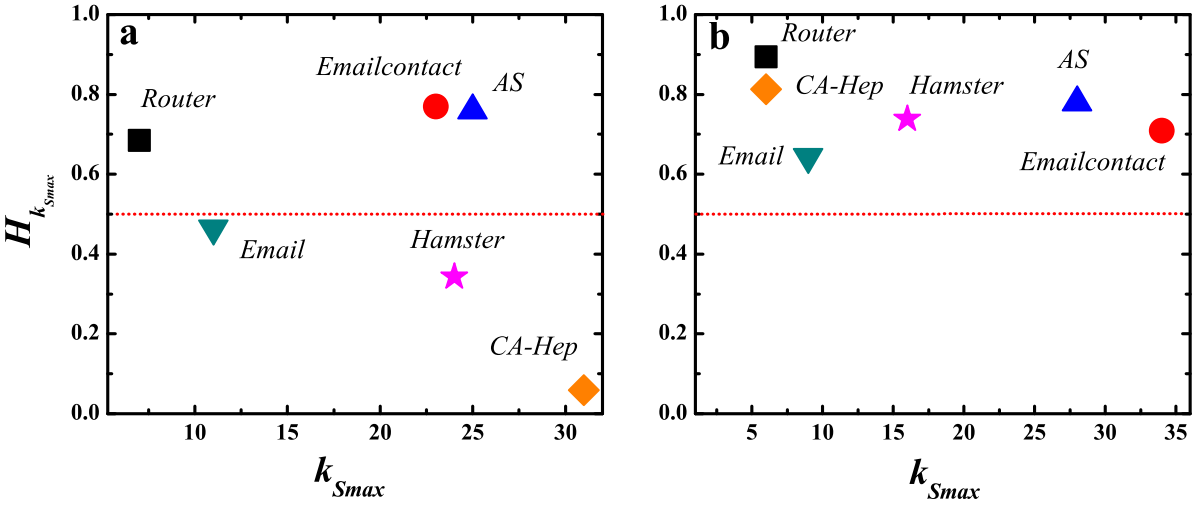

Figure 5 | Link entropy of the innermost core for the real networks and their randomized version. (a) Link entropy of the innermost core for the real networks. (b) Link entropy of the innermost core for the degree-preserving randomized networks. $k_{S_{max}}$ is the largest k_S value in the network. $H_{k_{S_{max}}}$ is the link entropy of the innermost core.

for all studied networks. High entropy corresponds to a more uniform link pattern, where the core is well-connected to the other parts of the network. Low entropy corresponds to a localized link pattern, where the core is densely connected within the shell. In fact, these false cores are not located in the central position of the networks, reflecting by the relatively low spreading efficiency, e.g. the 11-shell in Email, 31-shell in CA-Hep and 24-shell in Hamster, as shown in Figs. S9 and S10 in SI.

Locating the position of core-like groups throughout the networks. Uncovering locally connected core-like groups leads us to understand the imprecision of coreness centrality in the spreading process. We present the imprecision function of PGP, Netsci and Astro networks in Fig. 6 (a)–(c). The coreness performs very well at large k_S values, but then rises suddenly at certain shells. Specifically speaking, in PGP at the 22-shell and above, the imprecision of k_S is lower than that of k. However, there are sudden rises at the 21-shell, 16-shell and 15-shell. From the 10-shell, the k_S imprecision is lower than k again. In Netsci, the k_S imprecision is very low at the 8-shell. Then it rises up at the 7-shell and 6-shell. The k_S imprecision is worse than that of k until the 4-shell. In Astro, the k_S imprecision is low at the 51-shell and higher shells. Then it rises up at the 48-shell and then falls. The same phenomenon occurs at the 30-shell. The rise of k_S imprecision implies that the corresponding shells are core-like groups. Locating them by a dynamic spreading method requires time-consuming simulations.

According to Eq. (6), we calculate the link entropy of each shell in these networks. The shells outlined by hollow red circles in Fig. 6 (d)–(f) have relatively low entropy, which corresponds to locally connected core-like groups. This is reflected by the rise in k_S imprecision

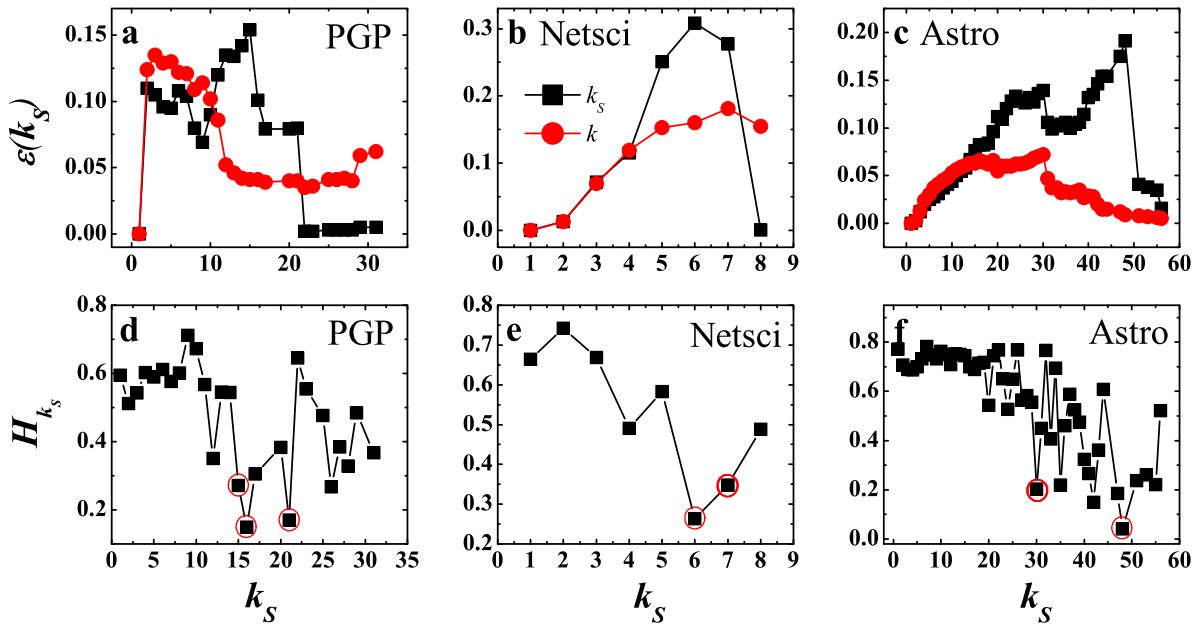

Figure 6 | Locating core-like groups in real networks by link entropy. (a)–(c) The imprecision of k_S and k as a function of k_S for three real networks. The k_S imprecision(black squares) and k imprecision(red circles) are compared. Link entropy of shells in three networks. H_{k_S} is the link entropy of k_S shell. Hollow red circles outline the shells which are densely connected core-like groups. These are 21-shell, 16-shell and 15-shell in PGP, 7-shell and 6-shell in Netsci and 48-shell and 30-shell in Astro. k_S ranges from the smallest k_S value to the largest k_S value in the network.

shown in Fig. 6 (a)–(c). The link patterns of the core-like groups, shown in Fig. S8, are similar to that of false cores in the second group of Email, CA-Hep and Hamster: a dense connection within the shell. The only difference is that these core-like groups locate in the outer shells of the network other than the innermost shell. These core-like groups have an obvious low spreading efficiency than their adjacent shells, which is also confirmed in Fig. S9 and S10. From the above, we see the link entropy provides a fast way to locate the position of core-like groups in the network without running a large amount of spreading simulations, which is very important in identifying key spreaders and controlling the spreading dynamics on networks.

Discussion

Analyzing and profiling the structures of real networks is an important step in understanding and controlling dynamic behaviors on networks. The *k*-shell decomposition is a powerful tool to profile the hierarchical structures of networks. The inner core corresponds to the shells of large k_S and the network periphery corresponds to the shells of small k_S. This makes k_S index an effective centrality measure to distinguish the spreading capability of nodes, which is validated in many real networks. However, there are circumstances where the *k*-shell decomposition is not able to identify influential spreaders, which leaves much space to explore. Here from the perspective of core's spreading efficiency, we discover that in some real networks, there exist core-like groups, which have high coreness but are in fact not located in the core of the network. By analyzing the *k*-core structure of real networks, we discover the distinct link patterns of true cores and core-like groups. For the true core of a network, it displays strong link diversity to other shells of the network, represented by a U-shape link curve. As for the core-like group, it has a very dense and local internal connection, represented as a slope-shape link curve. Based on the link pattern, we define a measure of link entropy to evaluate the link diversity of a shell to the remaining shells of the network. This provides a fast way to locate the core-like groups throughout the network from a structural perspective, which have a relatively low link entropy. We note that in Ref. 29 the authors calculated the entropy of each node to assess the heterogeneity of links. They use *k*-shell decomposition to assign each node a global feature and compute for each node an entropy as its global diversity, which is then combined with local feature to rank node influence. However, this entropy relates much to the degree of a node. A node with small degree has a limited number of layers it can connect to, even the connection is uniformly distributed. The node entropy is limited in the sense of statistics. Contrary to their work, we target to a group. We consider the link diversity of a shell, which consists of several nodes and these nodes have different degrees. By using the entropy for a shell, we can effectively locate the core-like groups, whose *k*-shell index is unable to reflect their global importance. This makes implication to the works that use the *k*-shell index in ranking node importance.

Uncovering these core-like groups is important in identifying key players and making control strategy for spreading dynamics. It is worth noticing that in the core-like groups, there may also exist some good spreaders. It implies that there should be new network analysis method which will effectively locate the nodes of different importance in core-like groups in the right hierarchical position. The new method should apply well in real networks with specific structures such as strong community structures.

Methods

The k-shell decomposition. The algorithm starts by removing all nodes with degree $k = 1$. After removing all nodes with $k = 1$, there may appear some nodes with only one link left. We should iteratively remove these nodes until there is no node left with $k = 1$. The removed nodes are assigned with an index $k_S = 1$ and are considered in the 1-shell. In a similar way, nodes with degree $k = 2$ are iteratively removed and assigned an index $k_S = 2$. This pruning process continues removing higher shells until all nodes are removed. As a result, each node is assigned a k_S index, and the network can be viewed as a hierarchical structure from the innermost shell to the periphery shell.

SIR Model. The Susceptible-Infected-Recovered (SIR) model is widely used for simulating the spreading process on networks. In the model, a node has three possible states: *S* (susceptible), *I* (infected) and *R* (recovered). An individual in the susceptible state does not have the disease yet but could catch it if they come into contact with someone who does. An individual in the infected state has the disease and can pass it to susceptible individuals. An individual in recovered state neither spread disease nor be infected by others. In the start of a spreading process, a single node is infected, considered as seed, and all other nodes are in susceptible states. At each time step, there are two stages. In the first stage, susceptible individuals become infected with probability λ when they have contacted with an infected neighbor. In the second stage, infected nodes recover or die (change to *R* state) with probability μ. Here we set $\mu = 1$ for generality. The spreading process stops when there is no infected node in the network. The proportion of recovered nodes defines the final infection population in a spreading process. We record the average infected population M_i originating at node *i* over 100 times of the spreading process to quantify the influence of node *i* in a SIR spreading.

As we take the final infected population to quantify the spreading efficiency of each node, the infection probability should be carefully considered. If it is too large, the effect of node position is not obvious and all nodes show almost identical spreading capabilities. If it is too small, the infection is very localized in the neighborhood, which cannot reflect the overall spreading influence of the nodes. So we first calculate the epidemic threshold of a network using the heterogeneous mean-filed method in Ref. 47. That is $\lambda_c = \langle k\rangle/(\langle k^2\rangle - \langle k\rangle)$. Then we chose an infection probability $\lambda > \lambda_c$[14,37], which makes the final infected population above the critical point, $M > 0$, and reaches a finite but small fraction of the network size for most nodes as spreading origins, in the range of 1%–20%[12]. In fact, we plot the infected population of a shell as an average over nodes belong to the shell when infection probability is 1–5 times of the threshold λ_c, as well as the infected population when infection probability is around the chosen infected probability λ. We find that, the relative spreading efficiency of shells is almost the same under different infection probabilities (See Fig. S9 and S10 in SI).

Rewiring Schemes. In the first rewiring scheme, we randomly choose two edges of the network, and label the ends of the first edge as A and B, and the ends of the second edge as C and D. Then we rewire the two edges, connecting end A and D as an edge, and connecting end B and C as another edge. We avoid multiple edge and self-edge in the rewiring process. This rewiring preserves the degree sequence of the original real network but destroys the degree correlations. In the second rewiring scheme, we randomly choose an edge and test the degree of one end, record as *k*. A second edge with an end having degree *k* is then chosen. We rewire the two edges as before and ensure that the end connecting to a node of degree *k* still connects to a node of degree *k* after rewiring. This scheme preserves both the degree sequence and the degree-degree correlations as the original real network.

Data Sets. The real networks studied in the paper are: (1) Router (the router level topology of the Internet, collected by the Rocketfuel Project)[48];(2) Email-contact (Email contacts at Computer Science Department of University college London)[12]; (3) AS (Internet at the autonomous system level)[49]; (4) Email (e-mail network of University at Rovira i Virgili, URV)[50];(5) CA-Hep (Giant connected component of collaboration network of arxiv in high-energy physics theory)[51]; (6) Hamster (friendships and family links between users of the website hamsterster.com)[52]; (7) PGP (an encrypted communication network)[53]; (8) Netsci (collaboration network of network scientists)[54]; (9) Astro physics (collaboration network of astrophysics scientists)[55].

1. Kempe, D., Kleinberg, J. & Tordos, E. Maximizing the spread of influence through a social network. *in Proc of the 9th ACM SIGKDD Int. Conf. on knowledge discovery and data mining (ACM, Washington, DC, USA, 2003), KDD 03* 137–146 (2003).
2. Gallos, L. K., Liljeros, F., Argyrakis, P., Bunde, A. & Havlin, S. Improving immunization strategies. *Phys. Rev. E* **75** 045101(R) (2007).
3. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
4. Freeman, L. C. Centrality in social networks conceptual clarification. *Social Networks* **1**, 215–239 (1978).
5. Freeman, L. C. A set of measures of centrality based upon betweenness. *Sociometry* **40**, 35–41 (1977).
6. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603 (1966).
7. Bonacich, P. & Floyd, P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **23**, 191–201 (2001).
8. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks* **30**, 107–117 (1998).
9. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM* **46**, 604–632 (1999).
10. Canright, G. S., & Engø-Monsen, K. Spreading on networks: a topographic view. *Complexus* **3**, 131–146 (2006).
11. Lü, L. Y., Zhang, Y. C., Yeung, C. H. & Zhou, T. Leaders in social networks, the delicious Case. *PLoS ONE* **6**, e21202 (2011).
12. Kitsak, M. *et al*. Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).

13. Chen, D. B., Lü, L. Y., Shang, M. S., Zhang, Y. C. & Zhou, T. Identifying influential nodes in complex networks. *Physica A* **391**, 1777–1787 (2012).
14. Macdonald, B., Shakarian, P., Howard, N. & Moores, G. Spreaders in the Network SIR mode: an empirical study. *arXiv e-print* 1208.4269v1 (2012).
15. Borge-Holthoefer, J., Rivero, A. & Moreno, Y. Locating privileged spreaders on an online social networks. *Phys. Rev. E* **85**, 066123 (2012).
16. Tanaka, G., Morino, K. & Aihara, K. Dynamical robustness in complex networks: the crucial role of low-degree nodes. *Sci. Rep.* **2**, 232 (2012).
17. Bolobás, B. *Graph Theory and Combinatorics: Proc. Cambridge Combinatorial Conf. in honor of P. Erdös* (Academic Press, NY, 1984).
18. Batagelj, V. & Zaversnik, M. An O(m) algorithm for cores decomposition of networks. *arXiv e-print* cs/0310049 (2003).
19. Alvarez-Hamelin, J. I., Asta, L. D., Barrat, A. & Vespignani, A. k-core decomposition: a tool for the visualization of large scale networks. *arXiv e-print* cs/0504107v2 (2005).
20. Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. & Shir, E. A model of Internet topology using k-shell decomposition. *Proc. Natl. Acad. Sci. USA* **104**, 11150–11154 (2007).
21. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. k-core organization of complex networks. *Phys. Rev. Lett.* **96**, 040601 (2006).
22. Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A. & Vespignani, A. K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases. *Networks and Hetero Media* **3**, 371–393 (2008).
23. Colomer-de-Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I. & Boguñá, M. Deciphering the global organization of clustering in real complex networks. *Sci. Rep.* **3**, 2517 (2013).
24. Holme, P. Core-periphery organization of complex networks. *Phys. Rev. E* **72**, 046111 (2005).
25. Zhang, G. Q., Zhang, G. Q., Yang, Q. F., Cheng, S. Q. & Zhou, T. Evolution of the Internet and its cores. *New J. Phys.* **10**, 123027 (2008).
26. Hébert-Dufresne, L., Allard, A., Young, J. & Dubé, L. J. Random networks with arbitrary k-core structure. *arXiv e-print* 1308.6537v1 (2013).
27. Castellano, C. & Pastor-Satorras, R. Competing activation mechanisms in epidemics on networks. *Sci. Rep.* **2**, 371 (2012).
28. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mech.* **12**, 12002 (2013).
29. Fu, Y.-H., Huang, C.-Y. & Sun, C.-T. Identifying super-spreader nodes in complex networks. *Mathematical Problems in Engineering* **2014**, 675713 (2014).
30. Garas, A., Argyrakis, P., Rozenblat, C., Tomassini, M. & Havlin, S. Worldwide spreading of economic crisis. *New J. Phys.* **12**, 113043 (2010).
31. Borge-Holthoefer, J. & Moreno, Y. Absence of influential spreaders in rumor dynamics. *Phys. Rev. E* **85**, 026116 (2012).
32. Reppas, A. I. & Lawyer, G. Low k-shells identify bridge elements critical to disease flow in small-world networks. *AIP Conf. Proc.* **1479**, 1426–1429 (2012).
33. Zeng, A. & Zhang, C. J. Ranking spreaders by decomposing complex networks. *Phys. Lett. A* **377**, 1031–1035 (2013).
34. Ren, Z. M., Zeng, A., Chen, D. B., Liao, H. & Liu, J. G. Iterative resource allocation for ranking spreaders in complex networks. *Europhys. Lett.* **106**, 48005 (2014).
35. Corominas-Murtra, B., Fuchs, B. & Thurner, S. Detection of the elite structure in a virtual multiplex social systems by means of a generalized k-core. *arXiv e-print* 1309.6740 (2013).
36. Liu, J. G., Ren, Z. M. & Guo, Q. Ranking the spreading influence in complex networks. *Physica A* **392**, 4154–4159 (2013).
37. Bae, J. & Kim, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A* **395**, 549–559 (2014).
38. Garas, A., Schweitzer, F. & Havlin, S. A k-shell decomposition method for weighted networks. *New J. of Phys.* **14**, 083030 (2012).
39. Eidsaa, M. & Almaas, E. S-core network decomposition: A generalization of k-core analysis to weighted networks. *Phys. Rev. E* **88**, 062819 (2013).
40. Miorandi, D. & Pellegrini, F. D. K-shell decomposition for dynamic complex networks. *Ad Hoc and Wireless Networks(WiOpt), 2010 Proc. of the 8th Int. Symp., IEEE* 499–507 (2010).
41. Azimi-Tafreshi, N., Gómex-Gardeñes, J. & Dorogovtsev, S. N. k-core percolation on multiplex networks. *arXiv e-print* 1405.1336v1 (2014).
42. Pei, S., Muchnik, L., Andrade, J. S., Zheng, Z. M. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Sci. Rep.* **4**, 5547 (2014).
43. Anderson, R. M. & May, R. M. *Infectious diseases in humans* (Oxford University Press, Oxford, 1991).
44. Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521–529 (2002).
45. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
46. Rombach, M. P., Porter, M. A., Fowler, J. H. & Mucha, P. J. Core-Periphery structure in networks. *SIAM J. Appl. Math.* **74**, 167–190 (2014).
47. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **105**, 218701 (2010).
48. Spring, N., Mahajan, R., Wetherall, D. & Anderson, T. Measuring ISP topologies with Rocketfuel. *IEEE/ACM Trans. Networking* **12**, 2–16 (2004).
49. Newman, M. E. J. Network data. (2006) Available at: http://www-personal.umich.edu/%7Emejn/netdata. (Accessed: 12/12/2012)
50. 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).
51. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph Evolution: Densification and Shrinking Diameters. *ACM Trans. on Knowledge Discovery from Data (ACM TKDD)* **1**, 1 (2007).
52. Kunegis, J. Hamsterster full network dataset - KONECT. (2014) Available at: http://konect.uni-koblenz.de/networks/petster-hamster. (Accessed: 01/03/2014)
53. Boguñá, M., Pastor-Satorras, R., Diaz-Guilera, A. & Arenas, A. Models of social networks based on social distance attachment. *Phys. Rev. E* **70**, 056122 (2004).
54. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
55. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404–409 (2001).

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11105025, 91324002, 61433014), Scientific Research Starting Project of Southwest Petroleum University (No. 2014QHZ024) and Youth Foundation of Southwest Petroleum University (Grant No. 285). Y. Do was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2010067).

Author contributions

Y.L., M.T. and T.Z. devised the research project. Y.L. performed numerical simulations. Y. L., M.T. and Y.H.D. analyzed the results. Y.L., M.T., T.Z. and Y.H.D. wrote the paper.

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: Liu, Y., Tang, M., Zhou, T. & Do, Y. Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Sci. Rep.* **5**, 9602; DOI:10.1038/srep09602 (2015).