Core-like groups resulting in invalidation of $k$-shell decomposition analysis

Ying Liu,1,2 Ming Tang,1 Tao Zhou,1,3 and Younghae Do4

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

(Dated: September 19, 2014)

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. Locating these influential nodes is important in improving the use of available resources and controlling the spread of information. 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, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality and its variance. Nodes with high centrality are considered more influential in the spreading process. Among these measures, degree centrality is a simple and effective way, although it is based only on local link information. The merit of degree is challenged by a recent study, 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. 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 $k$, representing its coreness in the network. A large $k$ value means a core position in the network, while a small $k$ value defines the periphery of the network. Because of a low computational complexity of $O(N + E)$, 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, analyze the core structure of networks, and capture the essential structural properties of real networks. Since the publication of Ref. 12, the coreness is widely used to quantify the importance of a node in a spreading process. For example, in an economic crisis network, nodes with the highest coreness are most likely to spread a crisis globally, while in a rumor spreading, nodes with high coreness act as firewalls to prevent the diffusion of a rumor to the whole system. 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. Many works extended the $k$-shell decomposition method, either modify it for a better ranking or generalize it to weighted networks, dynamical networks and multiplex networks.

In all these studies, the $k$-shell decomposition is used as a powerful tool to analyze the network structure and identify important nodes. Then a critical question is: 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 other networks 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, 36, 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 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 these core-like groups in real networks. These findings will help us in understanding the real network structures.
We first calculate the imprecision function of coreness and degree in the SIR spreading process and discover the true core and core-like group in real networks. We then analyze the structural characteristics of the true cores and core-like groups. Finally we successfully locate the position of 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 \cite{41, 42}, 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. \cite{12}. The imprecision function is defined as

\[
\epsilon_{k_S}(p) = 1 - \frac{M_{k_S}(p)}{M_{eff}(p)},
\]

where \( p \) is the fraction of network size \( N (p \in [0, 1]) \), \( M_{k_S}(p) \) 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 \( \epsilon_{k_S}(p) \) 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 consistently well in all networks. We divide them into three groups. In Router, Emailcontact and AS networks, the \( k_S \) imprecision is consistently 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 \( \leq p \leq 0.029 \), 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 Figs. 1(g), (h) and (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 \( p \leq 0.026 \), the imprecision of \( k_S \) is much higher than that of \( k \). There is a fast rise of the \( \epsilon_{k_S}(p) \) at \( p = 0.027 \), and at \( p = 0.05 \), \( k_S \) 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 the SIR spreading process 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

\[
\epsilon_{k_S}(k_S) = 1 - \frac{M_{k_S}(k_S)}{M_{eff}(k_S)},
\]

where \( M_{k_S}(k_S) \) is the average spreading efficiency of the nodes with coreness \( k_S \geq k_S \) (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 the number of nodes with coreness \( k_S \geq k_S \). To compare with \( k \) performance, we have

\[
\epsilon_{k}(k) = 1 - \frac{M_{k}(k_S)}{M_{eff}(k_S)},
\]

where \( M_{k}(k_S) \) is the average spreading efficiency of \( n \) nodes with highest degree, and \( n \) equals to the the number of nodes with coreness \( k_S \geq k_S \). The imprecision of \( k_S \) is supposed to be lower if the nodes in high shells are efficient spreaders. The results are shown in Fig. 2. In the first three networks (a)-(c), the \( \epsilon_{k_S}(k_S) \) is very low and much lower than the imprecision of degree \( \epsilon_k(k_S) \) 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 \( \epsilon_{k_S}(k_S) \) is much higher than the \( \epsilon_k(k_S) \) 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 dynamical perspective, we call the innermost core of the first three networks a true core, and presumably
FIG. 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.

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.

Investigating the impact of structural features on the coreness imprecision. In order to find out the reason for the distinct performances of coreness in the spreading process, we first look into the structural properties of the studied real networks. The structural features of the studied real networks are listed in Table I. From Table I we see that the degree heterogeneity of the first group is sufficiently larger than that of the second group. In Ref. [26] the authors pointed out that the maximum $k_S$ value, $k_{S_{\text{max}}}$, in a network scales approximately as the square root of maximum degree $\sqrt{k_{\text{max}}}$ in many real networks, which implies the larger the maximum degree of the network, the more shells. A broad distribution of degree with large $k_{\text{max}}$ results in a large degree heterogeneity $H_k$, which is the feature of the three networks in the first group. In addition, the degree assortativity $r$ of the first group is negative, which implies that nodes of large degrees are inclined to connecting 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 degree heterogeneity $H_k$ and assortativity $r$ have a direct impact on the performance of coreness, we randomize the networks using two rewiring schemes. In the first one, degrees of nodes are preserved after each single rewiring but correlations between the degrees of connected nodes are destroyed [43]. In the second scheme, the rewiring preserves both
FIG. 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 I: 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 ($k_{Smax}$), epidemic threshold ($\lambda_c$), infection probability in the SIR spreading in the main text ($\lambda$).

| Network   | $N$  | $E$  | $\langle k \rangle$ | $k_{max}$ | $H_k$ | $r$  | $C$ | $k_{Smax}$ | $\lambda_c$ | $\lambda$ |
|-----------|------|------|----------------------|-----------|-------|------|-----|------------|-------------|---------|
| Router    | 5022 | 6258 | 2.5                  | 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                  | 61.978    | -0.198| 0.230| 25  | 0.004      | 0.13        |
| Email     | 1133 | 5451 | 9.6                  | 1.942     | 0.078 | 0.220| 11  | 0.06       | 0.08        |
| CA-Hep    | 8638 | 24806| 5.7                  | 2.261     | 0.239 | 0.482| 31  | 0.08       | 0.12        |
| Hamster   | 2000 | 16097| 16.1                 | 2.719     | 0.023 | 0.540| 24  | 0.02       | 0.04        |
| PGP       | 10680| 24340| 4.6                  | 4.153     | 0.240 | 0.266| 31  | 0.06       | 0.19        |
| Netsci    | 379  | 914  | 4.8                  | 1.663     | -0.082| 0.741| 8   | 0.14       | 0.30        |
| Astro     | 14845| 119652| 16.1                | 2.820     | 0.228 | 0.670| 56  | 0.02       | 0.05        |

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. For each scheme, we rewire the real network by $100E$ times, where $E$ is the total number of edges. As is shown in Fig. S2 in SI, the $k_S$ imprecision is very low and basically lower than or close to the $k$ imprecision in degree-preserving networks, which indicates coreness an accurate measure for the spreading efficiency. The results for the degree-degree correlations preserved randomization in Fig. S3 are almost the same as that in Fig. S2, with only a slight difference in that the $k_S$ impression is lower than $k$ in high shells in Email and Hamster. The above results in the randomized networks imply that $H_k$ and $r$ are not key factors of the bad $k_S$ performance in high shells in the second group, and there should exist some specific structures in these real networks.

**Analyzing the connectivity between shells.** We then 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 index $k_S$ greater than...
the coreness of node \( i \), equal shell (the shell with index \( k_S \) equal to the coreness of node \( i \)) and lower shells (shells with index \( k_S \) smaller than the coreness of node \( i \)). We define the link strength of node \( i \) to its upper shells by the proportion function

\[
r_{i}^{u} = \frac{e_{i}^{u}}{k_{i}}
\]

(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

\[
R_{k_S}^{u(e,l)} = \frac{\sum_{i} \Gamma_{k_S} r_{i}^{u(e,l)}}{n_{k_S}}
\]

(5)

where \( \Gamma_{k_S} \) consists of nodes with coreness \( k_S \), \( n_{k_S} \) is the number of nodes with coreness \( k_S \), 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 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 \( k_S \) values, \( R_{k_S}^{l} \) 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, but 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 also 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. \( R_{k_S}^{e} \) 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 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 \( i \). \( R_{k_S}^{k_S} \) 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. In the second group, shown in Fig. 4 (b), the link of innermost core to all shells is different. 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. When a disease origins from core nodes, it is easily spreading through out the core, but is relatively difficult to spread system wide. This locally connected phenomena 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 an link entropy as

\[
H_{k_S} = -\frac{1}{\ln L} \sum_{k_S}^{k_{S_{max}}} r_{k_S} k_{S} \ln r_{k_S} k_{S},
\]

(6)

where \( r_{k_S} k_{S} \) is the average link strength of nodes of coreness \( k_S \) to the shell of coreness \( k_{S} \) and \( L \) is the number of shells. The formalized factor \( \ln L \) measures the entropy when links are uniformly distributed in all shells. This formalization 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. S5.
FIG. 3: **Link strength of shells for the real networks.** The link strength of each shell to its lower shells $R_{k_S}^l$ (black squares), equal shell $R_{k_S}^e$ (red circles) and upper shells $R_{k_S}^u$ (blue triangles) are represented. $k_S$ ranges from the smallest $k_S$ value to the largest $k_S$ value in the network.

FIG. 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.

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 randomized networks, which is above 0.5 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. 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 acts better 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 phenomena 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 shown in Fig. 5(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 these 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, in some real networks, there may exist core-like groups, which have high coreness but in fact are not influential spreaders. By analyzing the \( k \)-core structures of real networks, we discover the distinct link patterns of true cores and core-like groups. For
FIG. 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.

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 connections, 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 in the network.

Uncovering these core-like groups are important in identifying key players and making control strategy for spreading dynamics. It is worth noticing that in the core-like groups, except from nodes of low spreading efficiency, there may also exist some good spreaders. A better ranking strategy that is able to distinguish them need further investment. It implies that there should be new network decomposition 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 $\lambda$ when they have contacted with an infected neighbor. In the second stage, infected nodes recover or die (change to $R$ state) with probability $\mu$. Here we set $\mu = 1$ for generality. The spreading process stops when there is no infected node in the network. 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. [45]. That is $\lambda_c = (k) / ((k^2 - k))$. Then we chose an infection probability $\lambda > \lambda_c$ [14], 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 $\lambda_c$, as well as the infected population when infection probability is around the chosen infected probability $\lambda$. We find that, the relative spreading efficiency of shells is almost the same under different infection probabilities (See Fig. S9 and S10 in SI).

**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) [47]; (4) Email (e-mail network of University at Rovira i Virgili, URV) [48]; (5) CA-Hep (collaboration network of arxiv in high-energy physics theory) [49]; (6) Hamster (friendships and family links between users of the website hamsterster.com) [50]; (7) PGP (an encrypted communication network) [51]; (8) Netsci (collaboration network of network scientists) [52]; (9) Astro physics (collaboration network of astrophysics scientists) [53].

* Electronic address: tangminghuang521@hotmail.com

[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., Bundi, 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] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *J. Math. Social* **2**, 113-120 (1972).
[5] Freeman, L. C. A set of measures of centrality based upon betweenness. *Sociometry* **40**, (1977).
[6] Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, (1966).
[7] Bonacich,P. & Floyd, P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **23**, 191-201 (2001).
[8] Page, L., Brin, S., Motwani, R. & Winograd, T. Tech. Rep., Stanford Digital Library Technologies Project, (1998).
[9] Jeh, G. & Widom, J. in Proc the 12th Int. Conf. on World Wide Web (ACM, New York, NY, USA, 2003), WWW 03, 271-279 (2003).
[10] Geoffrey, S., Caraght & Engs-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., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E. & Makse, H. A. 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: Proceedings of the Cambridge Combinatorial Conference in Honor of P. Erdős **35**, (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/0504107 (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. A., 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., 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).
[29] Garas, A., Argyrakis, P., Rozenblat, C., Tomassini, M. & Havlin, S. Worldwide spreading of economic crisis. *New J. Phys.* **12**, 113043 (2010).
[30] Borge-Holthoefer, J. & Moreno, Y. Absence of influential spreaders in rumor dynamics. *Phys. Rev. E* **85**, 026116 (2012).
[31] 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).
[32] Zeng, A. & Zhang, C. J. Ranking spreaders by decomposing complex networks. *Phys. Lett. A* **377**, 1031-1035 (2013).
[33] Brown, P. E. & Feng, J. L. Measuring user influence on Twitter using modified k-shell decomposition. *arXiv e-print* 1308.6537v1 (2013).
[34] Liu, J. G., Ren, Z. M. & Guo, Q. Ranking the spreading influence in complex networks. *Physica A* **392**, 4154-4159 (2013).
[35] Bae, J. & Kim, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A* **395**, 549-559 (2014).
[36] Garas, A., Schweitzer, F. & Havlin, S. A k-shell decomposition method for weighted networks. *New J. of Phys.* **14**, 083030 (2012).
[37] Eidsaa, M. & Almas, E. S-core network decomposition: A generalization of k-core analysis to weighted networks. *Phys. Rev. E* **88**, 062819 (2013).
[38] 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).
[39] Azimi-Tafreshi, N., Gómez-Gardeñes, J. & Dorogovtsev, S. N. k-core percolation on multiplex networks. *arXiv e-print* 1405.1336v1 (2014).
[40] Anderson, R. M. & May, R. M. Infectious diseases in humans. Oxford University Press (1992).
[41] Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521-529 (2002).
[42] Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910-913 (2002).
[43] 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).
[44] Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.*, **105**, 218701 (2010).
[45] Newman, M. E. J. Network data. [http://www-personal.umich.edu/mejn/netdata](http://www-personal.umich.edu/mejn/netdata)
[46] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **68**, 065103(R) (2003).
[47] Guimerà, 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).
[48] Leskovec & Jure. Stanford network analysis project (snap). (2012).
[49] Hamsterster full network dataset - KONECT, March 2014. [http://konect.uni-koblenz.de/networks/petster-hamster](http://konect.uni-koblenz.de/networks/petster-hamster) (2014).
[50] 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).
[51] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
[52] Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404-409 (2001).

Acknowledgement

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

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

Supporting Information for

Core-like groups resulting in invalidation of $k$-shell decomposition analysis

TABLE S1: Proportion of nodes in high shells in the studied real networks. $d$ indicates the shell ranking difference of a shell from the highest shell. $d = 0$ corresponds to the highest shell, $d = 1$ corresponds to the nearest shell from the highest shell. For example, the highest shell of CA-Hep is 31-shell, while the second highest shell is 23-shell. Thus, $d = 0$ corresponds to 31-shell, and $d = 1$ correspond to 23-shell.

| Network | $d = 0$ | $d = 1$ | $d = 2$ | $d = 3$ | $d = 4$ | $d = 5$ | $d = 6$ | $d = 7$ | $d = 8$ | $d = 9$ |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Router  | 0.52%   | 0.80%   | 1.25%   | 2.15%   | 3.25%   | 6.53%   | 85.50%  | 0       | 0       | 0       |
| EmailContact | 0.33% | 0.05%   | 0.01%   | 0.03%   | 0.02%   | 0.02%   | 0.29%   | 0.10%   | 0.06%   | 0.26%   |
| AS      | 0.31%   | 0.03%   | 0.02%   | 0.03%   | 0.03%   | 0.04%   | 0.03%   | 0.03%   | 0.06%   | 0.06%   |
| Email   | 1.06%   | 9.62%   | 10.33%  | 9.80%   | 7.50%   | 8.83%   | 11.56%  | 7.33%   | 8.83%   | 11.47%  |
| CA-Hep  | 0.37%   | 0.28%   | 0.24%   | 0.22%   | 0.12%   | 2.07%   | 3.62%   | 6.44%   | 10.23%  | 13.36%  |
| Hamster | 1.25%   | 4.25%   | 2.75%   | 0.70%   | 2.75%   | 1.85%   | 1.30%   | 6.25%   | 2.90%   | 3.40%   |
| PGP     | 0.38%   | 0.02%   | 0.67%   | 0.06%   | 0.03%   | 0.02%   | 0.19%   | 0.02%   | 0.23%   | 0.01%   |
| Netsci  | 2.38%   | 4.22%   | 5.54%   | 6.07%   | 26.91%  | 24.80%  | 22.96%  | 7.12%   | 0       | 0       |
| Astro   | 0.38%   | 0.38%   | 0.35%   | 0.38%   | 0.33%   | 0.30%   | 0.28%   | 0.01%   | 0.29%   | 0.56%   |

Text S1. Explanation of the $k_S$ imprecision demonstrated in Fig. S1. In Router, the imprecision is under 0.1 at first, and then rises at around $p \approx 0.15$. As the number of nodes in 1-shell accounts for 85.5% of the network size, the rising results from the random selection of nodes in 1-shell. In EmailContact and AS, within in 20% of the network size, the spreading of nodes in high shells are consistently larger than nodes in low shells. Thus the imprecision is low in the range. In Fig.S1(d), (e), (f) The imprecision of $k_S$ decreases with $p$. In Email, the core-like groups, 11-shell, accounts for 1.1% of the network size. From 10-shell, the spreading efficiency decreases with shells. The number of nodes in 10-shell accounts for about 10% of the network size. This results in the sharp decrease of $k_S$ imprecision until $p \approx 0.11$. In CA-Hep, the 31-shell, 23-shell, 20-shell and 18-shell accounts for about 1% of network size, which corresponds to the high $k_S$ imprecision for $p < 0.01$ (See Fig. 1 in main text). Next to the 18-shell is the 9-shell. From the 9-shell, the spreading efficiency decreases with shells. The proportion of nodes in shells 9, 8, 7, 6 is 0.1%, 2%, 3.6%, 6.4% respectively. Thus at $p \approx 0.13$, the imprecision of $k_S$ goes to a low value. In Hamster, the innermost 24-shell accounts for 1.25% of the network size, which corresponds to the high $k_S$ imprecision before $p = 0.0125$ (See Fig. 1 in main text). Then the $k_S$ imprecision decreases with $p$ until $p \approx 0.9$, which corresponds to the high spreading efficiency of 22-shell and 21-shell. In Fig.S1(g), (h), (i) In PGP, from the 31-shell to 22-shell, accounting for 1.4% of network size, the spreading efficiency decrease from the inner shell to periphery shell, which corresponds to the low $k_S$ imprecision (See Fig. 1 in main text). Then from 21-shell to 11-shell, locally densely connected groups occurs. The proportion of nodes of these shells accounts for 0.021 of the total scale. Thus , at $p \approx 0.035$, the $k_S$ imprecision goes down. In Netsci, core nodes accounts for 0.024 of network size, that corresponds to the the low ks imprecision before $p = 0.024$. Then in shell 7 and 6 , where the proportion of nodes are 4.2%, 5.5% respectively, locally densely connected groups occurs, which corresponds to the highest ks imprecision at $p \approx 0.12$. The proportion of nodes in 5-shell is 6.1%, and since then, the ks imprecision begins to decrease. In Astro, the low $k_S$ imprecision before $p = 0.015$ (See Fig. 1 in main text) corresponds to high spreading of core to the 51-shell. The 48-shell is a local group, accounting for 0.3% of the the network size. This corresponds to the sharp rise at $p \approx 0.015$. Then $\varepsilon_{k_S}$ goes down.
FIG. S1: The imprecision of $k_S$ and $k$ as a function of $p$ for nine real networks. $p$ is the proportion of nodes calculated, ranging from 0.02 to 0.2.
FIG. S2: The imprecision of $k_S$ and $k$ as a function of $k_S$ for degree-preserving randomized networks. In all the networks shown in (a)-(f), the $k_S$ imprecision is very low, under the value of 0.07, and in most cases lower than $k$ imprecision. For the randomized networks of Email and Hamster, although the $k_S$ imprecision is slightly higher than that of degree in some shells, the absolute values are very low, under 0.025. This indicates that the $k$-shell strategy is more effective than or at least as well as degree in most cases.
FIG. S3: The imprecision of $k_S$ and $k$ as a function of $k_S$ for degree-degree correlation preserving randomized networks. The imprecision of $k_S$ is very low in high shells and is lower than that of $k$.

FIG. S4: Clustering coefficient of shells for the real networks. (a), (b), (c) The average clustering coefficient is smaller than 0.1 in the innermost core shell. (d), (e), (f) The average clustering coefficient is greater than 0.5 in the innermost core.
FIG. S5: Link strength of shells for degree-degree correlation preserving randomized networks. The link strength of each shell to its lower shells $R_{lS}^k$ (black squares), equal shell $R_{eS}^k$ (red circles), and upper shells $R_{uS}^k$ (blue triangles) in the degree-degree correlation preserving randomized networks are represented. (a), (b), (c) $R_{lS}^k$ is much larger than $R_{eS}^k$ in high shells. (d), (e), (f) $R_{lS}^k$ is strongly promoted and is always larger than $R_{eS}^k$ in high shells in CA-Hep (e) and Hamster (f), although in Email (d) there is no obvious promotion.

FIG. S6: Link strength of the second innermost shell to each shell of the network. (a) Similar to the core, the second innermost shell are well connected to other parts of the network in Router (black squares), Emailcontact (red circles) and AS (blue triangles). (b) The link ratio within the second highest shell is lower than 0.6 in Email (black squares) and Hamster (blue triangles), but is still close to 1.0 CA-Hep (red circles).
FIG. S7: 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-degree correlation preserving randomized networks. In all the randomized networks, the core entropy is above 0.5.

FIG. S8: Link strength of shells for three real networks. The link strength of each shell to its lower shells $R_{k_s}^l$ (black squares), equal shell $R_{k_s}^e$ (red circles) and upper shells $R_{k_s}^u$ (blue triangles) are represented. For those core-like groups, 21-shell, 16-shell and 15-shell in PGP, 7-shell and 6-shell in Netsci and 48-shell and 30-shell in Astro, $R_{k_s}^u$ is much higher than $R_{k_s}^l$. 
FIG. S9: Infected population of $k_S$ shells as a function of infection probability, which is $q$ times of the epidemic threshold $\lambda_c$. $q$ ranges from 1 to 5. (a), (b), (c) In the first group, high shells are consistently reaching a higher infection population than low shells. (d), (e), (f) In the second group, the innermost cores have a relatively low efficiency than low shells. (g), (h), (i) In the third group, the innermost cores have the highest spreading efficiency. But there exists some shells of high $k_S$ index that have a lower efficiency than adjacent shells. They 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.
FIG. S10: Infected population of $k_2$ shells as a function of infection probability, which is $q$ times of the infected probability $\lambda$, $q$ ranges from 0.9 to 1.5. The relative spreading efficiency of shells is the same as the spreading when the infection probability is around $\lambda_c$ as shown in Figure S9.