OPEN ACCESS Freely available online

PLOS ONE

Extracting the Globally and Locally Adaptive Backbone of Complex Networks

Xiaohang Zhang[1]*, Zecong Zhang[1], Han Zhao[1], Qi Wang[1], Ji Zhu[2]

1 School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China, **2** Department of Statistics, University of Michigan, Ann Arbor, Michigan, United States of America

Abstract

A complex network is a useful tool for representing and analyzing complex systems, such as the world-wide web and transportation systems. However, the growing size of complex networks is becoming an obstacle to the understanding of the topological structure and their characteristics. In this study, a globally and locally adaptive network backbone (GLANB) extraction method is proposed. The GLANB method uses the involvement of links in shortest paths and a statistical hypothesis to evaluate the statistical importance of the links; then it extracts the backbone, based on the statistical importance, from the network by filtering the less important links and preserving the more important links; the result is an extracted subnetwork with fewer links and nodes. The GLANB determines the importance of the links by synthetically considering the topological structure, the weights of the links and the degrees of the nodes. The links that have a small weight but are important from the view of topological structure are not belittled. The GLANB method can be applied to all types of networks regardless of whether they are weighted or unweighted and regardless of whether they are directed or undirected. The experiments on four real networks show that the link importance distribution given by the GLANB method has a bimodal shape, which gives a robust classification of the links; moreover, the GLANB method tends to put the nodes that are identified as the core of the network by the k-shell algorithm into the backbone. This method can help us to understand the structure of the networks better, to determine what links are important for transferring information, and to express the network by a backbone easily.

Citation: Zhang X, Zhang Z, Zhao H, Wang Q, Zhu J (2014) Extracting the Globally and Locally Adaptive Backbone of Complex Networks. PLoS ONE 9(6): e100428. doi:10.1371/journal.pone.0100428

Editor: Peter Csermely, Semmelweis University, Hungary

Received January 12, 2014; **Accepted** May 28, 2014; **Published** June 17, 2014

Funding: This work was partially supported by NSFC (Grant Nos: 71371034 and 71372194; www.nsfc.gov.cn), the National Basic Research Program of China (Grant No: 2012CB315805; www.973.gov.cn), the Program for NCET (Grant No: NCET-13-0687; www.1000plan.org/qrjh/channel/159), and the Youth Research and Innovation Program in Beijing University of Posts and Telecommunications (Grant No: 2012RC1006; www.bupt.edu.cn). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zhangxiaohang@bupt.edu.cn

Introduction

In recent years, complex networks have been investigated by scholars in many domains. The representation, analysis and modeling in complex network theory bring a new paradigm to research on some complex systems including the Internet, transportation systems, biological systems, and social systems [1]. One of the primary aims of complex network research is to reveal the structural characteristics of complex systems. Many emerging concepts, such as the small-world property [2], scale-free behavior [3], community structure [4], and fractality [5], form the basis of our understanding of complex network structure. Because the scales of networks are becoming larger, a more intuitive and efficient method is required to represent and analyze the complex networks. Reducing a large-scale network to an essential backbone can help to solve the conflicts between the large scale of the complex networks and the understanding of the network structure. The backbone of a network is a core component that is extracted by filtering redundant information from the network and preserving far fewer links and nodes from the original network.

The filtering methods for backbone extraction can be divided into two main categories: global methods and local methods. Some global methods use certain global measures to filter the links, such as the link betweenness-based method [6] and the link weight-based method [7]. These methods apply a global threshold on the weights or the betweenness of links in such a way that only those that exceed the threshold are preserved. These filters have been used in the study of functional networks that connect correlated human brain sites [8], food web resistance as a function of link magnitude [9], and mobile communications networks [7]. The link weight-based method, however, could neglect nodes that have a small strength (The strength of node i is defined as $s_i = \sum_{j \in \aleph_i} w_{ij}$, where w_{ij} is the weight of the link (i,j) and $\aleph_i$ is the set of neighbors of node i) because the introduction of a threshold induces a characteristics scale from the outside [10].

The link salience [11], another type of global method, defines the shortest-path tree $T(r)$ that summarizes the shortest paths from a reference node r to the remainder of the network and that is conveniently represented by a symmetric $N \times N$ matrix (N is the number of nodes in the network) that has the element $t_{ij}(r) = 1$ if the link (i,j) is part of at least one of the shortest paths and $t_{ij}(r) = 0$ if it is not. The central idea of the approach is based on the notion of the average shortest-path tree that is defined as $S = \langle T \rangle = \frac{1}{N} \sum_{r=1}^{N} T(r)$. The element $0 \le s_{ij} \le 1$ of the matrix S

quantifies the fraction of the shortest-path trees that the link (i,j) participates in and denotes the salience of the link (i,j). Link salience is a robust approach to classifying network elements because the distribution of s, the link salience, exhibits a characteristic bimodal shape on the unit interval in many kinds of networks [11]. Link salience, however, tends to give an higher evaluation to the links being adjacent to low-degree nodes that often lie in the periphery of networks than the links being adjacent to high-degree nodes. For example, in Figure 1, link (i,p) is a part of the shortest-path tree $T(r)$ for all of the reference nodes, i.e., $s_{ip}=1$, because (i,p) is the only path that connects node p to the remainder of the network. Thus, link (i,p) is always a part of the backbone extracted by the link salience method even though the link is meaningful only for node $p$to transfer information between it and the rest of the nodes.

The local methods use local measures to determine which links must be filtered, such as the disparity filter method [10] and the locally adaptive network sparsification (LANS) [12]. The disparity filter method introduces the normalized weight that corresponds to link (i,j) of a certain node i of degree k_i and is defined as $p_{ij}=w_{ij}/s_i$, where w_{ij} is the weight of the link, s_i is the strength of node i. The normalized weight is assumed to be produced by a random assignment from a uniform distribution; thus, the probability density function of p_{ij} is assumed to be $f(x;k_i)=(k_i-1)(1-x)^{k_i-2}$. The backbone will include those links whose normalized weights satisfy the relation $a_{ij}=1-(k_i-1)\int_0^{p_{ij}}(1-x)^{k_i-2}dx<\alpha$ or $a_{ji}=1-(k_j-1)\int_0^{p_{ji}}(1-x)^{k_j-2}dx<\alpha$, where α is a specified significance level. Here a_{ij} and a_{ji} denote significance of the link's normalized weight not following the uniform distribution. The local heterogeneity (Section 3.1) of a link's weight is the premise of the disparity filtering method [10].

The LANS method, for each node i and for any of its neighbors j, considers the fraction of non-zero links whose weights are less than or equal to p_{ij}, $\hat{F}(p_{ij})=\frac{1}{|\aleph_i|}\sum_{m\in\aleph_i}\mathrm{IND}\{p_{im}\le p_{ij}\}$, where $\mathrm{IND}\{\}$ is the indicator function, $|\aleph_i|$ is the number of neighbors of node i, and p_{ij} is the normalized weight of link (i,j). If $1-\hat{F}(p_{ij})$ is less than a predetermined significance level α, the link (i,j) is locally significant and is included in the backbone network.

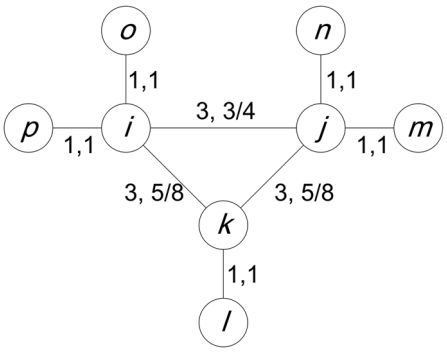

Figure 1. An undirected artificial network. The first number on the line is the value of the link weight, and the second number is the value of the link salience. Although the link (i,p) gets the largest value 1 of the link salience, it is only important for node p. The links (i,k) and (j,k) have the smallest value of the link salience, but they are in the core of the network.
doi:10.1371/journal.pone.0100428.g001

Although both of the local methods do not belittle some links that have small weights from a global view by considering the importance of the links in each specific node, we argue that they could ignore some links that have small weights with respect to the topological aspect. They assume that, for a certain node, its neighboring links (the links that connect to the node) with larger weights are more important. In many cases, however, local and global topological structures of a link determine how important the link is. For example, in Figure 2, although the weight of link (i,j) is greater than that of link (i,k), link (i,k) is more important than link (i,j) for node i because (i,k) is the path through which i can reach most of the other nodes. From the prospective of information transfer, link (i,k) can help node i send or receive information more effectively than link (i,j) can, because deleting link (i,k) could cause more damage than deleting link (i,j) for the information transfer of the network.

Because the local and global methods have advantages and disadvantages, in this study we aim to design a backbone extraction method that accounts for both the global and local topological structure of the networks. And the importance of links is synthetically determined by the weights of the links, the degree of the nodes, and the topological structure. The results of experiments on some real networks show that our propose method has some good characteristics.

Materials and Methods

In this study, we are inspired by the ideas of link salience and the disparity filter to propose a globally and locally adaptive network backbone (GLANB) extraction method. First, for each specific node, we compute the involvement of its neighboring links, which measures the fraction of the short paths connecting the node to the remainder of the network, which the links participate in. Second, we use a null hypothesis to determine whether each link is statistically important based on its involvement.

2.1 Link Involvement

We first consider a weighted, undirected and connected network. We define the length of the link (i,j) as $d_{ij}=1/w_{ij}$, with w_{ij} being the weight of link (i,j), which is consistent with definition of the link length in the link salience method. In most networks the link weights denote the connection strength between nodes. For example, in social networks the link weights often denote the communication frequency between people. Thus, we assume that the links with high weights are important in our case, and we invert the weights to compute the link length that measures the distance between nodes. In practice, the formula of measuring link

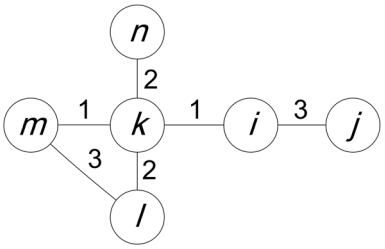

Figure 2. An undirected artificial network. The numbers on the lines denote the weights of the links. Although the weight of link (i,j) is greater than that of link (i,k), link (i,k) is more important for node i than link (i,j) is, because link (i,k) is the only path through which node i can reach the remainder of the network.
doi:10.1371/journal.pone.0100428.g002

length should depend on the meaning of the weights. The length of a path that connects two terminal nodes (n_1,n_T) and that consists of $T-1$ links by a sequence of intermediate nodes n_i, and the link weight $w_{n_in_{i+1}}>0$ is defined as $l=\sum_{i=1}^{T-1} d_{n_in_{i+1}}$. The shortest path minimizes the total distance l and can be interpreted as the most efficient route between its terminal nodes. The involvement I_{ij} of link (i,j) is defined as

$$I_{ij}=\frac{1}{N-1}\sum\nolimits_{s=1,s\neq i}^{N}\frac{g_{is}^{(i,j)}}{g_{is}}, \tag{1}$$

where N is the number of nodes in the network; $g_{is}^{(i,j)}$ is the number of shortest paths between node i and s that pass through the link (i,j); and g_{is} is the total number of shortest paths between node i and s. The involvement I_{ij} denotes how much the link (i,j) is involved in the most efficient connections between node i and the other nodes; thus, it can be a measure of the importance of link (i,j) for node i in the view of information transfer between node i and the remainder of the network. The larger the value of I_{ij} is, the more important link (i,j) is for node i. We can see that $\sum_{j\in\aleph_i} I_{ij}=1$, where $\aleph_i$ is the set of neighbors of node i.

The involvement is different from the betweenness centrality. The betweenness of link (i,j) depends on the shortest paths between all pairs of nodes, but the involvement I_{ij} only depends on the shortest paths between the node i and the rest of the nodes since the definition of involvement I_{ij} is based on the idea what proportion of the rest of the nodes can connect the node i through the link (i,j). The involvement is also different from the salience because the involvement considers the multiple shortest paths between each pair of nodes, but the salience assumes that only one shortest path exists between a pair of nodes. That is why the GLANB can also be applied to unweighted networks that often have multiple shortest paths between each pair of nodes.

2.2 Statistical Importance (*SI*) of Links

We find that the involvement of links that are around a single node is distributed heterogeneously (see Section 3.1). We are interested in the links that have a significant involvement at each given node. However, the local heterogeneity of involvement could simply be produced by random fluctuations. Similar to the disparity filter method, we adopt a null model to compute the random expectation for the distribution of the involvements that is associated with the links of a certain node. The null hypothesis is that the involvement I that corresponds to a connection of a certain node of degree k is produced by a random assignment from a probability density function of $f(x;k)$. Because the links that are adjacent to a certain node with the degree of k should have the same chance under the random condition to connect the node to the remainder of the network, the mean of the involvement must satisfy the condition

$$E(x;k)=\frac{1}{k},k\geq 2\text{and}x\in[0,1]. \tag{2}$$

Many probability density functions satisfy this condition and can be used to generate an involvement that is random and is based on specific assumptions. For example, if we assume that for each specific node that has the degree of k, its neighboring links independently participate in the shortest paths between the node and the remainder of the network with a probability of $1/k$; then, the involvement I obeys approximately the normal distribution that has a mean of $1/k$ and a variance of $\frac{1}{k}\left(1-\frac{1}{k}\right)/(N-1)$. Alternatively, we can assume that the involvement obeys the power law distribution $f(x)=\beta x^{\alpha}$ because for most complex networks, the degree and weight have been verified to follow power law distributions [1,3]. It is easy to obtain the probability density function $f(x;k)=\frac{1}{k-1}x^{-\frac{k-2}{k-1}},k\geq 2$. Moreover, the involvement can be assumed to follow a uniform distribution, which is similar to what the disparity filter method has performed for the normalized weights of the links [10] and has the probability density function of $f(x;k)=(k-1)(1-x)^{k-2}$.

The GLANB measures the statistical importance SI_{ij} of link (i,j) by using a null model to calculate the probability in such a way that its involvement I_{ij} is compatible with the null hypothesis. The statistical importance SI_{ij} of link (i,j) is defined as

$$SI_{ij}=1-\int_0^{I_{ij}} f(x;k_i)dx,k_i\geq 2, \tag{3}$$

where k_i is the degree of node i. In this study, the involvement is assumed to follow a uniform distribution, i.e., $f(x;k_i)=(k_i-1)(1-x)^{k_i-2}$; thus, $SI_{ij}=\left(1-I_{ij}\right)^{k_i-1},k_i\geq 2$. To control the impact of the degree on the statistical importance, we add a parameter $c\geq 0$ to the formula, as follows:

$$SI_{ij}=\left(1-I_{ij}\right)^{\left(k_i-1\right)^c},k_i\geq 2. \tag{4}$$

If $c=0$, then the statistical importance SI_{ij} is determined only by I_{ij} and is not affected directly by the degree (I_{ij} can be affected indirectly by k_i because the shortest paths to node i are affected by k_i). As c increases, the impact of the degree becomes larger. The experimental results show some interesting characteristics of the GLANB method under different values of c (see Section 3).

The smaller the value of SI_{ij} is, the more significantly the link (i,j) is not compatible with a random distribution; furthermore, the link (i,j) can be considered more important due to the network-organizing principles. The final statistical importance of an undirected link (i,j) is the minimum of SI_{ij} and SI_{ji}. In the case when a node i of degree $k_i=1$ is connected to a node j of degree $k_j>1$, the statistical importance of link (i,j) is SI_{ji}. The GLANB can identify a backbone of a network by setting the significance level α for the SI (a link is included in the backbone if its SI is less than α) based on the distribution of SI (see Section 3.3), or identify the hierarchical backbones by setting different significance levels since the backbone under high significance level will contain the backbone under low significance level. The backbone includes the links that are statistically important according to the specified significance level and their terminal nodes.

2.3 Unweighted and Directed Networks

The GLANB method can be easily applied in unweighted networks. In this case, the weights of all of the links are treated as being equal; thus, the length of a path is the number of links that lie in the path.

To be applied in directed networks, the GLANB must be modified. The directed link (i,j) from starting node i to ending node j is either an out-link for node i or an in-link for node j. Thus, we define the out (in) involvement $I_{ij}^{(out)}$ ($I_{ij}^{(in)}$) of the directed link (i,j) separately as

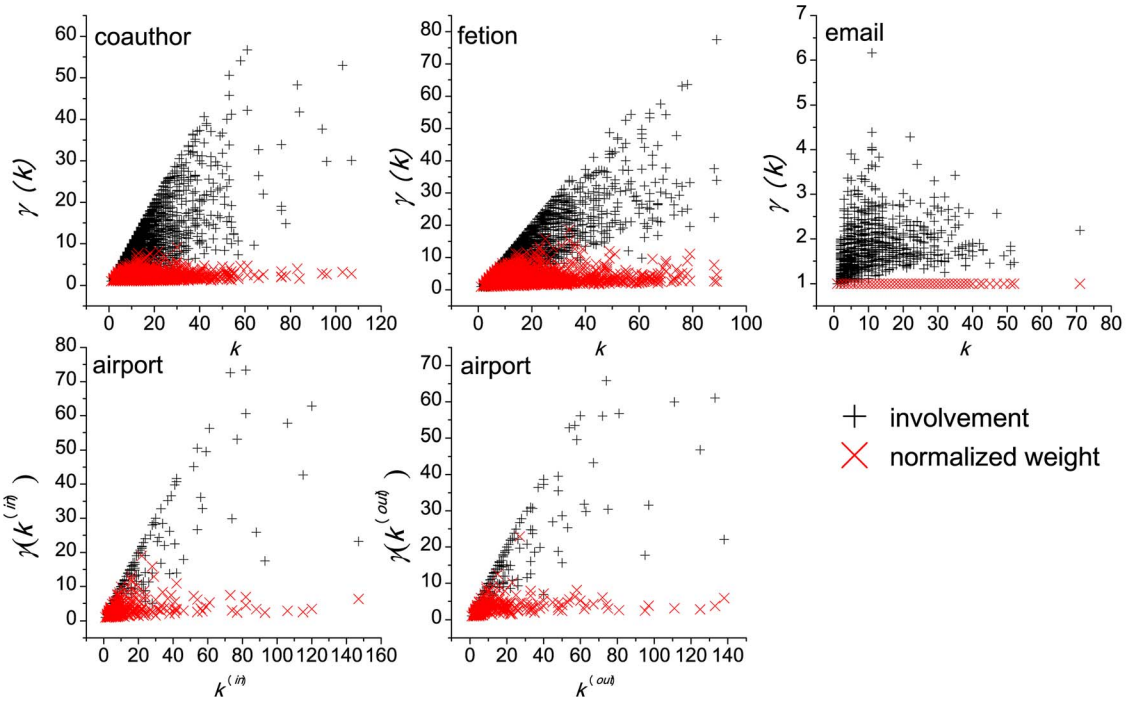

Figure 3. The local heterogeneity of the involvement and the normalized weight in four real networks. Each point in the figure denotes a node in the network. The local heterogeneity of the involvement for node i is defined as $\gamma_i(k_i)=k_i\sum_{j\in\aleph_i} I_{ij}^2$, where $\aleph_i$ is the set of neighbors of node i, k_i is the degree of node i, and I_{ij} is the involvement of link (i,j). The local heterogeneity of the normalized weight for node i is defined as $\gamma_i'(k_i)=k_i\sum_{j\in\aleph_i}\left(w_{ij}/s_i\right)^2$, where $s_i=\sum_{j\in\aleph_i} w_{ij}$ is the strength of node i. We can find that for all of the networks, the involvement is locally more heterogeneous than the normalized weight is.
doi:10.1371/journal.pone.0100428.g003

$$I_{ij}^{(out)}=\frac{1}{|\aleph_i^{(out)}|}\sum\nolimits_{s\in\aleph_i^{(out)}}\frac{g_{is}^{(i,j)}}{g_{is}}\ \text{and}$$

$$I_{ij}^{(in)}=\frac{1}{|\aleph_j^{(in)}|}\sum\nolimits_{s\in\aleph_j^{(in)}}\frac{g_{sj}^{(i,j)}}{g_{sj}}, \tag{5}$$

where $\aleph_i^{(out)}$ is the set of nodes that can be reached from node i through a directed path, and $\aleph_j^{(in)}$ is the set of nodes that can reach node j through a directed path; $|\aleph_i^{(out)}|$ denotes the size of $\aleph_i^{(out)}$; $g_{is}^{(i,j)}$ is the number of shortest paths from node i to s that pass through the link (i,j); and g_{is} is the total number of shortest paths from node i to s. The involvement $I_{ij}^{(out)}$ measures how much the link (i,j) is involved in the shortest paths from node i to the other nodes, and $I_{ij}^{(in)}$ measures how much the link (i,j) is involved in the shortest paths from the other nodes to node j.

The statistical importance of link (i,j) is composed of two parts, the in-importance $SI_{ij}^{(in)}$ and the out-importance $SI_{ij}^{(out)}$, which are defined from the viewpoint of the starting node i and the ending node j separately as

$$SI_{ij}^{(out)}=\left(1-I_{ij}^{(out)}\right)^{\left(k_i^{(out)}-1\right)^c},k_i^{(out)}\geq 2\ \text{and}$$

$$SI_{ij}^{(in)}=\left(1-I_{ij}^{(in)}\right)^{\left(k_j^{(in)}-1\right)^c},k_j^{(in)}\geq 2 \tag{6}$$

where $k_i^{(out)}$ is the out-degree of node i, $k_j^{(in)}$ is the in-degree of node j, and c is the control parameter. The final statistical importance of the directed link (i,j) is determined by the minimum of $SI_{ij}^{(out)}$ and $SI_{ij}^{(in)}$. Similar to the case of weighted and undirected networks, the GLANB can identify a backbone from unweighted or directed network by setting a significance level for SI based on the distribution of SI, or a hierarchical backbone by setting different significance levels for SI.

Results

To test the performance of the GLANB method, we apply it to four real-world networks, a collaboration network (coauthor) [13], an instant-message network (fetion), an email network (email) [14] and an airport traffic network (airport). We compare the obtained results with those obtained by the disparity filtering method and the link salience method. (1) The collaboration network is based on co-authorship of academic papers in the high-energy physics community from 1995–1999. Nodes represent individuals, and links measure the number of papers that were co-authored. The data are publicly available at http://www-personal.umich.edu/~mejn/netdata/. (2) The instant-message network is based on an instant-message tool, fetion, which is provided by Mobile Corporate. The nodes represent fetion users, and the links measure the number of messages sent between each pair of users. (3) The email network is an undirected and unweighted network. The nodes represent email users, and the links represent whether any communication exists between each pair of users. The email network data are available at http://deim.urv.cat/~aarenas/data/welcome.htm. (4) The airport traffic network is a weighted

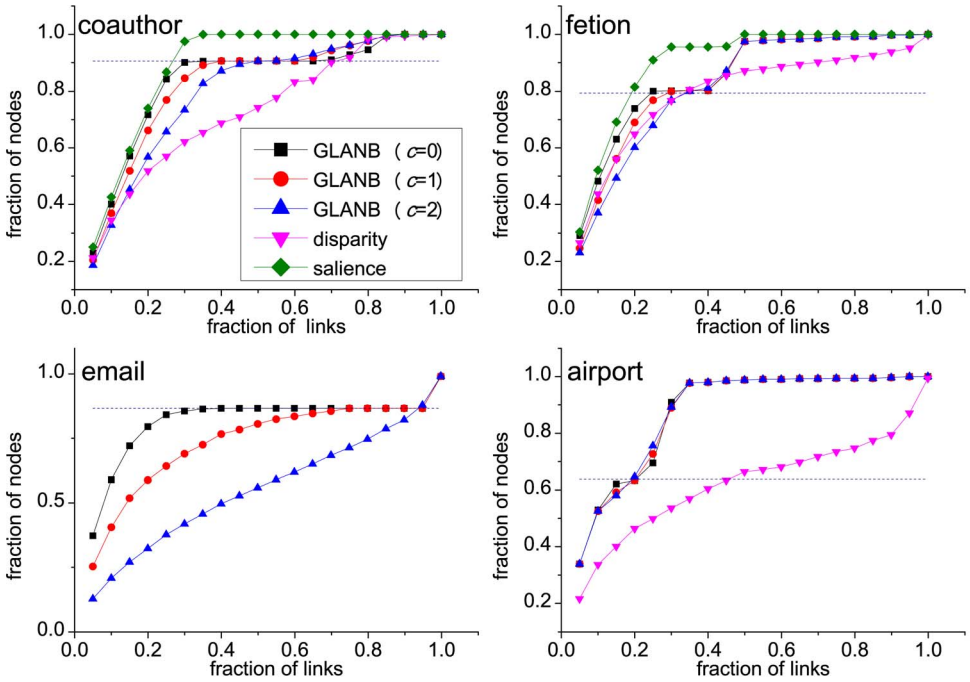

Figure 4. Fraction of nodes maintained in the backbones. The fraction of nodes is a function of the fraction of links retained by the filters. The dash lines correspond to the fraction of the nodes whose degree is greater than 1 in the networks.
doi:10.1371/journal.pone.0100428.g004

and directed network. It measures global air traffic that is based on flight data that is provided by OAG Worldwide Ltd. (http://www.oag.com), and it includes all of the scheduled commercial flights in the world in 2011. The nodes represent airports worldwide. The link weights measure the total number of passengers that travel between a pair of airports by direct flights per year. This network is well represented in the literature [15,16,17]. In the experiments only the largest connected subnetworks of each of the networks are used. The backbone includes the links that are significantly important according to the extraction methods and their terminal nodes. Because the authors in [11] do not mention how the salience method deals with the directed or unweighted networks, we do not apply the salience method to the email and the airport networks.

3.1 Local Heterogeneity of the Link Involvement

The condition under which the null model can perform well is that for each node, its links' involvement shows heterogeneity. If this condition is not satisfied, then it is difficult to identify the important links through the GLANB method. To assess the effect of heterogeneities in the links' involvements at the local level, for each node i of degree k_i, one can calculate the function [18,19]

$$\gamma_i(k_i) = k_i Y_i(k_i) = k_i \sum_{j \in \aleph_i} I_{ij}^2, \tag{7}$$

where $\aleph_i$ is the set of neighbors of node i and I_{ij} is the involvement of link (i,j).

As a standard indicator of measuring the concentration of data, the function $Y_i(k_i)$ has been extensively used in various domains, including ecology, economics, physics, and complex networks [10,19], where it is known as the disparity measure. Under perfect homogeneity, when all of the links share the same amount of the involvement of node i(i.e., $I_{ij} = 1/k_i$), $\gamma_i(k_i)$ equals 1 independently of k_i, while in the case of perfect heterogeneity, when only one of the links carries the whole involvement of the node, the function is $\gamma_i(k_i) = k_i$. In this way, this function can be used as a preliminary indicator of the presence of local heterogeneity. When local heterogeneity of involvements exists, the GLANB can be more useful than in the case of homogeneity because the GLANB aims to identify the links whose involvements are significantly higher than other neighboring links'. To compare the involvement with the weights of the links, we also compute the heterogeneity of the normalized weights [10] by

$$\gamma_i'(k_i) = k_i \sum_{j \in \aleph_i} \left(\frac{w_{ij}}{s_i}\right)^2, \tag{8}$$

where $s_i = \sum_{j \in \aleph_i} w_{ij}$ is the strength of node i. Figure 3 shows the local heterogeneity of the involvement and the normalized weight in the coauthor, fetion, email, and airport networks. We can find that for all of the networks, the involvement is locally more heterogeneous than the normalized weight is (Figure 3). These results indicate that applying the null model to the involvement can identify the statistically important links well.

3.2 Size of the Backbones

The main purpose of extracting backbones is to reduce the number of links in networks, while keeping more nodes. To measure the effects of these filtering methods on the extracted backbones, we analyze the relative sizes of the backbones as a function of the preserved fractions of the links when the network is filtered by the disparity filter, by the link salience and by the GLANB (Figure 4).

For the four real networks, the link salience method can preserve the largest fraction of nodes in the backbone, and the disparity filter method preserves the smallest (except when the

fraction of links is less than 0.4 for the fetion network) when the same fraction of links is maintained. The results of the GLANB methods fall in between the disparity and the salience methods. We must note that for the salience method, all of the links that are adjacent to the nodes with a degree of 1 have the largest salience of 1, and preserving these links can retain at least one node. Thus, the link salience method can preserve the largest fraction of the nodes when filtering the networks.

We also find that in the backbone of the coauthor and fetion networks identified by the GLANB method at the specified values of control parameter c, the fraction of nodes stays approximately unchanged for an interval of the fraction of links when the fraction of nodes reaches the threshold that is the fraction of nodes with a degree greater than 1. For the email networks, this phenomenon also exists when $c=0$ or $c=1$. For the airport networks, this phenomenon exists when $c=0$. The interval of keeping unchanged is the longest for all of the networks when $c=0$ (Figure 4). The reason of the phenomenon is that for the nodes that have a degree of 1, the value of SI of their neighboring links is very close to 1; thus, these nodes are difficult to include in the backbone when the fraction of links in the backbone is not sufficiently large. Moreover, as the control parameter c increases, the growth curves of the fraction of nodes become relatively flat (Figure 4), because high value of c prefers the links that correspond to the nodes that have a high degree, and these links have a low value of SI. Preserving these links in the backbone cannot increase the fraction of nodes proportionally because some other links that could have been preserved in the backbone are more likely to share the same terminal nodes with them. Thus, these results indicate that the parameter c can control the size of extracted backbone by impacting the degrees of the nodes on the value of the involvement.

3.3 Robust Classification of Links Based on the Statistical Importance

Similar to the link salience measure [11], the surprising feature of the statistical importance SI is that the distribution $p(SI)$ exhibits a characteristic bimodal shape on the unit interval (Figure 5). The networks' links naturally accumulate at the boundaries and have a small fraction at intermediate values. The statistical importance thus successfully classifies network links into two groups: important ($SI \approx 0$) or non-important ($SI \approx 1$). Because a small fraction of links fall into the intermediate range, the resulting classification is not significantly sensitive to an imposed threshold. This circumstance is fundamentally different from some link centrality measures, such as weight and betweenness, which possess broad distributions and which require external and often arbitrary threshold parameters to perform meaningful classifications. The distribution of links' statistical relevance when measured by the disparity filter method shows a unimodal shape in the coauthor network or a flat distribution in the fetion network (Figure 5), which has the result that choosing the appropriate significance level α to filter links becomes difficult. For the GLANB method, as the control parameter c increases, the number of links with high importance increases (Figure 5). The reason is that the GLANB method with a high value of c favors the links that correspond to the nodes that have the degree $k>1$, and these links occupy a large proportion of total links.

3.4 K-shell distribution of links

To deeply explore the hierarchy of links in the backbones that are extracted by the GLANB, disparity filter and salience methods, we use the k-shell decomposition method to compare the topological distribution of extracted links. The k-shell decomposition method is often used to identify the core and the periphery of the networks [20,21]. Although the k-shell method only takes into account the nodes' degree not the link weights, it provides a way to compare the backbone extraction methods from the view of topological structure. The process of the k-shell decomposition starts by removing all of the nodes that have one link (degree 1) only, until no more such nodes remain; then, it assigns them to the 1-shell. In the same manner, it recursively removes all of the nodes that have a degree of 2 (or less), creating the 2-shell. This process continues, increasing k until all of the nodes in the network have been assigned to one of the shells. The shells that have high indices lie in the core of the network. To assign all of the links to the shells, we define the shell index of a link as the minimum of its two terminal nodes' shell indices.

For the coauthor, fetion and email networks, we extract the top 10% important links based on the SI_{ij} of GLANB (from low to high), the a_{ij} of disparity filter (from low to high) and the s_{ij} of salience methods (from high to low) separately to analyze their distributions in terms of link-shells. Because the salience method ranks the links for which one terminal node has the degree of 1 as most important, and because both the disparity filter and the GLANB methods rank them as least important, we also exclude these links to extract the remaining top 10% important links based on the salience method (salience-E) to analyze the distribution. The distributions of the links in the range of the shell index are shown in Figure 6. We can see that compared with the disparity and salience methods, the GLANB ($c=2$) extracts more links that lie in the higher shells, i.e., the topological core of the networks. Especially for the salience method, most of the extracted links lie in the lower shells. This circumstance occurs because the links whose terminal nodes have a low degree tend to have a high salience. For example, the links that are adjacent to the nodes that have the degree of 1 have the highest salience of 1, which means that all of the links in the 1-shell are certain to be in the backbone that is extracted by the link salience method. For the salience-E method, most of the links still fall in the low shells, and the distribution almost coincides with that of the GLANB ($c=0$), which ignores the degree of the corresponding nodes in a similar way as the salience method. As the control parameter c increases, more links fall into the higher link-shells.

There are two reasons to explain why the GLANB ($c>0$) is more likely to extract links from the topological core of the networks than the other methods. One reason is that the backbone which is extracted by the GLANB ($c>0$) method does not include the links that are adjacent to the nodes that have a degree of 1. The second reason is that the null model depends on the degrees of the nodes. When I_{ij} stays unchanged, increasing the value of degree k_i can decrease the value of SI_{ij} in a power-law way (see formula 4). The larger the value of c is, the more greatly k_i affects SI_{ij}. Thus, some links that have a higher shell index would be in the backbone even though their involvement values are not very high. Furthermore, from Figure 3, we can see that the distribution of link involvements for the nodes that have a higher degree shows heterogeneity, which means that some links have both a high-degree terminal and a high involvement.

Discussion

The GLANB method accounts for both the global and local topology structure of the network when extracting the backbones. On the one hand, the involvement of each link is either a global measure (because it depends on the shortest paths that are determined by the global network structure and the link weights) or a local measure (because the sum of the involvements of the

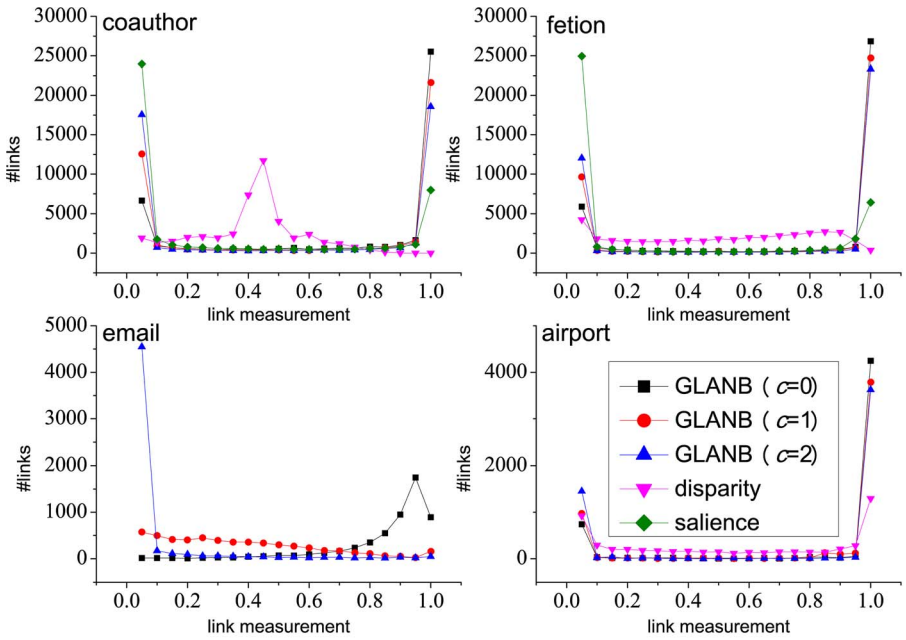

Figure 5. The distributions of the link salience, the link statistical importance and the disparity filtering importance. Link measurement refers to the values of the link salience, link statistical importance, and the disparity filtering importance that are given by the salience, GLANB and disparity methods separately. For the GLANB and disparity methods, the smaller values mean higher importance. For the salience method, the larger values mean higher importance.
doi:10.1371/journal.pone.0100428.g005

links that are adjacent to any certain node has the value of 1). On the other hand, the null model that is adopted in GLANB is based on a local view because the probability density function depends on the degree of each certain node. Thus, the GLANB determines the importance of the links by synthetically considering the topological structure, the weights of the links and the degrees of the nodes. In this method, the links that have a small weight but are important from the view of structure are not belittled.

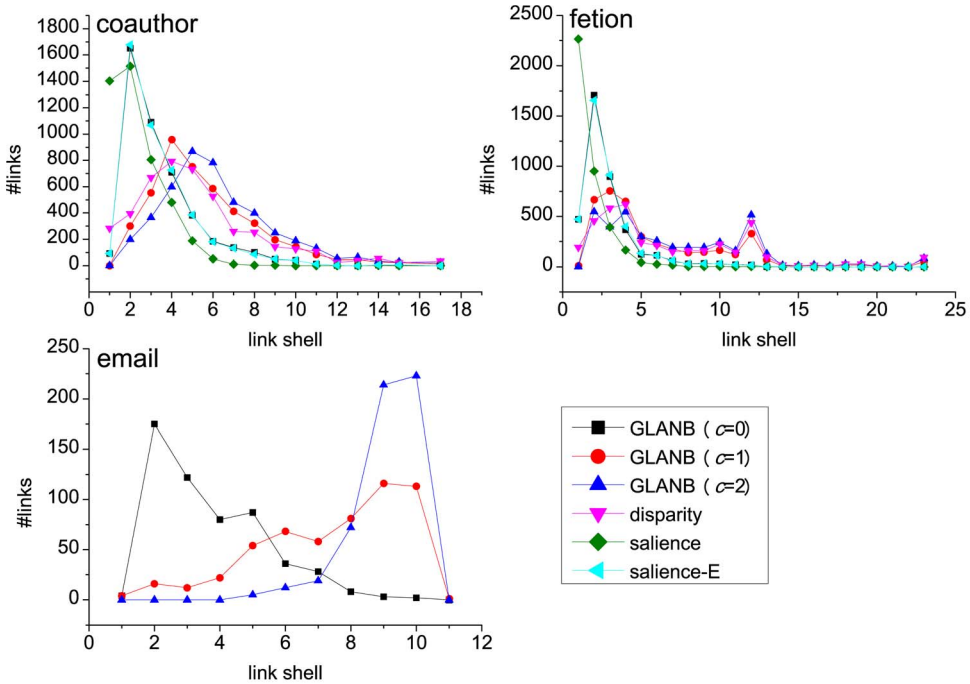

Figure 6. The distribution of links in link-shells. For the coauthor, fetion and email networks, we extract the top 10% important links, based on the GLANB, disparity filter and salience methods separately, to analyze their distributions in terms of link-shells. In addition, we also exclude the links that have degree of 1 to extract the remaining top 10% important links based on the salience method (salience-E) to analyze the distribution.
doi:10.1371/journal.pone.0100428.g006

Furthermore, introducing the control parameter c into GLANB provides a way to adjust the impacts of the node degrees on the extracted backbones, which makes the backbone adaptive to the global structure and the local structure by changing the value of c. When $c \to 0$, the backbone mainly concentrates on the global structure. When the value of c becomes larger, the backbone is affected more greatly by the local structure. Another advantage is that the GLANB method can be applied to all types of networks regardless of whether they are weighted or unweighted and regardless of whether they are directed or undirected.

The computational complexity of the GLANB method is determined by the computation of the involvement and the statistical importance of the links. To compute the involvement, we must find all of the shortest paths between each pair of nodes, which results in the computational complexity being $O(NL+N^2\ln(N))$ [22], where N is the number of nodes and $L$the number of links in the network. The computation of the statistical importance must scan all of the links to compute the degrees of the nodes and the SI of the links; thus, the computational complexity is $O(L)$. Because $L<N^2$, the computational complexity of GLANB is $O(NL+N^2\ln(N))$. When the size of the network is very large, GLANB is not adaptable if it is executed on only a single computer. However, the computational environment has recently been changing dramatically. Parallel computation platforms are being used pervasively because of their low implementation costs and high performance. Because the GLANB method is based on each single node to measure the involvement and statistical importance of their neighboring links, it is easy to implement GLANB on a parallel platform.

The experiments on the real-world networks show some interesting results. First, the link involvements show local heterogeneity that arises from the topological structure of the networks and from the heterogeneous weight distributions because the shortest paths are determined by those two aspects. Moreover, the involvement is more heterogeneous in the weighted network than in the unweighted network (Figure 3). Second, the link importance distribution, which shows a bimodal shape, gives a robust classification of the links. The bimodal distribution comes from both the local heterogeneity of involvement and the null model that is adopted in GLANB. Third, as the fraction of links in the backbones increases, the size of the backbones that are extracted by the GLANB method first increases rapidly and then becomes almost unchanged, and at last, increases again. The GLANB method assesses the links that are adjacent to the nodes that have a degree of 1 as the least important; thus, as the number of links in the backbone increases, the size of the extracted backbone becomes unchanged for an interval when only the nodes with a degree of 1 are not included in the backbone. Fourth, the control parameter c can affect the size of the backbones. A larger value of c decreases the growth rate of the size of the backbone, because the links that are adjacent to the nodes that have a larger degree are favored, and they cannot efficiently add more nodes into the backbone. Fifth, the GLANB method tends to give more importance to the nodes that are in the core of the network than the other methods do. Especially as the control parameter c increases, more nodes in the core are included in the backbone. In practice, the choice of c value depends on what backbone is needed. The larger the value of c is, more likely the backbone includes the links that are adjacent to nodes with high degree and that are in the core of network from the view of k-shells, and more likely includes less nodes at preserving the same proportion of links.

The GLANB method aims to extract backbones from networks by filtering unimportant links, which can decrease the size of the network greatly. Thus, this method can help us to understand the structure of the networks better, to determine what links are important to transferring information, to express the network by a graph picture easily, and to control the network densities.

Author Contributions

Conceived and designed the experiments: XZ QW JZ. Performed the experiments: XZ ZZ HZ. Analyzed the data: XZ HZ ZZ. Contributed reagents/materials/analysis tools: XZ ZZ HZ. Wrote the paper: XZ QW JZ.

References

1. Newman M, Barabasi A-L, Watts DJ (2006) The structure and dynamics of networks. Princeton: Princeton University Press.
2. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
3. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science (Washington D C) 286: 509–516.
4. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821–7826.
5. Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. Nature 433: 392–395.
6. Goh K-I, Salvi G, Kahng B, Kim D (2006) Skeleton and fractal scaling in complex networks. Phys Rev Lett 96.
7. Zhang X, Zhu J (2013) Skeleton of weighted social network. Physica A 392: 1547–1556.
8. Eguiluz VM, Chialvo DR, Cecchi GA, Baliki M, Apkarian AV (2005) Scale-Free Brain Functional Networks. Phys Rev Lett 94: 018102.
9. Allesina S, Bodini A, Bondavalli C (2006) Secondary extinctions in ecological networks: Bottlenecks unveiled. Ecol Model 194: 150–161.
10. Serrano MA, Boguna M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106: 6483–6488.
11. Grady D, Thiemann C, Brockmann D (2012) Robust classification of salient links in complex networks. Nat Commun 3: 864.
12. Foti NJ, Hughes JM, Rockmore DN (2011) Nonparametric Sparsification of Complex Multiscale Networks. PLoS ONE 6.
13. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98: 404–409.
14. Ruimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68: 065103.
15. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci USA 101: 3747–3752.
16. Guimera R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci USA 102: 7794–7799.
17. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439: 462–465.
18. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427: 839–843.
19. Barthelemy M, Gondran B, Guichard E (2003) Spatial structure of the internet traffic. Physica A 319: 633–642.
20. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. Proc Natl Acad Sci USA 104: 11150–11154.
21. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, et al. (2010) Identification of influential spreaders in complex networks. Nat Phys 6: 888–893.
22. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25: 163–177.