Principal Component Analysis Based Measure of Structural Holes

Shiguo Deng, Wenqing Zhang, Huijie Yang
Biz School, University of Shanghai for Science and Technology, Shanghai 200093, China
E-mail: hjyang@ustc.edu.cn

Abstract. Based upon principal component analysis, a new measure called compressibility coefficient is proposed to evaluate structural holes in networks. This measure incorporates a new effect from identical patterns in networks. It is found that compressibility coefficient for Watts-Strogatz small-world networks increases monotonically with the rewiring probability and saturates to that for the corresponding shuffled networks. While compressibility coefficient for extended Barabasi-Albert scale-free networks decreases monotonically with the preferential effect and is significantly large compared with that for corresponding shuffled networks. This measure is helpful in diverse research fields to evaluate global efficiency of networks.

1. Introduction
Structural hole [1] is a novel concept for structural analysis of social networks. Let us consider two nodes or clusters between which there exist not direct linkage or indirect redundancy. This kind of structural pattern will induce resistance to information exchange between the two parts. This hinder effect is called structural hole. As shown in Fig.1, three rectangles and triangle Y connect directly with triangle X, while the three rectangles and Y do not link with each other, namely, X is the single route by which information exchange can occur. X is much more competitive in collecting and controlling information. There exists a typical structural hole. As for the triangle Y, it has an equivalent position with X, but the direct linkages between the three rectangles that are connected with Y introduce redundancy edges besides the routes bridged by Y, which reduces significantly the control ability of Y. Structural hole does not exist here.

Burt [1] introduces the concept of network constraint index to measure quantitatively structural hole. Denoting with $P_{i,j}$ the cost of node $i$ on maintaining its relation with node $j$, the network constraint index of $i$ reads,

$$H_i = \sum_{j=1}^{N} (W_{i,j} P_{i,j} + \sum_{q=1}^{N} W_{i,q} P_{i,q} P_{q,j} W_{q,j})^2,$$

(1)

where $W$ is the adjacency matrix of a social network, namely, $W_{i,j} = W_{j,i} = (0)$ if node $i$ and node $j$ are (dis-)connected, respectively. A large value of $H_i$ means that the node $i$ is positioned in a dense local pattern, which leads to a low degree of structural hole. But this measure is limited to the contributions from the nearest and the secondary neighbors. Especially, the redundancy from linkages of its neighbors is not included explicitly.

Published under licence by IOP Publishing Ltd
Figure 1. Illustration of structural hole. X is the only way for its neighbor’s exchanging information. X is positioned at a typical hole. As for Y, linkages between its neighbors induce redundancy. Y is not positioned at a structural hole.

Figure 2. A typical curve for percent explained versus N/N. N = 2000. Each pair of nodes is linked with probability 1/2000.

Figure 3. Compressibility coefficients for WSSW networks. N = 2000.

Figure 4. Compressibility coefficients for EBA networks. N = 2000.

Besides the contributions of direct linkages, indirect redundant linkages, and linkages between neighbors, there exists actually another important effect from identical patterns in networks. In a social network, optimal strategies in one node is widely duplicated by others. And common environments may force similar nodes to employ the same strategies. Strategies of a node determine its relationships with other nodes. Hence, there may exist some structural patterns occurring with high frequencies. This kind of symmetry may decrease significantly the cost of a node, because the node can use the same strategies in keeping its relationships with structurally identical groups. Once one group obtains information from other nodes, its response action will also be duplicated by the other groups with the same patterns. By this way, information is transferred to a large size in the social network. As a result, degree of structural hole will decrease. This effect is not included in the original definition of structural hole. In the present paper, by means of principal component analysis (PCA) we take into account this effect.

2. Methods and Materials

2.1. Compressibility Coefficients of Networks

Principal component analysis (PCA) [2] is a powerful tool to eliminate redundancy in data, by using of which raw data is compressed. Let us consider a social network described with adjacency
matrix $A$ with size $N$. The element $A_{i,j}$ is 1 and 0 if $i$ and $j$ are connected or disconnected, respectively. Redundant edges between two nodes, $i$ and $j$ means the two nodes share many nodes as their neighbors. At the same time, if the neighbors of node $i$ are densely connected, they will share much more nodes as neighbors, also. These two kinds of redundancy induce high degree of symmetry of the nodes in the social network. The corresponding data can be compressed much more effectively. Hence, compressibility of adjacency matrix can be used to measure structural holes of the social network.

From adjacency matrix $A$, one can calculate the covariance matrix $\text{Cov}$, whose elements read,

$$\text{Cov}(i,j) = \frac{\sum_{k=1}^{N} (A_{k,i} - \frac{1}{N} \sum_{s=1}^{N} A_{s,i}) \cdot (A_{k,j} - \frac{1}{N} \sum_{s=1}^{N} A_{s,j})}{\sqrt{\sum_{k=1}^{N} (A_{k,i} - \frac{1}{N} \sum_{s=1}^{N} A_{s,i})^2 \cdot \sum_{k=1}^{N} (A_{k,j} - \frac{1}{N} \sum_{s=1}^{N} A_{s,j})^2}},$$

$i, j = 1, 2, \ldots, N.$

Ranking the eigenvalues of matrix $\text{Cov}$ in ascending order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, percent explained by the first $n$ eigenvalues reads,

$$P(n) = \frac{\sum_{s=1}^{n} \lambda_s}{\sum_{s=1}^{N} \lambda_s} \times 100\%.$$

A typical curve for $P(n)$ versus $\frac{n}{N}$ is illustrated in Fig.2. Compressibility coefficient of the social network is defined as,

$$G \equiv \frac{S_{a\rightarrow b\rightarrow c\rightarrow d\rightarrow a}}{S_{a\rightarrow b\rightarrow c\rightarrow e\rightarrow a}} = 2S_{a\rightarrow b\rightarrow c\rightarrow d\rightarrow a},$$

where $S_{a\rightarrow b\rightarrow c\rightarrow d\rightarrow a}$ and $S_{a\rightarrow b\rightarrow c\rightarrow e\rightarrow a}$ are the surface areas of the domains surrounded by curves $a \rightarrow b \rightarrow c \rightarrow d \rightarrow a$ (filled with grey) and $a \rightarrow b \rightarrow c \rightarrow e \rightarrow a$, respectively. A larger value of $G$ tells us that to reach the same percent explained, much more nodes are required to be included. Consequently the corresponding network has low level of redundancy and there exist much more structural holes.

2.2. Data

We consider the Watts-Strogatz small-world (WSSW) model [3]. One can construct initially a regular lattice with periodic boundary condition, in which each node is linked with its right-handed $d$ neighbors. Randomness is then introduced by rewiring with probability $p_r$ destination of each link to another randomly selected node. Double or degeneration of linkages are strictly forbidden.

We consider also an extended version of Barabasi-Albert scale-free (EBA) model [4]. Starting from several connected nodes as seed, at each step a new node is connected to the existing network with $e_{\text{new}}$ edges. The probability for an existing node $i$ being linked with the new comer is proportional to $k_i^\alpha$, where $k_i$ is the degree of node $i$. For the case of $\alpha = 1$, the EBA model degenerates to the standard Barabasi-Albert scale-free model, while for $\alpha = 0$ new comer will be attached randomly to the existing network without any preferential effect.

Detailed calculations show that the results are robustness for changing of size. Results for $N = 2000$ are shown in present paper. The structural properties are obtained by average over 100 realizations. All the networks are shuffled by means of the so-called edge-exchange algorithm [5].
3. Results

Fig. 3 shows the compressibility coefficients for WSSW networks with $d = 2, 3$ and 4. One can find that with the increase of rewiring probability $p_r$, $G$ increases monotonically. When $p_r$ is large enough (e.g., $p_r \approx 0.5$), $G$ saturates to values very close to that for the corresponding shuffled networks. While with the increase of $p_r$, $G$ for shuffled networks decreases monotonically but slightly. At the same time, increase of $d$ leads to significant decrease of $G$. The significant low degree of structural hole comes from non-trivial patterns rather than degree distribution. On the initial regular lattice, each node has $2d$ duplications and the two sets of $d$ duplications separated by it form two sub-complete graphs. Tightness and high symmetry of structure induces low degree of structural hole.

Fig. 4 presents the compressibility coefficients for EBA scale-free networks. One can find that with the increase of $\alpha$, i.e., increase of preferential effect, $G$ decreases monotonically. For $\epsilon_{new} = 1$, with the increase of $\alpha$ the difference between $G$ values for the initial and the shuffled networks becomes nondistinctive. While for $\epsilon_{new} = 2, 3, 4$, the difference keeps significant large with the increase of $\alpha$. $G$ values for original networks are large compared that for the shuffled ones. Hence, the high degree of structural hole comes from nontrivial patterns in original networks.

4. Conclusions

In the present work we propose the principal component analysis to evaluate degree of structural hole in a global way. For WSSW networks with the increase of rewiring probability degree of structural hole increases monotonically and saturates to a value very close to that for corresponding shuffled networks. While increase of connection density decreases significantly degree of structural hole. For EBA networks, with the increase of $\alpha$, namely the preferential effect, degree of structural hole decreases monotonically. For $\epsilon_{new} = 1$, $G$ tends to a value very close to that for corresponding shuffled networks, while for $\epsilon_{new} > 1$, $G$ keeps significant larger compared with that for shuffled networks.

PCA-based measure of structural hole provides a powerful tool to evaluate in a global way social networks. Comparing with the initial definition of structural hole, it takes account of a new effect in generating mechanism of structural hole, namely, identical patterns occurring in networks.

Acknowledgements

The work is supported by the National Science Foundation of China under Grant Nos. 10975099 and 50974052, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Innovation Program of Shanghai Municipal Education Commission under Grant No.13YZ072, and the Shanghai leading discipline project under grant No.S30501. One of the authors (W. Zhang) thank the support of the Innovation Fund Project For Graduate Student of Shanghai under Grant No. JWCXSL1102.

References

[1] R. S. Burt, The social structure of competition (Cambridge: Harvard University Press,1992).
[2] K. Pearson, Philosophical Magazine 2,559(1901).
[3] D.J. Watts, S.H. Strogatz, Nature 393,409(1998).
[4] A.-L. Barabasi, and R. Albert, Science 286(5439),509(1999).
[5] S. Maslov, K. Sneppen, Science 296,910(2002).