Recognition of Important Subgraphs in Collaboration Networks

Chun-Hua Fu, Yue-Ping Zhou, Xiu-Lian Xu, Hui Chang, Ai-Xia Feng, Jian-Jun Shi, and Da-Ren He*

College of Physics Science and Technology, Yangzhou University, Yangzhou, 225002, China
darendo10@yahoo.com.cn

Abstract. We propose a method for recognition of most important subgraphs in collaboration networks. The networks can be described by bipartite graphs, where basic elements, named actors, are taking part in events, organizations or activities, named acts. It is suggested that the subgraphs can be described by so-called k-cliques, which are defined as complete subgraphs of two or more vertices. The k-clique act degree is defined as the number of acts, in which a k-clique takes part. The k-clique act degree distribution in collaboration networks is investigated via a simplified model. The analytic treatment on the model leads to a conclusion that the distribution obeys a so-called shifted power law $P(q) \propto (q+\alpha)^{-\gamma}$ where α and γ are constants. This is a very uneven distribution. Numerical simulations have been performed, which show that the model analytic conclusion remains qualitatively correct when the model is revised to approach the real world evolution situation. Some empirical investigation results are presented, which support the model conclusion. We consider the cliques, which take part in the largest number of acts, as the most important ones. With this understanding we are able to distinguish some most important cliques in the real world networks.

Keywords: subgraph, collaboration network, bipartite graph, clique, shifted power law.

1 Introduction

Complex networks have attracted attentions in recent years [1,2]. Among the studies, the investigations on social networks become increasingly more attractive. An obvious feature of social networks is the community structure of the basic elements ("vertices" or "actors"). Inside communities the connections ("edges" or "ties" or "arcs") between actors are much denser than the connections between the communities [3]. This indicates that the active or self-determining actors get together and form groups. In addition to social networks, Zhang, Chang, Su, Fu and cooperators suggested considering the networks, in

* Corresponding author.

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 210–219, 2009.

which the basic elements were not active, however, they might be influenced by some "manipulators" and showed indirect activity [4-7]. We may call such networks "quasi-social". Examples of quasi-social networks may be mentioned as transportation networks (with stations as the quasi-actors), language networks (with languages as the quasi-actors), and other man-made or man-collected systems.

One of the most interesting questions in social network studies is whether communities can be divided into some basic subgraphs. The basic subgraphs are composed of several (2 to k) actors, which are always together in performing a network function. For example, if a movie actor and an actress always perform sweet heart couple in many famous movies, the audience thinks them as a fixed unit. They will strongly unsatisfy if one of them suddenly performs couple with a different partner. In social networks, the pair of actors can be defined as a "dyad", which is "an unordered pair of actors and the arcs that exist between the two actors in the pair" (Ref. [3], page 510). In this paper we shall only discuss mutual dyads, in which both of the actors has a directed tie to the other (Ref. [3], page 511), and use an ordinary un-directed tie to express the two directed ties. A "triad" is similarly defined as a triple of actors with the ties between them (Ref. [3], page 559). We also only consider mutual triads where all the three pairs of actors are connected by un-directed ties. In social networks "a clique is defined as a maximal complete subgraph of three or more vertices" (Ref. [3], page 254). If we revise the definition to "a complete subgraph of two or more vertices", we can address the mutual dyads as "2-cliques" and mutual triads as "3-cliques". We can similarly define "k-cliques".

Interesting research achievements have been published about social collaboration networks, including Hollywood actor collaboration network and scientist collaboration network [8-11]. The collaboration networks can be described by bipartite graphs. In these graphs the vertices can be divided into two sets [3]. One type of the vertices is "actors" taking part in some activities, organizations or events. The other type of vertices is the activity, organization or event named "acts". When we note only the collaboration relationship between actors, we may project the bi-graph onto the actor-vertices, and obtain a unipartite graph. In the projected graph each act is represented by a complete subgraph where edge is connected between every pair of vertices. Each act complete subgraph can be divided into some smaller complete subgraphs, or cliques. Different act complete subgraphs may share some cliques. A clique is more important if it takes part in more acts. The number of vertices in an act complete subgraph is addressed as "act size" and denoted by T. The number of acts, in which an actor takes part, is addressed as "act degree" of the actor vertex and denoted by h [4-7]. Zhang, Chang, Su and Fu and cooperators investigated some quasi-social collaboration networks [4-7].

In this article we shall concentrate on the recognition of most important cliques in social and quasi-social collaboration networks. The most important cliques must take part in the largest number of acts, thus they must show the largest k-clique act degree (defined as the number of acts, in which the k-clique

takes part). The k-clique act degree distribution then becomes the most important property in the current study. The article will be organized as follows. In section 2 we shall develop a model describing the evolution of social and quasi-social collaboration networks in a very ideal and simplified situation. By analysis of the model we can show the general function form of the k-clique act degree distribution in the simplified case. For the situations nearer to practical, we shall show, by some numerical investigation, that the general function form probably is qualitatively correct. In section 3 we shall present empirical investigation results in three real world quasi-social collaboration networks as proofs, which show good agreement with the model conclusion. We also shall recognize most important 2-cliques and 3-cliques in these real world networks. The empirical investigation results on eight different real world collaboration networks shall also be mentioned. In the last section the text will be summarized and some discussions will be presented.

2 The Model

2.1 The Definition of k-Clique Act Degree

Now we introduce accurate definition of k-clique act degree. In a bipartite graph the act degree of a 2-clique (a mutual dyad) is defined as $D_{i,j} = \sum_m a_{i,m}a_{j,m}$, where i, j denotes two different actors, and m denotes an act. $a_{i,m}$ is the element of the bipartite graph adjacency matrix. It is defined as $a_{i,m} = 1$ if actor i takes part in act m (they are connected by a bipartite graph edge); and $a_{i,m} = 0$ otherwise. Similarly, the act degree of a 3-clique (a mutual triad) is defined as $Tr_{i,j,k} = \sum_m a_{im}a_{jm}a_{km}$, where i, j, k denotes three different actors, m denotes an act. One can then write the general definition of k-clique act degree as $q_{i_1,i_2,\cdots,i_k} = \sum_m a_{i_1,m}a_{i_2,m}\cdots a_{i_k,m}$. A k-clique act degree distribution $P(q)$ is defined as the probability of a k-clique with act degree q, stands for the number of k-cliques with act degree q in the network [12].

2.2 The Simplified Model

The ideal and simplified situation we consider firstly is that the act size (the number of vertices in an act), T, is a constant, also, it takes a value, $k \times n$ where n is an integer, so that each act can just include n "legal k-cliques" where k is a constant. All the vertex actors unite together and form legal k-cliques. There are just $k \times N$ vertex actors in the network therefore every actor is in a legal k-clique. There is no single vertex remaining in the network. During the evolution of the network, the legal k-cliques will not disband, and each vertex actor can only join one of them. The legal k-cliques perform as a fixed unit in the network evolution. At each time step, a new legal k-clique joins network and select $n-1$ old legal k-cliques by certain rule to form a new act. Of course, when a new act is formed, because all the edges between every pair of vertex actors must be connected, certainly some "illegal k-cliques", which share actor vertices with

legal k-cliques, should appear. However, for the simplified model considered in this subsection we shall only count the act degree of legal k-cliques.

Firstly we consider the rule of selecting $n-1$ old legal k-cliques with a probability proportional to the k-clique act degree q of each old legal k-clique. This can be addressed as a "k-clique act-degree linear preference rule". We can, similar to what BA did [1,9], get a conclusion that the legal k-clique act degree distribution $P(q)$ takes exact power functions.

Secondly, we consider the rule of selecting $n-1$ old legal k-cliques randomly. We can write down, following the references [1,4,5,9], the evolution equation of the legal k-clique act degree and analytically obtain the conclusion that the legal k-clique act degree distribution $P(q)$ takes exact exponential functions.

To interpolate between the above two extreme cases [4,5,13,14], we consider the rule of selecting the $n-1$ old legal k-clique randomly with a probability p, and using legal k-clique act degree linear preference rule with probability $1-p$. Similarly, we have (when t is large)

$$\frac{\partial q_i}{\partial t} = p\frac{n-1}{t} + (1-p)\frac{(n-1)q_i}{nt}. \tag{1}$$

This equation can be written as

$$\frac{\partial q_i}{\partial \ln(t)} = p\frac{T-k}{k} + (1-p)\frac{T-k}{T}q_i. \tag{2}$$

This can be solved to give

$$q_i = C_i t^{(T-k)(1-p)/T} - \frac{Tp}{k(1-p)}, \tag{3}$$

where C_i is the integration constant, which can be determined using the condition $q_i(t=t_i)=1$. Let $\alpha = Tp/[k(1-p)]$ and $\eta = T/[(T-k)(1-p)]$. Now we have

$$P(q_i < q) = P(t_i > t(\frac{q+\alpha}{1+\alpha}))^{-\eta}. \tag{4}$$

The legal k-clique act degree distribution is then given by

$$P(q) = \frac{dP(q_i<q)}{dq} = \frac{\eta}{1+\alpha}(\frac{q+\alpha}{1+\alpha})^{-\eta-1}. \tag{5}$$

The legal k-clique act degree distribution function is called "Shifted Power Law" (SPL) [5]. Now let's check the limiting case. For $p=0$, $\alpha=0$ and $\eta = T/(T-k)$; it is easy to see that

$$P(q) \propto q^{-\frac{kT-k}{T-k}}. \tag{6}$$

For $p \to 1$, $\alpha \to \infty$, and $\eta \to k\alpha/(T-k)$,

$$P(q) \propto e^{k(1-q)/(T-k)}. \tag{7}$$

So the distribution we obtained for $0 < p < 1$ interpolates between the power-law distribution and the exponential distribution. When the parameter p continuously changes from 0 to 1, $P(q)$ continuously varies from a power-law distribution to an exponential distribution.

2.3 The Models Approaching the Real World Network Evolution

In the real world network evolution one usually cannot distinguish the legal k-cliques and illegal k-cliques, therefore it is unreasonable to ignore the illegal k-clique act degree. Also, the act size (the number of vertices in an act), T, in general cannot just take the value $k \times n$ therefore the acts may include some "isolated vertices". In this subsection we shall consider both the situations. However, the act size, T, will be considered still as a constant since our previous investigation results showed that this simplification could be accepted for many real world network studies [4,5].

It is difficult to solve the more complex model analytically, we have to discuss it numerically. The results show that the conclusions are qualitatively the same. Several figures showing the numerical investigation results for $k = 2$ have been already published in Ref. [15]. Some numerical investigations have been done for $k = 3$, which show similar results. We can therefore expect a model conclusion that in collaboration networks 2-clique and 3-clique act degree distribution obeys SPL functions. Considering the simplified model prediction we believe that at least in a large potion of real world collaboration networks all the k-clique act degree distributions obey SPL functions. SPL means a very uneven k-clique act degree distribution. Only a few k-cliques take part in a lot of acts, they can be viewed as fixed units and the most important subgraphs in the network. Most of the k-cliques take part only in a few acts. The vertices in the cliques join different k-cliques with different partners when perform cooperation function in different acts. They cannot be thought of as fixed units, neither as important subgraphs.

3 Empirical Investigations on Some Real World Collaboration Networks

The above mentioned model analysis and simulations only discuss one fixed k value although the discussion is rather general for all k. In the real world collaboration networks, however, many k-cliques (with different k) coexist. If we could get empirical investigation results, which show that the k-clique act degree distributions for several values of k obey SPL functions, we can believe that the model conclusion is correct at least for a large portion of real collaboration networks. In this section we present such empirical investigations.

3.1 Empirical Investigation on World Language Distribution Network

World language distribution has been of research interests [16]. We propose a network description on world language distribution. The actor vertices of the network are defined as languages. The acts are defined as countries or regions where the languages are spoken. Two vertices are connected by an edge if they are coexisting in a common region. The data were downloaded from the fifteenth

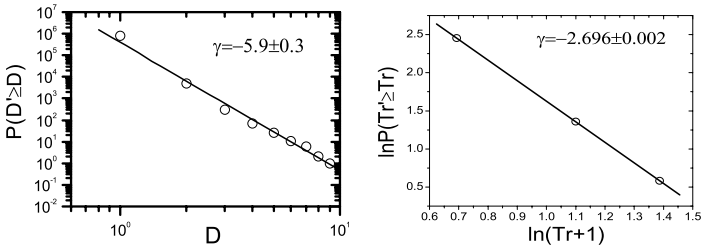

Fig. 1. Left: The cumulative dyad (2-clique) act degree distribution of the language network; Right: The cumulative triad (3-clique) act degree distribution of the language network.

edition of Ethnologue [17], published in 2005, which lists 6142 languages and 228 countries plus the 8 regions.

Figure 1 shows that the cumulative dyad (2-clique) act degree distribution of the language network, which can be described with a power law, $P(D' \geq D) \propto D^{-5.9}$. With an approximation that the data number is large and quasi-continuous, one can easily prove that the original dyad act degree distribution obeys $P(D) \propto D^{-4.9}$. This is an extreme case of a SPL distribution. Figure 1 also shows the cumulative triad (3-clique) act degree distribution of the language network, which can be described with a SPL function, $P(Tr' \geq Tr) \propto (Tr + 1)^{-2.696}$.

The most important 2-cliques can be listed as: 1) Izora and Mangas (appear together in 9 regions); 2) Mbulungish and Pular (appear together in 8 regions); 3) Mann and Mbulungish (appear together in 7 regions); 4) Mbulungish and Bainouk-Gunyu Adobe (appear together in 7 regions); 5) Mann and Pular (appear together in 7 regions).

The most important 3-cliques can be listed as: Biali, Gude and Majera; Biali, Gyele and Wawa; Giziga(South), Majera and Mofu(North); Giziga(South), Majera and Wawa; and Giziga(South), Mofu(North) and Wawa. All these 3-cliques take part in 3 acts.

3.2 Empirical Investigation on Mixed Drink Network

The acts of mixed drink network are defined as mixed drinks. The actor vertices are defined as drink ingredients. Two vertices are connected by an edge if they are coexisting in a common act. The data were downloaded from the website http://www.drinknation.com, which lists 1501 drink ingredients and 7804 mixed drinks.

Figure 2 shows the cumulative dyad (2-clique) act degree distribution of the mixed drink network, which can be described with a SPL, $P(D' \geq D) \propto (D + 20)^{-2.81}$. Figure 2 also shows the cumulative triad (3-clique) act degree distribution of the mixed drink network, which can be described with a SPL, $P(Tr' \geq Tr) \propto (Tr + 5)^{-8.4}$.

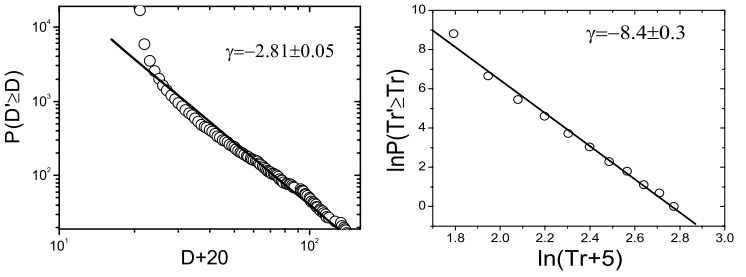

Fig. 2. Left: The cumulative dyad (2-clique) act degree distribution of the mixed drink network; Right: The cumulative triad (3-clique) act degree distribution of the mixed drink network

The most important 2-cliques can be listed as: 1) Vodka and OrangeJuice (appear together in 273 drinks); 2) OrangeJuice and PineappleJuice (appear together in 201 drinks); 3) OrangeJuice and Grenadine (appear together in 195 drinks); 4) Vodka and TripleSec (appear together in 191 drinks); 5) Vodka and GranberryJuice (appear together in 177 drinks).

The most important 3-cliques can be listed as: 1) PineappleJuice, Rumdark and Rum; 2) OrangeJuice, PineappleJuice and Grenadine; 3) OrangeJuice, PineappleJuice and Grenadine; 4) OrangeJuice, Amaretto and SouthernComfort; 5) IrishCreamBaileys, Kahlua and Vodka; 6) Amaretto, SouthernComfort and SloeGin.

3.3 Empirical Investigation on Information Technology Product Network in China

The data are downloaded from the websites: http://www.pcpop.com/ and http://www.it168.com/. In the IT product network the companies are defined as the actor vertices, and the IT products are defined as the acts. An edge between two actor vertices represents that the two companies produce at least one common IT product and thus compete in the market. This network includes 265 IT products and 2121 IT companies.

Figure 3 shows the cumulative dyad (2-clique) act degree distribution of the IT product network, which can be described with a power law, $P(D' \geq D) \propto D^{-3.7}$. Figure 3 also shows the cumulative triad (3-clique) act degree distribution of the mixed drink network, which can be described with a SPL, $P(Tr' \geq Tr) \propto (Tr+3)^{-10.2}$.

The most important 2-cliques can be listed as: 1) BIGBUFFALO (HEDY company) and MASTER (produce 16 common products); 2) Kangguan ViewTech Center and Unika (produce 15 common products); 3) SAMSUNG and NEWGRAND (produce 14 common products); 4) SAMSUNG and Adobe (produce 13 common products); 5) MASTER and SAMSUNG (produce 13 common products).

The most important 3-cliques can be listed as: 1) BIGBUFFALO, MASTER and TESSM (produce 10 common products); 2) BIGBUFFALO, MASTER and

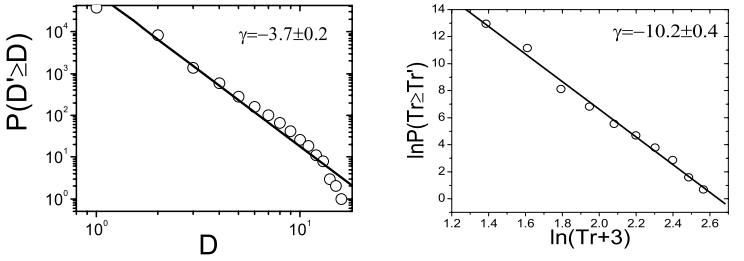

Fig. 3. Left: The cumulative dyad (2-clique) act degree distribution of the IT product network; Right: The cumulative triad (3-clique) act degree distribution of the IT product network

SAMSUNG (produce 10 common products); 3) BIGBUFFALO, MASTER and NEWGRAND (produce 9 common products); 4) Kangguan ViewTech Center, Unika and Oneforall (produce 9 common products); 5) MASTER, SAMSUNG and NEWGRAND (produce 9 common products).

3.4 Some Other Empirical Investigated Collaboration Networks

We have investigated another 9 collaboration networks. 8 of them also show SPL 2-clique and 3-clique act degree distributions:

(1) The undergraduate elective network of Yangzhou University (YZU) [7]: Vertices: 121 general support courses, Edge: two vertices are belong to a common scientific subject, Acts: 78 scientific subjects, Data: provided by University Academic Affairs office.

(2) The Chinese professional training organization network: Vertices: 2674 training courses, Edge: two vertices are provided by a common training organization, Acts: 398 training organizations, Data: http://www.ot51.com/, http://www.00100.cc/, http://philosophy.cass.cn/, and http://www.people.com.cn/ et al..

(3) China mainland movie network [18]: Vertices: 3085 movies, Edge: a common movie actor performs in these two movies, Acts: 920 movie actors, Data: http://soft6.com/, http://www.mtime.com/movie/.

(4) 2004 Olympic game network: Vertices: 4496 athletes, Edge: two athletes take part in a common sports item, Acts: 229 sports items,

Data: http://2004.sina.com.cn/results/summary/.

(5) Traditional Chinese herb prescription formulation network [4,5]: Vertices: 681 herbs, Edge: two herbs included in a prescription, Acts: 1536 prescriptions, Data: Refs. [19,20].

(6) Beijing bus route network [4,5]: Vertices: 4199 bus stations, Edge: two stations in a common route, Acts: 1572 bus routes, Data: http://www.bjbus.com/.

(7) The Travel Route Network of China [4,5]: Vertices: 171 scenic spots, Edge: two scenic spots in a travel route, Acts: 240 routes, Data: http://www.cnta.com/8-ssls/lyqd.asp.

(8) Network of Huai-Yang recipes of Chinese cooked food [4,5]: Vertices: 242 foods, Edge: two foods form a dish, Acts: 329 recipes, Data: Ref. [21].

In our empirically investigated networks, the only exception is the Fruit Nutritive Factor Network [22]: Vertices: 45 nutritive factors, Edge: one fruit contains these two nutritive factors, Acts: 151 fruits, Data: http://www.fumuqin.com/. Both the 2-clique and 3-clique act degree distributions follow normal distribution functions.

Therefore our empirical investigations strongly support the model prediction.

4 Conclusion and Discussion

We show, with a very simplified network evolution model, that k-clique act degree distribution probably always obey SPL functions in collaboration networks. Some empirical proofs are presented, which have been obtained in some real world systems and show SPL 2-clique and 3-clique act degree distributions. This indicates that small complete subgraphs do widely exist in collaboration networks, which may include some non-social networks. However, only a few of them take part in many collaboration acts so that they can be considered as important basic units of the networks. It is worth noting that some other real world collaboration networks show k-clique act degree distributions, which do not obey SPL distribution. The evolution mechanism in these collaboration networks must be basically different and thus deserves a further investigation.

Acknowledgement

The research is supported by the Chinese National Natural Science Foundation under the grant numbers 10635040 and 70671089.

References

1. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex network. Rev. Mod. Phys. 74, 47–97 (2002)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–225 (2003)
3. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge Univ. Press, Cambridge (1994)
4. Zhang, P.P., Chen, K., He, Y., et al.: Model and empirical study on some collaboration networks. Physica A 360, 599–616 (2006)
5. Chang, H., Su, B.-B., Zhou, Y.-P., He, D.-R.: Assortativity and act degree distribution of some collaboration networks. Physica A 383, 687–702 (2007)
6. Su, B.-B., Chang, H., Chen, Y.-Z., He, D.-R.: A game theory model of urban public traffic networks. Physica A 379, 291–297 (2007)
7. Fu, C.-H., Zhang, Z.-P., Chang, H., et al.: A kind of collaboration-competition networks. Physica A 387, 1411–1420 (2008)

8. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)
9. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
10. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E 64, 016131 (2001); Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths. weighted networks and Centrality. Phys. Rev. E 64, 016132 (2001)
11. Ramasco, J.J., Dorogovtsev, S.N., Pastot-Satorras, R.: Self-organization of collaboration networks. Phys. Rev. E 70, 036106 (2004)
12. Krapivsky, P.L., Redner, S.: Rate equation approach for growing networks. In: P-Satorras, R., Rubi, M., D-Guilera, A. (eds.) Statistical Mechanics of Complex networks, p. 4. Springer, Heidelberg (2003)
13. Liu, Z., Lai, Y.-C., et al.: Connectivity distribution and attack tolerance of general networks with both preferential and random attachments. Phys. Lett. A 303, 337–344 (2002)
14. Li, X., Chen, G.: A local world evolving network model. Physica A 328, 274–286 (2003)
15. Chang, H., He, D.-R.: General collaboration networks. In: Guo, L., Xu, X.-M. (eds.) Complex Networks, pp. 166–186. Shanghai Scientific and Educational press, Shanghai (2006) (in Chinese)
16. Gomes, M.A.F., Vasconcelos, G.L., Tsang, I.J., Tsang, I.R.: Scaling relations for diversity of languages. Physica A 271, 489–495 (1999)
17. `http://www.ethnologue.com/`
18. Liu, A.-F., Fu, C.-H., Chang, H., He, D.-R.: An empirical statistical investigation on Chinese mainland movie network. Complex Sys. and Complexity Sci. 4, 10–16 (2007) (in Chinese)
19. Zhu, Y.-X.: Chinese Herb Prescription Manual, 2nd edn. Jindun Publishing House, Beijing (1996) (in Chinese)
20. Liu, D.-L., et al.: Most Frequently Used Chinese Medication Handbook. People's Military Medication Publisher, Beijing (1996) (in Chinese)
21. Compiling group of Beijing Nationality Restaurant.: Huai-Yang Bill of Fare. Chinese Travel Publisher, Beijing (1993) (in Chinese)
22. Qu, Y.Q., Jiang, Y.M., He, D.-R.: Fruit Nutritive Factor Network. Jrl. Syst. Sci. and Complexity 22, 150–158 (2009)