Research Article

Social Network Analysis Based on Network Motifs

Xu Hong-lin, Yan Han-bing, Gao Cui-fang, and Zhu Ping

School of Science, Jiangnan University, Wuxi, Jiangsu 214122, China

Correspondence should be addressed to Zhu Ping; zhuping@jiangnan.edu.cn

Based on the community structure characteristics, theory, and methods of frequent subgraph mining, network motifs findings are firstly introduced into social network analysis; the tendentiousness evaluation function and the importance evaluation function are proposed for effectiveness assessment. Compared with the traditional way based on nodes centrality degree, the new approach can be used to analyze the properties of social network more fully and judge the roles of the nodes effectively. In application analysis, our approach is shown to be effective.

1. Introduction

A large number of systems in the real world exist as networks, such as social networks (coauthor network, criminal networks, etc.), biological networks (protein interaction networks, metabolic networks, etc.), and technology networks (electricity networks, the Internet, etc.) [1–12]. In order to reveal their structure and principle, Milo et al. first proposed the concept of "network motifs," which can be defined as patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized network [13]. Later, research on network motifs has been developed extensively. Kim et al. defined biological network motifs as biologically significant subgraphs [14]. Farina et al. identified regulatory network motifs from gene expression data, and they proposed the corresponding algorithm [15]. In order to specify network motifs, Ohnishi et al. analyzed an interfirm network consisting of about one million firms and four million directed links [16].

The study of social networks has always been a hot research topic. In order to judge the importance of nodes, the staple methods of traditional social network analysis are basing on the calculation of the centrality of nodes in network, [17, 18]. In recent years, various new methods are introduced into social network analysis; network motif is an important kind of them [19, 20]. Analyzing motifs for the large social networks derived from email communication firstly, Juszczyszyn found that the distribution of motifs in all analyzed real social networks is similar and can be treated as the network fingerprint. This property is most distinctive for stronger human relationships [21, 22].

In this paper, we introduce network motifs to develop a set of network analysis methods, which is different from the traditional social network analysis, and also illustrate its application.

2. Research Methods

2.1. Directed Graph and Point Centrality. A network with n nodes is denoted by $G = (V, E)$, where $V = \{v_i \mid i = 1, 2, \ldots, n\}$ is node set, $E = \{\langle v_i, v_j \rangle \mid v_i$ is the start node, v_j is the end node, and $i, j = 1, 2, \ldots, n\}$ is edge set. $A = (a_{ij})_{n \times n}$ is adjacency matrix of G, of which the elements are as follows:

$$a_{ij} = \begin{cases} 1, & \langle v_i, v_j \rangle \in E, \\ 0, & \langle v_i, v_j \rangle \notin E, \end{cases} \tag{1}$$

where $i, j = 1, 2, \ldots, n$.

The centrality analysis is the staple method of traditional social network analysis [17, 18]. In a network, if there are direct links between an actor and other actors, this actor resides in the centre of the network, having more "power" [17]. The importance of a node, point centrality, can be measured by the number of contacted nodes [18]. Based on

the adjacency matrix, the formula of point centrality of node v_i is as follows:

$$C(v_i) = c_i = \sum_{j=1}^{n} a_{ij} + \sum_{k=1}^{n} a_{ki}, \qquad (2)$$

where $i = 1, 2, \ldots, n$.

2.2. Network Motifs Finding

2.2.1. Frequent Graph. Frequent subgraph mining is an important method of network information mining. Frequent subgraph mining algorithms are divided into breadth-first search (BFS) algorithm and depth-first search (DFS) algorithm based on subgraph search path [23]. As a breadth-first search algorithm, Apriori graph mining (AGM) algorithm is an early adopter of Apriori idea. AGM algorithm takes an adjacency matrix to represent the graph. Then it generates code based on adjacency matrix and takes minimum coding as unique identification for the graph in order to solve NP problem of subgraph isomorphism [24].

The graph which is constituted by K node set $V_s = \{v_{s1}, v_{s2}, \ldots, v_{sk}\}$ and edge set $E_s = \{\langle v_{s_i}, v_{s_j} \rangle \mid v_{s_i}, v_{s_j} \in V_s\}$ is a subgraph of G. Let $SG_s = G(V_s) = (V_s, E_s)$. Furthermore, the node set of subgraph SG_s is denoted as $V_s = V(SG_s)$.

Based on the adjacency matrix of a subgraph, the maximum encoding is obtained as unique identification of the subgraph. AGM algorithm is used to mine frequent subgraphs based on the maximum encoding.

2.2.2. Random Network Model. In typical network motifs finding algorithms, random network model maintains the degree sequence of the real network very well [25]. Exchange algorithm is an algorithm for generating random network according to degree sequence, which is as follows [26].

Algorithm A.

> Input: degree sequence
>
> Output: random network
>
> Step 1: Construct a network according to degree sequence.
>
> Step 2: Randomly select a pair of edges (e.g., $A \to B$, $C \to D$).
>
> Step 3: Carry out the Monte Carlo exchange ($A \to D$, $C \to B$).
>
> Step 4: Cancel the exchange if the exchange has led to multiple edges or loops.
>
> Step 5: Repeat until reaching the target number of times.

In this way, a set of random networks with the same degree sequence as G can be obtained.

2.2.3. Statistical Significance of Network Motifs. Network motifs are frequent subgraphs with special statistical significance, which have some special functions in the network.

Network motifs satisfy the following conditions: occurrence of the subgraph in real network is not less than a minimum and is significantly higher than their occurrence in random network [13, 27].

The statistical significance of network motifs is denoted by Z-score:

$$Z = \frac{N_{\text{real}} - \langle N_{\text{rand}} \rangle}{\sigma_{\text{rand}}}, \qquad (3)$$

where N_{real} denotes the occurrences of a subgraph in real network and $\langle N_{\text{rand}} \rangle$ and σ_{rand} denote mean and standard deviation of the occurrences of the subgraph in random networks.

2.3. Frequency Matrix.

The nodes of a network always have a lot of roles, such as teacher and student. Most social networks can be simulated by role network model (RNM). The role set is denoted by $R = \{R_i = 1, 2, \ldots, m\}$, where m is the number of roles. The set of nodes whose role is R^k ($k = 1, 2, \ldots, m$) is denoted by V_k^R, so the set of V_k^R is denoted by $V^R = \{V_1^R, V_2^R, \ldots, V_m^R\}$. MSG denotes the set of subgraphs whose structure is the same as the network motifs in the network.

In order to determine the role tendentiousness of the unknown role nodes, we can count the frequency of the unknown role nodes occurring in different network motifs, respectively, through the composition of nodes in network motifs, which contain different known role nodes.

Based on network motifs, frequency matrix $F = (f_{ij})_{n \times m}$ is obtained. The elements f_{ij} of F denotes the total occurrences of node v_i in the network motifs that contain known role nodes, whose role is R_j.

The algorithm for calculating frequency matrix is as follows.

Algorithm B.

> Input: V, V^R, MSG
>
> Output: F
>
> Step 1: For any $i \in \{1, 2, \ldots, n\}$, for any $j \in \{1, 2, \ldots, n\}$ $f_{ij} \Leftarrow 0$.
>
> Step 2: For any $MSG_i \in MSG$, for any $V_j^R \in V^R$, if $V(MSG_i) \cap V_j^R \neq \emptyset$, then for any $v_{i_k} \in MSG_i$, $f_{i_k j} = f_{i_k j} + 1$.
>
> Step 3: Output $F = (f_{ij})_{n \times m}$.
>
> Step 4: End.

2.4. Evaluation Function

2.4.1. Tendentiousness Evaluation Function (TEF). Based on frequency matrix F, the tendentiousness of node v_i with respect to role R_j is evaluated by TEF.

TABLE 1: Point centrality.

| Name	c_i	Name	c_i	Name	c_i	Name	c_i	Name	c_i
Gretchen	29	Lois	14	Dwight	11	Cory	6	Tran	3
Sherri	26	Donald	14	Marion	11	Jia	6	Quan	3
Jerome	23	Neal	14	Malcolm	11	Shelley	6	Olina	3
Neal	22	Wayne	14	Karen	11	Wesley	6	Phille	2
Julia	22	Marian	14	Yao	10	Chris	6	Cole	2
Franklin	21	Crystal	14	Douglas	10	Este	5	Bariol	2
Paige	21	Kristine	14	Claire	10	Hark	5	Le	2
Elsie	20	Beth	14	Jean	10	Mai	5	Darol	2
Darlene	18	Sandy	14	Christina	9	Lars	5	Lao	2
Patricia	18	Dolores	14	Priscilla	9	Gerry	4	Cha	2
Paul	17	Patrick	14	Reni	8	Ellin	4	Vind	2
Kristina	17	William	13	Ulf	8	Kim	4	Carina	1
Alex	16	Francis	13	Harvey	8	Seeni	3	Sheng	1
Hazel	16	Beth	12	Erica	8	Fanti	3	Chara	1
Marcia	15	Stephanie	12	Jerome	8	Gard	3	Dayi	0
Eric	15	Katherine	11	Han	7	Andra	3		
Gretchen	15	Elsie	11	Louis	7	Melia	3		
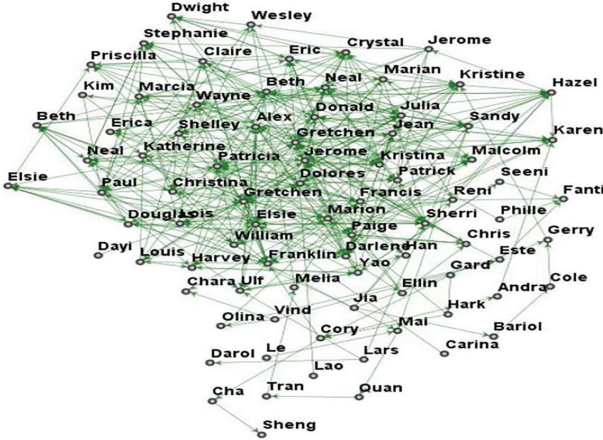
FIGURE 1: Visualized of the network model of the 83 people (nodes) and 400 messages between these people (links).

Definition 1. The TEF value p_{ij} is defined by

$$
p_{ij} = P\left(v_i, R_j\right)
$$
$$
= \begin{cases}
1, & R_{(v_i)} = R_j, \\
0, & R_{(v_i)} \in R, \ R_{(v_i)} \neq R_j, \\
\dfrac{f_{ij}}{\sum_{k=1}^n f_{kj}} \times \dfrac{f_{ij}}{\sum_{k=1}^m f_{ik}} & R_{(v_i)} \notin R,
\end{cases} \quad (4)
$$

where $0 \leq p_{ij} \leq 1$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$. Obviously, the greater p_{ij}, the greater the tendentiousness of node v_i with respect to role R_j.

2.4.2. Importance Evaluation Function (IEF). Based on point centrality and TEF, the importance of node v_i with respect to role R_j in the network is evaluated by IEF.

Definition 2. The IEF value S_{ij} is defined by

$$
S_{ij} = S\left(v_i, R_j\right) = c_i \times p_{ij} = \frac{f_{ij}}{\sum_{k=1}^n f_{kj}} \times \frac{f_{ij}}{\sum_{k=1}^m f_{ik}}. \quad (5)
$$

The normalized form is

$$
\begin{aligned}
x_{ij} &= X\left(v_i, R_j\right) \\
&= \frac{S\left(v_i, R_j\right) - \min_{1 \leq i \leq n}\left(S\left(v_i, R_j\right)\right)}{\max_{1 \leq i \leq n}\left(S\left(v_i, R_j\right)\right) - \min_{1 \leq i \leq n}\left(S\left(v_i, R_j\right)\right)},
\end{aligned} \quad (6)
$$

where $0 \leq x_{ij} \leq 1$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$. The greater x_{ij}, obviously, the greater the importance of node v_i with respect to role R_j in the network.
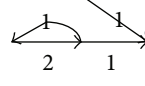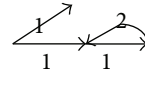
3. Application Analysis
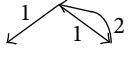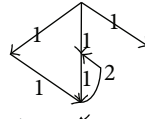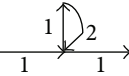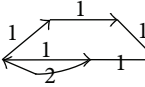
The Intergalactic Crime Modelers (ICM) is investigating a conspiracy to commit a criminal act. The case involves 83 members and 400 messages between these people, as shown in Figure 1. As priorly known in [28], Jean, Alex, Elsie, Paul, Ulf, Yao, Harvey are conspirators, Darlene, Tran, Jia, Ellin, Gard, Chris, Paige, Este are nonconspirators.

Now, we analyze the set of prior conspirator and the set of prior non-conspirator by using the theory and methods of network motifs.

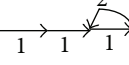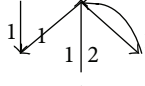Firstly, let R_1 be "conspirator" and let R_2 be "non-conspirator". Then the links are divided into two categories, of which daily topic is denoted by topic 1, and conspiracy topic is denoted by topic 2.

Based on adjacency matrix of the network, point centrality of nodes is calculated by using formulas (2) as shown in Table 1. The point centrality reflects the influence of a node in the network, which means the larger point centrality

TABLE 2: The sample of frequent subgraph and network motifs.

Scale	ID	Structure	N_{real}	Z	Scale	ID	Structure	N_{real}	Z
2	S1		230	1.1609	4	M4		30	5.4641
2	S2		10	3.6995	4	M5		30	5.7152
3	M1		15	5.8795	5	M6		136	5.7783
4	M2		38	9.7047	5	M7		126	9.4614
4	M3		36	8.1592	5	M8		102	8.3725

Note of figure: S1-2 are not network motifs ($Z < 5$); M1-8 are network motifs ($Z > 5$), which are structure module with special features (the minimum of Z-score is 5).

TABLE 3: Priority list of criminal tend.

Name	p_{i1}	p_{i2}	Name	p_{i1}	p_{i2}	Name	p_{i1}	p_{i2}
Jean	1	0	Donald	0.0069	0.0064	Gerry	0.0001	0.0038
Alex	1	0	Shelley	0.0065	0.0051	Franklin	0.0001	0.0060
Elsie	1	0	Dwight	0.0061	0.0100	Phille	0	0
Paul	1	0	Louis	0.0057	0.0005	Cole	0	0
Ulf	1	0	Beth	0.0054	0.0015	Bariol	0	0
Yao	1	0	Karen	0.0054	0.0231	Cory	0	0
Harvey	1	0	Kim	0.0051	0.0007	Quan	0	0
Marcia	0.0316	0.0059	Kristina	0.0042	0.0000	Mai	0	0
Stephanie	0.0229	0.0178	Han	0.0033	0.0003	Le	0	0
Jerome	0.0210	0.0068	Lars	0.0031	0.0000	Darol	0	0
Crystal	0.0196	0.0197	Erica	0.0031	0.0014	Sheng	0	0
Priscilla	0.0173	0.0016	Wayne	0.0030	0.0249	Cha	0	0
Neal	0.0153	0.0100	Carina	0.0027	0.0001	Olina	0	0
William	0.0141	0.0014	Elsie	0.0027	0.0159	Chara	0	0
Sherri	0.0140	0.0072	Gretchen	0.0025	0.0075	Vind	0	0
Douglas	0.0129	0.0054	Marion	0.0024	0.0052	Dayi	0	0
Patrick	0.0111	0.0005	Katherine	0.0023	0.0049	Seeni	0	0.0005
Gretchen	0.0098	0.0180	Kristine	0.0023	0.0303	Lao	0	0.0018
Dolores	0.0096	0.0024	Francis	0.0016	0.0001	Fanti	0	0.0023
Julia	0.0096	0.0033	Eric	0.0014	0.0007	Darlene	0	1
Patricia	0.0089	0.0106	Sandy	0.0007	0.0022	Tran	0	1
Jerome	0.0088	0.0001	Melia	0.0004	0	Jia	0	1
Christina	0.0086	0.0048	Wesley	0.0004	0	Ellin	0	1
Neal	0.0085	0.0006	Claire	0.0003	0.0012	Gard	0	1
Reni	0.0084	0.0010	Marian	0.0003	0.0026	Chris	0	1
Lois	0.0080	0.0055	Hark	0.0002	0.0002	Paige	0	1
Hazel	0.0078	0.0084	Malcolm	0.0001	0.0006	Este	0	1
Beth	0.0074	0.0523	Andra	0.0001	0.0014			

Table 4: Priority list for key monitoring.

ID	Name	x_{i1}	ID	Name	x_{i1}	ID	Name	x_{i1}
8	Jean	1.0000	2	Donald	0.0036	78	Gerry	0.0000
44	Alex	0.8500	17	Shelley	0.0035	73	Franklin	0.0000
22	Elsie	0.8000	29	Dwight	0.0034	82	Phille	0
68	Paul	0.5000	83	Louis	0.0034	81	Cole	0
19	Ulf	0.5000	6	Beth	0.0030	80	Bariol	0
55	Yao	0.4000	30	Karen	0.0021	79	Cory	0
50	Harvey	0.4000	47	Kim	0.0020	77	Quan	0
35	Marcia	0.0241	36	Kristina	0.0020	76	Mai	0
28	Stephanie	0.0237	5	Han	0.0019	75	Le	0
4	Jerome	0.0182	20	Lars	0.0016	72	Darol	0
33	Crystal	0.0143	38	Erica	0.0015	69	Sheng	0
31	Priscilla	0.0138	14	Wayne	0.0013	66	Cha	0
21	Neal	0.0137	43	Carina	0.0012	65	Olina	0
32	William	0.0107	40	Elsie	0.0012	64	Chara	0
16	Sherri	0.0106	70	Gretchen	0.0012	63	Vind	0
18	Douglas	0.0094	12	Marion	0.0010	62	Dayi	0
51	Patrick	0.0092	34	Katherine	0.0010	60	Seeni	0
45	Gretchen	0.0080	23	Kristine	0.0010	59	Lao	0
37	Dolores	0.0078	61	Francis	0.0008	58	Fanti	0
7	Julia	0.0078	13	Eric	0.0005	57	Darlene	0
11	Patricia	0.0067	27	Sandy	0.0002	56	Tran	0
41	Jerome	0.0064	74	Melia	0.0001	54	Jia	0
9	Christina	0.0062	26	Wesley	0.0001	53	Ellin	0
46	Neal	0.0056	24	Claire	0.0001	52	Gard	0
42	Reni	0.0048	25	Marian	0.0001	49	Chris	0
39	Lois	0.0045	10	Hark	0.0001	3	Paige	0
48	Hazel	0.0039	67	Malcolm	0.0001	1	Este	0
15	Beth	0.0038	71	Andra	0.0000			

a criminal node have, the bigger negative impact on the network will occur. Therefore, ICM should focus on high-ranking members in Table 1.

According to the method in Section 2.2, frequent subgraphs and network motifs of the network are obtained, as shown in Table 2 (example).

Depending on the network motifs, frequency matrix is obtained by applying Algorithm B. Based on the evaluation functions TEF and IEF, the priority list of criminal tend and the priority list for key monitoring are obtained, as shown respectively, in Tables 3 and 4.

By comparing and analyzing, the "network motifs" offer a much more comprehensive way to analyze social networks; the high-ranking members having both higher point centrality and more criminal tend in Tables 3 and 4 are more suspect than others, so ICM should monitor them. Our research confirmed that the method is suitable in social network and the results are reliable.

4. Conclusions

In this paper, we developed a set of network analysis methods based on the theory and methods of social network analysis,

frequent subgraph mining, network motifs, and so forth, which is different from the traditional social network analysis. In application analysis, a series of priority lists are obtained based on the evaluation functions. The priority lists reflect network information effectively, which is of great reference value for ICM.

Based on the study on node connection relationship of social networks, a follow-up study will involve more attention to structural relationship with more practical value.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

References

[1] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[2] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, Article ID 208701, 4 pages, 2002.

[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[4] T. N. Friemel, "Dynamics of social networks," *Procedia-Social and Behavioral Sciences*, vol. 22, pp. 2–3, 2011.

[5] J. Xu and H. Chen, "Criminal network analysis and visualization," *Communications of the Association for Computing Machinery*, vol. 48, no. 6, pp. 100–107, 2005.

[6] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guénoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," *Genome Biology*, vol. 5, no. 1, pp. 1474–1760, 2003.

[7] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[8] K. Musiał and P. Kazienko, "Social networks on the Internet," *World Wide Web*, vol. 16, no. 1, pp. 31–72, 2013.

[9] C. E. Hutchins and M. Benham-Hutchins, "Hiding in plain sight: criminal network analysis," *Computational and Mathematical Organization Theory*, vol. 16, no. 1, pp. 89–111, 2010.

[10] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[11] T. Velden, A. Haque, and C. Lagoze, "A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation," *Scientometrics*, vol. 85, no. 1, pp. 219–242, 2010.

[12] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[14] W. Kim, M. Li, J. Wang, and Y. Pan, "Biological network motif detection and evaluation," *BMC Systems Biology*, vol. 5, Supplement 3, article S5, 2011.

[15] L. Farina, A. Germani, G. Mavelli, and P. Palumbo, "Identification of regulatory network motifs from gene expression data," *Journal of Mathematical Modelling and Algorithms*, vol. 9, no. 3, pp. 233–245, 2010.

[16] T. Ohnishi, H. Takayasu, and M. Takayasu, "Network motifs in an inter-firm network," *Journal of Economic Interaction and Coordination*, vol. 5, no. 2, pp. 171–180, 2010.

[17] Q. H. Zhu and L. Li, "Social network analysis method and application study in information science," *Information Study*, vol. 31, no. 2, pp. 179–183, 2008.

[18] P. Chen and H. Y. Yuan, "Social network analysis of crime organization structures," *Journal of Tsinghua University*, vol. 51, no. 8, pp. 1097–1101, 2011.

[19] D. Gómez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo, and M. Saboyá, "The cohesiveness of subgroups in social networks: a view from game theory," *Annals of Operations Research*, vol. 158, no. 1, pp. 33–46, 2008.

[20] X. G. Wang, "Discovering critical nodes in social networks based on cooperative games," *Computer Science*, vol. 40, no. 4, pp. 155–159, 2013.

[21] K. Juszczyszyn, P. Kazienko, and K. Musial, "Local topology of social network based on motif analysis," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 5178 of *Lecture Notes in Computer Science*, pp. 97–105, 2008.

[22] K. Juszczyszyn and K. Musial, "Motif analysis and the periodic structural changes in an organizational email-based social network," *International Journal of Virtual Communities and Social Networking*, vol. 1, no. 2, 14 pages, 2009.

[23] X.-T. Li, J.-Z. Li, and H. Gao, "Efficient frequent subgraph mining algorithm," *Journal of Software*, vol. 18, no. 10, pp. 2469–2480, 2007.

[24] H. M. Lu, B. Q. Feng, and Q. B. Song, "Survey of frequent subgraph mining research," *Microelectronics & Computer*, vol. 26, no. 3, pp. 156–161, 2009.

[25] G.-M. Qin, L. Gao, and J.-L. Hu, "A review on algorithms for network motif discovery in biological networks," *Acta Electronica Sinica*, vol. 37, no. 10, pp. 2258–2265, 2009.

[26] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Neuman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," Cond-mat/0312028, 2003.

[27] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.

[28] ICM, "Modeling for crime busting," The Interdisciplinary Contest in Modeling, 2012.