Uncovering Overlap Community Structure in Complex Networks using Particle Competition

Fabricio Breve, Liang Zhao, and Marcos Quiles

Institute of Mathematics and Computer Science, University of São Paulo, São Carlos SP 13560-970, Brazil,
{fabricio,zhao,quiles}@icmc.usp.br

Abstract. Identification and classification of overlap nodes in communities is an important topic in data mining. In this paper, a new clustering method to uncover overlap nodes in complex networks is proposed. It is based on particles walking and competing with each other, using random-deterministic movement. The new community detection algorithm can output not only hard labels, but also continuous-valued output (soft labels), which corresponds to the levels of membership from the nodes to each of the communities. Computer simulations were performed with synthetic and real data and good results were achieved.

Keywords: overlap community structure, particle competition, complex networks community detection

1 Introduction

In the last years, the advances and the convergence of computing and communication has rapidly increased our capacities of generating and collecting data. However, most of this data is in its raw form, and it is not useful until it is discovered and articulated. Data Mining is the process of extracting the implicit potentially useful information from the data. It is a multidisciplinary field, drawing works from areas including statistics, machine learning, artificial intelligence, data management and databases, pattern recognition, information retrieval, neural networks, data visualization, and others [1–5].

Community Detection is one of the data mining problems that arose with the advances in computing and the increasingly interest in complex networks, which studies large scale networks with non-trivial topological structures, such as social networks, computer networks, telecommunication networks, transportation networks, and biological networks [6–8]. Many of these networks are found to be divided naturally into communities or modules, therefore discovering of these communities structure became an important topic of study [9–13]. Recently, a particle competition approach was successfully applied to detect communities modeled in networks [14].

The notion of communities in networks is straightforward, they are defined as a subgraph whose nodes are densely connected within itself but sparsely

connected with the rest of the network. However, in practice there are common cases where some nodes in a network can belong to more than one community. For example: in a social network of friendship, individuals often belong to several communities: their families, their colleagues, their classmates, etc. These nodes are often called overlap nodes, and most known community detection algorithms cannot detect them. Therefore, uncovering the overlapping community structure of complex networks becomes an important topic in data mining [15–17].

In this paper we present a new clustering technique, based on particle walking and competition. We have extended the model proposed in [14] to output not only hard labels, but also a fuzzy output (soft labels) for each node in the network. The continuous-valued output can be seen as the levels of membership from each node to each community. Therefore, the new model is able to uncover the overlap community structure in complex networks.

The rest of this paper is organized as follows: Section 2 describes the model in details. Section 3 shows some experimental results from computer simulations, and in Section 4 we draw some conclusions.

2 Model Description

The model we propose in this paper is an extension of the particle competition approach proposed by [14], which is used to detect communities in networks. Some particles walk in a network, competing with each other for the possession of nodes, while rejecting intruder particles. After a number of iterations, each particle will be confined within a community of the network, so the communities can be divided by examining the nodes ownership. The new model is not only suitable to detect community structure, but it can also uncover overlap community structure. In order to achieve that, we have changed the nodes and particles dynamics, and introduced a few new variables, among other details that will follow.

Let the network structure be represented as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, with $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$, where each node v_i is an element from the network. An adjacency matrix $\mathbf{W}$ defines which network nodes are interconnected:

$$W_{ij} = 1, \quad \text{if there is an edge between nodes } i \text{ and } j, \tag{1}$$

$$W_{ij} = 0, \quad \text{otherwise}, \tag{2}$$

and $W_{ii} = 0$.

Then, we create a set of particles $\mathbf{P} = (\rho_1, \rho_2, \ldots, \rho_c)$, in which each particle corresponds to a different community. Each particle ρ_j has a variable $\rho_j^{\omega}(t) \in [\omega_{\min} \quad \omega_{\max}]$ is the particle potential characterizing how much the particle can affect a node at time t, in this paper we set the constants $\omega_{\min} = 0$ and $\omega_{\max} = 1$.

Each node v_i has two variables $v_i^{\omega}(t)$, and $v_i^{\lambda}(t)$. The first variable is a vector $v_i^{\omega}(t) = \{v_i^{\omega_1}(t), v_i^{\omega_2}(t), \ldots, v_i^{\omega_c}(t)\}$ of the same size of $\mathbf{P}$, where each element $v_i^{\omega_j}(t) \in [\omega_{\min} \quad \omega_{\max}]$ corresponds to the instantaneous level of ownership by particle ρ_j over node v_i. The sum of the levels of ownership of each node is

always a constant, because a particle increases its own ownership level and, at the same time, decreases the other particles ownership levels. Thus, the following equations always holds:

$$\sum_{j=1}^{c} v_i^{\omega_j} = \omega_{\max} + \omega_{\min}(c-1). \tag{3}$$

The second variable is also a vector $v_i^{\lambda}(t) = \{v_i^{\lambda_1}(t), v_i^{\lambda_2}(t), \ldots, v_i^{\lambda_c}(t)\}$ of the same size of $\mathbf{P}$ and it also represents ownership levels, but unlike $v_i^{\omega}(t)$ which denotes the instantaneous ownership levels, $v_i^{\lambda_j}(t) \in [0 \quad \infty]$ rather denotes long term ownership levels, accumulated through the whole process. The particle with higher ownership level in a given non-overlap node after the last iteration of the algorithm is usually the particle which have visited that node more times, but that does not always apply to overlap nodes, in which sometimes the dominant particle could easily change in the last iterations, and thus it would not correspond to the particle which have dominated that node for more instants of time. Therefore, the new variable $v_i^{\lambda}(t)$ was introduced in order to define the ownership of nodes considering the whole process. Using a simple analogy, we can say that now the champion is not the one who have won the last games, but rather the one who have won more games in the whole championship. Notice that the long term ownership levels only increases and their sum is not constant, they are normalized only at the end of the iterations.

We begin the algorithm by setting the initial level of instantaneous ownership vector v_i^{ω} by each particle ρ_j as follows:

$$v_i^{\omega_j}(0) = \omega_{\min} + (\frac{\omega_{\max} - \omega_{\min}}{c}), \tag{4}$$

which means that all nodes starts with all particles instantaneous ownership levels equally set. Meanwhile, the long term ownership levels $v_i^{\lambda}(t)$ are all set to zero:

$$v_i^{\lambda_j}(0) = 0. \tag{5}$$

The initial position of each particle $\rho_j^v(0)$ is set randomly, to any node in $\mathbf{V}$, and the initial potential of each particle is set to its minimum value, as follows:

$$\rho_j^{\omega}(0) = \omega_{\min}. \tag{6}$$

Each particle will choose a neighbor to visit based in a random-deterministic rule. At each iteration, each particle will chose between *random walk* or *deterministic walk*, where *random walk* means the particle will try to move to any neighbor randomly chosen, i.e., the particle ρ_j will try to move to any node v_i chosen with the probabilities defined by:

$$p(v_i|\rho_j) = \frac{W_{ki}}{\sum_{q=1}^{n} W_{qi}}, \tag{7}$$

where k is the index of the node node being visited by particle ρ_j, so $W_{ki} = 1$ if there is an edge between the current node and v_i, and $W_{ki} = 0$ otherwise.

The *deterministic walk* means that the particle will try to move to a neighbor with probabilities according to the nodes instantaneous ownership levels, i.e., the particle ρ_j will try to move to any neighbor v_i chosen with probabilities defined by:

$$p(v_i|\rho_j) = \frac{W_{ki} v_i^{\omega_j}}{\sum_{q=1}^{n} W_{qi} v_i^{\omega_j}}, \tag{8}$$

again, k is the index of the node stored being visited by particle ρ_j.

At each iteration, each particle has probability p_{det} of taking deterministic movement and probability $1 - p_{\text{det}}$ of taking random movement, with $0 \leq p_{\text{det}} \leq 1$. Once the random movement or deterministic movement is chosen, a target neighbor $\rho_j^{\tau}(t)$ will be randomly chosen with probabilities defined by Eq. 7 or Eq. 8 respectively.

Regarding the node dynamics, at time t, each instantaneous ownership level $v_i^{\omega_k}(t)$ of each node v_i, which was chosen by a particle ρ_j as its target $\rho_j^{\tau}(t)$, is defined as follows:

$$v_i^{\omega_k}(t+1) = \begin{cases} \max\{\omega_{\min}, v_i^{\omega_k}(t) - \frac{\Delta_v \rho_j^{\omega}(t)}{c-1}\} & \text{if } k \neq j \\ v_i^{\omega_k}(t) + \sum_{q \neq k} v_i^{\omega_q}(t) - v_i^{\omega_q}(t+1) & \text{if } k = j \end{cases}, \tag{9}$$

where $0 < \Delta_v \leq 1$ is a parameter to control the changing rate of the instantaneous ownership levels. If Δ_v takes a low value, the node ownership levels change slowly, while if it takes a high value, the node ownership levels change quickly. Each particle ρ_j will increase their corresponding instantaneous ownership level $v_i^{\omega_j}$ of the node v_i they are targeting, while decreasing the instantaneous ownership levels (of this same node) that corresponds to the other particles, always respecting the conservation law defined by Eq. 3.

Regarding the particle dynamics, at time t, each particle potential $\rho_j^{\omega}(t)$ is set as:

$$\rho_j^{\omega}(t+1) = \rho_j^{\omega}(t) + \Delta_{\rho}(v_i^{\omega_j}(t+1) - \rho_j^{\omega}(t)) \tag{10}$$

where $v_i(t+1)$ is the node rho_j is targeting, $0 < \Delta_{\rho} \leq 1$ is a parameter to control the particle potential changing rate. Therefore, every particle ρ_j have their potential ρ_j^{ω} set to approximate the value of instantaneous ownership level $v_i^{\omega_j}$ from the node it is currently targeting. In this sense, a particle gets stronger when it is visiting a node with higher ownership level of its own, but it will be weakened if it tries to invade a node dominated by other particle.

The long term ownership levels are adjusted only when the particle selects the random movement. This rule is important because although the deterministic movement is useful to prevent particles from abandoning their neighborhood, which would let it susceptible to other particles attack, it is also a mechanism that makes a node gets more visits from the particle that currently dominates it. We consider only when the *random movement* was chosen because, in this case, particles will choose a target node based only in their current neighborhood, and not in their instantaneous ownership levels that are important for community detection, but are too volatile in overlap nodes. Therefore, for each particle

selected in *random movement* by a particle ρ_j, the long term ownership levels $v_i^{\lambda_j}$ are update as follows:

$$v_i^{\lambda_j}(t+1) = v_i^{\lambda_j}(t) + \rho_j^{\omega}(t), \tag{11}$$

where v_i is the node ρ_j is targeting. The increase will always be proportional to the current particle potential, which is a desirable feature because the particle will probably have a higher potential when it is arriving from its own neighborhood, while it will have a lower potential when it is arriving from a node from other particles neighborhoods.

It should be noted that a particle really visits a target node only if its ownership level in that node is higher than the others; otherwise, a shock happens and the particle stays at the current node until next iteration.

At the end of the iterations, the degrees of membership $f_i^j \in [0 \quad 1]$ for each node v_i are calculated using the long term ownership levels, as follows:

$$f_i^j = \frac{v_i^{\lambda_j}(\infty)}{\sum_{q=1}^{c} v_i^{\lambda_q}(\infty)} \tag{12}$$

where f_i^j represents the final membership level of the node v_i to community j.

In summary, our algorithm works as follows:

1. Build the adjacency matrix $\mathbf{W}$ by using Eq. 1,
2. Set nodes ownership levels by using Eq. 4 and Eq. 5,
3. Set particles initial positions randomly and their potentials by using Eq. 6,
4. Repeat steps 5 to 8 until convergence or for a pre-defined number of steps,
5. Select the target node for each particle by using Eq. 8 or Eq. 7 for deterministic movement or random movement respectively,
6. Update nodes ownership levels by using Eq. 9,
7. If the random movement was chosen, update the long term ownership levels by using Eq. 11,
8. Update particles potentials by using Eq. 10,
9. Calculate the membership levels (fuzzy classification) by using Eq. 12.

3 Computer Simulations

In order to test the overlap detection capabilities of our algorithm, we generate a set of networks with community structure using the method proposed by [13]. Here, all the generated networks have $n = 128$ nodes, split into four communities containing 32 nodes each. Pairs of nodes which belongs to the same community are linked with probability p_{in}, whereas pairs belonging to different communities are joined with probability p_{out}. The total average node degree k is constant and set to 16. The value of p_{out} is taken so the average number of links a node has to nodes of any other community, z_{out}, can be controlled. Meanwhile, the value of p_{in} is chosen to keep the average node degree k constant. Therefore,

z_{out}/k defines the mixture of the communities, and as z_{out}/k increases from zero, the communities become more diffuse and harder to identify. In each of these generated networks we have added a 129th node and created 16 links between the new node and nodes from the communities, so we could easily determine an expected "fuzzy" classification for this new node based on the count of its links with each community.

The networks were generated with $z_{out}/k = 0.125$, 0.250, and 0.375 and the results are shown in Tables 1, 2, and 3 respectively. The first column of these tables shows the number of links the 129th node has to communities A, B, C, and D, respectively. Notice that in each configuration the 129th node has different overlap levels, varying from the case where it fully belongs to a single community up to the case where it belongs to the four communities almost equally. From 2nd to 5th column we have the fuzzy degree of membership of the 129th node relative to communities A, B, C, and D respectively, obtained by our algorithm. The presented values are the average of 100 realizations with different networks. For these simulations, the parameters were set as follows: $p_{\text{det}} = 0.5$, $\Delta_v = 0.4$ and $\Delta_\rho = 0.9$.

Table 1. Fuzzy classification of a node connected to network with 4 communities generated with $z_{out}/k = 0.125$

Connections	**Fuzzy Classification**			
A-B-C-D	**A**	**B**	**C**	**D**
16-0-0-0	0.9928	0.0017	0.0010	0.0046
15-1-0-0	0.9210	0.0646	0.0079	0.0065
14-2-0-0	0.8520	0.1150	0.0081	0.0248
13-3-0-0	0.8031	0.1778	0.0107	0.0084
12-4-0-0	0.7498	0.2456	0.0032	0.0014
11-5-0-0	0.6875	0.3101	0.0016	0.0008
10-6-0-0	0.6211	0.3577	0.0111	0.0101
9-7-0-0	0.5584	0.4302	0.0011	0.0103
8-8-0-0	0.4949	0.4944	0.0090	0.0017
8-4-4-0	0.5025	0.2493	0.2461	0.0021
7-4-4-1	0.4397	0.2439	0.2491	0.0672
6-4-4-2	0.3694	0.2501	0.2549	0.1256
5-4-4-3	0.3144	0.2491	0.2537	0.1828
4-4-4-4	0.2512	0.2506	0.2504	0.2478

The results shown that the method was able to accurately identify the fuzzy communities of the overlap nodes. The accuracy gets lower as z_{out}/k increases, this was expected, since a higher z_{out}/k means that the communities are more diffuse and the observed node can be connected to nodes that are overlap nodes themselves.

Table 2. Fuzzy classification of a node connected to network with 4 communities generated with $z_{out}/k = 0.250$

Connections	**Fuzzy Classification**			
A-B-C-D	**A**	**B**	**C**	**D**
16-0-0-0	0.9912	0.0027	0.0024	0.0037
15-1-0-0	0.9318	0.0634	0.0026	0.0023
14-2-0-0	0.8715	0.1219	0.0023	0.0044
13-3-0-0	0.8107	0.1827	0.0036	0.0030
12-4-0-0	0.7497	0.2437	0.0044	0.0022
11-5-0-0	0.6901	0.3036	0.0034	0.0029
10-6-0-0	0.6298	0.3654	0.0020	0.0028
9-7-0-0	0.5584	0.4360	0.0026	0.0030
8-8-0-0	0.4952	0.4985	0.0027	0.0036
8-4-4-0	0.5060	0.2485	0.2427	0.0028
7-4-4-1	0.4442	0.2477	0.2429	0.0652
6-4-4-2	0.3762	0.2465	0.2514	0.1259
5-4-4-3	0.3178	0.2500	0.2473	0.1849
4-4-4-4	0.2470	0.2518	0.2489	0.2523

Based on this data, we have created an overlap measure in order to easily illustrate the application of the algorithm in more complex networks with lots of overlap nodes. Therefore, the overlap index o_i for a node v_i is defined as follow:

$$o_i = \frac{f_i^{j**}}{f_i^{j*}} \tag{13}$$

where $j* = \arg\max_j f_i^j$ and $j** = \arg\max_{j,j\neq j*} f_i^j$, and $o_i \in [0 \quad 1]$, where $o_i = 0$ means completely confidence that the node belongs to a single community, while $o_i = 1$ means the node is completely undefined among two or more communities.

Then, we have applied the algorithm to a problem with 1000 elements, split into four communities with 250 elements each. There are four gaussian kernels in a two dimensional plane and the elements are distributed around them. To build the network, each element is transformed into a network node. Two elements i and j are connected if their Euclidean distance $d(i,j) < 1$. The algorithm parameters were set as follows: $p_{\text{det}} = 0.6$, $\Delta_v = 0.4$ and $\Delta_\rho = 0.9$. In Figure 1 the overlap index of each node is indicated by their colors. It is easy to realize that the closer to the communities frontier the nodes are, the higher are their respective overlap indexes.

Finally, the algorithm was applied to the famous Zachary's Karate Club Network [18] and the results are shown in Figure 2. The algorithm parameters were set as follows: $p_{\text{det}} = 0.6$, $\Delta_v = 0.4$ and $\Delta_\rho = 0.9$. Again, the overlap index of each node is indicated by their colors. In Table 4 the fuzzy classification of all the nodes on this network are shown.

Table 3. Fuzzy classification of a node connected to network with 4 communities generated with $z_{out}/k = 0.375$

Connections	**Fuzzy Classification**			
A-B-C-D	**A**	**B**	**C**	**D**
16-0-0-0	0.9709	0.0092	0.0108	0.0091
15-1-0-0	0.9160	0.0647	0.0093	0.0101
14-2-0-0	0.8571	0.1228	0.0104	0.0097
13-3-0-0	0.8008	0.1802	0.0100	0.0090
12-4-0-0	0.7422	0.2385	0.0095	0.0098
11-5-0-0	0.6825	0.2958	0.0123	0.0093
10-6-0-0	0.6200	0.3566	0.0111	0.0123
9-7-0-0	0.5582	0.4181	0.0128	0.0109
8-8-0-0	0.4891	0.4846	0.0130	0.0133
8-4-4-0	0.5045	0.2437	0.2406	0.0113
7-4-4-1	0.4397	0.2461	0.2436	0.0705
6-4-4-2	0.3797	0.2471	0.2445	0.1287
5-4-4-3	0.3175	0.2439	0.2473	0.1913
4-4-4-4	0.2462	0.2494	0.2549	0.2495

4 Conclusions

This paper presents a new clustering technique using combined random-deterministic walking and competition among particles, where each particle corresponds to a class of the problem. The algorithm outputs not only hard labels, but also soft labels (fuzzy values) for each node in the network, which corresponds to the levels of membership from that node to each community. Computer simulations were performed in both synthetic and real data, and the results shows that our model is a promising mechanism to uncover overlap community structure in complex networks.

Acknowledgements

This work is supported by the State of São Paulo Research Foundation (FAPESP) and the Brazilian National Council of Technological and Scientific Development (CNPq).

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. 2 edn. Morgan Kaufmann (2006)
2. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2 edn. Morgan Kauffman (2005)
3. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press (2001)
4. Weiss, S.M., Indurkhya, N.: Predictive Data Mining: A Practical Guide. Morgan Kaufmann (1998)

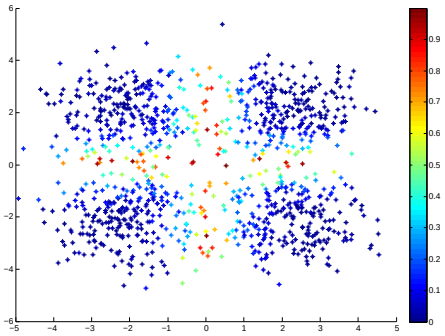

Fig. 1. Problem with 1000 elements split into four communities, colors represent the overlap index from each node, detected by the proposed method.

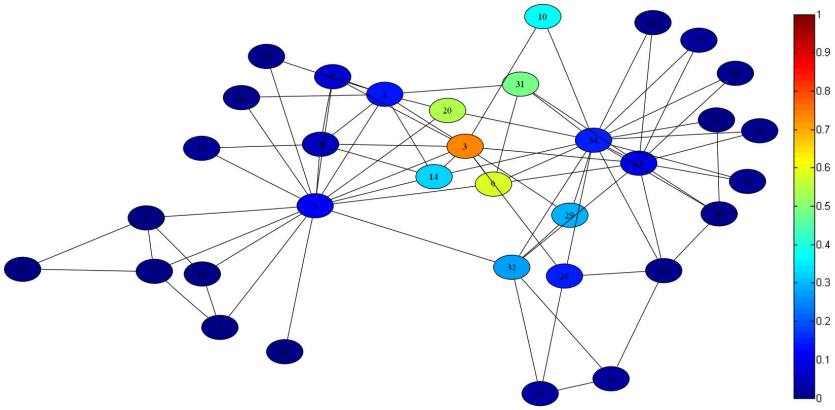

Fig. 2. The karate club network, colors represent the overlap index from each node, detected by the proposed method.

5. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley (2005)
6. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
7. Dorogovtsev, S., Mendes, F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press (2003)
8. Bornholdt, S., Schuster, H.: Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH (2006)
9. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69** (2004) 026113
10. Newman, M.: Modularity and community structure in networks. In: Proceedings of the National Academy of Science of the United States of America. Volume 103. (2006) 8577–8582
11. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Physical Review E **72** (2006) 027104

Table 4. Fuzzy classification of the Zachary's Karate Club Network achieved by the proposed method.

Node	Community A	Community B	Node	Community A	Community B
1	0,8934	0,1066	**18**	0,9861	0,0139
2	0,8744	0,1256	**19**	0,0126	0,9874
3	0,5727	0,4273	**20**	0,6452	0,3548
4	0,9470	0,0530	**21**	0,0140	0,9860
5	0,9960	0,0040	**22**	0,9877	0,0123
6	0,9979	0,0021	**23**	0,0168	0,9832
7	0,9979	0,0021	**24**	0,0106	0,9894
8	0,9282	0,0718	**25**	0,0275	0,9725
9	0,3703	0,6297	**26**	0,0262	0,9738
10	0,2750	0,7250	**27**	0,0050	0,9950
11	0,9968	0,0032	**28**	0,1319	0,8681
12	0,9957	0,0043	**29**	0,2298	0,7702
13	0,9791	0,0209	**30**	0,0123	0,9877
14	0,7510	0,2490	**31**	0,3293	0,6707
15	0,0238	0,9762	**32**	0,2188	0,7812
16	0,0100	0,9900	**33**	0,0878	0,9122
17	1,0000	0,0000	**34**	0,1342	0,8658

12. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters **93**(21) (2004) 218701
13. Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment **9** (2005) P09008
14. Quiles, M.G., Zhao, L., Alonso, R.L., Romero, R.A.F.: Particle competition for complex network community detection. Chaos **18**(3) (2008) 033107
15. Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A Statistical Mechanics and its Applications **374** (January 2007) 483–490
16. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043) (2005) 814–818
17. Zhang, S., Wang, R.S., Zhang, X.S.: Uncovering fuzzy community structure in complex networks. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) **76**(4) (2007) 046103
18. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research **33** (1977) 452–473