Community Detection for Multiplex Social Networks Based on Relational Bayesian Networks

Jiuchuan Jiang and Manfred Jaeger

Department of Computer Science, Aalborg University, Denmark
{jiuchuan,jaeger}@cs.aau.dk

Abstract. Many techniques have been proposed for community detection in social networks. Most of these techniques are only designed for networks defined by a single relation. However, many real networks are multiplex networks that contain multiple types of relations and different attributes on the nodes. In this paper we propose to use relational Bayesian networks for the specification of probabilistic network models, and develop inference techniques that solve the community detection problem based on these models. The use of relational Bayesian networks as a flexible high-level modeling framework enables us to express different models capturing different aspects of community detection in multiplex networks in a coherent manner, and to use a single inference mechanism for all models.

Keywords: Community detection; Multiplex networks; Relational Bayesian networks; Statistical relational learning.

1 Introduction

Social networks like Facebook, Twitter, Flickr, or Youtube have prospered in recent years. People in an online society here can communicate and interact with each other. Community structure is one of the most important characteristics for social networks [1]. Within a community, the connections between nodes are very dense but they are sparse in between communities. In a social network, a community can be a friend group which has close relations, a group of people with similar interests, a group of people in a same workplace, and so on. Community detection, therefore, has received significant attention in the research of social networks [2][3].

Most existing methods have been developed to analyze single relation networks, where there is only one type of relation between nodes. However, in the real world, social networks may often appear as multiplex networks, in which there exist different types of nodes, which are connected by different types of links [4][5]. For example, in Facebook, the relations between two users could be friends, common interests, alumni, and so on; and the users in Facebook could be humans, companies, or organizations, which makes the characteristics of users different from each other.

T. Andreasen et al. (Eds.): ISMIS 2014, LNAI 8502, pp. 30–39, 2014.

A few studies have investigated community detection for multiplex networks, but mostly these are characterised by strong simplifications that reduce community detection in multiplex networks to community detection in single relation networks. On the other hand, in Machine Learning the field of *statistical relational learning (SRL)* is specifically concerned with statistical models for multi-relational data. Since probabilistic models are a powerful tool for clustering in general, and community detection in networks in particular, it is natural to apply SRL techniques to the community detection problem in multiplex networks. In this paper we give an initial report on the application of the SRL modeling framework of *Relational Bayesian Networks (RBNs)* [6] to this task. The RBN framework provides an expressive and flexible representation language in which a variety of probabilistic community detection models can be specified. The different models are supported by a common set of generic inference algorithms. Such a general representation and inference framework provides a good basis to explore different aspects of communities in multiplex networks using different models, without the need to re-design inference methods for each model.

The rest of this paper is organized as follows. In Section 2, we compare our work with the related work on the subject; in Section 3, we present the problem description; in Section 4, we model the community detection in multiplex networks; in Section 5, we provide the experimental results; finally, we conclude our paper in Section 6.

2 Related Work

Community Detection in Single Relation Networks. Girvan and Newman [2] published the seminal paper on discovering the community structure in networks. After that, a lot of community detection methods for single networks has been developed. Typical methods include graph partitioning [7], hierarchical clustering [3], partitional clustering [8] and spectral methods [9]. A good review of community detection for single networks can be found in [10].

Community Detection in Multiplex Networks. Some researchers divide a multiplex network into single network layers and do community detection on these single networks, such as [4][5]. Yang et al. [11] extended traditional random walk algorithm to detect communities in signed networks which includes positive and negative relations. Breiger et al. [12] describe a hierarchical clustering algorithm that is based on an iterative transformation of incidence matrices into a block form, and which can be simultaneously applied to matrices representing multiple relations.

SRL Methods. An early paper that considered clustering in graphs as a possible application of SRL techniques is [13]. However, their model representation framework only allowed for the modeling of random attributes of nodes in a network, and not the modeling of random link structures, which are essential for

natural probabilistic community detection models. *Markov Logic Networks* are a currently popular SRL framework, which in [14] also was applied to clustering in multi-relational data, though no application to a community detection problem was presented. Xu et al. [15] proposed to use the *Infinite Hidden Relational Model (IHRM)* for social network analysis. The resulting probabilistic model is quite similar to the basic RBN-based community detection model we will describe below. A main difference between [15] and our work lies in the fact that Xu et al. work with a fixed graphical model, whereas we are using a higher-level representation language that allows us to experiment with different clustering models without the need to re-design the update equations for inference in the underlying graphical model. On the other hand, a key concern of the IHRM is to support a nonparametric Bayesian approach for automatically finding the 'right' number of clusters or communities, whereas we are currently taking the simpler approach of requiring the number of communities to be a user-defined input.

3 Problem Description

In this paper, we will use probabilistic models to analyze community structure problems. Graphs will be used to represent social networks. The vertices and arcs in graphs correspond to nodes and relations in social networks, respectively. Considering only single relation networks first, we can give the problem description as follows.

Given a directed graph $G =< V, A >$, with vertices V and arcs A, we want to find the partition Γ with maximal probability

$$P_\theta(\Gamma|A) = \frac{P_\theta(A, \Gamma)}{P_\theta(A)} \tag{1}$$

Here P_θ is the underlying probabilistic model, which defines for the given set V a joint distribution over partitions and arcs. The model can depend on unknown parameters θ. Since $P_\theta(A)$ does not depend on Γ, the community detection problem therefore amounts to computing

$$\arg\max_{\Gamma} \max_{\theta} P_\theta(A, \Gamma) = \arg\max_{\Gamma} \max_{\theta} P_\theta(A|\Gamma) P_\theta(\Gamma) \tag{2}$$

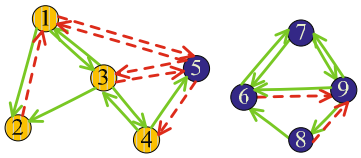

Fig. 1. A small example of multiplex network

We now turn to the generalization of this probabilistic community detection approach to multiplex networks. Multiplex networks are networks with more

than one type of relations between nodes, and a node may have multiple attributes. Figure 1 gives an example of a multiplex network. This small network includes 9 nodes, two types of relations (green line and red dash), and two types of attributes (yellow and blue). The green line relation is assumed to be a positive relation and the red dash relation is a negative relation. Positive relations represent "attraction", such as "like", "friend of"; negative relations represent "repulsion", such as "dislike", "objector of". Two nodes tend to belong to a community if they are connected with positive relations, and belong to different communities if they are connected with negative relations. We can see that a reasonable community structure is given by two communities that consist of the left 4 nodes and right 5 nodes, respectively. However, node 5 would be in the left community if we only considered the green relation, but appears "closer" to the right community when also considering the red relation and the node attributes.

The information from all relations and attributes can be integrated into a probabilistic model; we only need to generalize (2) as follows.

• All types of relations should be considered. Therefore, if a network with p types of relations, $\boldsymbol{A} = \{A_1, \ldots, A_p\}$, then (2) becomes

$$\arg\max_{\Gamma}\max_{\theta} P_\theta(\boldsymbol{A}, \Gamma) = \arg\max_{\Gamma}\max_{\theta} P_\theta(\boldsymbol{A}|\Gamma)P_\theta(\Gamma) \tag{3}$$

• The attributes of the nodes can also play important roles in the community structure. The attributes of nodes within a same community should be as same as possible. Therefore, if a network with p types of relations and q types of attributes, $\boldsymbol{At} = \{At_1, \ldots, At_p\}$, equation (3) should be extended as (4):

$$\arg\max_{\Gamma}\max_{\theta} P_\theta(\boldsymbol{A}, \boldsymbol{At}, \Gamma) = \arg\max_{\Gamma}\max_{\theta} P_\theta(\boldsymbol{A}, \boldsymbol{At}|\Gamma)P_\theta(\Gamma) \tag{4}$$

4 RBN Models for Community Detection

4.1 Model Specification

We now briefly describe the elements of the RBN framework that we need for our community detection models. We begin by introducing a few basic concepts that are common for most SRL frameworks.

Given a set of relations $\boldsymbol{A}$, a set of attributes, $\boldsymbol{At}$, and a set of entities (vertices) V, we define following:

Definition 1. *A ground atom is an expression of the form $A_i(v_{j_1}, v_{j_2})$ or $At_i(v_j)$, where A_i is a relation, At_i an attribute, and the v_j are elements from V. In a non-ground atom, one can also have variables that range over all vertices as arguments for the A_i, At_i.*

A *probabilistic relational model* defines a joint probability distribution for all ground atoms $A_i(v_{j_1}, v_{j_k})$ or $At_i(v_j)$ as Boolean random variables [16]. Community membership will be represented by special attributes, so that also the joint distribution $P_\theta(\boldsymbol{A}, \boldsymbol{At}, \Gamma)$ is covered by these definitions. RBNs are a formal

representation language for probabilistic relational models. It is based on *probability formulas* that for each relation and each attribute specify the probability distributions for the ground atoms in a single declaration of the form

$$P(A_i(n_j, n_k) = true) \leftarrow ProbabilityFormula(n_j, n_k).$$

Here n_j, n_k are variables that range over vertices. The RBN language provides a small number of simple syntax rules with which complex probability formulas can be inductively constructed. The inductive definition is grounded by the base constructs of constants, parameters, and ground atoms. In this paper we will construct complex formulas only using the *convex combination* construct, which can be understood as a probabilistic if-then-else rule: if *PF1,PF2,PF3* are probability formulas, then *(PF1:PF2,PF3)* is a new formula. This formula evaluates to a mixture of the values returned by *PF2,PF3*, with *PF1* defining the mixture weights. In the particular case that *PF1* is purely Boolean (i.e., evaluates to 0 or 1), this means that the formula evalutes to *PF2* if *PF1* returns *true*, and to *PF3* if *PF1* returns *false*. In this paper we will only make use of the convex combination construct in this special form. Full definitions of syntax and semantics of RBNs can be found in [16].

We now turn to concrete encodings of community detection models using RBNs. For the time being, we only consider the case of two communities, and we use two special attributes $c1, c2$ to represent community membership. We then first define the prior distribution $P_\theta(\Gamma)$ using the two formulas

$$c1([Node]n) \leftarrow 0.5; \tag{5}$$

$$c2([Node]n) \leftarrow (c1(n) : 0, 1); \tag{6}$$

The first probability formula specifies that the probability of a node belongs to community $c1$ is 0.5; the second one specifies that a node with probability 0 belongs to cluster $c2$ if it belongs to community $c1$, and with probability 1 belongs to cluster $c2$ if it does not belong to community $c1$. This specification implies that the two communities form a partition of the nodes. By replacing the constants 0,1 in (6) with non-extreme values, the model will also allow overlapping communities, and nodes belonging to no community.

The following are probability formulas that define the model $P_\theta(\boldsymbol{A}, \boldsymbol{At}|\Gamma)$ for the relations and attributes of the graph in Figure 1:

$$\begin{aligned} link\_green([Node]n1, [Node]n2) \leftarrow (c1(n1) : (c1(n2) : \theta_1, 0.01), \\ (c2(n1) : (c2(n2) : \theta_2, 0.01), 0.5)); \end{aligned} \tag{7}$$

$$\begin{aligned} link\_red([Node]n1, [Node]n2) \leftarrow (c1(n1) : (c1(n2) : 0.01, \theta_3), \\ (c2(n1) : (c2(n2) : 0.01, \theta_4), 0.5)); \end{aligned} \tag{8}$$

$$attribute\_yellow([Node]n) \leftarrow (c1(n) : \theta_5, \theta_6); \tag{9}$$

$$attribute\_blue([Node]n) \leftarrow (c2(n) : \theta_7, \theta_8); \tag{10}$$

The θ_i in these formulas are free parameters that are estimated in the optimization process. The nested if-then-else construct of formula (7) can be expanded as follows:

$$P(link\_green(n_1, n_2) = true) = \begin{cases} \theta_1 & if\ n_1 \in c_1 \wedge n_2 \in c_2 \\ 0.01 & if\ n_1 \in c_1 \wedge n_2 \notin c_2 \\ \theta_2 & if\ n_1 \notin c_1 \wedge n_1 \in c_2 \wedge n_2 \in c_2 \\ 0.01 & if\ n_1 \notin c_1 \wedge n_1 \in c_2 \wedge n_2 \notin c_2 \\ 0.5 & if\ n_1 \notin c_1 \wedge n_1 \notin c_2 \end{cases}$$

The placement of the constants 0.01 here encodes that *link_green* is a positive relation. Formula (8) is structurally similar, but parameterized differently in order to enforce that *link_red* is considered a negative relation. Formulas (9) and (10) represent in a generic manner the dependency of these attribute values on the community membership.

In the model given by (5)-(10) all attributes and relations are independent given the community membership. A dependency of the *green* on the *red* relation could be modeled by the formula

$$\begin{aligned} link\_green([Node]n1, [Node]n2) = (c1(n1) : \\ (c1(n2) : (link\_red(n1, n2) : \theta_9, \theta_{10}), \theta_{11}), \\ (c2(n2) : (link\_red(n1, n2) : \theta_{12}, \theta_{13}), \theta_{14})); \end{aligned} \tag{11}$$

4.2 Inference

To solve the inference problem (4) when P_θ is given by an RBN, we extend a datastructure and inference techniques that were introduced for RBN parameter learning in [17].

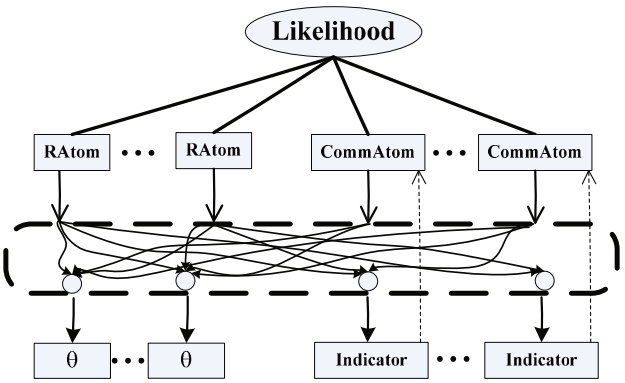

Fig. 2. The architecture of *MAP*-inference and parameter learning module

First, given the general RBN model and a concrete multiplex network, a representation of $P_\theta(\boldsymbol{A}, \boldsymbol{At}, \Gamma)$ in the form of a *likelihood graph* is constructed. Figure 2 illustrates the structure of this graph. Below the root *likelihood* node, the graph contains for each ground atom a node that represents the contribution of this ground atom to the overall likelihood. These nodes are inserted both for the atoms corresponding to observed relations and attributes in the network (denoted *RAtom* in Figure 2), and for atoms representing the unobserved community attributes (*CommuAtom*).

The leaves in the graph are the variables in the optimization problem (4): the free parameters θ of the model, and the truth settings for the community attributes, which jointly define Γ (*Indicator* nodes in Figure 2). Intermediate nodes in the graph (indicated by the dashed box in Figure 2) represent intermediate values that are obtained from probability sub-formulas in the recursive evaluation of the formulas associated with the ground atoms.

Using the likelihood graph as the common inference structure we compute (4) by alternating between maximization for θ and Γ. For maximizing θ we use the general parameter learning method of [17]. Maximization over Γ is a *Maximum A Posteriori Probability (MAP)* inference problem. This problem is intractable to solve optimally, and we use a local search procedure that combines greedy, random, and lookahead elements to obtain an approximate solution. In this process, the likelihood graph is used in two ways: first, it is used to compute the likelihood values of candidate solutions Γ, and second, the dependency structure of the likelihood function on the different atoms encoded in the graph structure is exploited to identify in the lookahead search candidate community membership atoms whose truth values might be changed to improve the likelihood.

5 Experiments

We first test the feasibility of our approach on a standard single relational network, the Zachary's karate club [18] network, which is a well-known benchmark for testing community detection algorithms. The network consists of 34 nodes as

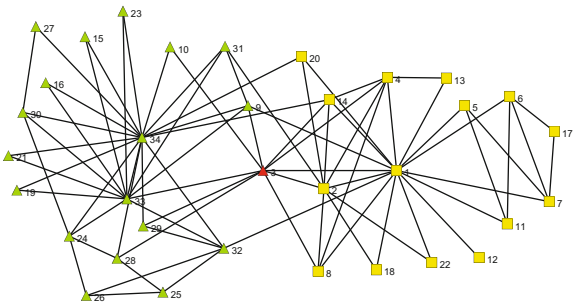

Fig. 3. The communities of Zacharys karate club network. Node shapes (and colors) indicate the community memberships of nodes.

members of the Karate club and 78 edges as friendships between members. We use probability formulas (5) and (6) to define $P_\theta(\Gamma)$. However, since the hard constraints on $c1$ and $c2$ are very difficult for MAP inference, we relax the model slightly by replacing 0 with 0.001, and 1 with 0.999. Since the relation in this network is positive, we model it using formula (7). The optimization terminates with parameter setting $\theta_1 = 0.259$,and $\theta_2 = 0.253$, and the communities shown in Fig.3. Yellow square and green triangle nodes represent nodes for which in the MAP solution for the community atoms exactly one of $c1$ and $c2$ was set to true. For node 3 (red triangle) in the middle both $c1$ and $c2$ were set to true in our solution. Apart from this ambiguous membership of node 3, our solution corresponds with the solution of [19].

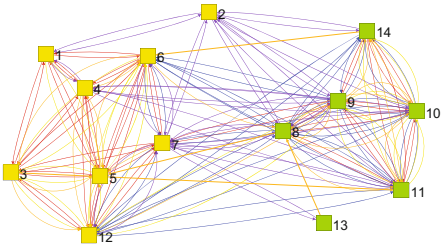

Fig. 4. The communities of bank wiring room network. Node colors indicate the community memberships of nodes.

Table 1. Parameters of the model for bank wiring dataset

	θ_1	θ_2	θ_3	θ_4
playing games together	0.500	0.563	-	-
friendship	0.276	0.219	-	-
helping	0.277	0.141	-	-
arguments	-	-	0.164	0.167
antagonism	-	-	0.272	0.271

We next conduct an experiment on the 'bank wiring room' multiplex network introduced in [12]. This network includes 14 employees work in a single wiring room. Relations between them are quite complex, and include *playing games together*(red arcs in Fig.4), *friendship*(yellow), *helping*(orange), *arguments about whether to open window*(blue) and *antagonism*(purple). The first three relations are positive and the latter two are negative. We define $P_\theta(\Gamma)$ as for the Zachary network. Each relation is modeled using the formulas of (7) and (8), depending of whether it is a positive or negative relation. The network and the computed communities are shown in Figure 4, and the learned parameters in Table 1. In this case, all nodes were assigned uniquely to one of the two communities (yellow and green squares in Figure 4), and the results coincide exactly with the structure suggested in [12].

In the preceding experiment the model encoded explicit information on which relations are positive, and which are nega tive. We next investigate the applicability of our approach when this distinction is not provided a-priori. For this, we use the network of Figure 5 (c), which contains two relations that are also separately drawn in Figure 5 (a) and (b). The network (a) shows an obvious community structure, whereas (c) shows no clear structure. We define a model with $P_\theta(\Gamma)$ as before, and for each of the two relations a formula of the form

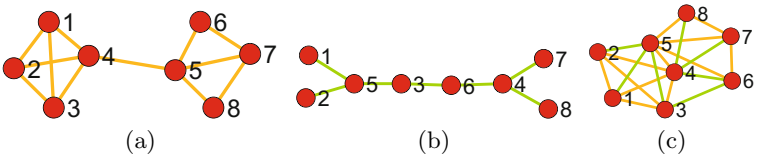

Fig. 5. A network which is a summation of two networks: (a) the network with obvious community structure; (b) the network with unobvious community structure; (c) summation of network (a) and (b).

$$link([Node]n1, [Node]n2) \leftarrow (c1(n1) : (c1(n2) : \theta_{14}, \theta_{15}), \\ (c2(n1) : (c2(n2) : \theta_{16}, \theta_{17}), 0.5)); \tag{12}$$

In these formulas all parameters are free, and therefore no prior bias is imposed on whether a relation is to be seen as positive or negative. The optimization terminates with parameters setting $\theta_{14} = 0.749$, $\theta_{15} = 0.063$,$\theta_{16} = 0.625$, $\theta_{17} = 0.063$ for the yellow relation (a), and $\theta_{14} = 1.2E-8$, $\theta_{15} = 0.312$,$\theta_{16} = 0.1.98E-6$, $\theta_{17} = 0.313$ for the green relation (b). From the parameters we can see that our model considers the yellow relation as positive and the green relation as negative. The results are clusters $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$.

For comparison, we also apply Newman's Edge Betweenness(EB)[19] and Ruan's Spectral Clustering (SC) method [9] to the network (c) without distinguishing the two relations. We obtain the clusters $\{1, 2, 3, 4, 5\}, \{6, 7, 8\}$ from SC, and $\{1\}, \{2, 3, 4, 5, 6, 8, 8\}$ from EB. Neither of these two clusterings appear very meaningful, and it is clear that the clustering evidence provided by the relation (a) is not strong enough for single-relational methods to detect the two clusters, when it is masked by relation (b).

6 Conclusions

We addressed the problem of community detection in multiplex networks using RBNs as a high-level and flexible specification language for probabilistic models. The main benefit of this approach is that we can use a single coherent framework with uniform inference techniques to experiment with different models that can capture different aspects and objectives that arise in the context of multiplex networks. Even though our results are quite preliminary at this point, they already demonstrate that using this coherent framework we can easily reconstruct results that have previously been obtained using very different techniques (graph cut techniques for the Zachary network, and matrix permutation for the Wiring Room). Our current system can handle networks up to approximately 300 nodes with arbitrary link structure. Future work will be directed towards exploring additional aspects such as multiple and overlapping community detection, and the application to bigger datasets.

References

1. Newman, M.: Communities, Modules and Large-Scale Structure in Networks. Nature Physics 8(1), 25–31 (2011)
2. Girvan, M., Newman, M.E.: Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
3. Newman, M.E.: Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems 38(2), 321–330 (2004)
4. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science 328(5980), 876–878 (2010)
5. Cai, D., Shao, Z., He, X., Yan, X., Han, J.: Mining Hidden Community in Heterogeneous Social Networks. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 58–65. ACM (2005)
6. Jaeger, M.: Relational Bayesian Networks. In: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, pp. 266–273. Morgan Kaufmann Publishers Inc. (1997)
7. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.M.: Self-Organization and Identification of Web Communities. Computer 35(3), 66–70 (2002)
8. Rattigan, M.J., Maier, M., Jensen, D.: Graph Clustering with Network Structure Indices. In: Proceedings of the 24th International Conference on Machine Learning, pp. 783–790. ACM (2007)
9. Ruan, J., Zhang, W.: An Efficient Spectral Algorithm for Network Community Discovery and its Applications to Biological and Social Networks. In: Seventh IEEE International Conference on Data Mining, pp. 643–648. IEEE (2007)
10. Fortunato, S.: Community Detection in Graphs. Physics Reports 486(3), 75–174 (2010)
11. Yang, B., Cheung, W.K., Liu, J.: Community Mining from Signed Social Networks. IEEE Transactions on Knowledge and Data Engineering 19(10), 1333–1348 (2007)
12. Breiger, R.L., Boorman, S.A., Arabie, P.: An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling. Journal of Mathematical Psychology 12(3), 328–383 (1975)
13. Taskar, B., Segal, E., Koller, D.: Probabilistic Classification and Clustering in Relational Data. In: International Joint Conference on Artificial Intelligence, vol. 17, pp. 870–878 (2001)
14. Kok, S., Domingos, P.: Statistical Predicate Invention. In: Proceedings of the 24th International Conference on Machine Learning, pp. 433–440. ACM (2007)
15. Xu, Z., Tresp, V., Yu, S., Yu, K.: Nonparametric Relational Learning for Social Network Analysis. In: KDD 2008 Workshop on Social Network Mining and Analysis (2008)
16. Jaeger, M.: Complex Probabilistic Modeling with Recursive Relational Bayesian Networks. Annals of Mathematics and Artificial Intelligence 32(1-4), 179–220 (2001)
17. Jaeger, M.: Parameter Learning for Relational Bayesian Networks. In: Proceedings of the 24th International Conference on Machine Learning. pp. 369–376. ACM (2007)
18. Zachary, W.W.: An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research 33(4), 452–473 (1977)
19. Newman, M.E., Girvan, M.: Finding and Evaluating Community Structure in Networks. Physical Review E 69(2), 026113 (2004)