Front. Phys. 12(6), 128903 (2017)
DOI 10.1007/s11467-017-0657-y

RESEARCH ARTICLE

# Modularity-like objective function in annotated networks

**Jia-Rong Xie, Bing-Hong Wang**†

*Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China*
*Corresponding author. E-mail: †bhwang@ustc.edu.cn*
*Received November 4, 2016; accepted January 5, 2017*

We ascertain the modularity-like objective function whose optimization is equivalent to the maximum likelihood in annotated networks. We demonstrate that the modularity-like objective function is a linear combination of modularity and conditional entropy. In contrast with statistical inference methods, in our method, the influence of the metadata is adjustable; when its influence is strong enough, the metadata can be recovered. Conversely, when it is weak, the detection may correspond to another partition. Between the two, there is a transition. This paper provides a concept for expanding the scope of modularity methods.

**Keywords** community structure, annotated networks, modularity, objective function

**PACS numbers** 89.75.Hc, 02.50.Tt



## 1 Introduction

Community structure, a partition of nodes in which the density of edges within groups is denser than that between groups, is an important large-scale structure in complex networks, and has attracted significant attention in recent years [1–3]. Many methods have been proposed for detecting community structure. Here, we focus on two: statistical inference [4–6] and modularity-based methods [7]. Statistical inference is flexible; it can be used for different purposes, such as detecting generalized communities [8] or estimating group number [9]. Additionally, statistical inference can be used for detecting annotated networks, in which annotations or metadata that describe the attributes of nodes (such as the age, gender, or ethnicity of individuals in a social network) accompany the network structure [10]. Newman–Girvan modularity [7] is the most popular measure of the quality of a partition. Several modifications have been proposed for measuring different unannotated network structures, including weighted [11], directed [12], bipartite [13] and multiplex networks [14]. However, modularity in annotated networks has not been defined. In the paper, we focus on the objective function in these networks and its relation to Newman–Girvan modularity.

The equivalence between modularity optimization and maximum likelihood [15, 16] may inspire us to our goal. However, this derivation is for unannotated networks. In the statistical inference method, the model of a network with community structure is defined and then fit to observed network data. In most cases, the model parameters are estimated by likelihood maximization; for different considerations or data types, the likelihoods are different. The likelihood in annotated networks differs from (though is similar to) that of unannotated networks. Herein, we ascertain the modularity-like objective function whose optimization is equivalent to the maximum likelihood in annotated networks. We demonstrate that the modularity-like objective function is a linear combination of modularity and conditional entropy. In contrast with the statistical inference method, we set a variable parameter that controls the influence of the metadata. Our results, in both synthetic and real-world networks, demonstrate that if the parameter is strong enough, the metadata can be recovered; however, if it is weak, our method may recover another partition that is more evident, instead of the metadata. Between the two, we find a transition from the more evident partition to the metadata.

## 2 Method

To illuminate our method, we first provide a brief introduction to the likelihood of statistical inference in annotated networks [10]. In this paper, we consider only

*Special Topic: Soft-Matter Physics and Complex Systems (Ed. Zhi-Gang Zheng). arXiv: 1701.04241.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2017

the case in which the metadata is a classification or a partition of nodes, $\boldsymbol{x} = \{x_i\}$. In this method, a degree-corrected stochastic block model is defined to a network. The probability, or likelihood, that the model generates a particular network $\boldsymbol{A}$ and group assignment $\boldsymbol{s}$ with $q$ groups is

$$\begin{aligned} P(\boldsymbol{A}, \boldsymbol{s}|\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{x}) &= P(\boldsymbol{A}|\boldsymbol{\Theta}, \boldsymbol{s}) P(\boldsymbol{s}|\boldsymbol{\Gamma}, \boldsymbol{x}) \\ &= \prod_{i<j} p_{ij}^{A_{ij}} (1-p_{ij})^{1-A_{ij}} \prod_i \gamma_{s_i x_i}, \end{aligned} \quad (1)$$

where $\gamma_{sx}$ is the probability that a node is assigned to group $s$ given its metadata $x$; $\boldsymbol{\Gamma}$ denotes the matrix of parameters $\gamma_{sx}$; $p_{ij} = k_i k_j \theta_{s_i s_j}$ is the probability of node $i$ connecting to $j$, where $k_i$ ($k_j$) is degree of node $i$ ($j$) and $\theta_{st}$ are parameters indicate the strength of connection between groups; and $\boldsymbol{\Theta}$ denotes the matrix of parameters $\theta_{st}$.

The likelihood maximization is equivalent to the maximization of the logarithm

$$\begin{aligned} \log P(\boldsymbol{A}, \boldsymbol{s}|\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{x}) \sim & \sum_i \log \gamma_{s_i x_i} + \frac{1}{2} \sum_{ij} A_{ij} \log(k_i k_j \theta_{s_i s_j}) \\ & + \frac{1}{2} \sum_{ij} \log(1 - k_i k_j \theta_{s_i s_j}) \\ \sim & \sum_x \sum_s N_{sx} \log \frac{N_{sx}}{N_x} + \frac{1}{2} \sum_{ij} A_{ij} \log \theta_{s_i s_j} \\ & - \frac{1}{2} \sum_{ij} k_i k_j \theta_{s_i s_j}, \end{aligned} \quad (2)$$

where $N_{sx}$ is the number of nodes assigned to group $s$ with annotation $x$ and $N_x$ is the number of nodes with annotation $x$. The first term is

$$\begin{aligned} \sum_x \sum_s N_{sx} \log \frac{N_{sx}}{N_x} &= N \sum_x \sum_s p(s,x) \log \frac{p(s,x)}{p(x)} \\ &= N \sum_x p(x) \left( \sum_s p(s|x) \log p(s|x) \right) \\ &= -N H(S|X), \end{aligned} \quad (3)$$

where $N$ is the number of nodes in the network and $H(S|X)$ is the conditional entropy. The second and third terms induce the modularity [16]. The planted partition model [17] is a special case of the stochastic block model in which the parameters $\theta_{st}$ describing the community structure take only two different values:

$$\theta_{st} = \begin{cases} \theta_{in} & \text{if} \quad s = t \\ \theta_{out} & \text{if} \quad s \neq t \end{cases}. \quad (4)$$

Eq. (4) implies that

$$\theta_{st} = (\theta_{in} - \theta_{out}) \delta_{st} + \theta_{out}, \quad (5)$$

$$\log \theta_{st} = (\log \theta_{in} - \log \theta_{out}) \delta_{st} + \log \theta_{out}. \quad (6)$$

Thus, the second and third terms of Eq. (2) are [16]

$$\begin{aligned} & \frac{1}{2} \sum_{ij} A_{ij} \log \theta_{s_i s_j} - \frac{1}{2} \sum_{ij} k_i k_j \theta_{s_i s_j} \\ & \sim M \log \frac{\theta_{in}}{\theta_{out}} \frac{1}{2M} \sum_{ij} \left( A_{ij} - \frac{2M(\theta_{in} - \theta_{out})}{\log \theta_{in} - \log \theta_{out}} \frac{k_i k_j}{2M} \right) \delta_{s_i s_j}, \end{aligned} \quad (7)$$

in which some constants have been dropped. The maximization of Eq. (2) is equivalent to the maximization of

$$\frac{1}{2M} \sum_{ij} \left( A_{ij} - \gamma \frac{k_i k_j}{2M} \right) \delta_{s_i s_j} - \alpha H(S|X) = Q(\gamma) - \alpha H, \quad (8)$$

where $\gamma = \frac{2M(\theta_{in} - \theta_{out})}{\log \theta_{in} - \log \theta_{out}}$ and $\alpha = \frac{N}{M(\log \theta_{in} - \log \theta_{out})}$, which can be estimated. In this paper, we set $\gamma = 1$ and treat $\alpha$ as a variable parameter to control the balance between the structure and metadata. High values of $\alpha$ drag the result to the metadata, though the principle for selecting the appropriate value of $\alpha$ is still unknown. We emphasize that our goal is to determine how metadata can be recovered, so the number of groups of detected partitions is equals to that of the metadata in most case. Eq. (8) is the modularity-like objective function, which is a linear combination of modularity and conditional entropy. We have demonstrated that the optimization of Eq. (8) is equivalent to the maximum likelihood of Eq. (1). As the modularity-like objective function is known, we use simulated annealing [18] for optimization with a fixed $q$.

## 3 Results

Our first example is a network generated by a stochastic block model (SBM). In SBM, nodes are randomly assigned to one of $q$ groups and the probability that any pair of nodes connects depends on the node memberships, $p_{ij} = \omega_{s_i s_j}$. In this case, we set $q = 4$ and

$$\boldsymbol{\omega} = \frac{4c}{N(1+\epsilon_1)(1+\epsilon_2)} \begin{pmatrix} 1 & \epsilon_2 & \epsilon_1 & \epsilon_1\epsilon_2 \\ \epsilon_2 & 1 & \epsilon_1\epsilon_2 & \epsilon_1 \\ \epsilon_1 & \epsilon_1\epsilon_2 & 1 & \epsilon_2 \\ \epsilon_1\epsilon_2 & \epsilon_1 & \epsilon_2 & 1 \end{pmatrix}, \quad (9)$$

where $c$ is the average degree in the network. It is also a special case of a nested SBM [19], in which $L$ (here $L = 2$) community structures are coupled. In the first partition $\boldsymbol{s}$, original groups 1 and 2 are merged into one group, and the remaining two original groups are merged into another group. In partition $\boldsymbol{s}'$, original groups 1 and

3 are merged and the left original groups are merged. $\epsilon_1$ and $\epsilon_2$ denote the strength of the two planted structures. In this case, we set $N = 2000$, $c = 3$, $\epsilon_1 = 0.1$, $\epsilon_2 = 0.15$ and metadata $\boldsymbol{x} = \boldsymbol{s}'$. $\boldsymbol{s}'$ is much weaker than $\boldsymbol{s}$, so that with this metadata, the method in Ref. [10] recovers $\boldsymbol{s}$ rather than $\boldsymbol{s}'$. However, by adjusting the influence of the metadata with parameter $\alpha$, our method can recover $\boldsymbol{s}$ in an appropriate range (see Fig. 1).

Figure 1 shows that the modularity-link function looks like a broken line with three segments. There is a transition at $\alpha_c = 0.052$. Below this transition, $\alpha$ is small enough that structure plays a leading role. Optimization of the objective function finds the partition with the highest modularity. In this case, $Q(\boldsymbol{s}) > Q(\boldsymbol{s}')$, so $\boldsymbol{s}$ is recovered. Above $\alpha_c$, the value of overlaps (i.e., the fraction of nodes correctly detected) with the two structures exchanges. If $\alpha$ is not high, both the structure and metadata play important roles in detection. The metadata provides all information of $\boldsymbol{s}'$, $H(\boldsymbol{s}'|\boldsymbol{x}) = 0$; while it provides no information to $\boldsymbol{s}$, $H(\boldsymbol{s}|\boldsymbol{x})$ is high. Thus, the metadata drags the detection to it. However, the landscape has a smooth valley surrounding $\boldsymbol{s}'$ [20]. Due to fluctuation, there are some partitions that are correlated with $\boldsymbol{s}'$ (i.e., the Hamming distance to $\boldsymbol{s}'$ is low) with higher modularity-like objective functions than those of $\boldsymbol{s}'$. Optimization methods will recover one of them, so the overlap between the detected partition and metadata is high but not equal to 1. Only when $\alpha$ is high enough, metadata plays crucial role and can be recovered absolutely.

Our second example is a network generated by a planted partition model, which is a special case of SBM with edge probabilities $p_{in}$ and $p_{out}$ for within-group and between-group edges. We generated node metadata that matched the true planted assignments, but with an error rate of $\rho = 0.2$ to indicate random noise. Without metadata, or if $\alpha = 0$, the approximate planted structure

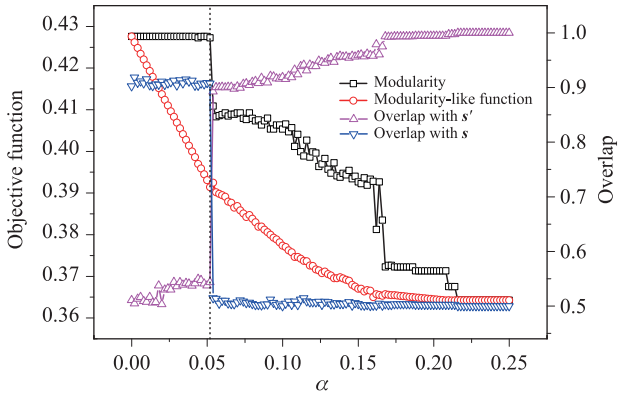


**Fig. 1** The objective functions and overlap of a network generated by the SBM in Eq. (9), with $N = 2000$, $c = 3$, $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.15$.

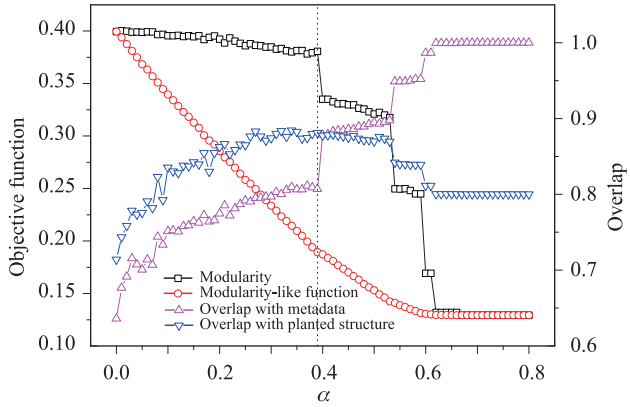


**Fig. 2** The objective functions and overlap of a network generated by a planted partition model with $N = 2000$, $q = 2$, $c = 3$, and $\epsilon = p_{out}/p_{in} = 0.2$.

can be recovered. As $\alpha$ increases, detection is gradually dragged to the metadata (see Fig. 2). If $\alpha$ is high enough, the metadata is recovered absolutely and the overlap with the planted structure was $1-\rho$. The transition in Fig. 2 is not as strong as that in Fig. 1; the overlap with the planted structure in Fig. 2 changes continuously at $\alpha_c$. The planted structure was recovered best at an $\alpha$ value of about 0.34. In Ref. [10], the strength of the metadata is fixed and may be not the best choice.

Our third example is a network of students drawn from the US National Longitudinal Study of Adolescent to Adult Health [21]. This network consists of a high school (US grades 9 to 12) and its feeder middle school (grades 7 and 8). The annotations of high/middle school and ethnicity construct two possible partitions [see Figs. 3(a) and (b)]. Between the two, the school is more evident than ethnicity; thus, we treat ethnicity as the metadata. The ethnicity annotation is so weak that with this metadata, the method in [10] recovers the school level rather than ethnicity. However, with $\alpha$, our method can recover ethnicity in an appropriate range [see Figs. 3(d)–(f) and Fig. 4]. Here, we use the normalized mutual information (NMI) [24] rather than overlap to measure how the detected partition matches the annotation, because the detection may have a different group number than the annotations.

## 4 Conclusion and discussion

In this paper, we ascertain the modularity-like objective function whose optimization is equivalent to the maximum likelihood in annotated networks. We demonstrate that the modularity-like objective function is a linear combination of modularity and conditional entropy, with a variable scale $\alpha$ that indicates the influence of the metadata. Unlike in the statistical inference method, our

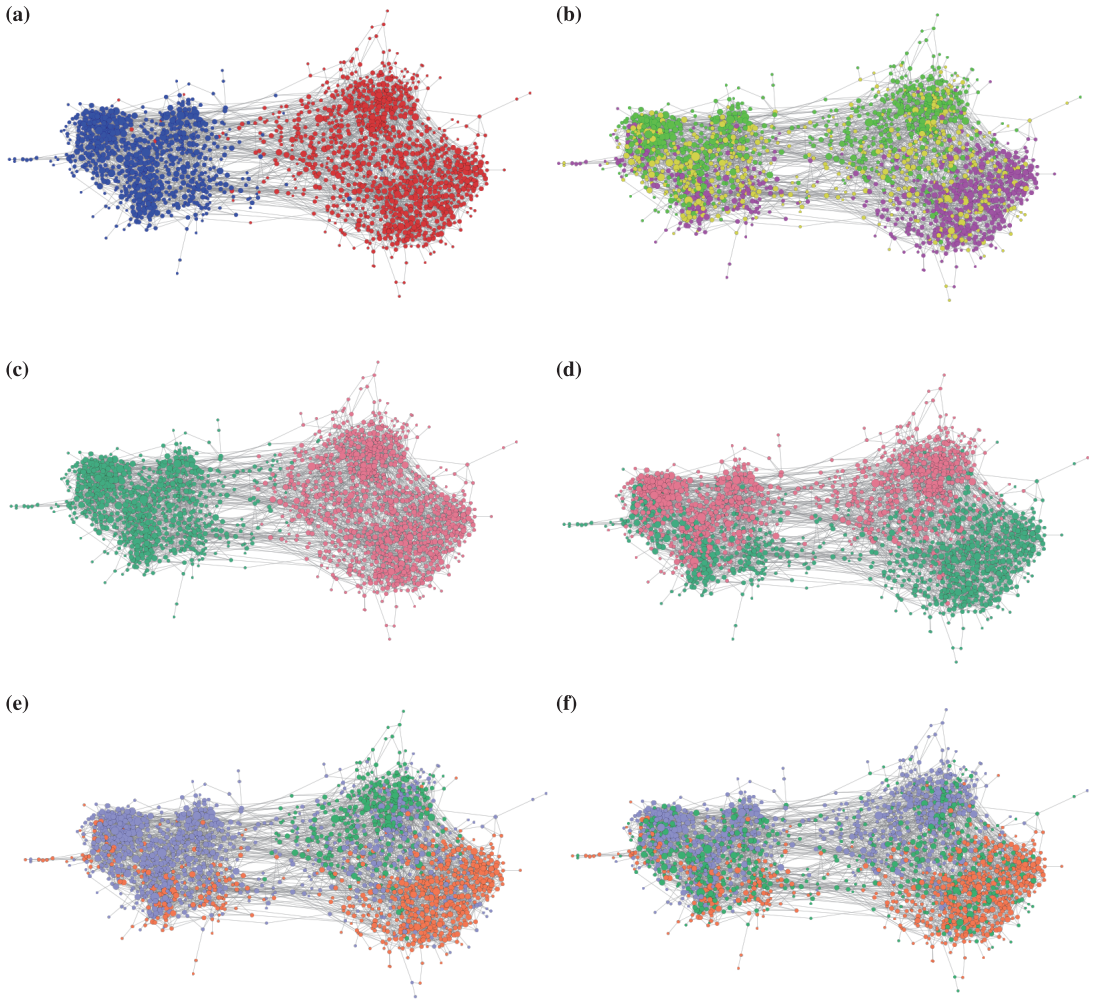


**Fig. 3** The ground-truth and detected partitions in the network of students. **(a)** The classifications of middle (blue) and high (red) school. **(b)** Ethnicity metadata: purple for White, green for Black, and yellow for others. **(c)** The detected partition recovers high/middle school, $q = 2$, $\alpha = 0.1$. (d)–(f) The detected partitions recover ethnicity. **(d)** $q = 2$, $\alpha = 0.3$, **(e)** $q = 3$, $\alpha = 0.5$ and **(f)** $q = 3$, $\alpha = 0.75$. The figures are drawn with the Gephi network visualization software [22] and ForceAtlas2 layout algorithm [23].

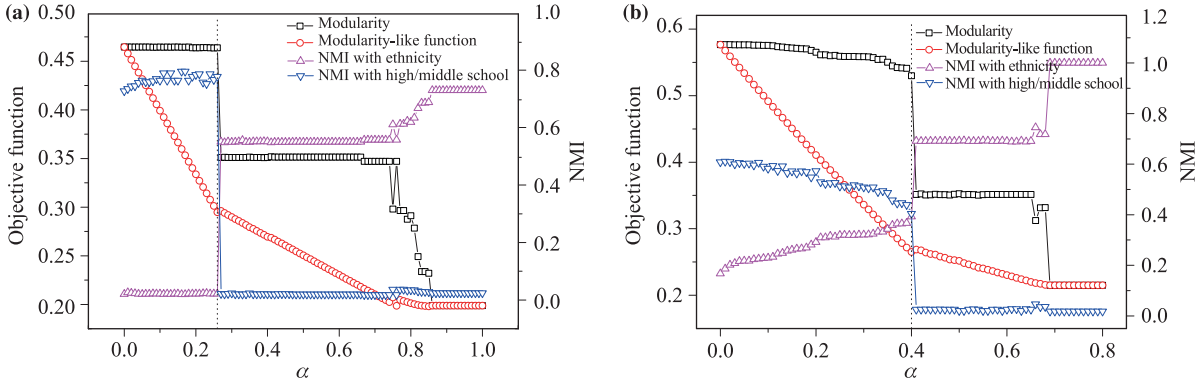


**Fig. 4** The objective functions and NMI of the network of students. **(a)** With detected group number $q = 2$. **(b)** $q = 3$.

method allows us to adjust the influence of the metadata. Examples in synthetic and real-world networks show that for an appropriate range of $\alpha$ (in which the influence is sufficiently strong), the metadata can be recovered. However, when $\alpha$ is low, another partition may be detected. Between the two values, there is a transition phase.

The statistical inference method is flexible, and it can be used to detect generalized communities [8] and estimate group number [9]. It is therefore interesting to find the corresponding modularity-like objective functions. In this paper, we optimized the modularity-like objective function by simulated annealing. Other optimization algorithms, such as belief propagation [15], are left for future work.

**Acknowledgements** This work was funded by the National Natural Science Foundation of China (Grant Nos. 11275186, 91024026, and FOM2014OF001).

## References and notes


1. S. Fortunato, Community detection in graphs, *Phys. Rep.* 486(3–5), 75 (2010)
2. M. E. J. Newman, Communities, modules and large-scale structure in networks, *Nat. Phys.* 8(1), 25 (2011)
3. S. Fortunato and D. Hric, Community detection in networks: A user guide, *Phys. Rep.* 659, 1 (2016)
4. A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Inference and phase transitions in the detection of modules in sparse networks, *Phys. Rev. Lett.* 107(6), 065701 (2011)
5. A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, *Phys. Rev. E* 84(6), 066106 (2011)
6. B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* 83(1), 016107 (2011)
7. M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* 69(2), 026113 (2004)
8. M. E. J. Newman and T. Peixoto, Generalized communities in networks, *Phys. Rev. Lett.* 115(8), 088701 (2015)
9. M. E. J. Newman and G. Reinert, Estimating the number of communities in a network, *Phys. Rev. Lett.* 117(7), 078301 (2016)
10. M. E. J. Newman and A. Clauset, Structure and inference in annotated networks, *Nat. Commum.* 7, 11863 (2016)
11. M. E. J. Newman, Analysis of weighted networks, *Phys. Rev. E* 70(5), 056131 (2004)
12. E. A. Leicht and M. E. J. Newman, Community structure in directed networks, *Phys. Rev. Lett.* 100(11), 118703 (2008)
13. M. J. Barber, Modularity and community detection in bipartite networks, *Phys. Rev. E* 76(6), 066102 (2007)
14. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J. P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, *Science* 328(5980), 876 (2010)
15. P. Zhang and C. Moore, Scalable detection of statistically significant communities and hierarchies, using message passing for modularity, *Proc. Natl. Acad. Sci. USA* 111(51), 18144 (2014)
16. M. E. J. Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection, *Phys. Rev. E* 94(5), 052315 (2016)
17. A. Condon and R. M. Karp, Algorithms for graph partitioning on the planted partition model, *Random Structures Algorithms* 18(2), 116 (2001)
18. R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, *Phys. Rev. E* 70, 025101(R) (2004)
19. T. P. Peixoto, Hierarchical block structures and high-resolution model selection in large networks, *Phys. Rev. X* 4(1), 011047 (2014)
20. B. H. Good, Y. A. de Montjoye, and A. Clauset, Performance of modularity maximization in practical contexts, *Phys. Rev. E* 81(4), 046106 (2010)
21. This research uses data from Add Health, a program project directed by K. M. Harris and designed by J. R. Udry, P. S. Bearman, and K. M. Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due R. R. Rindfuss and B. Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.
22. M. Bastian, S. Heymann, and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in: International AAAI Conference on Weblogs and Social Media, 2009
23. M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software, *PLoS One* 9(6), e98679 (2014)
24. L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, Comparing community structure identification, *J. Stat. Mech.* 09, P09008 (2005)