Overlapping Community Detection in Social Networks Based on Stochastic Simulation

Hadi Zare*, Mahdi Hajiabadi

University of Tehran Faculty of New Sciences and Technology

Received 6 February 2016; accepted 4 April 2016

Abstract

Community detection is a task of fundamental importance in social network analysis. Community structures enable us to discover the hidden interactions among the network entities and summarize the network information that can be applied in many applied domains such as bioinformatics, finance, e-commerce and forensic science. There exist a variety of methods for community detection based on different metrics and domain of applications. Most of these methods are based on the existing of the non-overlapping or sparse overlapping communities. Moreover, the experimental analysis showed that, overlapping areas of communities become denser than non-overlapping area of communities. In this paper, significant methods of overlapping community detection are compared according to well-known evaluation criteria. The experimental analyses on artificial network generation have shown that earlier methods of community detection will not discover overlapping communities properly and we offered suggestions for resolving them.

Keywords: dense overlapping communities; community detection; Social networks; artificial networks; Conductance.

1. Introduction

One of the most significant tasks in the network analysis is identifying the network communities [1]. Fundamentally, communities allow us to discover groups of interacting objects (i.e. nodes) and the relations between them. Moreover, experimental analysis have shown that the average distance between any pairs of American users in the Facebook is 4.3 and between any pairs of users is 4.7 [2]. So, social networks are playing an important role on advertising and marketing and detecting communities in these networks will be exploited as a tool for recommender systems, leading of consumer's behavior and marketing [1]. There are several definitions about communities, but generally, community is a group of nodes with dense interactions within the community and sparse interactions with other communities. There are many methods for finding communities in networks, which are divided into three groups. Graph-based methods optimization based methods and machine learning based methods [1, 3, and 4].

Most of the community detection methods are based on the non-overlapping community detection methods. In these methods one node would be belong to one community [5], however, in the real world networks one node would be belong to many communities. Recently, studying on the real world

* Corresponding author. Email: h.zare@ut.ac.ir>

networks have been revealed that overlapping regions of communities are denser than non-overlapping regions of communities [6]. For example, who attended in the same class and came from the same town are more likely to form link between them. Figure 1 shows the probability of having links between any pair of nodes in Youtube, Amazon, Livejournal, DBLP, Freiendster and Orkut social networks according to number of sharing communities. According to this point, earlier methods which are based on the sparse overlapping communities are not enable to discover overlapping regions of communities accurately. So, researchers are trying to present a new method for discovering dense overlapping community detection in networks.

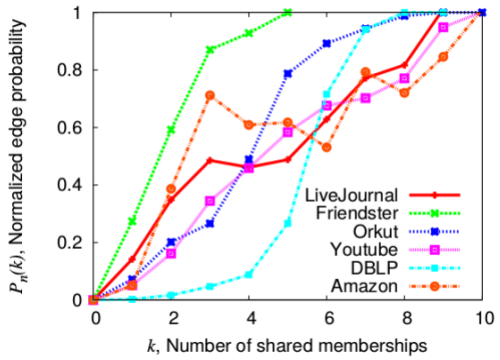

Fig. 1. Probability of having links according to number of shared communities

In this paper we exploit the artificial network generation approach to find the weakness of earlier approach to discover the overlapping communities in networks and we offered suggestions to resolve them.

In Section 2, some basic definitions about the community detection methods through stochastic generative approaches are presented. Then the well-known probabilistic overlapping community detection approaches are explained and analyzed in Section 3. Dataset description will be in the Section 4. The results are explained in Section 5 along with the standard evaluation metrics including the F1-score and the conductance measure. Finally, the conclusion and future works on this interesting field of research are discussed in Section 6.

2. Basic Definitions

According to the previous section, community is a group of nodes with high interactions with the community and the low interaction with nodes belong to other communities [1]. One of the fundamental question of community detection problems is finding the seed sets for communities. There are several approaches for finding the seed sets of communities [7, 8, 9]. Recently, researchers believe that the conductance is a good approach for finding the seed sets of communities [10, 11]. The conductance measure is related to the cut size of the community and the internal density of community. In the following equation, conductance level of community is calculated,

$$\Phi(S) = \frac{cut(S)}{min(Vol(S), vol(\bar{S}))} \quad (1)$$

Conductance is a measure for finding the quality level of community. In [12] the conductance measure is calculated according to the community size of network. Also all of the communities have a specific behavior according to the conductance, where the conductance level of these communities began to increase and gradually they decreased respect to the number of nodes in the given community. Figure 2 depicts the conductance level of network according to the community size.

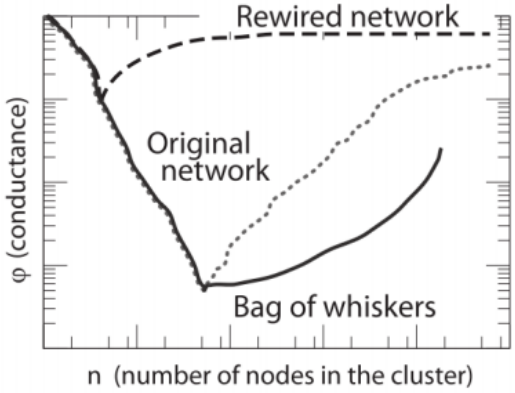

Fig. 2. Conductance level of community according to the community size [12]

3. Algorithms Description

In this section, two significant models for overlapping community detection in networks will be elaborated. First of all, we present Mixed Membership Stochastic Block-Models (MMSB) and finally, Affiliation Graphical Models (AGM) will be explained.

3.1. MMSB

Stochastic Block-Models provide a rich probabilistic framework for modeling relational data which each object effected on his neighbors objects [13]. Discovering communities according to MMSB method are done in two steps of local and global steps.

Global step: In this step, the probability memberships of each node to each community are updated. Probability membership of each node is calculated according to Dirichlet distribution. Dirichlet distribution is based on the rich-get-richer phenomenon. For example, if community one has 60 nodes, community two has 40 nodes and community has 20 nodes, consequently the probability membership of community one is higher than other communities. Here is the probability membership of each node to each community according to Dirichlet distribution:

$$p(\pi \mid \alpha_1, \alpha_2, ..., \alpha_n) = \frac{\Gamma(\alpha_1 + \alpha_2 + ... + \alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_n)} \quad (2)$$

In equation (2), π shows the mixing coefficients for each community. This parameter is obtained according to the Dirchlet distribution.

$$\pi_p \sim Dirichlet(\vec{\alpha}) \quad (3)$$

In equation (3), α represents the overlapping rates between communities, the increasing value of α leads to communities become more overlaps and the decreasing value of α resulted in communities will become more non-overlap respectively [13].

Local Step: Every node in the social networks follows the local behavior. High level of clustering coefficients, great number of cliques and power-law rule for degree distributions are showing the local behavior of each node in the social networks [14]. So, the community membership of each node would be determined from the communities of their neighbors. MMSB have two parameters for detecting communities of each node according to their neighbors. Zp→q denotes the group membership of node p when it contacts with node q. Zp←q shows the group membership of node q when it connects to node p. The probabilistic graphical model of the MMSB method is shown in Figure 3.

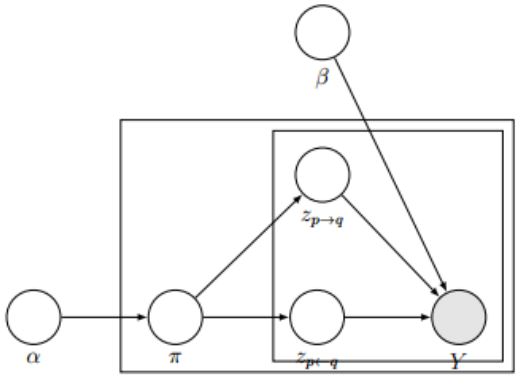

Fig. 3. Probabilistic graphical model of the MMSB

Where β is a community interaction matrix and according to the general definition of a community, since the interaction community matrix become more diagonal then community will detect more accurate [13]. Mixed membership stochastic Block-Models exploit the variational inference framework based on an auxiliary function to approximate the posterior of the communities as the latent variables in this model,

$$q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} \mid \vec{\gamma}_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}) = \prod_{p} q_1(\vec{\pi}_p \mid \vec{\gamma}_p) \prod_{p,q} (q_2(Z_{p\rightarrow q} \mid \phi_{p\rightarrow q}) q_2(Z_{p\leftarrow q} \mid \Phi_{p\leftarrow q})) \quad (4)$$

Where q_1 is a Dirichlet distribution, q_2 is a multinomial distribution and $\Delta = (\vec{\pi}\, 1:N, \Phi\rightarrow, \Phi\leftarrow)$ represents the set of latent variables should be estimated in the distribution. The procedure of minimizing the Kulback-Leibler divergence between the approximation distribution and the original posterior distribution are presented in the following algorithms,

Algorithm 1 Mixed Membership Stochastic Blockmodels

initialize $\vec{\gamma}^{0}_{pk} = \frac{2N}{K}$ for all p,k.
repeat
  for $p = 1$ *to* N **do** ?
    for $q = 1$ *to* N **do**
      get **variational** $\vec{\phi}^{t+1}_{p\rightarrow q}$ *and* $\vec{\phi}^{t+1}_{p\leftarrow q} = f(Y(p,q), \vec{\gamma}^{t}_{p}, \vec{\gamma}^{t}_{q}, B^{t})$
      Partially update $\vec{\gamma}^{t+1}_{p}$, $\vec{\gamma}^{t+1}_{q}$ *and* B^{t+1}
until convegence

Algorithm 2 f function of previous algorithm

initialize $\phi^{0}_{p\rightarrow q,g} = \phi^{0}_{p\leftarrow q,h} = \frac{1}{K}$ for all g,h.
repeat
for $g = 1$ *to* K **do** ?
  update $\phi^{s+1}_{p\rightarrow q} \propto f_1(\vec{\phi}^{s}_{p\leftarrow q}, \vec{\gamma}_{p}, B)$
normalize $\vec{\phi}^{s+1}_{p\rightarrow q}$ to sum to 1.
for $h = 1$ *to* K **do**
  update $\phi^{s+1}_{p\leftarrow q} \propto f_2(\vec{\phi}^{s}_{p\rightarrow q}, \vec{\gamma}_{q}, B)$
  normalize $\vec{\phi}^{s+1}_{p\leftarrow q}$ to sum to 1.
until *convergence*

MMSB exploits an iterative step for discovering the overlapping communities in a network. One of the main challenges of the MMSB is selecting the proper amount for Dirichlet hyper-parameter which strongly impact on the overlapping level of communities. Increasing value of Dirichlet hyper-parameter near to one will lead to communities become densely overlap and decreasing value of Dirichlet hyper-parameter will lead to more non-overlap communities.

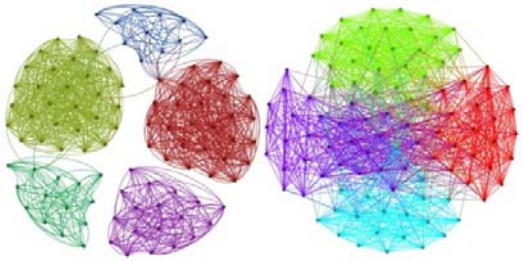

Fig. 4. Impact of the Dirichlet hyper-parameter of the overlapping level between communities

Figure (4) depicts the impact of Dirichlet hyper-parameter on the simulated network. The left subfigure has sparse overlapping communities with Dirichlet hyper-parameter 0.0001 and the right subfigure shows the dense overlapping communities with Dirichlet hyper-parameter 0.02.

In the following subsection, another outstanding method for finding overlapping communities in networks are presented.

3.2. Affiliation Graphical Models

Understanding and modeling of communities has been developed over time [1]. Controversial researchers think of networks as consisting of modular or dense communities that are linked by a small number of ties [15]. On the other hand, empirical observation of ground-truth networks lead that the probability of the pair of nodes sharing an edge, depends on the number of common communities which they are shared together [6]. A direct consequence of this claim is that the parts of the networks with overlapping community structures tend to more densely connected than their non-overlap parts of the network [6]. Due to this reason, earlier overlapping community detection works are not able to detect communities properly. Recently, different statistical methods are built up for detecting the dense overlapping communities in networks and we present one of these methods in the following subsection.

3.3. AGM

A new method for finding overlapping community detection in networks is presented in [16]. The assumption underlying is that the probability of forming a link between any pair of nodes depends on the common communities which they shared together. On the other hand, increasing the number of common communities between any pair of nodes will lead to raise the probability of sharing an edge between any pair of nodes. These hypotheses are evaluated on Figure (1) and the experimental results show that this

mode is very similar to the real world networks [9, 11, and 17].

Let B (V, C, M) be a bipartite graph where V is a set of nodes, C is a set of communities and an (u,v) ϵ M means that node u ϵ V belongs to community c ϵ C. Let also $\{P_c\}$ be a set of probabilities for all c ϵ C. Give model generates a graph G(V,E) by creating an edge (u,v) between the pair of nodes u,v ϵ V with probability p(u,v).

$$p(u,v) = 1 - \prod_{k \in C_{uv}} (1 - P_k) \qquad (5)$$

Where C_{uv} is a set of communities that u and v belong to them. Figure (5) shows generative model of network that is using the affiliation graph model.

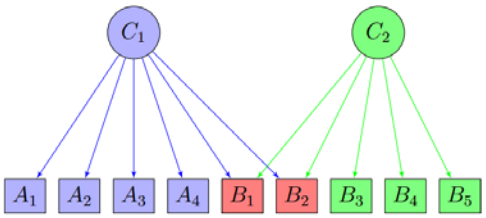

Fig. 5. Generative model of AGM

Figure (5) shows that we can generate a network with community membership of each node and the probability of sharing an edge between any pair of nodes according to Equation (6). So, having an adjacency matrix and exploiting iterative steps the community memberships of nodes will achieve. In each step, each node has three strategies in order to maximize the likelihood probability of network generation. These strategies include *leaving the community*, *appending to the new community and switching to another community* [16]. The iterative step repeat until convergence. The likelihood function of this model has a following form:

$$argmax\ ll = \prod_{(u,v)\in E} p(u,v) \times \prod_{(u,v)\notin E} (1 - p(u,v)) \qquad (6)$$

In the following section the aforementioned methods are compared on simulated and real-world network datasets.

4. Dataset Description

In order to generate the artificial networks satisfying in a wide variety of situations, two well-known approaches are exploited here, the Mixed Membership Stochastic Block-Models approach, MMSB and the LFR method [18]. Table (1) shows the general table of the simulated networks characteristics.

Table 1

General Table

Method	Type(dense/sparse)	Parameters	Nodes. No
MMSB	Sparse - Dense	Dirichlet hyperparameter α	100-500
LFR	Sparse - Dense	Mixing Parameters and mean degree	100-500

The MMSB procedure for network generation is built upon probabilistic approach such that the link formation between two nodes p and q within a network, denoted by Y (p, q), are assumed to be distributed as,

$$Y(p,q) \sim Bernoulli(\overrightarrow{Z^T}_{p\to q} B \vec{Z}_{q\to q}) \qquad (7)$$

According to Section (3.A) β is a community interaction matrix and Z is a multinomial distribution. The parameter α controls the overlapping behavior of communities. While the decreasing the value of α near to 0 resulted in formation of networks with sparse overlapping behavior of community structures, the increasing value of α to one tends to formation of dense overlapping community structures within a network. Due to complexity of different behavior of alpha, the modularity metric is applied to formally categorize the levels of overlapping behavior among the communities. We assume, the network has sparse overlapping communities, if the modularity level of network is greater than 0.5 and in the dense

overlapping communities if a modularity level of network is less than 0.4. In Table (2), characteristics of MMSB networks generator are shown. There are 3 different classes of networks with 100, 200 and 500 nodes and with sparse or dense overlapping communities.

Table 2

The details of artificial networks generation based on MMSB method

Nodes. No	Links. No	Communities. No	Modularity	Type(dense/sparse)	Hyperparameter
100	1021	5	0.71	Sparse	0.003
100	1277	5	0.37	Dense	0.03
200	4635	11	0.26	Dense	0.05
200	2224	11	0.71	Sparse	0.002
500	6074	32	0.32	Dense	0.02
500	4277	32	0.81	Sparse	0.001

We use another method for network and communities generation [18]. Table (3) clarifies the parameters of this method. We want to generate two types of network with LFR method, sparse overlapping communities and dense overlapping communities.

Table 3

LFR parameters descriptions

Parameters	Descriptions
-N	Number of Nodes
-k	Average Degree
-maxk	Maximum Degree
-mu	Mixing Parameters
-t1	Minus Exponent for the Degree Sequence
-t2	Minus Exponent for the Community Size Distribution
-minc	Minimum for the Community Size
-maxc	Maximum for the Community Size
-on	Number of Overlapping Nodes
-om	Number of Memberships of the Overlapping Nodes

For generating LFR simulated networks, we should discover the impact of each parameter on the network and the communities which are generated. One of the most important parameter on this model is the mixing parameter. This parameter manages the interactions between communities and greater level of mixing parameter will decrease the modularity value of the network. Another significant parameter is average degree. Increasing value of the average degree will lead to raise the interactions between communities will raise and overlapping communities will become more dense [18].

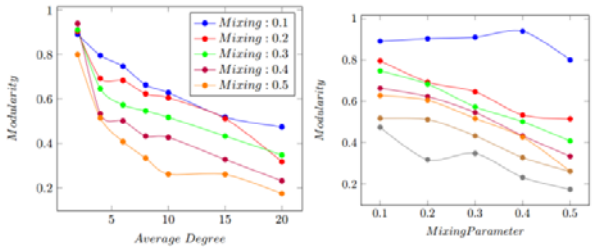

Fig. 6. Impact of the average degree and mixing parameter on the modularity

The details of network generation through the LFR approach are given in Table (4).

Table 4

The characterization of LFR procedure for network generation

Nodes. No	Links. No	Communities. No	Modularity	Type	Mixing	Overlapping of nodes
100	963	6	0.336	Dense	0.3	20
100	727	4	0.605	Sparse	0.1	10
200	1924	10	0.277	Dense	0.5	40
200	1503	11	0.6	Sparse	0.2	20
500	6123	16	0.306	Dense	0.5	100
500	6224	16	0.61	Sparse	0.2	50

The typical network generated through the LFR approach is shown in the following Figure. The left figure related to the first row of Table (4) and the right figure visualizes the second row of Table (4).

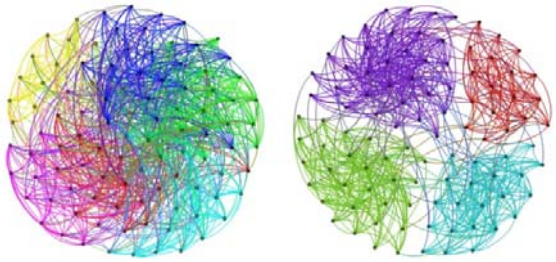

Fig. 7. LFR network generator

5. Results

The performances of the applied algorithms in our study are investigated based on two well-known evaluation criteria, the F1-score and the conductance measure. The F1-score measures the correctly

classified members in each community based on the ground-truth information. The conductance measure is related to the cut size of the communities and it is presented in the Section (2) [10]. Figure (8) shows the F1-score results on the benchmark algorithms. The top left subfigure shows the results for MMSB sparse overlapping network, top right related to the MMSB dense overlapping networks.

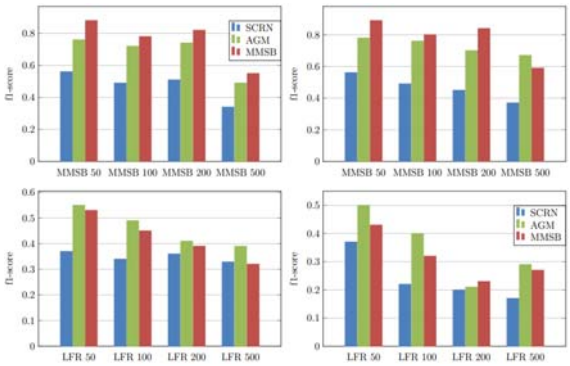

Fig. 8. F1-Score for comparing the benchmark methods

Left bottom and right bottom shows the sparse overlapping and dense overlapping communities of LFR respectively. MMSB dominated on the AGM on both of the LFR and MMSB networks simulated. We compare the AGM and MMSB methods according to conductance measure and the conductance results are shown in Figure (9). In Figure (9) conductance value of each method based on the number of communities are calculated. MMSB method are weakly dominated the AGM method on the conductance results.

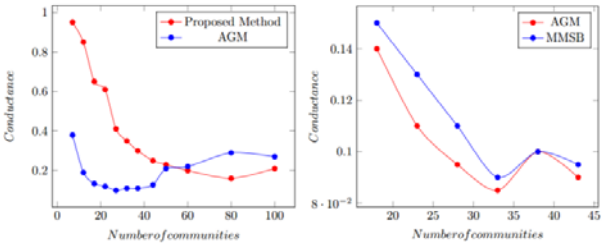

Fig. 9. Conductance level of each method according to the number of communities. The left subfigure is the result of LFR method and the right one is the result of MMSB

Finally, AGM and MMSB are compared on the three real-world networks. American Football network [20], Dolphins network [21] and the political book network. F1-score measure of this networks are shown in the following Table.

Table 5

F1-score for comparing the AGM and MMSB algorithms

Network	Nodes	Edges	Communities	f1-score	Methodology
American Football	115	616	12	0.43	MMSB
American Football	115	616	12	0.3	AGM
Dolphins	62	159	2	0.41	MMSB
Dolphins	62	159	2	0.86	AGM
Polbooks	105	441	3	0.31	MMSB
Polbooks	105	441	3	0.5	AGM

6. Conclusion

In this paper the most important probabilistic methods on overlapping community detection approaches in networks are compared through the stochastic simulation approach. Exploiting of the stochastic simulation approach leads us to gain a deeper insight with these methods and become familiar with the challenges of the earlier approach. Although AGM method has a better results on the real-world networks but on the simulated networks, AGM is worse than MMSB and neither of these methods reached to a reliable solution on finding the truth overlapping community detection in networks. The results revealed that the basic assumption on dense overlapping communities or sparse overlapping communities would not be resulted in to an effective community detection approach on the real-world network. Indeed it seems that an integrated approach based on overlapping and non-overlapping assumption for community detection would be more appropriate and reliable technique to extract the correct hidden community structures from the networks.

References

[1] S. Fortunato, "Community detection in graphs," Physics reports, vol.486, no.3, pp.75–174, 2010.

[2] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," arXiv preprint arXiv: 1111.4503, 2011.

[3] J. Xie, S. Kelley and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," Acm computing surveys (csur), vol.45, no.4, p.43, 2013.

[4] S. E. Schaeffer, "Graph clustering," Computer Science Review, vol.1, no.1, pp.27–64, 2007.

[5] M. E. Newman, "Modularity and community structure in networks," Proceedings of the national academy of sciences, vol.103, no.23, pp.8577–8582, 2006.

[6] J. Leskovec, K. J. Lang, A. Dasgupta and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in Proceedings of the 17th international conference on World Wide Web, pp.695–704, ACM, 2008.

[7] K. H. Lim and A. Datta, "A seed-centric community detection algorithm based on an expanding ring search," in Proceedings of the First Australasian Web Conference-Volume 144, pp.21–25, Australian Computer Society, Inc., 2013.

[8] I. M. Kloumann and J. M. Kleinberg, "Community membership identification from small seed sets," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1366–1375, ACM, 2014.

[9] J. J. Whang, D. F. Gleich and I. S. Dhillon, "Overlap-ping community detection using seed set expansion," in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp.2099–2108, ACM, 2013.

[10] D. F. Gleich and C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for lo-cal community methods," in Proceedings of the 18th ACM SIGKDD international conference on Knowl-edge discovery and data mining, pp.597–605, ACM, 2012.

[11] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in Proceedings of the sixth ACM international conference on Web search and data mining, pp.587–596, ACM, 2013.

[12] J. Leskovec, K. J. Lang and M. Mahoney, "Empirical comparison of algorithms for network community detection," in Proceedings of the 19th international conference on World wide web, pp.631–640, ACM, 2010.

[13] E. M. Airoldi, D. M. Blei, S. E. Fienberg and E. P. Xing, "Mixed membership stochastic block models," in Advances in Neural Information Processing Systems, pp.33–40, 2009.

[14] M. E. Newman, "The structure and function of complex networks," SIAM review, vol.45, no.2, pp.167– 256, 2003.da

[15] Y.-Y. Ahn, J. P. Bagrow and S. Lehmann, "Link communities reveal multiscale complexity in net-works," Nature, vol.466, no.7307, pp.761–764, 2010.

[16] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in Data Mining (ICDM), 2012 IEEE 12th International Conference on, pp.1170–1175, IEEE, 2012.

[17] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in Advances in neural information processing systems, pp.539–547, 2012.

[18] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical review E, vol.78, no.4, p.046110, 2008.

[19] L. Tang and H. Liu, "Scalable learning of collective behavior based on sparse social dimensions," in Pro-ceedings of the 18th ACM conference on Information and knowledge management, pp.1107–1116, ACM, 2009.

[20] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol.99, no.12, pp.7821–7826, 2002.

[21] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology, vol.54, no.4, pp.396–405, 2003.