COMMUNITY DETECTION GAME

Xuanyu Cao, Yan Chen†, and K. J. Ray Liu**

*Dept. of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA
†School of Electronic Engineering, University of Electronic Science and Technology of China, China

ABSTRACT

Real-world networks are often cluttered and hard to organize. Recent studies show that most networks have the *community structure*, i.e., nodes with similar attributes form a certain community, which enables people to better understand the constitution of the networks. Hitherto, various community detection methods have been proposed in the literature yet none of them takes the strategic interactions among nodes into consideration. Additionally, many real-world observations of networks are noisy and incomplete, i.e., with some missing links or fake links, due to either technology constraints or privacy regulations. In this work, a game-theoretic framework of community detection is established, where nodes interact and produce links with each other in a rational way based on mutual benefits. Given the proposed game-theoretic generative models for communities, we use expectation maximization (EM) algorithm to detect communities. Simulations on synthetic networks and experiments on real-world networks demonstrate that the proposed detection method outperforms the state-of-the-art.

Index Terms— Community detection, game theory, noisy networks

1. INTRODUCTION

Nowadays, networks are ubiquitous and often cluttered, leading to difficulties for recognizing patterns and mining knowledge from them. The first step to the understanding of the network structures is to arrange the networks in an organized manner: identifying nodes with similar attributes or functions and combining them together as a community. Given the importance of community structure, various community detection approaches have been proposed in the literature to identify meaningful communities in networks [1]. Existing community detection methods can be categorized into two classes: graph-theoretic approaches [2–7] and probabilistic generative models [8–10].

In a real-life network, nodes form links with each other through intelligent interactions. Users are rational in forming their social networks, in other words, when deciding whether to form a link or not, a user will judge if the benefit of this link is worthy of its cost (efforts and time spent in the relation). Hitherto, such strategic interactions among nodes have not been considered in community detection yet. Moreover, most real-world observations of networks are noisy and incomplete, i.e., there are missing links and fake links in the observed graph, due to technological constraints or privacy regulations. So far, no existing work has studied the community detection problem in noisy networks.

In this paper, we propose a game-theoretic framework to model the interactions among rational nodes in a network with community structure. The network can be either noiseless or noisy. The proposed link formation game connects the observed network structure with the hidden community structure. The Nash equilibrium (NE) of the noiseless network game and the subgame perfect equilibrium (SPE) of the noisy network game are derived. With these equilibria, a game-theoretic generative model of networks is obtained, according to which we use expectation maximization (EM) algorithm to detect communities. To the best of our knowledge, this is the first work on community detection taking noise effect into account. The effectiveness of the proposed detection algorithm is validated through simulations on synthetic networks and experiments on real-world networks.

2. GAME-THEORETIC GENERATIVE MODEL OF THE NETWORKS

Game theory is a mathematical tool used to study the strategic interactions among multiple rational decision makers [11]. In a network, each node (e.g., users in a social network) can be modeled as a rational player. The nodes interact with each other to form links, generating the graph structure that we observe. The utilities of the interactions depend on the community affiliations of the nodes.

In the following, we present our proposed game-theoretic generative models for both noiseless networks and noisy networks. Consider a network with N nodes and K communities. For each user $u \in \{1, 2, ..., N\}$, we denote the nonnegative vector $\mathbf{x}_u \in \mathbb{R}^K$ as its community affiliation strength vector, whose k-th component represents the strength of node u's affiliation to community k. The larger a certain entry of $\mathbf{x}_u$, the stronger the affiliation of node u to the corresponding community.

Table 1. The utility table of the game for noiseless networks.

u \ v	Link	Not Link
Link	1, 1	$f_2(\mathbf{x}_u, \mathbf{x}_v), f_1(\mathbf{x}_u, \mathbf{x}_v)$
Not Link	$f_1(\mathbf{x}_u, \mathbf{x}_v), f_2(\mathbf{x}_u, \mathbf{x}_v)$	0, 0

2.1. Game for Noiseless Networks

Each pair of nodes interacts with each other to decide whether to form a link or not. Specifically, when two nodes u, v interact, they play the following game:

- Pure strategies: {Link, Not Link}.
- Mixed strategies: $[0, 1]$, the probability of Link.
- Utility functions:
  1. If both nodes choose Link, then each one gets utility 1.
  2. If both nodes choose Not Link, then each one gets utility 0.
  3. From node u's perspective, i) if it chooses Not Link but its opponent v chooses Link, then it may get some one-shot information sharing or benefits from v, and thus gets utility $f_1(\mathbf{x}_u, \mathbf{x}_v)$; ii) if it chooses Link but its opponent v chooses Not Link, then it may have spent some efforts on trying to make this connection and thus gets (possibly negative) utility $f_2(\mathbf{x}_u, \mathbf{x}_v)$. We assume that f_1 and f_2 are symmetric functions, i.e., $f_i(\mathbf{x}_u, \mathbf{x}_v) = f_i(\mathbf{x}_v, \mathbf{x}_u), i \in \{1, 2\}$ so that the utility structure of the pair $\{u, v\}$ is symmetric. The utility functions are summarized in Table 1.

We note that the above proposed game contains two general functions f_1 and f_2. Different choices for these two functions lead to different games, and hence different game-theoretic generative models of the networks. For general f_1, f_2, the Nash equilibrium (NE) of the proposed game is identified in the following proposition.

Proposition 1. *In the proposed game for noiseless networks, suppose* $f_1(\boldsymbol{x}_u, \boldsymbol{x}_v) < 1, f_2(\boldsymbol{x}_u, \boldsymbol{x}_v) < 0$ *or* $f_1(\boldsymbol{x}_u, \boldsymbol{x}_v) > 1, f_2(\boldsymbol{x}_u, \boldsymbol{x}_v) > 0$*, then choosing the strategy* Link *with probability:*

$$p^\star(\boldsymbol{x}_u, \boldsymbol{x}_v) = \frac{f_2(\boldsymbol{x}_u, \boldsymbol{x}_v)}{f_1(\boldsymbol{x}_u, \boldsymbol{x}_v) + f_2(\boldsymbol{x}_u, \boldsymbol{x}_v) - 1} \tag{1}$$

is a symmetric mixed-strategy NE.

We assume that two nodes will link with each other if and only if both of them choose the strategy Link. Hence, at the NE, the link probability of the node pair (u, v) is:

$$H(\mathbf{x}_u, \mathbf{x}_v) \triangleq p^\star(\mathbf{x}_u, \mathbf{x}_v)^2 = \left(\frac{f_2(\mathbf{x}_u, \mathbf{x}_v)}{f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1} \right)^2. \tag{2}$$

Table 2. Utility table of the second stage in the game for noisy networks.

(a) When u, v are linked in the first stage.

u \ v	Truth-telling	Not Truth-telling
Truth-telling	1, 1	$g_2(\mathbf{x}_u, \mathbf{x}_v), g_1(\mathbf{x}_u, \mathbf{x}_v)$
Not Truth-telling	$g_1(\mathbf{x}_u, \mathbf{x}_v), g_2(\mathbf{x}_u, \mathbf{x}_v)$	0, 0

(b) When u, v are not linked in the first stage.

u \ v	Truth-telling	Not Truth-telling
Truth-telling	1, 1	$g_4(\mathbf{x}_u, \mathbf{x}_v), g_3(\mathbf{x}_u, \mathbf{x}_v)$
Not Truth-telling	$g_3(\mathbf{x}_u, \mathbf{x}_v), g_4(\mathbf{x}_u, \mathbf{x}_v)$	0, 0

Different utility functions $f_1()$ and $f_2()$ lead to different link probability function $H()$. Two examples of such functions that satisfy the assumption of Proposition 1 are listed as follows.

- When $f_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \exp(-\mathbf{x}_u^\top \mathbf{x}_v)}$ and $f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v)$, the link probability function is $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^\top \mathbf{x}_v)$, which coincides with the affiliated graph model (AGM) proposed in [9, 12].
- When $f_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{\frac{\mathbf{x}_u^\top \mathbf{x}_v}{1 + \mathbf{x}_u^\top \mathbf{x}_v}}$ and $f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v)$, the link probability function is $H(\mathbf{x}_u, \mathbf{x}_v) = \frac{\mathbf{x}_u^\top \mathbf{x}_v}{1 + \mathbf{x}_u^\top \mathbf{x}_v}$.

2.2. Game for Noisy Networks

The game-theoretic generative process of the noisy networks consists of two stages since, in addition to the generative process for the noiseless networks, we need another stage to take the generation of noise into consideration. The first stage is to determine whether to form a link or not while the second stage is to decide whether to report the truth about the link state. The overall utility is the sum of the utilities obtained in the two stage games. The first stage is the same as the game for the noiseless networks. Thus, we just focus on the second stage, which is specified for a node pair (u, v) as follows.

- Pure strategies: Truth-telling and Not Truth-telling
- Mixed strategies: $[0, 1]$, the probability of Truth-telling
- Outcome: The true linking state is reported if and only if both nodes adopt strategy Truth-telling.
- Utility functions: If u, v are linked in the first stage, the utility functions of all possible circumstances are listed in Table 2(a). Similarly, if u, v are not linked in the first stage, the utility functions are listed in Table 2(b). The utility functions $g_i()$ are all symmetric functions, i.e., $g_i(\mathbf{x}_u, \mathbf{x}_v) = g_i(\mathbf{x}_v, \mathbf{x}_u), i \in \{1, 2, 3, 4\}$.

We denote the overall strategy of the formulated two-stage dynamic game as $\langle p, (q_1, q_2) \rangle$ where p is probability of the strategy Link in the first stage and (q_1, q_2) are the probability

of the strategy `Truth-telling` in the second stage given that a link between u, v is formed or not formed in the first stage, respectively.

Proposition 2. *In the proposed dynamic game for noisy networks, $\langle p^\star, (q_1^\star, q_2^\star)\rangle$ is a symmetric mixed-strategy subgame perfect equilibrium (SPE), where*

$$q_1^\star(x_u, x_v) = \frac{g_2(x_u, x_v)}{g_1(x_u, x_v) + g_2(x_u, x_v) - 1}$$
$$q_2^\star(x_u, x_v) = \frac{g_4(x_u, x_v)}{g_3(x_u, x_v) + g_4(x_u, x_v) - 1}$$
$$p^\star(x_u, x_v) = \frac{f_2(x_u, x_v)}{f_1(x_u, x_v) + f_2(x_u, x_v) - 1 - g_1(x_u, x_v)q_1^\star(x_u, x_v) + g_3(x_u, x_v)q_2^\star(x_u, x_v)}$$

provided that $0 \le p^\star(\boldsymbol{x}_u, \boldsymbol{x}_v), q_1^\star(\boldsymbol{x}_u, \boldsymbol{x}_v), q_2^\star(\boldsymbol{x}_u, \boldsymbol{x}_v) \le 1$.

Denote $Y(u,v), \hat{Y}(u,v)$ the binary variable representing the true link state and the observed noisy link state between nodes u, v respectively, i.e., "1" represents the presence of a link while "0" represents no link. Then, at the SPE $\langle p^\star, (q_1^\star, q_2^\star)\rangle$, the link probability of nodes u, v is $H(\mathbf{x}_u, \mathbf{x}_v) = p(\mathbf{x}_u, \mathbf{x}_v)^2$ while the fake link and missing link probabilities are:

$$\epsilon_1(\mathbf{x}_u, \mathbf{x}_v) \triangleq \mathbb{P}\left(\hat{Y}(u,v) = 1 | Y(u,v) = 0\right) = 1 - q_2(\mathbf{x}_u, \mathbf{x}_v)^2, \quad (3)$$
$$\epsilon_2(\mathbf{x}_u, \mathbf{x}_v) \triangleq \mathbb{P}\left(\hat{Y}(u,v) = 0 | Y(u,v) = 1\right) = 1 - q_1(\mathbf{x}_u, \mathbf{x}_v)^2. \quad (4)$$

Thus, different utility functions lead to different link probabilities and link error probabilities. Specifically, for any link probability function $H()$, any fake link probability ϵ_1 and any missing link probability ϵ_2, we can achieve them by setting the utility functions in the game model as follows:

$$f_1(\mathbf{x}_u, \mathbf{x}_v) = (1 + \epsilon_1 - \epsilon_2)\sqrt{H(\mathbf{x}_u, \mathbf{x}_v)}, \quad (5)$$
$$g_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \epsilon_2},\ g_3(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \epsilon_1}, \quad (6)$$
$$f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v), \quad (7)$$
$$g_2(\mathbf{x}_u, \mathbf{x}_v) = -g_1(\mathbf{x}_u, \mathbf{x}_v),\ g_4(\mathbf{x}_u, \mathbf{x}_v) = -g_3(\mathbf{x}_u, \mathbf{x}_v). \quad (8)$$

Thus, by properly tuning the utility functions as above, the game-theoretic framework can model a general class of generative processes of networks with community structure.

3. A GENERAL COMMUNITY DETECTION ALGORITHM FOR NOISY NETWORKS

In this section, we briefly discuss how to detect communities in the proposed game-theoretic model. Since noiseless networks simply correspond to noisy networks with $\epsilon_1 = \epsilon_2 = 0$, we only focus on community detection in noisy networks from now on in this section. We assume that the link error probabilities ϵ_1 and ϵ_2 are constants independent of the affiliation strength $\mathbf{x}_u$. A graphical representation of the proposed game-theoretic generative model for noisy networks is shown in Fig. 1. For each pair of users u, v with community affiliation strength $\mathbf{x}_u, \mathbf{x}_v$, a link between them is formed with probability $H(\mathbf{x}_u, \mathbf{x}_v)$. The link state $Y(u,v)$ can be either '1' (linking) or '0' (not linking), with linking probability

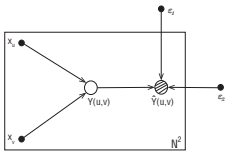


Fig. 1. Graphical illustration of the proposed game-theoretic generative model.

$H(\mathbf{x}_u, \mathbf{x}_v)$, i.e.,

$$Y(u,v) \sim \texttt{Bernoulli}(H(\mathbf{x}_u, \mathbf{x}_v)). \quad (9)$$

Afterwards, noise is added in so that the link state $Y(u,v)$ is flipped with fake link probability ϵ_1 and missing link probability ϵ_2 to generate the observed link state $\hat{Y}(u,v)$, i.e.,

$$\hat{Y}(u,v) \sim \texttt{Bernoulli}\left(\epsilon_1^{1-Y(u,v)}(1-\epsilon_2)^{Y(u,v)}\right). \quad (10)$$

We assume that the link error probabilities ϵ_1, ϵ_2 are known. Our goal is to infer the unknown community affiliation strength $\mathbf{X} \triangleq \{\mathbf{x}_u\}_{u=1}^N$, based on which we can do community detection.

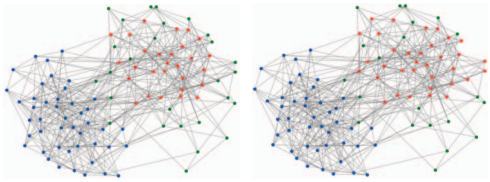
Due to the existence of the latent variables $\mathbf{Y}$ (the true network), direct maximum likehood estimation is intractable. We thus resort to the expectation maximization (EM) algorithm [13], an efficient algorithm iterating between two steps, i.e., the expectation step (E-step) and the maximization step (M-step). Because of space limitation, detailed algorithm is not presented here.

4. SIMULATIONS AND REAL DATA EXPERIMENTS

In this section, synthetic data based simulations as well as real data based experiments are conducted to validate the proposed community detection algorithm for the game-theoretic generative model.

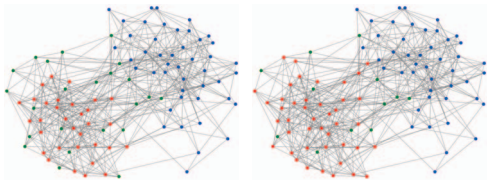
To implement simulations, we synthesize networks with N nodes and K communities according to the following procedure. Partition all nodes into K non-overlapping equal groups of nodes so that each group has N/K nodes. For each group, randomly pick $\eta N/K$ nodes outside of the group and add these nodes into the group, where $0 < \eta < 1$ is a user-defined parameter. Each group is defined to be a community. Choose some community affiliation strength for nodes in the community. This strength will influence the edge density of the networks. Generate the links according to the chosen link probability function $H(\mathbf{x}_u, \mathbf{x}_v)$. Add noise into the network according to the link error probabilities ϵ_1, ϵ_2.

The networks generated in this way have overlapping community structure. Actually, on average, for each community, a proportion of $2\eta/(1+\eta)$ nodes in the community also belong to other communities. The parameter setup for the simulation is as follows. We set $N = 100$, $K = 2, 3,$



(a) Detection of community 1 with AGM

(b) Detection of community 1 with the proposed noise-aware game-theoretic algorithm



(c) Detection of community 2 with AGM

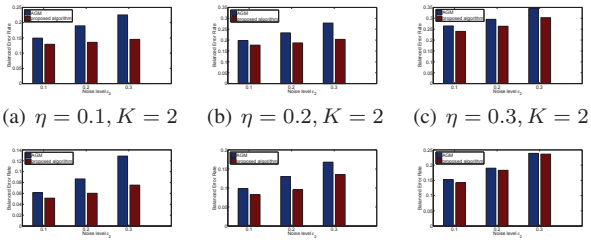
(d) Detection of community 2 with the proposed noise-aware game-theoretic algorithm

Fig. 2. Synthetic network with missing link probability $\epsilon_2 = 0.3$: comparison of the two detected communities with the ground-truth by using the proposed noise-aware game-theoretic algorithm and the AGM in [9], respectively. Red nodes: belonging to the community and detected as in the community; blue nodes: not belonging to the community and detected as not in the community; green nodes: belonging to the community but detected as not in the community; black nodes: not belonging to the community but detected as in the community. There happens to be no black node in this network instance.

$\eta = 0.1, 0.2, 0.3$. For link error probabilities, we select $\epsilon_1 = 0.005$ and $\epsilon_2 = 0.1, 0.2, 0.3$. The reason is that in practical networks, most of the link errors are missing links (incomplete graphs) instead of fake links. For link probability function, we choose $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^{\mathrm{T}}\mathbf{x}_v)$ and compare the performance with that of the affiliated graph model (AGM) proposed in [9]. A visualization of the community detection results of the proposed method and AGM, a state-of-the-art community detection algorithm with brilliant performance, for a synthetic network is presented in Fig. 2. There are two communities in the network, i.e., community 1 and community 2, whose detection results are shown respectively. We observe that the proposed method outperforms AGM, especially in community 2 where many undetected nodes (green nodes) of AGM becomes detected (red nodes) in the proposed approach.

For a detected community $\mathcal{C}$ and a ground-truth community $\bar{\mathcal{C}}$, the `Balanced Error Rate (BER)` between the two communities is defined to be:

$$\texttt{BER}(\mathcal{C}, \bar{\mathcal{C}}) = \frac{1}{2}\left(\frac{|\mathcal{C}\backslash\bar{\mathcal{C}}|}{|\mathcal{C}|} + \frac{|\bar{\mathcal{C}}\backslash\mathcal{C}|}{|\bar{\mathcal{C}}|}\right). \tag{11}$$



(a) $\eta = 0.1, K = 2$ (b) $\eta = 0.2, K = 2$ (c) $\eta = 0.3, K = 2$

(d) $\eta = 0.1, K = 3$ (e) $\eta = 0.2, K = 3$ (f) $\eta = 0.3, K = 3$

Fig. 3. Comparison between the proposed noise-aware game-theoretic community detection algorithm and the AGM method in [9].

Table 3. Relative enhancement of the proposed noise-aware game-theoretic algorithm over the AGM on real-world datasets.

Dataset \ Noise level ϵ_2	0.1	0.2	0.3	0.4
Facebook ego-network dataset	4.08 %	7.09 %	9.42 %	16.93 %
DBLP dataset	3.90 %	7.55 %	11.49 %	14.07 %

For every detected community $\mathcal{C}$, we calculate $\min_{\bar{\mathcal{C}}} \texttt{BER}(\mathcal{C}, \bar{\mathcal{C}})$. For every ground-truth community $\bar{\mathcal{C}}$, we calculate $\min_{\mathcal{C}} \texttt{BER}(\mathcal{C}, \bar{\mathcal{C}})$. Then, the performance metric is the average of all these minimum `BER`'s. The simulation results for different number of communities and different community overlapping extent are shown in Fig. 3, where we compare the proposed noise-aware game-theoretic algorithm with the AGM in [9]. We find that the proposed algorithm always outperforms the AGM, and the performance enhancement increases with the noise level ϵ_2 (except for networks in Fig. 3-f).

For real data experiments, we consider two datasets: the Facebook ego-networks dataset [14] and the DBLP collaboration network dataset [15]. Both networks have well-defined ground-truth communities. The relative improvement of the proposed noise-aware game-theoretic algorithm over the AGM is listed in Table 3. Again, the proposed algorithm always outperforms the AGM and the performance improvement increases with the noise level ϵ_2.

5. CONCLUSION

A game-theoretic analysis of the community detection problem in both noiseless networks and noisy networks has been presented, which takes nodes' rational decision making into account. The equilibria of the formulated game lead to a probabilistic generative model of networks with community structure. Based on the game-theoretic model, we propose a general community detection algorithm by using an EM algorithm. The effectiveness of the proposed algorithm is validated by simulations as well as real data experiments.

6. REFERENCES

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

[2] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[3] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA., 1967.

[4] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM Journal of Research and Development*, vol. 17, no. 5, pp. 420–425, 1973.

[5] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[6] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

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

[9] 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.

[10] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1348–1356, ACM, 2012.

[11] S. Tadelis, *Game theory: an introduction*. Princeton University Press, 2013.

[12] 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.

[13] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 4. springer New York, 2006.

[14] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Neural Information Processing Systems*, 2012.

[15] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.