

RESEARCH

Open Access

# Community detection in networks: a game-theoretic framework



Yan Chen<sup>1\*</sup> , Xuanyu Cao<sup>2</sup> and K. J. Ray Liu<sup>3</sup>

## 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 and thus gain more insights into the complicated networks. Strategic nodes belonging to different communities interact with each other to decide mutual links in 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, i.e., maximizing their own utility functions when forming a community. Given the proposed game-theoretic generative models for communities, we present a general community detection algorithm based on expectation maximization (EM). Simulations on synthetic networks and experiments on real-world networks demonstrate that the proposed detection method outperforms the state of the art.

**Keywords:** Community detection, Game theory, Noisy networks, EM algorithm

## 1 Introduction

<sup>1</sup>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 group or cluster. In fact, most real-world networks are empirically observed to possess the *community structure* [2–4], where nodes with analogous properties compose functional modules in networks. For instance, in online social networks, users form groups according to common experiences, affiliations, or hobbies; in biological networks, cells with similar functions constitute tissues; in research networks, researchers with similar interests comprise research fields or disciplines. Revealing the hidden

community structure can significantly simplify the representations of networks and facilitate the comprehension of networks.

Given the importance of community structure, various community detection approaches have been proposed in the literature to identify meaningful communities in networks [2]. Existing community detection methods can be categorized into two classes: graph-theoretic approaches and probabilistic generative models. In traditional graph-theoretic approaches, general clustering methods such as hierarchy clustering [5], *k*-means clustering [6], and spectral clustering [7] are applied to detect communities in networks. In the recent decade, various novel graph-theoretic methods were proposed. Divisive methods (e.g., Newman and Girvan [8]) iteratively deleted the identified inter-community links to separate the entire network into isolated communities. Newman [9] initiated the concept of modularity to assess the quality of the detected communities while the optimization of modularity led to a series of detection algorithms. Clique percolation method by Palla et al. [10] detected overlapping cohesive clusters, or cliques in networks. The graph-theoretic community

\*Correspondence: [eecyan@uestc.edu.cn](mailto:eecyan@uestc.edu.cn)

<sup>1</sup>School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China  
Full list of author information is available at the end of the article

<sup>1</sup>A preliminary version of this work [1] was presented in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016.

detection methods utilize various graph-theoretic measures to identify cohesive groups of nodes in the networks. Additionally, the performance limits of community detection in various random graph models are investigated in [11–13].

In probabilistic generative models, the observed network is regarded as the ramification of a community structure-related probabilistic generative process and the community detection problem is posed as a statistical inference problem. Among this category, Airoldi et al. presented the mixed membership stochastic blockmodel in [14], where each node's community affiliation strengths among all the communities are regarded as a probability distribution. A variational EM algorithm was proposed to infer the community affiliations efficiently. Yang and Leskovec proposed the affiliated graph model (AGM) in [15]. In AGM, total community affiliation strengths are allowed to vary from node to node, leading to more degrees of freedom in modeling overlapping communities. Sun et al. investigated community detection in heterogeneous information networks [16], where different types of nodes, e.g., authors, papers, and venues for citation networks, are present.

In real-life networks such as the Facebook friendship network and the DBLP collaboration network, nodes form links with each other through intelligent interactions. For example, in a social network, users interact with each other on common hobbies, experiences, affiliations, and finally decide whether to make connections (friends), i.e., to form a link, or not. Our hypothesis is that 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. On the other hand, game theory, originating from microeconomics, is a mathematical tool that has been applied to various engineering problems to model the strategic interactions among rational players [17–22]. The outcome of the mutual interactions between rational players can be predicted by using game theory. This motivates us to resort to game theory to investigate the interactions among nodes in a network with community structure.

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. For instance, observations in social networks are often incomplete because of the privacy policy of social websites and the flaws of the acquisition approaches, not to mention the social network data cannot track all the interactions among users. So far, no existing work has studied the behavior of community detection algorithms under a generative model of noise in observed

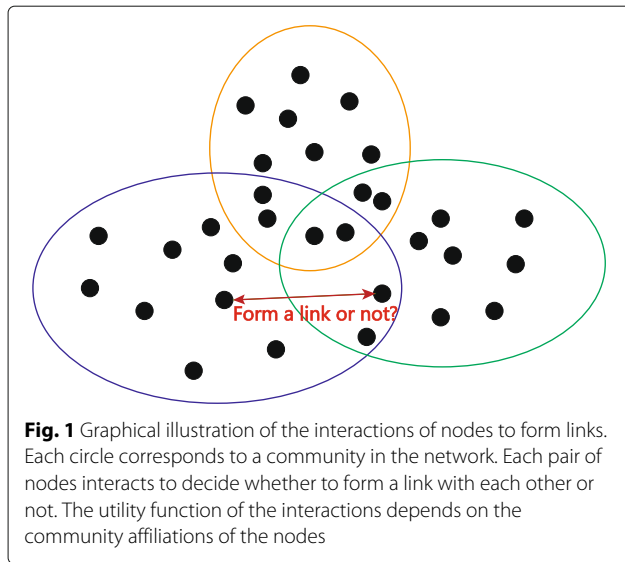
networks. This motivates us to consider the community detection problem in both noiseless networks and noisy networks. We find that the link errors or noise can be well absorbed into the proposed game theoretic framework. The main contributions of this paper are summarized as follows.

- We propose a game-theoretic framework to model the interactions among strategic 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, which enables community detection in both noiseless networks and noisy networks.
- According to the proposed game-theoretic generative model, we derive a general community detection algorithm based on expectation maximization (EM) for both noiseless networks and noisy networks. The effectiveness of the proposed detection algorithm is validated through simulations on synthetic networks and experiments on real-world networks.

The roadmap of the rest of this paper is as follows. In Section 2, we elaborate the game-theoretic model and present the equilibrium analyses. In Section 3, an EM-based general community detection algorithm is presented according to the proposed game-theoretic model. In Section 4, simulations as well as real-world dataset experiments are conducted. In Section 5, we conclude this work.

## 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 [23]. A game consists of (i) a set of players, (ii) a set of actions for each player, and (iii) a set of utilities for each player given the actions of all the players. Outcomes of the games can be obtained by resorting to solution concepts such as Nash equilibrium and subgame perfect equilibrium, which will be discussed later. 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 interactions can be illustrated as in Fig. 1. The utilities of the interactions depend on the community affiliations of the nodes. The fundamental hypothesis of this work is that, for two users in a social network, if they both belong



to a certain group, then they prefer to be friends, because forming a link will result in higher utilities for both parties. The interactions among nodes in a network can be analogous to the interactions among players in a game. Therefore, game theory is indeed an ideal tool to model and understand the community detection problem. Moreover, due to the acquisition errors, the networks acquired from the real-world data may be noisy and incomplete, i.e., many true links can be missed and lots of spurious links may be formed. In this paper, we will show that the proposed game theoretic framework can tackle such an issue, and therefore, we will consider community detection over both noiseless networks and noisy networks.

In the following, we first present our proposed game-theoretic generative models for both noiseless networks and noisy networks. Specifically, given the nodes' utility functions, which depend on their community affiliations, we derive the equilibrium of the pairwise link probability and based on which we propose generative models of the networks.

Consider a network with  $N$  nodes and  $K$  communities. For each user  $u \in \{1, 2, \dots, N\}$ , 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.

## 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. We consider two regions for the utility function  $f_1$  and  $f_2$ :  $f_1(\mathbf{x}_u, \mathbf{x}_v) < 1, f_2(\mathbf{x}_u, \mathbf{x}_v) < 0$  and  $f_1(\mathbf{x}_u, \mathbf{x}_v) > 1, f_2(\mathbf{x}_u, \mathbf{x}_v) > 0$ . We explain these two regions of utilities from node  $u$ 's perspective as follows. In the first region, if it selects Not Link while the opponent node  $v$  selects Link, then it gets some one-shot benefits from  $v$  such as sharing of information but loses long-term potential benefits from the potential connection so that its utility  $f_1(\mathbf{x}_u, \mathbf{x}_v)$  is smaller than 1. If it selects Link while node  $v$  selects Not Link, then it may have spent some efforts on trying to establish the connection and therefore lose some utility, i.e., its utility  $f_2(\mathbf{x}_u, \mathbf{x}_v)$  is less than 0. In the second region, if it selects Not Link while the node  $v$  selects Link, then it gets a one-shot benefit  $f_1(\mathbf{x}_u, \mathbf{x}_v)$  larger than 1 since it does not need to pay any efforts on establishing the connection. If it selects Link while node  $v$  selects Not Link, then though the connection is not established, it can make node  $v$  to know better about it or advertise itself. Hence, it still gets some positive utility.

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

**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

$$p^*(\mathbf{x}_u, \mathbf{x}_v) = \frac{f_2(\mathbf{x}_u, \mathbf{x}_v)}{f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1} \quad (1)$$

is a symmetric mixed-strategy NE.

*Proof* It is conspicuous that  $p^*(\mathbf{x}_u, \mathbf{x}_v) \in (0, 1)$ , i.e.,  $p^*(\mathbf{x}_u, \mathbf{x}_v)$  is a proper probability. Suppose node  $v$  selects the strategy `Link` with probability  $p^*(\mathbf{x}_u, \mathbf{x}_v)$ . Then, if node  $u$  selects `Link`, its utility is  $p^*(\mathbf{x}_u, \mathbf{x}_v) + (1 - p^*(\mathbf{x}_u, \mathbf{x}_v))f_2(\mathbf{x}_u, \mathbf{x}_v) = f_1(\mathbf{x}_u, \mathbf{x}_v)f_2(\mathbf{x}_u, \mathbf{x}_v)/(f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1)$ . If node  $u$  selects `Not Link`, its utility is  $p^*f_1(\mathbf{x}_u, \mathbf{x}_v) = f_1(\mathbf{x}_u, \mathbf{x}_v)f_2(\mathbf{x}_u, \mathbf{x}_v)/(f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1)$ . Consequently, node  $u$  is indifferent between the two strategies. Due to the symmetric structure of the game, node  $v$  is also indifferent as long as node  $u$  is playing the mixed strategy  $p^*(\mathbf{x}_u, \mathbf{x}_v)$ . Hence,  $p^*(\mathbf{x}_u, \mathbf{x}_v)$  is a symmetric mixed-strategy NE.  $\square$

**Remarks 1** We note that, besides the mixed-strategy NE mentioned in Proposition 1, there also exist other pure strategy NEs. For instance,  $(\text{Link}, \text{Link})$  and  $(\text{Not Link}, \text{Not Link})$  are NEs when  $f_1(\mathbf{x}_u, \mathbf{x}_v) < 1, f_2(\mathbf{x}_u, \mathbf{x}_v) < 0$ , while  $(\text{Link}, \text{Not Link})$  and  $(\text{Not Link}, \text{Link})$  are NEs when  $f_1(\mathbf{x}_u, \mathbf{x}_v) > 1, f_2(\mathbf{x}_u, \mathbf{x}_v) > 0$ . However, since our aim is to obtain a non-degenerated link probability for the generative model, we only focus on the 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:

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

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  $f_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \exp(-\mathbf{x}_u^T \mathbf{x}_v)}$  and  $f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v)$ , the link probability function is  $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^T \mathbf{x}_v)$ , which coincides with the affiliated graph model (AGM) proposed in [15, 24]. The AGM becomes a special case of our game-theoretic model if we choose the link probability function  $H$  is in this form.
- When  $f_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{\frac{\mathbf{x}_u^T \mathbf{x}_v}{1 + \mathbf{x}_u^T \mathbf{x}_v}}$  and  $f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v)$ , the link probability function is  $H(\mathbf{x}_u, \mathbf{x}_v) = \frac{\mathbf{x}_u^T \mathbf{x}_v}{1 + \mathbf{x}_u^T \mathbf{x}_v}$ .

The above two link probability functions are intuitively reasonable: if nodes  $u, v$  share a lot of community affiliations in common, the inner product  $\mathbf{x}_u^T \mathbf{x}_v$  is large, and so is the link probability  $H(\mathbf{x}_u, \mathbf{x}_v)$ . The differences of these

two link probability functions lie in their increasing speed with respect to  $\mathbf{x}_u^T \mathbf{x}_v$ . Different networks may be suitable for different link probability functions. After every pair of nodes finishes the game and decides whether to form a link or not, the entire network is constructed. Hence, the proposed game-theoretic model is a generative model of the networks.

## 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.

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

u	v	Truth-telling	Not Truth-telling
(a) When $u, v$ are linked in the first stage.			
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	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.			
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	Truth-telling	$g_3(\mathbf{x}_u, \mathbf{x}_v), g_4(\mathbf{x}_u, \mathbf{x}_v)$	0, 0

**Proposition 2** In the proposed dynamic game for noisy networks,  $\langle p^*, (q_1^*, q_2^*) \rangle$  given in (3), (4), and (5) is a symmetric mixed-strategy subgame perfect equilibrium (SPE)

$$q_1^*(\mathbf{x}_u, \mathbf{x}_v) = \frac{g_2(\mathbf{x}_u, \mathbf{x}_v)}{g_1(\mathbf{x}_u, \mathbf{x}_v) + g_2(\mathbf{x}_u, \mathbf{x}_v) - 1} \quad (3)$$

$$q_2^*(\mathbf{x}_u, \mathbf{x}_v) = \frac{g_4(\mathbf{x}_u, \mathbf{x}_v)}{g_3(\mathbf{x}_u, \mathbf{x}_v) + g_4(\mathbf{x}_u, \mathbf{x}_v) - 1} \quad (4)$$

$$p^*(\mathbf{x}_u, \mathbf{x}_v) = \frac{f_2(\mathbf{x}_u, \mathbf{x}_v)}{f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1 - g_1(\mathbf{x}_u, \mathbf{x}_v)q_1^*(\mathbf{x}_u, \mathbf{x}_v) + g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v)} \quad (5)$$

provided that  $0 \leq p^*(\mathbf{x}_u, \mathbf{x}_v), q_1^*(\mathbf{x}_u, \mathbf{x}_v), q_2^*(\mathbf{x}_u, \mathbf{x}_v) \leq 1$ .

*Proof* According to Proposition 1, the mixed-strategy in the second stage, i.e.,  $q_1^*$  and  $q_2^*$ , is an NE at the second stage. To show that  $\langle p^*, (q_1^*, q_2^*) \rangle$  is also a NE at the first stage, we assume that node  $v$  uses the strategy  $\langle p^*, (q_1^*, q_2^*) \rangle$ . Thus, if node  $u$  chooses Link in the first stage, regardless of its strategy in the second stage, its total utility is given in (6).

$$\begin{aligned} & p^*(\mathbf{x}_u, \mathbf{x}_v) + (1 - p^*(\mathbf{x}_u, \mathbf{x}_v))f_2(\mathbf{x}_u, \mathbf{x}_v) + p^*(\mathbf{x}_u, \mathbf{x}_v)g_1(\mathbf{x}_u, \mathbf{x}_v)q_1^*(\mathbf{x}_u, \mathbf{x}_v) \\ & + (1 - p^*(\mathbf{x}_u, \mathbf{x}_v))g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v) \\ & = g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v) \\ & + \frac{f_1(\mathbf{x}_u, \mathbf{x}_v)f_2(\mathbf{x}_u, \mathbf{x}_v)}{f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1 - g_1(\mathbf{x}_u, \mathbf{x}_v)q_1^*(\mathbf{x}_u, \mathbf{x}_v) + g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v)}. \end{aligned} \quad (6)$$

If node  $u$  chooses Not Link in the first stage, regardless of its strategy in the second stage, its total utility is given in (7).

$$\begin{aligned} & p^*(\mathbf{x}_u, \mathbf{x}_v)f_1(\mathbf{x}_u, \mathbf{x}_v) + g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v) \\ & = g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v) \\ & + \frac{f_1(\mathbf{x}_u, \mathbf{x}_v)f_2(\mathbf{x}_u, \mathbf{x}_v)}{f_1(\mathbf{x}_u, \mathbf{x}_v) + f_2(\mathbf{x}_u, \mathbf{x}_v) - 1 - g_1(\mathbf{x}_u, \mathbf{x}_v)q_1^*(\mathbf{x}_u, \mathbf{x}_v) + g_3(\mathbf{x}_u, \mathbf{x}_v)q_2^*(\mathbf{x}_u, \mathbf{x}_v)}. \end{aligned} \quad (7)$$

Thus, at first stage, node  $u$  is indifferent among all the pure strategies. We see that  $\langle p^*, (q_1^*, q_2^*) \rangle$  is also an NE at the first stage and hence a SPE of the entire dynamic game.  $\square$

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  $\langle p^*, (q_1^*, q_2^*) \rangle$ , 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:

$$\begin{aligned} \epsilon_1(\mathbf{x}_u, \mathbf{x}_v) & \triangleq \mathbb{P}(\hat{Y}(u, v) = 1 | Y(u, v) = 0) \\ & = 1 - q_2^*(\mathbf{x}_u, \mathbf{x}_v)^2, \\ \epsilon_2(\mathbf{x}_u, \mathbf{x}_v) & \triangleq \mathbb{P}(\hat{Y}(u, v) = 0 | Y(u, v) = 1) \\ & = 1 - q_1^*(\mathbf{x}_u, \mathbf{x}_v)^2. \end{aligned}$$

Thus, different utility functions lead to different link probabilities and link error probabilities. Specifically, for any link probability function  $H()$ , any fake link probability  $\epsilon_1$  and any missing link probability  $\epsilon_2$ , we can achieve them by setting the utility functions in the game model as follows:

$$f_1(\mathbf{x}_u, \mathbf{x}_v) = (1 + \epsilon_1 - \epsilon_2)\sqrt{H(\mathbf{x}_u, \mathbf{x}_v)}, \quad (8)$$

$$g_1(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \epsilon_2}, \quad g_3(\mathbf{x}_u, \mathbf{x}_v) = \sqrt{1 - \epsilon_1}, \quad (9)$$

$$f_2(\mathbf{x}_u, \mathbf{x}_v) = -f_1(\mathbf{x}_u, \mathbf{x}_v), \quad (10)$$

$$g_2(\mathbf{x}_u, \mathbf{x}_v) = -g_1(\mathbf{x}_u, \mathbf{x}_v), \quad (11)$$

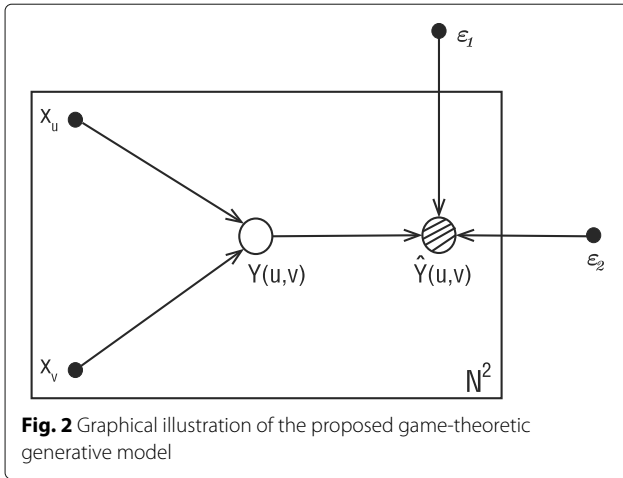
$$g_4(\mathbf{x}_u, \mathbf{x}_v) = -g_3(\mathbf{x}_u, \mathbf{x}_v). \quad (12)$$

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. In the game-theoretic model, each node pair links with each other with probability  $H(\mathbf{x}_u, \mathbf{x}_v)$  and then each link state  $Y(u, v)$  flips with probability  $\epsilon_1$  and  $\epsilon_2$ , producing the observed networks  $\hat{Y}(u, v)$ .

### 3 A general community detection algorithm for noisy networks

In this section, a community detection algorithm for the game-theoretic generative model is derived. 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. The game-theoretic model of noisy networks can be represented by three elements: the link probability function  $H(\mathbf{x}_u, \mathbf{x}_v)$ , the fake link probability  $\epsilon_1$ , and the missing link probability  $\epsilon_2$ , i.e., a triple  $\langle H(\mathbf{x}_u, \mathbf{x}_v), \epsilon_1, \epsilon_2 \rangle$ . We assume that the link error probabilities  $\epsilon_1$  and  $\epsilon_2$  are constants independent of the affiliation strength  $\mathbf{x}_u$ . The reason of this assumption is that the link error probabilities are related to the accuracy of the data acquisition technology, which is independent of the community structure of the networks.

A graphical representation of the proposed game-theoretic generative model for noisy networks is shown in Fig. 2. 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



Afterwards, noise is added in so that the link state  $Y(u, v)$  is flipped with fake link probability  $\epsilon_1$  and missing link probability  $\epsilon_2$  to generate the observed link state  $\hat{Y}(u, v)$ , i.e.,

$$\hat{Y}(u, v) \sim \text{Bernoulli} \left( \epsilon_1^{1-Y(u,v)} (1 - \epsilon_2)^{Y(u,v)} \right). \tag{14}$$

We assume that the link error probabilities  $\epsilon_1, \epsilon_2$  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.

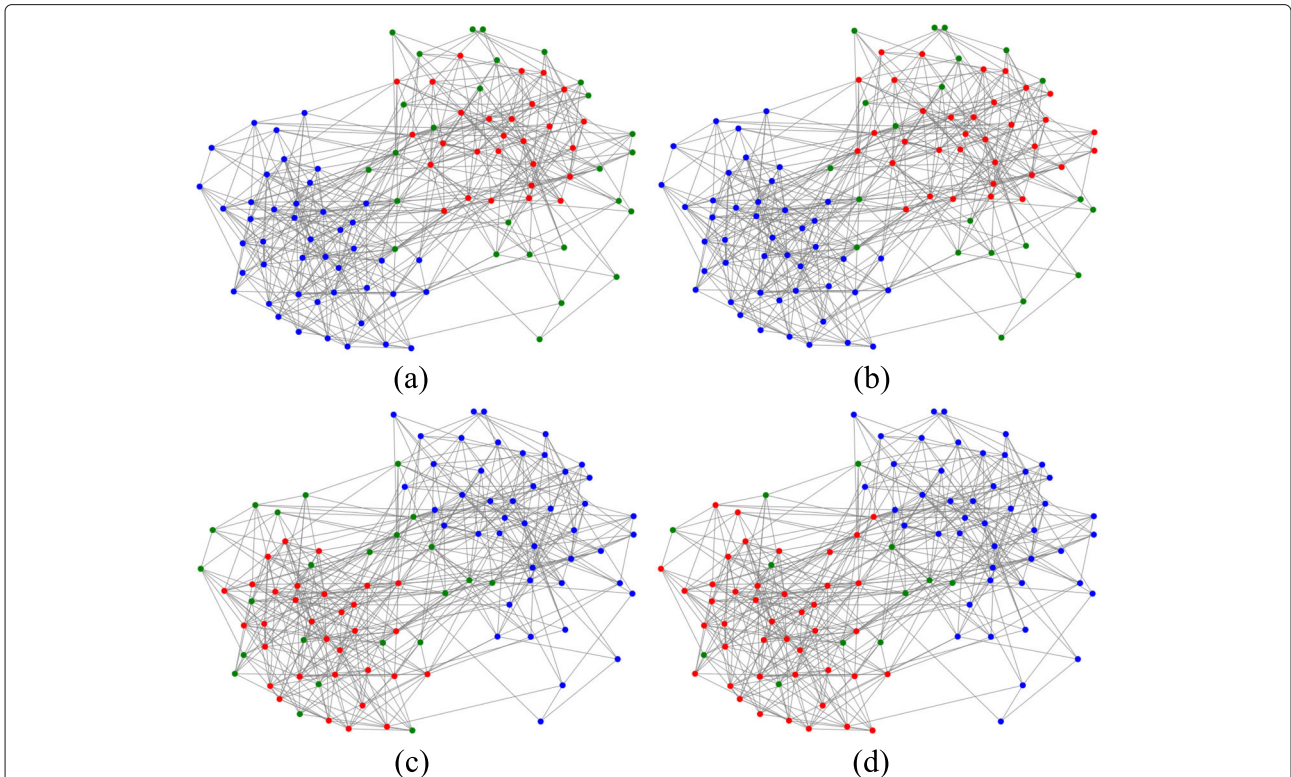
According to the generative model, the joint probability distribution function (PDF) of the true network  $\mathbf{Y} \triangleq \{Y(u, v)\}_{u,v=1, u < v}^N$  and the observed noisy network  $\hat{\mathbf{Y}} \triangleq \{\hat{Y}(u, v)\}_{u,v=1, u < v}^N$  is:

“1” (linking) or “0” (not linking), with linking probability  $H(\mathbf{x}_u, \mathbf{x}_v)$ , i.e.,

$$p(\mathbf{Y}, \hat{\mathbf{Y}} | \mathbf{X}) = \prod_{u < v} \left[ p(\hat{Y}(u, v) | Y(u, v)) p(Y(u, v) | \mathbf{x}_u, \mathbf{x}_v) \right], \tag{15}$$

$$Y(u, v) \sim \text{Bernoulli}(H(\mathbf{x}_u, \mathbf{x}_v)). \tag{13}$$

while the marginal PDF of the observation  $\hat{\mathbf{Y}}$ :



**Fig. 3** 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 [15], 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. **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

$$\begin{aligned}
& p(\hat{\mathbf{Y}}|\mathbf{X}) \\
&= \prod_{u < v} \sum_{Y(u,v) \in \{0,1\}} p(\hat{Y}(u,v)|Y(u,v)) p(Y(u,v)|\mathbf{x}_u, \mathbf{x}_v).
\end{aligned} \tag{16}$$

Hence, the maximum likelihood estimate (MLE) of the community affiliation strength parameter  $\mathbf{X}$  can be calculated as:

$$\mathbf{X}_{\text{ML}} = \arg \max_{\mathbf{X}} \log p(\hat{\mathbf{Y}}|\mathbf{X}). \tag{17}$$

However, due to the existence of the latent variables  $\mathbf{Y}$  (the true network), the maximization problem for the MLE is hard to solve: there is summation (marginalization) inside the logarithm, which cannot be operated directly onto the joint distribution. We thus resort to the expectation maximization (EM) algorithm [25], an efficient algorithm iterating between two steps, i.e., the expectation step (E-step) and the maximization step (M-step). Now, we proceed to derive an EM algorithm for the proposed generative model.

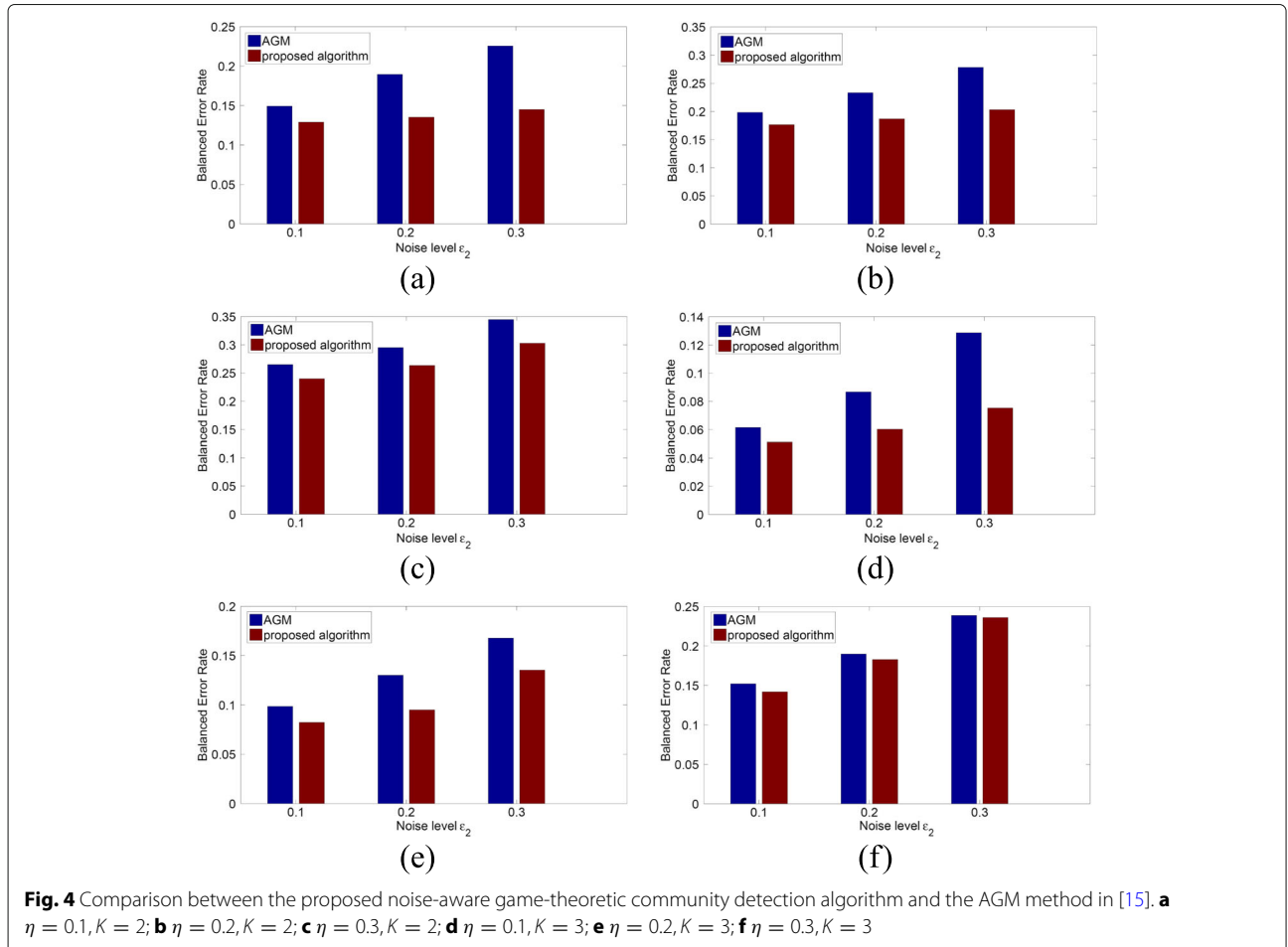
### 3.1 Derivation of the e-step

The joint PDF of the true link state  $Y(u, v)$  and the observed noisy link state  $\hat{Y}(u, v)$  is:

$$\begin{aligned}
& p(Y(u, v), \hat{Y}(u, v) | \mathbf{x}_u, \mathbf{x}_v) \\
&= p(Y(u, v) | \mathbf{x}_u, \mathbf{x}_v) p(\hat{Y}(u, v) | Y(u, v)) \\
&= \epsilon_1^{(1-Y(u,v))\hat{Y}(u,v)} (1 - \epsilon_1)^{(1-Y(u,v))(1-\hat{Y}(u,v))} \\
&\quad \epsilon_2^{Y(u,v)(1-\hat{Y}(u,v))} (1 - \epsilon_2)^{Y(u,v)\hat{Y}(u,v)} \\
&\quad H(\mathbf{x}_u, \mathbf{x}_v)^{Y(u,v)} (1 - H(\mathbf{x}_u, \mathbf{x}_v))^{1-Y(u,v)}
\end{aligned} \tag{18}$$

Suppose we have an estimate of the community affiliation strength matrix  $\mathbf{X}^{\text{old}}$  which we would like to update. Based on (18), the posterior distribution of the latent variable  $Y(u, v)$  is given as (19).

$$\begin{aligned}
q_{uv} &\triangleq p(Y(u, v) = 1 | \hat{Y}(u, v), \mathbf{x}_u^{\text{old}}, \mathbf{x}_v^{\text{old}}) \\
&= \frac{p(Y(u, v) = 1, \hat{Y}(u, v) | \mathbf{x}_u^{\text{old}}, \mathbf{x}_v^{\text{old}})}{p(Y(u, v) = 1, \hat{Y}(u, v) | \mathbf{x}_u^{\text{old}}, \mathbf{x}_v^{\text{old}}) + p(Y(u, v) = 0, \hat{Y}(u, v) | \mathbf{x}_u^{\text{old}}, \mathbf{x}_v^{\text{old}})}
\end{aligned} \tag{19}$$



Thus, we can derive the objective function in the M-step, i.e., the expected complete data log-likelihood, as follows:

$$\begin{aligned}
Q(\mathbf{X}, \mathbf{X}^{\text{old}}) &= \mathbb{E}_{\mathbf{Y}|\hat{\mathbf{Y}}, \mathbf{X}^{\text{old}}} \left[ \log p(\mathbf{Y}, \hat{\mathbf{Y}} | \mathbf{X}) \right] \\
&= \sum_{1 \leq u < v \leq N} \mathbb{E}_{\mathbf{Y}|\hat{\mathbf{Y}}, \mathbf{X}^{\text{old}}} \left[ \log p(Y(u, v), \hat{Y}(u, v) | \mathbf{x}_u, \mathbf{x}_v) \right] \\
&= \sum_{1 \leq u < v \leq N} \left\{ q_{uv} \left[ (1 - \hat{Y}(u, v)) \log \epsilon_2 + \hat{Y}(u, v) \log(1 - \epsilon_2) \right] \right. \\
&\quad \left. + \log H(\mathbf{x}_u, \mathbf{x}_v) \right\} + (1 - q_{uv}) \left[ \hat{Y}(u, v) \log \epsilon_1 \right. \\
&\quad \left. + (1 - \hat{Y}(u, v)) \log(1 - \epsilon_1) + \log(1 - H(\mathbf{x}_u, \mathbf{x}_v)) \right\}. \tag{20}
\end{aligned}$$

### 3.2 Derivation of the m-step

In the M-step, we maximize the expected complete-data log likelihood. In other words, we want to solve the following optimization problem:

$$\begin{aligned}
&\text{maximize } Q(\mathbf{X}, \mathbf{X}^{\text{old}}) \\
&\text{s.t. } \mathbf{X} \geq \mathbf{0}, \tag{21}
\end{aligned}$$

where the matrix inequality stands for componentwise inequalities. We note that only two terms in the objective function (20) depend on the optimization variable  $\mathbf{X}$ . So, the problem can be equivalently written as:

$$\begin{aligned}
&\text{maximize } J(\mathbf{X}) \triangleq \sum_{1 \leq u < v \leq N} \left[ q_{uv} \log H(\mathbf{x}_u, \mathbf{x}_v) \right. \\
&\quad \left. + (1 - q_{uv}) \log(1 - H(\mathbf{x}_u, \mathbf{x}_v)) \right] \tag{22} \\
&\text{s.t. } \mathbf{X} \geq \mathbf{0}.
\end{aligned}$$

The gradient of  $J$  with respect to  $\mathbf{x}_u$  is:

$$\nabla_{\mathbf{x}_u} J = \sum_{v \neq u} \frac{q_{uv} - H(\mathbf{x}_u, \mathbf{x}_v)}{H(\mathbf{x}_u, \mathbf{x}_v)(1 - H(\mathbf{x}_u, \mathbf{x}_v))} \nabla_{\mathbf{x}_u} H(\mathbf{x}_u, \mathbf{x}_v). \tag{23}$$

A projected coordinate ascent algorithm is utilized to solve the optimization problem (22). Each time we only optimize  $J$  with respect to one single vector  $\mathbf{x}_u$  using gradient ascent while keeping other vectors  $\mathbf{x}_v$  ( $v \neq u$ ) fixed. After each iteration, we project the updated  $\mathbf{x}_u$  onto the nonnegative orthant to meet the nonnegative constraint.

The EM iterations are known to converge to some locally maximum point of the likelihood function [25]. As such, we iterate between the E-step and the M-step until convergence. After the estimate of the community affiliation strength  $\mathbf{X}$  is obtained, a threshold is needed to decide the hard community affiliation, i.e., whether a node belongs to a community or not. Denote  $\mathbf{Z} \in \{0, 1\}^{K \times N}$  the community affiliation matrix, whose  $(k, n)$  entry is 1 if node  $n$  belongs to community  $k$ . Denote  $\mathbf{e}_1 \in \mathbb{R}^K$  the vector with first entry equal to 1 and remaining entries

equal to 0. One reasonable threshold is the solution  $t$  of the equation  $H(t\mathbf{e}_1, t\mathbf{e}_1) = \alpha$ , where  $\alpha$  is the background edge probability, i.e., the total number of links in the graph divided by the total number of links in a complete graph with  $N$  nodes. We judge that node  $u$  belongs to community  $k$ , i.e.,  $z(k, n) = 1$ , if  $\mathbf{x}_u(k)$  is larger than  $t$ . The overall community detection algorithm is summarized in Algorithm 1. From Algorithm 1, we can see that the computational complexity of the E-step and M-step is  $O(N^2)$ . We note that the proposed algorithm is general in the sense that we have not specified the concrete form of the link probability function  $H(\mathbf{x}_u, \mathbf{x}_v)$  yet. Several possible forms of the link probability function are listed as follows:

$$H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp \left\{ -(\mathbf{x}_u^T \mathbf{x}_v)^\theta \right\}, \theta > 0, \tag{24}$$

$$H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^T \mathbf{R} \mathbf{x}_v), \tag{25}$$

$$H(\mathbf{x}_u, \mathbf{x}_v) = \frac{\mathbf{x}_u^T \mathbf{x}_v}{1 + \mathbf{x}_u^T \mathbf{x}_v}, \tag{26}$$

where  $\mathbf{R}$  is some symmetric and nonnegative matrix. Both (24) and (26) are to detect cohesive communities, where intra-community link density is much higher than the inter-community one. The link probability function (25) can model more flexible community structures, e.g.,

---

**Algorithm 1** A general community detection algorithm for the proposed game-theoretic generative model for noisy networks

---

**Inputs:**

- The observed noisy network  $\hat{\mathbf{Y}} \in \{0, 1\}^{N \times N}$ .
- The number of communities  $K$ .
- The link probability function  $H()$ .

**Outputs:**

- The community affiliation matrix  $\mathbf{Z} \in \{0, 1\}^{K \times N}$

- 1: Initialize  $\mathbf{X}^{\text{old}} \in \mathbb{R}_+^{K \times N}$ .
  - 2: **repeat**
  - 3: E-step: Based on the old estimate  $\mathbf{X}^{\text{old}}$ , compute the posterior distributions  $q_{uv}$  according to (18) and (19),  $\forall 1 \leq u < v \leq N$ .
  - 4: M-step: Solve the optimization problem (22) using projected coordinate ascent to obtain the new estimate  $\mathbf{X}^{\text{new}}$ .
  - 5:  $\mathbf{X}^{\text{old}} \leftarrow \mathbf{X}^{\text{new}}$ .
  - 6: **until** convergence.
  - 7: Solve the equation  $H(t\mathbf{e}_1, t\mathbf{e}_1) = \alpha$  for the threshold  $t$ .
  - 8:  $\forall 1 \leq k \leq K, 1 \leq n \leq N$ , set  $z(k, n) = 1$ , if  $\mathbf{x}_u(k) > t$ , and 0 otherwise.
-



intra-community link density is lower than the inter-community one or some communities link with each other more often while some not. The shortcoming of (25) is we need a priori knowledge about the community structure in order to determine the structure of the matrix  $\mathbf{R}$  in (25).

#### 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.

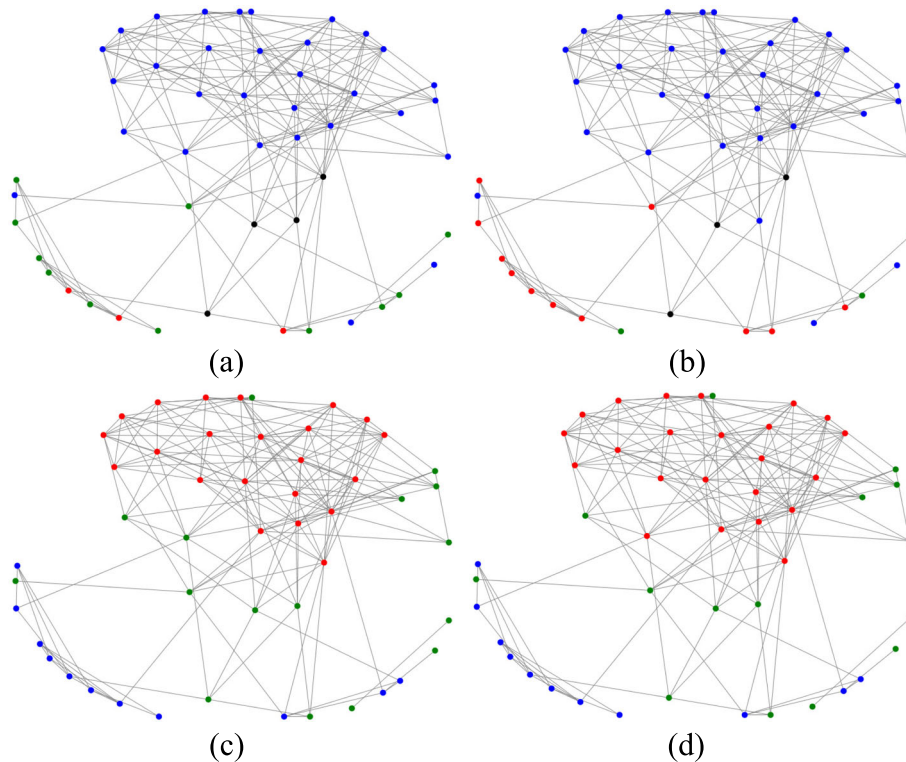
##### 4.1 Simulations

To implement simulations, we synthesize networks with  $N$  nodes and  $K$  communities according to the following procedure:

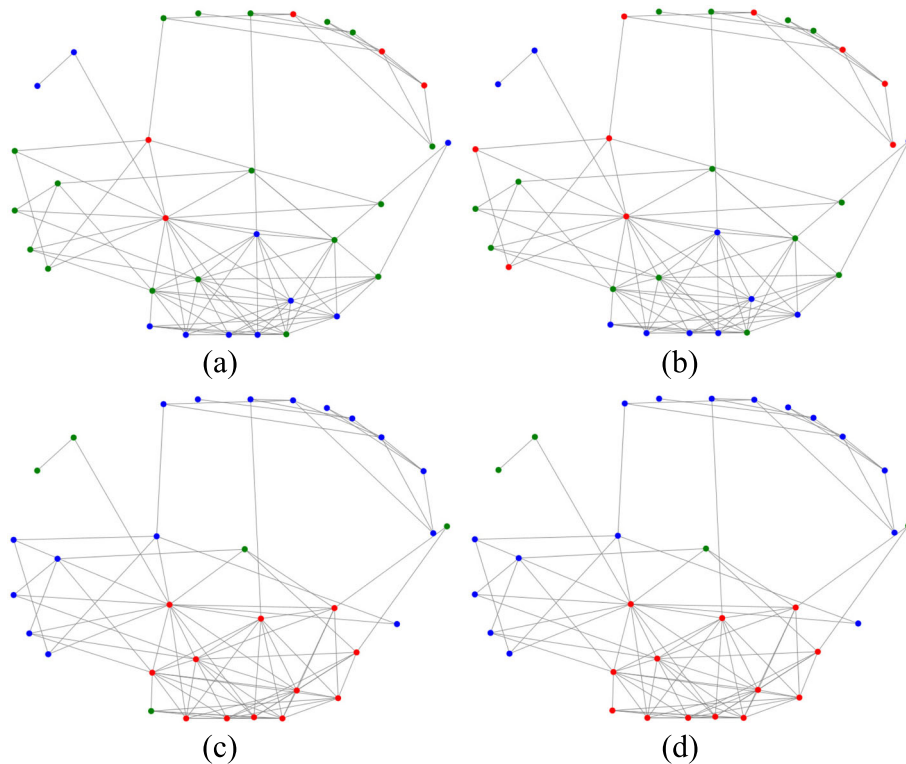
1. Partition all nodes into  $K$  non-overlapping equal groups of nodes so that each group has  $N/K$  nodes.
2. For each group, randomly pick  $\eta N/K$  nodes outside of the group and add these nodes into the group, where  $0 < \eta < 1$  is a user-defined parameter.

3. 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.
4. Generate the links according to the chosen link probability function  $H(\mathbf{x}_u, \mathbf{x}_v)$ .
5. Add noise into the network according to the link error probabilities  $\epsilon_1, \epsilon_2$ .

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  $N = 100$ ,  $K = 2, 3$ ,  $\eta = 0.1, 0.2, 0.3$ . 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  $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^T \mathbf{x}_v)$  and compare the performance with that of the affiliated graph model (AGM) proposed in [15]. A visualization of the community detection results of the proposed method and AGM, a state-of-the-art community



**Fig. 5** Facebook ego-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 [15], respectively. The nodes' colors have the same meaning as in Fig. 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. 6** DBLP 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 [15], respectively. The nodes' colors have the same meaning as in Fig. 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

detection algorithm with brilliant performance, for a synthetic network is presented in Fig. 3. 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:

$$\text{BER}(\mathcal{C}, \bar{\mathcal{C}}) = \frac{1}{2} \left( \frac{|\mathcal{C} \setminus \bar{\mathcal{C}}|}{|\mathcal{C}|} + \frac{|\bar{\mathcal{C}} \setminus \mathcal{C}|}{|\bar{\mathcal{C}}|} \right). \quad (27)$$

For every detected community  $\mathcal{C}$ , we calculate  $\min_{\bar{\mathcal{C}}} \text{BER}(\mathcal{C}, \bar{\mathcal{C}})$ . For every ground-truth community  $\bar{\mathcal{C}}$ , we calculate  $\min_{\mathcal{C}} \text{BER}(\mathcal{C}, \bar{\mathcal{C}})$ . 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. 4, where we compare the proposed noise-aware game-theoretic algorithm with the AGM in [15]. We find that the proposed algorithm always outperforms the

AGM, and the performance enhancement increases with the noise level  $\epsilon_2$  (except for networks in Fig. 4f).

#### 4.2 Real-data experiments

For real-data experiments, we consider two datasets: the Facebook ego-networks dataset [26] and the DBLP collaboration network dataset [27]. Both networks have well-defined ground-truth communities. The detailed statistics about the datasets are listed as follows.

- Facebook ego-networks: number of nodes = 4039, number of edges = 88234. Each node is a Facebook user. Two users are linked if they are Facebook friends. The ground-truth communities are identified by humans manually.

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

Dataset	Noise level $\epsilon_2$	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%

**Table 4** Relative improvement of the community detection accuracy with different link probability functions  $H(\mathbf{x}_u, \mathbf{x}_v)$

Dataset	Link Prob. function	$1 - \exp\{-\mathbf{x}_u^\top \mathbf{x}_v\}^{1.2}$	$\mathbf{x}_u^\top \mathbf{x}_v / (1 + \mathbf{x}_u^\top \mathbf{x}_v)$	$1 - \exp(-\mathbf{x}_u^\top \mathbf{R} \mathbf{x}_v)$
Facebook ego-network dataset		0.42%	0.49%	- 13.69%
DBLP dataset		2.13%	- 1.31%	- 35.34%

- DBLP collaboration network: number of nodes = 31708, number of edges = 1049866. Each node is an author. Two authors are linked if they have co-authored at least one paper together. The publication venue defines the ground-truth communities.

To control the size of the input network to the community detection algorithm, we sample the original network to obtain smaller subnetworks, on which we perform the community detection [15]. Specifically, we randomly select one node belonging to at least two communities and the subnetwork consists of all nodes with at least one common community with the selected node. Furthermore, we add noise onto the networks with  $\epsilon_1 = 0.005, \epsilon_2 = 0.1, 0.2, 0.3, 0.4$ . For link probability function, we still choose  $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^\top \mathbf{x}_v)$ . Visualizations of the detection results of a Facebook ego-network and a DBLP network are shown in Figs. 5 and 6, respectively. Both networks have two communities. Similar to the synthetic network in Fig. 3, we remark that the proposed approach still outperforms AGM especially for the community 1 in Facebook and DBLP networks. 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  $\epsilon_2$ .

We further investigate the impact of the selection of link probability function  $H(\mathbf{x}_u, \mathbf{x}_v)$  on the performance. The performance of the function  $H(\mathbf{x}_u, \mathbf{x}_v) = 1 - \exp(-\mathbf{x}_u^\top \mathbf{x}_v)$  serves as a benchmark. Additionally, we select three different link probability functions and study their relative community detection accuracy improvements on the Facebook and DBLP datasets. The results are shown in Table 4, where the matrix  $\mathbf{R}$  in the last function is the block diagonal matrix  $\mathbf{R} = \text{diag}(\mathbf{R}_0, \mathbf{R}_0, \dots, \mathbf{R}_0)$ , with

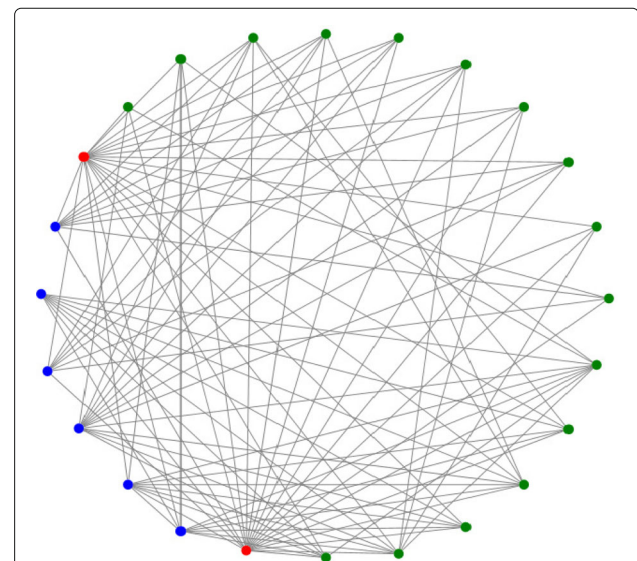
$$\mathbf{R}_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{28}$$

We note this link probability function is suitable for detecting community structure with inter-community links denser than intra-community links. The results indicate that different choices of link probability function lead to different performances and the performance variations depend on the datasets. Specifically, the performance degradation of using the link probability function (25) suggests that in Facebook and DBLP dataset, the

intra-community links are denser than inter-community links. We also study the Chesapeake and Florida Bay foodweb network [28], where the inter-community links are denser than intra-community links (there are lots of links between a group of predators and the corresponding prey group). Thus, we utilize the link probability function in (25), where the matrix  $\mathbf{R}$  is set to be  $\mathbf{R}_0$ . Then, the link density between the two detected communities, i.e., the ratio between the number of links and the number of possible links in a complete graph, is 0.590, while that of the entire graph is only 0.223. The two detected communities are depicted in Fig. 7. We observe that there are lots of links between the two communities while only few links exist within each community. So, the detected community structure correctly characterizes the predator-prey relationship in the network.

### 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



**Fig. 7** The detected two communities in the Chesapeake and Florida Bay foodweb network. Blue nodes and green nodes represent two communities, respectively. The red nodes correspond to the intersection of the two communities

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. We hope that this paper can open a new direction to look at the community detection problem from the microeconomic perspective.

#### Abbreviations

AGM: Affiliated graph model; BER: Balanced error rate; EM: Expectation maximization; NE: Nash equilibrium; SPE: Subgame perfect equilibrium

#### Acknowledgements

Not applicable.

#### Authors' contributions

All authors equally contribute to the problem formulation, theoretic derivation, simulations, and experiments, as well as paper writing. All authors read and approved the final manuscript.

#### Funding

This paper is supported by the National Key Research and Development Program of China (2017YFB1400100).

#### Availability of data and materials

Not applicable.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. <sup>2</sup>Coordinated Science Lab, University of Illinois, Urbana-Champaign, Urbana 61801, IL, USA. <sup>3</sup>Department of Electrical and Computer Engineering, University of Maryland, College Park 20742, MD, USA.

Received: 19 September 2018 Accepted: 22 November 2019

Published online: 16 December 2019

#### References

- X. Cao, Y. Chen, K. J. R. Liu, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Community detection game (IEEE, 2016), pp. 6220–6224
- S. Fortunato, Community detection in graphs. *Phys. Rep.* **486**(3), 75–174 (2010)
- S. Fortunato, D. Hric, Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44 (2016)
- E. Abbe, Community detection and stochastic block models: recent developments. arXiv preprint arXiv:1703.10146 (2017)
- T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **27**(2), 83–85 (2005)
- J. MacQueen, et al., in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Some methods for classification and analysis of multivariate observations, vol. 1, (Oakland, 1967), pp. 281–297
- W. E. Donath, A. J. Hoffman, Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.* **17**(5), 420–425 (1973)
- M. E. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E.* **69**(2), 026113 (2004)
- M. E. Newman, Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**(23), 8577–8582 (2006)
- G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature.* **435**(7043), 814–818 (2005)
- B. E. Hajek, Y. Wu, J. Xu, in *COLT*. Computational lower bounds for community detection on random graphs, (2015), pp. 899–928
- E. Abbe, C. Sandon, in *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*. Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery (IEEE, 2015), pp. 670–688
- C. Bordenave, M. Lelarge, L. Massoulié, in *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs (IEEE, 2015), pp. 1347–1357
- E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing, in *Advances in Neural Information Processing Systems*. Mixed membership stochastic block models, (2009), pp. 33–40
- J. Yang, J. Leskovec, in *IEEE 12th International Conference on Data Mining (ICDM)*. Community-affiliation graph model for overlapping network community detection (IEEE, 2012), pp. 1170–1175
- Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, X. Yu, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks (ACM, 2012), pp. 1348–1356
- B. Wang, Y. Wu, K. R. Liu, Game theory for cognitive radio networks: an overview. *Comput. Netw.* **54**(14), 2537–2561 (2010)
- C. Jiang, Y. Chen, K. Liu, Y. Ren, Network economics in cognitive networks. *IEEE Commun. Mag.* **53**(5), 75–81 (2015)
- Z. Han, Z. J. Ji, K. Liu, Fair multiuser channel allocation for OFDMA networks using NASH bargaining solutions and coalitions. *IEEE Trans. Commun.* **53**(8), 1366–1376 (2005)
- C. Jiang, Y. Chen, K. R. Liu, Graphical evolutionary game for information diffusion over social networks. *IEEE J. Sel. Top. Signal Process.* **8**(4), 524–536 (2014)
- C. Jiang, Y. Chen, K. R. Liu, Distributed adaptive networks: a graphical evolutionary game-theoretic view. *IEEE Trans. Signal Process.* **61**(22), 5675–5688 (2013)
- Y. Yuan, A. Alabdulkareem, A. S. Pentland, An interpretable approach for social network formation among heterogeneous agents. *Nat. Commun.* **9**(4074) (2018)
- S. Tadelis, *Game Theory: an Introduction*. (Princeton University Press, 2013)
- J. Yang, J. Leskovec, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. Overlapping community detection at scale: a nonnegative matrix factorization approach (ACM, 2013), pp. 587–596
- C. M. Bishop, et al., *Pattern Recognition and Machine Learning*. (Springer, New York, 2006)
- J. J. McAuley, J. Leskovec, in *Neural Information Processing Systems*. Learning to discover social circles in ego networks, (2012)
- J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **42**(1), 181–213 (2015)
- R. E. Ulanowicz, D. L. DeAngelis, Network analysis of trophic dynamics in south florida ecosystems. *US Geol. Surv. Program S. Fla. Ecosyst.* **114** (2005)

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.