Physica A 392 (2013) 1287–1301

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Phase transition model for community detection

Jianshe Wu [a,*], Rui Lu [a], Licheng Jiao [a], Fang Liu [a], Xin Yu [a], Da Wang [a], Bo Sun [b]

[a] *Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an 710071, China*

[b] *ZTE Corporation, China*

ARTICLE INFO

Article history:
Received 20 February 2012
Received in revised form 18 November 2012
Available online 28 November 2012

Keywords:
Complex network
Phase transition
Clustering dynamics
Community detection

ABSTRACT

Motivated by social and biological interactions, a novel type of phase transition model is provided in order to investigate the emergence of the clustering phenomenon in networks. The model has two types of interactions: one is attractive and the other is repulsive. In each iteration, the phase of a node (or an agent) moves toward the average phase of its neighbors and moves away from the average phase of its non-neighbors. The velocities of the two types of phase transition are controlled by two parameters, respectively. It is found that the phase transition phenomenon is closely related to the topological structure of the underlying network, and thus can be applied to identify its communities and overlapping groups. By giving each node of the network a randomly generated initial phase and updating these phases by the phase transition model until they reach stability, one or two communities will be detected according to the nodes' stable phases, confusable nodes are moved into a set named O_f. By removing the detected communities and the nodes in O_f, another one or two communities will be detected by an iteration of the algorithm, ... In this way, all communities and the overlapping nodes are detected. Simulations on both real-world networks and the LFR benchmark graphs have verified the efficiency of the proposed scheme.

1. Introduction

Many real-world networks often feature an organization of communities, with dense connections within each community and only sparse connections between them [1,2], e.g. coauthor network [3], metabolic network [4,5], protein network [6], etc. Different communities usually correspond to different functionalities of the network and the overlapping nodes of two communities usually act as their bridge [7], thus detecting the communities has great significance in understanding both the organization and function of the network [8]. Many research works have been done on community detection in recent years [9–19]. Among them, the most popular technique was proposed by Newman and Girvan, which is to maximize the measure function called modularity of the network [9], and a variety of algorithms have been established based on the modularity (see Refs. [10–12], just to name a few). However, some authors observed that modularity maximization met the resolution limit problem [20–22]. Comparative analysis of community detection algorithms can be found in Refs. [23,24], comparative analysis of spectral methods for community detection can be found in Ref. [19], and a detailed review of the algorithms was recently given in Ref. [25].

As for the detection of overlapping communities, by giving a precise definition of community in mathematics, named a k-clique community, Palla et al. provide a popular method known as clique percolation by locating all cliques and constructing the clique–clique overlap matrix [8,26]. The algorithm is exponential but it can be improved to polynomial.

* Corresponding author. Tel.: +86 29 88202279; fax: +86 29 88201023.
E-mail address: jshwu@mail.xidian.edu.cn (J. Wu).

0378-4371/$ – see front matter
doi:10.1016/j.physa.2012.11.032

Reichardt and Bornholdt present a q-state Potts model for community detection, in which different local minima of the Hamiltonian correspond to different community structures [27]. By employing a simple Monte Carlo single spin flip heat-bath algorithm, the ground state of the system can be approximated. With a set of several runs of the algorithm and collecting the number of coappearances of nodes in each community, the overlapping community structure of the network can be revealed. The proposed algorithm is fast but it requires many runs accordingly. Zhang et al. define a generalized modularity function and utilize the fuzzy c-means algorithm for overlapping community detection by embedding the nodes into an n-dimensional space using spectral mapping [28], but the eigenvector calculation made it expensive for computation complexity. To avoid the need for spectral embedding, Nepusz et al. proposed a new goal function for fuzzy community detection based on vertex similarities [29]. Both the algorithms in Refs. [28,29] need to know the number of communities, which is usually a difficult problem though a method is given to evaluate it in Ref. [29].

Roughly speaking, there are two types of algorithms for the problem of community detection. One type uses only topological information of the network, namely, information of the nodes and connections. Except for the topological information of the network, it has been shown that the dynamical or functional information can also be applied to reveal the modular structure [30–35]. Using only topological information, it is found that the overlapping nodes, which lie on the border between two communities, are difficult to be detected [36,37]. Cheng and Shen have investigated the diffusion dynamics of network and demonstrated that the community structure can be revealed through optimization of the conductance of the network [38]. By using dynamical information, the overlapping nodes can be easily detected because their dynamical behaviors are quite different [39,40]. Sendiña-Nadal et al. extended the methods to detect overlapping nodes in Refs. [39,40] to be able to apply even if the information about the adjacency matrix of the network is incomplete or even wrong [41]. In Ref. [42], by introducing a negative coupling strength, the Kuramoto phase oscillator model is modified to be able to detect both the communities and their overlapping groups. Very recently, Anderson et al. studied the multiscale dynamics in an oscillator network of several communities; the coupling between oscillators in the same community is attractive, which promotes synchronization; whereas the coupling between oscillators in different communities is repulsive, which repels synchronization [43].

The common point of the research in Refs. [30–34,39–43] is that they all apply the Kuramoto phase oscillator to represent the dynamics of the nodes, which runs at arbitrary intrinsic frequencies, and couples through the sine of their phase differences. But intuitively, it is not so suitable to treat the persons in a social network or the animals in a biological system as oscillators; it is also not so suitable to treat their interactions as sine functions of their phase differences. Suppose in a social network, if two persons can enhance their friendship after an interaction, they may exist in the same community; otherwise, if their friendship is weakened or their animosity is enhanced after an interaction, they should be in different communities. This phenomenon indicates that the community organization is formed by the dynamical interactions (attractive or repulsive) between nodes other than the dynamics of the nodes themselves. Inspired by this observation, in this paper, a novel discrete-time phase transition model for community detection is provided, which includes two types of interactions for each node. The effect of one type of interactions is to make a node's phase approach the average phase of its neighbors and the other type is to move apart from the average of its non-neighbors. Clustering phenomena of the phase transition model on modular networks are observed, which can be applied to detect both the communities and their overlapping groups. The proposed scheme is verified on both computer-generated and real-world networks.

Compared with previous methods, the greatest advantage of our scheme is that two (maybe one) of the communities will be detected in an iteration of the algorithm in general. Then as the iteration of the algorithm, the problem size becomes smaller and smaller until all communities are detected.

Vicsek et al. have presented a famous phase transition model in Ref. [44] and its dynamics has been carefully studied (see Refs. [45,46], just to name a few). In short, the differences between the Vicsek model and that in this paper are as follows. (1) The topology of the Vicsek model is switching, whereas the topology of the network model in this paper is not switching, thus the modular structure of a network can be revealed by our model. (2) There is only one type of interaction in the Vicsek model, which drives the phase of a node to be equal to the average of its neighbors [45]. There are two types of interactions in our model, and the effects of them are different: one is to make the phase of a node approach the average phase of its neighbors and the other is to make the phase of a node move away from the average phase of its non-neighbors. Thus the network of the Vicsek model can reach a consensus if the union of the topologies is connected, whereas the network of our model can realize clustering on modular structure. (3) The velocities of the phase transition are controlled by two parameters in our model (see Section 2.1), whereas there is no velocity control in the Vicsek model.

The rest of this paper is organized as follows. In Section 2, the phase transition model is introduced and its dynamics is also analyzed. Based on the phase transition model, the algorithm to detect communities and their overlapping groups is introduced in Section 3. Simulations are provided in Section 4. Finally, Section 5 concludes the paper.

2. Phase transition model

2.1. The model

A network is usually described by an undirected graph $G(V, E)$ with V representing the set of the nodes and E the edges. For each node $i \in V(i = 1, \ldots, N)$, let $S_i(i = 1, \ldots, N)$ be the set of its neighbors, that is, $j \in S_i$ iff $(i, j) \in E$. The set of non-neighbors of i is denoted as $\overline{S}_i(i = 1, \ldots, N)$, that is, $j \in \overline{S}_i$ iff $(i, j) \notin E$. Thus, we have $|S_i| + |\overline{S}_i| + 1 = N$, where $|S_i|$