MASTER'S THESIS

Discovering communities by information diffusion and link density propagation

Chen, Weidong

Date of Award:
2012

Discovering Communities by Information Diffusion and Link Density Propagation

CHEN Weidong

A thesis submitted in partial fulfillment of the requirements

for the degree of

Master of Philosophy

Principal Supervisor : Prof. Jiming Liu

Hong Kong Baptist University

November 2011

Abstract

Community structure is one of the ubiquitous topological characteristics in networks, where nodes of a community are more densely connected to nodes inside than the rest of the network. Those dense subgraphs can represent a set of functional units, which provide insights into the analysis of interactions between units. In the last decade, a large variety of algorithms have been developed to solve the community detection problem. However, most existing algorithms require either global information or iteratively add one node to a community at a time. Those algorithms are infeasible for large-scale real networks due to their high computational or memory complexity.

Community structure can emerge through the self-organizing process of community entities. This is, each entity updates and discovers its community membership in a decentralized fashion. This observation allows us not only to understand the emergence of community structure, but also enables us to process large-scale distributed real networks. There are no much studies on decentralized computing and self-organized computing for community detection. To the best of our knowledge, there are two categories of truly decentralized algorithms for the community detection problem: multi-agent-based approaches [13] [14] and label propagation-based approaches [64]. In this thesis, we focus on developing decentralized community detection algorithms which will be applicable to large-scale distributed networks. For simplicity, we only consider community detection for unweighted and undirected networks. The two proposed algorithms are briefly summarized as follows.

The first algorithm is based on the concept of information diffusion in social networks. The algorithm has differences from traditional diffusion models. The algorithm is based on the structural similarity of nodes and simulates how people form a group by exchanging information with their neighbors. The emerging "dominant" information of nodes can be used to determine node community memberships. The information diffusion-based method is suitable for discovering "critical" nodes with respect to community structure. An influential node and its neighbors form a clique-like structure. The influence maximization problem is to find a fixed

number of influential nodes to maximize influence in the propagation. Unlike the influence maximization problem, we do not restrict the number of the most influential nodes. This provides some insights on influential nodes with respect to community structure.

The second algorithm discovers communities by the propagation of link density. In contrast, node membership assignments are determined in terms of link membership assignments. The algorithm does not rely heavily on any parameters, and it performs locally and asynchronously. Link clustering-based methods are less sensitive to noise than node clustering-based methods, meanwhile they are able to find overlapping communities naturally. This is, a node may belong to multiple communities, whereas a link can be assigned to only one community [62] [76]. We believe that link density reflects the natural "cores" in a community. We introduce the concept of disturbance, which can be described by an aggregation process of links. "Cores" in a community will be resilient to disturbances from the rest of the network.

Processing large dynamically evolving networks requires to handle new data or incremental changes over time. The link density propagation-based method is suitable for mining dynamic networks. What is more, the algorithm does not rely heavily on parameters. It is more flexible to be further extended to different applications.

Experiments on various networks, including real-world networks and synthetic networks, show that the performance of the two proposed algorithms are comparable to the state-of-the-art decentralized and centralized algorithms, in terms of their accuracy and stability.

Table of Contents