

MASTER'S THESIS

Discovering communities by information diffusion and link density propagation

Chen, Weidong

Date of Award:
2012

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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

Declaration	i
Abstract	ii
Acknowledgements	iv
List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Community Structure and Community Detection	2
1.2 Motivation and Contribution	5
1.3 Problem Statement	7
1.4 Thesis Outline	8
2 Related Work	9
2.1 Multi-Agent-Based Approaches	9
2.2 Label Propagation-Based Approaches	11
2.3 Two Centralized Approaches	13
2.4 Link Clustering-Based Approaches	14
3 Discovering Communities by Information Diffusion	16
3.1 Node Structure and Local Environment	18
3.1.1 Structure of Nodes	19
3.1.2 Structure-Based Similarity Measures	20
3.1.3 Interactions and Positive Feedback	22

3.2	Description of Behaviors	23
3.2.1	Information Pull	25
3.2.2	Transmission Probability Update	26
3.2.3	Community Membership Evaluation	26
3.2.4	Community Membership Reselection	27
3.3	Stopping Criterion	28
3.3.1	Oscillation	28
3.3.2	Stopping Criterion	29
3.4	Summary of the Information Diffusion-Based Method	30
3.5	Experiments	31
3.5.1	Benchmark Networks	33
3.5.1.1	Zachary Karate Club Network	33
3.5.1.2	American College Football Network	34
3.5.1.3	Bottleneck Dolphin Network	34
3.5.2	Synthetic Networks	36
3.5.2.1	GN Synthetic Networks	36
3.5.2.2	LFR Synthetic Networks	39
3.5.2.3	Erdős-Rényi Random Networks	40
3.5.3	Real-World Networks with Large Sizes	40
3.5.4	Time Complexity Analysis	44
3.6	Discussions	45
4	Discovering Communities by Link Density Propagation	46
4.1	Description of the Algorithm	47
4.1.1	Structure-Based Link Similarity Measure	47
4.1.2	Link Density-Based Propagation	48
4.1.3	Community Membership Evaluation	49
4.1.4	Boundary Refinement	49
4.2	Experiments	51
4.2.1	Benchmark Networks and Real-World Networks	51
4.2.2	Synthetic Networks	56
4.2.2.1	GN Synthetic Networks	56

4.2.2.2	LFR Synthetic Networks	56
4.3	Incremental LinkDP Method for Mining Dynamic Networks	58
4.3.1	Problem Definition	59
4.3.2	Incremental LinkDP Method	59
4.4	Discussions	61
5	Conclusion and Future Work	62
5.1	Conclusion	62
5.2	Future Work	63
5.2.1	Information Diffusion-Based Method	64
5.2.2	Link Density Propagation-Based Method	64
	Bibliography	66
	Curriculum Vitae	74