THE EXISTENCE AND DISCOVERY OF OVERLAPPING COMMUNITIES IN LARGE-SCALE NETWORKS

By

Stephen Kelley

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Mark Goldberg, Thesis Adviser
Malik Magdon-Ismail, Member
William Wallace, Member
Mohammed Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

November 2009
(For Graduation December 2009)

ABSTRACT

The identification of communities, also known as clusters, modules, and coalitions, has long been an important part of any network analysis. Accurate groupings can offer unique insight into large, complex systems which defy manual comprehension. As such there is constant development in the field of community detection.

Traditionally, methods used to identify communities have produced partitionings of the vertex set of networks being studied. Such partitionings produce groups which often attempt to maximize some global measure such as *modularity*. Recently, researchers have begun to develop methods which produce non-disjoint groups, allowing vertices to be members of one, zero, or many communities at the same time. However, despite what appears to be a gradual shift toward formulating methods which account for community overlap, there is a general lack of consensus as to what should formally qualify as a community.

In this thesis, an axiomatic definition of a community is given. The axioms given are minimal; they enumerate intuitive traits which all all reasonable communities should have. However, surveying the landscape of overlapping community detection, current methods seem to fall short of even these simple axioms. As such, it becomes necessary to formulate new methods with these criteria in mind.

Connected Iterative Scan is a local optimization algorithm which has been developed to satisfy the axiomatic definition of a community. As a result of its local nature, it is a "group-centric" approach; it processes groups independently and attempts to construct each group such that its quality is maximized. The algorithm is presented relative to various disjoint and overlapping benchmarks and performs well. In addition, various parameters of the algorithm are fully tested, giving potential users a set of observations which can aid in fine-tuning the parameters relative to a specific network or group structure.

This thesis also attempts to show that allowing groups to overlap is a natural and essential part of social network analysis. In previous literature, much of this justification is limited to intuition or small, toy graphs. Because of this, the amount

of overlap within real networks has been largely unexplored. Since allowing groups to overlap greatly increases the number of possible communities and possibly the computational requirements of discovery methods, the lack of a quantification of the significance of overlap in various networks has led many researchers to continue usage of disjoint grouping methods. This text examines a large social network and attempts to quantify the significance of the group overlap within. The results show that usage of a disjoint method will fail to capture many of the associations within the data, clearly demonstrating the need for methods which account for overlap.