ALGORITHMS FOR FINDING HIDDEN GROUPS AND THEIR STRUCTURE FROM THE STREAMING COMMUNICATION DATA

By

Mykola Hayvanovych

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:

   Malik Magdon-Ismail, Thesis Adviser
   Mark K. Goldberg, Member
   William A. Wallace, Member
   Mohammed J. Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

December 2009
(For Graduation December 2009)

ABSTRACT

A *planning hidden group* is a set of individuals planning an activity over a communication medium without announcing their existence. In order to plan, hidden groups need to communicate regularly, possibly in a streaming manner (*streaming hidden groups*). The hidden group's communication patterns exhibit structure, which differentiates these communications from random background communications. Here we propose efficient algorithms for identifying streaming hidden group structure by isolating the hidden group's non-random, planning-related communications from the random background communications. We validate our algorithms on real data (the Enron email corpus and Blog communication data). Analysis of the results reveals that our algorithms extract meaningful hidden group structures.

We also present a software system SIGHTS (Statistical Identification of Groups Hidden in Time and Space), which employs the hidden group algorithms and can be used for the discovery, analysis, and knowledge visualization of social coalitions in communication networks such as Blog-networks. The evolution of social groups reflects information flow and social dynamics in social networks. The goal of SIGHTS is to be an assistant to an analyst in identifying relevant information.

One of the uses of group detection algorithms is to monitor group dynamics. We develop algorithms to measure similarity between clusterings (sets of sets) which we use to quantify the rate of group evolution. We apply these comparison algorithms to the groups discovered by our algorithms.

Trust is an important aspect of groups, and we extend our algorithms to develop two measures of trust which can be used to analyze the "behavioral" trust relationship between people in a social network. We use Twitter network communications for our experimentation and validation of our proposed measures.

All the work in this thesis is based on purely statistical analysis of the data, not requiring semantic analysis. This is especially useful in social networks because the volume of information makes semantic analysis intractable. Further, it means that our algorithms are language independent.