ALGORITHMS FOR DISCOVERING HIDDEN GROUPS IN COMMUNICATIONS

By

Jeffrey Baumes

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:

Dr. Mark Goldberg, Thesis Adviser
Dr. Malik Magdon-Ismail, Member
Dr. William (Al) Wallace, Member
Dr. Mohammed Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

August 25, 2006
(For Graduation December 2006)

ABSTRACT

Technologies such as email, chatrooms and web logs allow individuals to communicate in a number of new ways, and new forms of communications are continually appearing. As vast amounts of digital communication data are collected, communication logs may potentially contain millions of pieces of information, which must be pre-processed in order to present an intelligence analyst with a set of communication groups. We call these groups “hidden groups” since their membership is not explicit, but is implicitly implied in their communication structure. We present a strategy for discovering hidden groups within a communication network; through searching for *significant correlations*, which may be temporal or structural. The temporal algorithms discover hidden groups connected over any range of time in the dataset. We find that a temporal hidden group must choose between two options: trust individuals outside the group to relay information, or else impose a structure upon themselves which is more easily detected by our algorithms. For structural hidden groups, we propose a novel approach to the problem of graph clustering, where we allow clusters to overlap by letting them extend to local optima with respect to a generic density metric. We compare and contrast different algorithms for finding clusters in this way, and show that a particular algorithm combination, LA→IS, performs well on both simulated and real-world data. The software system SIGHTS incorporates these algorithms into a single framework, along with associated visualizations.