

© 2011 Wei Dong

MUTUAL INFORMATION: INFERRING TIE STRENGTH AND PROXIMITY IN
BIPARTITE SOCIAL NETWORK DATA WITH NON-METRIC ASSOCIATIONS

BY

WEI DONG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Human Factors
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Advisor:

Professor Alex Kirlik

ABSTRACT

The current study was a first step exploration of a new method that used mutual information-based measures to represent tie strength or proximity between individuals from bipartite social network data with non-metric associations. Unlike network datasets with explicit links between nodes, bipartite networks provide only implicit indications of the probable existence of connections. Therefore, as a measure of the amount of information shared between these two random variables, mutual information can be used to infer social network structure in bipartite network data. A literature review found surprisingly low utilization of mutual information in social network analysis, although it was widely used in other areas of network analysis. Two studies in the current thesis showed that mutual information can be effectively used to infer tie strength and proximity from bipartite social network data with non-metric associations. Other social network analysis techniques such as graph theory-based centrality measures and hierarchical cluster analysis can then be applied to the mutual information-based measures to further investigate the underlying social network structure such as detecting members of subgroups and detecting important nodes that centered the network. Advantages and potential disadvantages of using mutual information-based measures in social network analysis and future directions in ways of improving this method were discussed.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Alex Kirlik, for his remarkable thoughtfulness, support, and advice in my research in the area of Human Factors as well as in my process of personal growth through self-reflection. It was reassuring to know that I could rely on such a brilliant, insightful and gifted mentor. This thesis would not have been possible without the instruction of Professor Alex Kirlik.

Thank you to Professor Wai-Tat Fu. I was fortunate to have him as my co-advisor, also as an incredible teacher and resource. He was exceedingly generous with his time and his thoughts, and provided a tremendous level of support that helped me transit smoothly from one discipline to another.

I would like to express special thanks to my fellow lab group members. Thank you to Sven Bertel, who provided a refreshing perspective on my data analysis and helped me on gathering potential datasets. Thank you to Carolyn Ratcliffe, whose thoughtful insights in lab discussions served as a major source of inspiration. Thank you to Jennifer Tsai, for her valuable feedback and readiness to to lend an ear for questions, problems, and brainstorm. They each truly encapsulate the meaning of scholar and co-worker, and my work benefited greatly from discussions with them. Together, this collection of researchers provided me with astonishing resources and a memorable experience.

I would also like to thank Draper Laboratory for supporting this research, both financially and intellectually. Their feedback on the intermediate reports throughout the process of this research was very insightful.

Thank you to my parents, not only for their support of my graduate studies as I completed this work, but for their dedication to my education from the beginning as well. Without their guidance and encouragement I would not have accomplished as much as I have today.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1 INTRODUCTION	1
1.1 Overview of Current Study	3
1.2 Basic Concepts in Network Analysis	4
1.3 Problem Definition	7
1.3.1 Bipartite Networks	7
1.3.2 Related Works in Bipartite Social Network Analysis	8
1.3.3 Bipartite Networks with Non-metric Associations	12
1.4 Information Theory in Network Analysis	14
1.4.1 Entropy and Mutual Information in Information Theory	14
1.4.2 Application of Information Theory in Network Analysis	18
1.5 Hierarchical Cluster Analysis for Network Structure Inference	19
CHAPTER 2 THE CURRENT STUDY	21
2.1 Study 1: The Southern Women Network	21
2.1.1 The Dataset	21
2.1.2 Data Analysis	23
2.1.2.1 Mutual Information-Based Association Matrix	23
2.1.2.2 Network Structure Visualization	28
2.2 Study 2: Social Network from Cell Phone Call Records	31
2.2.1 The VAST 2008 Challenge Cell Phone Call Records Data	31
2.2.2 Analysis with Original Data	34
2.2.3 Data Reconstruction	37
2.2.4 Inferring Network Structure from Mutual Information-Based Measures	39
2.2.4.1 Mutual Information-Based Association Measures	39
2.2.4.2 Mutual Information-Based Proximity Measures	43
CHAPTER 3 DISCUSSION AND FUTURE DIRECTIONS	49
3.1 Mutual Information in Social Network Analysis	50
3.2 Limitations and Future Directions	53
3.3 Conclusion	55
REFERENCES	56

LIST OF TABLES

1.1	Adjacency matrix of the network in Figure 1.1	5
1.2	Adjacency matrix of the bipartite network in Figure 1.2	8
2.1	Normalized measure of association (U^2) between each pair of ladies in Southern Women Network.	24
2.2	Contingency table of the association patterns between two pairs of women .	25
2.3	Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from original data	36
2.4	Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from mutual information-based tie strength estimation (U^2)	40
2.5	Contingency table of association patterns and weighting matrix	41
2.6	Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from weighted mutual information-based tie strength estimation ($U^2_{weighted}$)	42

LIST OF FIGURES

1.1	A small example graph representation of a network with 8 nodes and 14 undirected edges with the same weight	4
1.2	An example of a bipartite network	8
1.3	Entropy $H(X)$ as a function of $p_{(x)}$ for a binary variable	15
1.4	Relationship between entropy, joint entropy and mutual information.	17
2.1	Women-by-event matrix in the bipartite Southern Women Network (from A. Davis, Gardner, & Gardner, 1941)	22
2.2	Clique membership of the Southern Women Network (from A. Davis et al., 1941).	22
2.3	Visualization of association (U^2) matrix before and after weighting in Southern Women Network	26
2.4	Dendrogram representations of the Southern Women Network structure derived from hierarchical clustering analysis using proximity matrices	29
2.5	Structure of the 5 core members in the Catalano/Vidro social network summarized from award-winning contest submissions in VAST 2008 Challenge	33
2.6	Social network structure of the 5 core members from original data	35
2.7	Visualization of minute-by-minute and day-by-day cell phone call activities for all 400 nodes	38
2.8	Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using proximity matrix ($D = 1 - U^2$) from day 1 to day 7	44
2.9	Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using proximity matrix ($D = 1 - U^2$) from day 8 to day 10	45
2.10	Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using weighted proximity matrix ($D_{weighted} = 1 - U_{weighted}^2$) from day 1 to day 7	46
2.11	Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using weighted proximity matrix ($D_{weighted} = 1 - U_{weighted}^2$) from day 8 to day 10	47

CHAPTER 1

INTRODUCTION

Analysis of networks has been widely used in a great number of areas to understand relationships between different entities in a system, as well as behavior of a system as a whole due to the interactions between entities in the system. Researchers have conducted observations and experiments, developed a variety of network analysis techniques including graphical visualization, statistical inference and computational algorithms, and built a number of mathematical models in an effort to understand and predict the behavior of network systems (for a review, see Newman, 2003).

A bipartite network is a type of network constituted by entities of two distinct types, with connections that can only exist between entities of *different* types (Wasserman & Faust, 1994). Unlike networks with all entities belonging to the same category, in which a connection between two entities is explicit and definitive, bipartite networks provide only implicit indications of “probable existence” of a connection between entities of the same type when their connections with entities in the other type covaries (for example, when they are both connected with a same group of entities in the other type). Thus, inferring network structures among entities of the same type from bipartite networks can be particularly challenging.

Researchers in social network analysis have developed a variety of techniques to detect subgroups or other patterns of connections from bipartite (or even tripartite) social network

data (Borgatti, 2009; Borgatti & Everett, 1997; Borgatti & Halgin, 2011; Breiger, 1974; Ghosh, Kane, & Ganguly, 2011; Liu & Murata, 2009). However, most of the techniques are based on the assumption that the bipartite connections are of only two forms (exist vs. non-exist), or are metric in nature (i.e., the connection has a value representing the tie strength between the two entities of different types). As a result, these techniques cannot be easily generalized to another type of bipartite network, in which the connections are of multiple forms that are only categorically different from each other.

One way to represent tie strength between two entities of the same type in a bipartite network with non-metric associations is to calculate the mutual information between the two entities based on information theory (Cover & Thomas, 2006; Kvalseth, 1987; Shannon, 1948). This method has been frequently used in analyzing a number of network types, such as calculating word associations in linguistic networks (e.g., Church & Hanks, 1990; P. Li & Church, 2007; Seretan & Wehrli, 2006) and associations between genes in bioinformatics (e.g., Butte & Kohane, 2000; Dawy et al., 2006). To our knowledge, mutual information has not been used in social network analysis to represent tie strength between social entities. Instead, researchers analyzing bipartite social networks have been mainly focusing on co-occurrence of the same type of connections based on the assumption of homophily in social networks, which means "similarity breeds connection" (McPherson, Smith-Lovin, & Cook, 2001). However, co-occurrence of different types of connections can also be highly informative of a strong tie, which might be different in nature than the ties formed as a result of homophily. Therefore, mutual information can be a more general measure of tie strength between social entities in bipartite networks due to its capability of capturing co-occurrence of any type.

1.1 Overview of Current Study

The current study is a first step exploration to apply the method of using mutual information to represent tie strength in bipartite social networks with non-metric associations. This thesis is organized as follows. First, I will introduce the basic concepts and measures used to analyze networks based on graph theory. Then I will define the type of networks to which I am going to apply the methodology of mutual information in the current study, and summarize existing methods for analyzing this type of networks in the context of social network analysis. The next section will introduce the basic concepts and mathematics of calculating entropy and mutual information of discrete random variables in information theory, followed by a summary of empirical studies that employ the measure of mutual information in network analysis, such as studies in bioinformatics and natural language processing. The final section in the introduction introduces a method of visualizing network structures from proximity matrices, which can be calculated from mutual information. In the main part of this thesis, I will apply the method of using mutual information to represent tie strength and proximity onto the analyses of two social network datasets, a small one with 18 nodes and a medium sized one with 400 nodes. Then I will apply a number of methods to infer and create visualizations of the underlying network structure from mutual information-based measures. Effectiveness of the new method can then be investigated by comparing the inferred network structures with the ones discovered in previous studies. In the discussion section, I will discuss potential issues when applying this method to social network data such as the scope of its applications, the advantages and disadvantages of this method and possible modifications that can be made to this method.

1.2 Basic Concepts in Network Analysis

Graphs and matrices are typically used to represent networks so they can be studied mathematically (Newman, 2003; Wasserman & Faust, 1994). This section introduces the basic concepts and notations in graph theory that are applicable in the context of network analysis.

A *graph*, $G = (V, E)$ is comprised of two sets of information: a set of *vertices* or *nodes*, $V = \{v_1, v_2, \dots, v_n\}$, each corresponding to an entity in a network, and a set of *edges* or *lines* between vertices, $E = \{e_1, e_2, \dots, e_n\}$, $e_k = (v_i, v_j)$, representing a connection between two entities in the network. Researchers in different areas have differing preferences of using the terms. Figure 1.1 illustrates an example of a small network with 8 nodes and 14 edges.

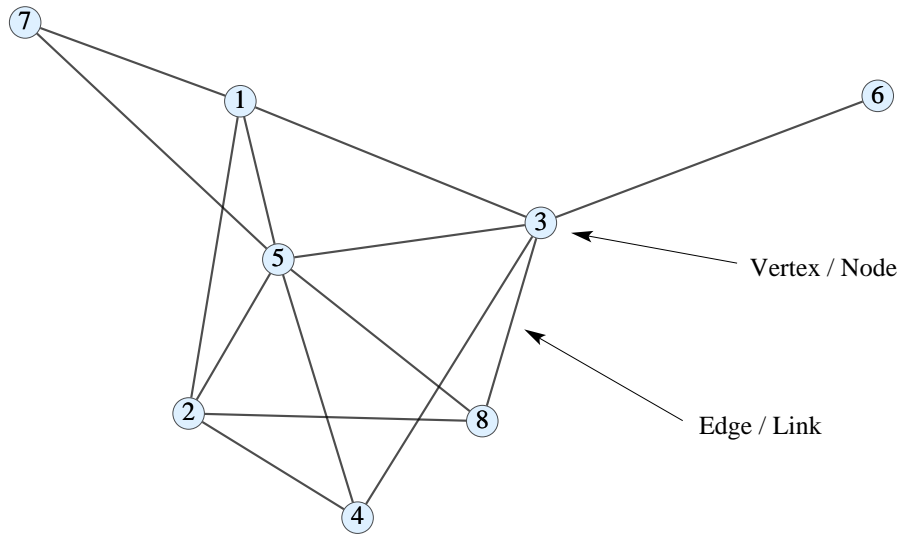


Figure 1.1: A small example graph representation of a network with 8 nodes and 14 undirected edges with the same weight

An edge can be *undirected* (i.e., $e_k = (v_i, v_j) = (v_j, v_i)$) or *directed* (i.e., $e_k = (v_i, v_j)$ and $e_l = (v_j, v_i)$ are different). Edges can also carry *edge weights*, representing a quantitative property such as tie strength, similarity, or distance between two nodes. In Figure 1.1, all edges are undirected and carrying the same weight.

Matrices are widely used to represent network data too. An $n \times n$ *adjacency matrix* (X) can be used to represent a network with n nodes ($V = \{v_1, v_2, \dots, v_n\}$), with a value 1 in the $(i, j)^{th}$ cell ($x_{ij} = 1$) if a connection exists between node v_i and v_j , and a value 0 in that cell ($x_{ij} = 0$) if the connection does not exist. Table 1.1 is the corresponding adjacency matrix of the network illustrated in Figure 1.1. When a network only contains undirected edges, the adjacency matrix is *symmetric*. *Asymmetric* adjacency matrices are used to represent networks with directed edges.

	1	2	3	4	5	6	7	8
1	0	1	1	0	1	0	1	0
2	1	0	0	1	1	0	0	1
3	1	0	0	1	1	1	0	1
4	0	1	1	0	1	0	0	0
5	1	1	1	1	0	0	1	1
6	0	0	1	0	0	0	0	0
7	1	0	0	0	1	0	0	0
8	0	1	1	0	1	0	0	0

Table 1.1: Adjacency matrix of the network in Figure 1.1

A number of measures can be generated from a graph to describe different properties of a network, the relative position of a particular node and the relationship between a subgroup of nodes in a network. One of the primary goals in network analysis is to identify the “most important” node in a network. *Centrality* is one of the most frequently used measures to describe the prominence of a node’s location in a network (Freeman, 1979). The following three node-based centrality measures (degree, closeness and betweenness) will be used in this thesis.

The *degree* of a node is the number of edges connected to it. Degree is a basic measure of the extent to which a node is connected with other nodes in a network and indicates the

node’s potential communication activity with other nodes (Freeman, 1979). A node has both an *in-degree* (i.e., number of incoming edges) and an *out-degree* (i.e., number of outgoing edges) in networks with directed edges. When edges carry weights, a node’s degree will be calculated with edge weights taken into account depending on the meaning of edge weights.

Both closeness and betweenness centrality measures are based on the concept of *geodesic* between two nodes. The *geodesic* between two nodes in a network, g_{ij} is the shortest path through the network from one node to the other. Sometimes there could be more than one geodesics between two given nodes. The length of a geodesic is the *geodesic distance*, $d(v_i, v_j)$, between two nodes. The geodesic path from v_i to v_j is the same as the one from v_j to v_i in a network with undirected edges. Whereas the two geodesic paths, $d(v_i, v_j)$ and $d(v_j, v_i)$, can be different in networks with directed edges.

The *closeness centrality* of a node (v_i) is a measure that aggregates the geodesic distance between v_i and all other nodes in the network. The Sabidussi’s (1966) index of closeness centrality takes the inverse of the sum of all the geodesic distances from node v_i , as illustrated in Equation 1.1. The closeness centrality measures the inverse of how far a node is from all other nodes in a network. Therefore, the larger the number of $C_{closeness}(v_i)$, the more “centered” the node v_i is located in the network. However, this measure is meaningful only when all other nodes are reachable from node v_i , since the geodesic distance between two unconnected nodes is infinity.

$$C_{closeness}(v_i) = \left[\sum_{j=1}^n d(v_i, v_j) \right]^{-1} \quad (1.1)$$

The *betweenness centrality* of a node measures the probability of a node occurring on the geodesics of all other pairs of nodes. It is calculated by taking the proportion of all geodesics between v_j and v_k that pass through node v_i , and then summing all the proportions together (see Equation 1.2). High values of $C_{betweenness}(v_i)$ means that the node v_i is more likely to be strategically located on the communication paths linking pairs of other nodes, thus influential in the information transmission within the network (Freeman, 1979).

$$C_{betweenness}(v_i) = \sum_{j < k} \frac{g_{jk}(v_i)}{g_{jk}} \quad (1.2)$$

1.3 Problem Definition

The focus type of networks of the current study is bipartite social networks with non-metric associations. The goal is to explore the method of using mutual information to represent tie strength and proximity between nodes in this type of networks. This mutual information-based measure can be a more general way of revealing various types of social ties than the traditional methods that only considers one association pattern.

1.3.1 Bipartite Networks

A *bipartite network* is a network consisting of two different types of nodes ($V = \{v_1, v_2, \dots, v_m\}$ and $W = \{w_1, w_2, \dots, w_n\}$) and edges that exist only between nodes from different types ($e_k = (v_i, w_j)$). Figure 1.2 is an example of a bipartite network with nodes 1 through 9 belonging to one type and nodes A through F belonging to the other. Note that no direct link exists between the nodes of the same type. The adjacency matrix of a bipartite network is

of the form $A = \begin{pmatrix} B & 0 \\ 0 & B^T \end{pmatrix}$. Therefore, the *2-mode adjacency matrix* B is sufficient to represent the structure of a bipartite network. Table 2 is the corresponding 2-mode adjacency matrix of the network illustrated in Figure 1.2.

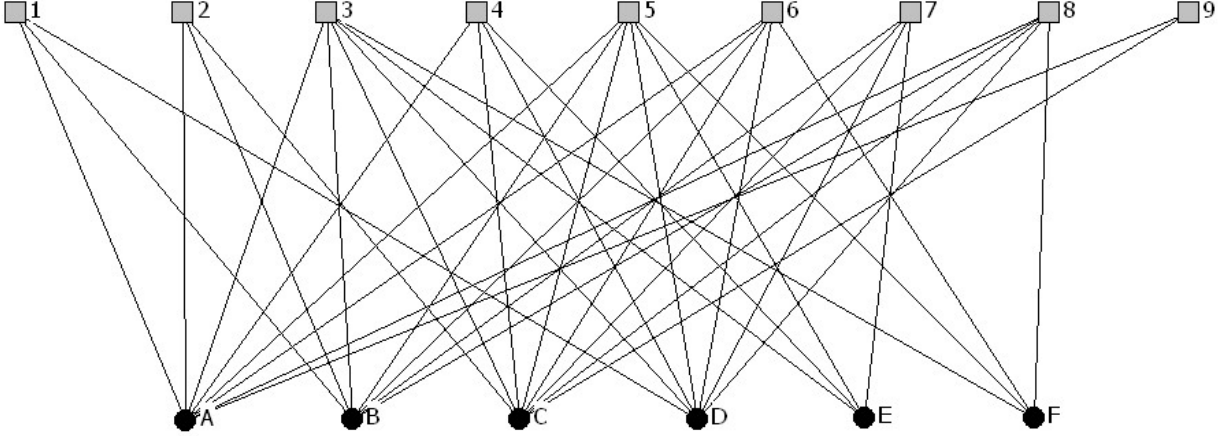


Figure 1.2: An example of a bipartite network

	1	2	3	4	5	6	7	8	9
A	1	1	1	1	1	1	0	1	1
B	1	1	1	0	1	1	1	1	0
C	0	1	1	1	1	1	1	1	1
D	1	0	1	1	1	1	1	1	0
E	0	0	1	1	1	0	1	0	0
F	0	0	1	0	1	1	0	1	0

Table 1.2: Adjacency matrix of the bipartite network in Figure 1.2

1.3.2 Related Works in Bipartite Social Network Analysis

In the context of social network analysis, bipartite networks are also known as *affiliation networks* (Wasserman & Faust, 1994). For example, let all the numbers in Figure 1.2 represent a group of individuals and all the letters represent a list of groups. Then in this affiliation network, we only know that a subset of individuals are members of a particular

group. But we do not know whether these individuals directly interacted with each other or not. However, it is reasonable to infer that being members of the same group might indicate a higher probability that these two individuals know each other.

Researchers in social network analysis often take the assumption that common group membership indicates a link. Based on this assumption, researchers have studied various types of affiliation networks to understand their properties and underlying dynamics. For example, quite a number of studies have examined the structure and dynamics of scientific collaboration networks by analyzing co-authorship in published scientific articles (e.g., Barabasi et al., 2002; Batagelj & Mrvar, 2000; Melin & Persson, 1996; Moody, 2004; Newman, 2001, 2004). Corporate elite networks in which a link between two company directors is assumed if they belong to the board of the same company are also studied to understand the nature of corporate governance and the diffusion of corporate practices, strategies, and structures (G. F. Davis & Greve, 1997; G. F. Davis, Yoo, & Baker, 2003; Mariolis, 1975). Online tagging system is another type of commonly examined bipartite network in which links between users have been identified through the overlap of tags they use to describe online documents such as webpages, pictures, music, and so on (Ghosh et al., 2011; Schifanella, Barrat, Cattuto, Markines, & Menczer, 2010). A number of data mining algorithms have also been developed to identify social ties between individuals from their co-location information, either self-reported or as provided by geolocation devices such as global positioning systems (GPS) and cellphone bluetooth devices (Crandall et al., 2010; Huang, Pei, & Xiong, 2006; Lin & Lim, 2008; Mardenfeld et al., 2010; Yoo, Shekhar, Smith, & Kumquat, 2004). The major purposes of these analyses on affiliation networks include social tie inference (e.g.,

Crandall et al., 2010; Schifanella et al., 2010), community or subgroup detection (e.g., Ghosh et al., 2011; Liu & Murata, 2009; Mardenfeld et al., 2010), network structure reconstruction (e.g., G. F. Davis & Greve, 1997; Moody, 2004), and so on.

However, the assumption of common group membership indicating a link does not hold for all types of affiliation networks. The nature of different types of affiliation networks might influence the likelihood of common group membership leading to direct interaction between individuals. For example, being on the board of the same company at the same time probably has a very strong indication that two company directors should know each other and have been working with each other on company governance issues. On the other hand, people encounter with strangers all the times in their everyday activities. Hence, co-location of two individuals does not always mean acquaintance. Moreover, the probability of two online users knowing each other given their usage of the same tag in describing a piece of document might be even lower.

One good example is the study done by Crandall and colleagues (2010), in which they examined the relationship between spatial-temporal co-occurrences and the existence of real social ties among users on Flickr.com, an online image sharing community. The authors used geo-tags and time stamps (either user generated or automatically recorded by the photo taking devices) of photos uploaded by users to infer whether two users were at the same location within a certain period of time. Their results suggested that as the number of spatial-temporal co-occurrences increased, the probability of two users actually having a social tie as indicated in Flickr’s public social network increased drastically. However, although two users were three hundred times more likely to know each other if they had

3 co-occurrences (within 1 latitude-longitude degree on any given day) than chance level, the actual probability level (5%) is still low. Therefore, it might not be appropriate to assume existence of a social tie based on every occurrence of common group membership in every type of affiliation networks. Instead, an aggregated measure of the probability of existence of a social tie between two individuals taking all occurrences of their common group membership into account might be a better way to represent the inferred relationship between these two individuals in a bipartite social network.

On the other hand, common group membership is one of the four possible association patterns of how two individuals can be associated with the same entity in the other node type. That is, $(e_{v_i, w_k}, e_{v_j, w_k}) = (1, 1)$ in all four association patterns: $(1, 1)$, $(1, 0)$, $(0, 1)$ and $(0, 0)$. It is not surprising that association co-existence receives most attention in social network analysis because numerous studies on *homophily* have provided strong evidence that similarity in various aspects of our daily life such as ethnicity, gender, age, religion, education, occupation, behavior and attitude breeds the formation of social ties (for a review, see McPherson et al., 2001). However, in some affiliation networks, the other three association patterns can be informative of revealing the relationship between two individuals too. The nature of the social ties indicated by the other three association patterns might be fundamentally different from the ties identified through association co-existence. For example, when the association patterns of $(1, 0)$ and $(0, 1)$ dominate the association pairs of two individuals in an affiliation network, that is, whenever one is associated with a node of a different type, the other is always not associated with the same node, or vice versa, these two individuals might be purposefully avoiding each other instead of behav-

ing independently. This “avoidance” can be a result of different psycho-social reasons and thus indicate different types of social ties. Two individuals might be avoiding each other if they dislike each other, which indicates a negative tie. But two individuals might also be strategically avoiding each other for other purposes, such as optimizing allocation of limited resources. For example, two individuals might strategically choose to attend different sessions during a multi-session conference to maximize coverage of the conference content. This “avoidance” behavioral pattern then indicates a strong social tie since it foreshadows a higher probability of communication between these two individuals due to heightened need of information exchange. Therefore, a more general measure that takes the distribution of all four types of association patterns into account is needed in order to capture social ties of various types, especially the ones established due to reasons other than homophily. This need of utilizing a more generalized measure can be particularly high when the associations in a bipartite social network can take more than two forms.

1.3.3 Bipartite Networks with Non-metric Associations

Affiliation networks usually assumes that a link between the two nodes of different types either exists or not. In a more general form of bipartite networks, a link can be one of multiple states, which are different from each other either quantitatively or qualitatively. A *bipartite network with non-metric associations* contains links of qualitatively different forms, which can be more than dichotomous. For example, in a bipartite network with a group of individuals attending a number of events, in addition to the information of whether each person attended a certain event or not, we can also know whether this individual attended

the event as an organizer, a presenter or an attendee. Another example of this general form of bipartite network is time resolved human movement data. With a group of individuals forming one type of nodes, and a list of time stamps forming the other, this type of network has a number of locations as the links between the two node groups. Therefore, if we take a revisit at the Flickr data analyzed in Crandall and colleagues' (2010) study, this is an example of a bipartite network with non-metric associations¹.

As the number of possible associations (n) between two node types increase in this type of bipartite networks, the number of association patterns of how two nodes in one type are linked to the same node in the other type grows quadratically (n^2). However, only $1/n$ link patterns are taken into account if we want to discover social ties solely based on the assumption of homophily, which is, by only considering association patterns with two individuals linked to the same node in the other type in exactly the same way. For example, some data mining tools might only consider the cases when two individuals are at the same location at the same time when inferring social ties, without taking the cases when two individuals are at different locations (which might be systematically different) at the same time into account. Therefore, a large proportion $((n - 1)/n)$ of all association patterns is ignored and this proportion keeps increasing as n increases. As discussed earlier, these ignored association patterns can indicate social ties that are different in nature than the ones formed as a result of homophily. Thus, an increased number of different types of non-metric associations in a bipartite network calls for a heightened need of a more general way to infer social ties.

¹However, it is an *incomplete*, and actually very *sparse* dataset of a bipartite network with non-metric associations. Due to its incompleteness, it is not feasible to apply the method of calculating mutual information to the dataset in Crandall and colleagues' (2010) study. This limitation will be discussed later.

The measure of mutual information between two nominal variables takes all association patterns into account when estimating the extent to which the two variables covary with each other. Therefore, this mutual information-based measure probably is a more general way of inferring tie strength in bipartite networks with non-metric associations. The current study is a first step exploration of using mutual information to represent social tie strength in this type of networks.

1.4 Information Theory in Network Analysis

1.4.1 Entropy and Mutual Information in Information Theory

Entropy is a measure of uncertainty and unpredictability of a random variable in information theory (Cover & Thomas, 2006; Shannon, 1948). For a discrete random variable X , let $p(x)$ denote the probability mass function of x , with $p(x_i)$ denoting the probability of $x = x_i$. The *entropy* of this discrete random variable X is defined by Equation 1.3. Entropy is measured in *bits* if the base of the logarithm is 2 and in *nats* if the base of the logarithm is e .

$$H(X) = - \sum_i p(x_i) \log p(x_i) \quad (1.3)$$

For the simple case of a binary variable X , with $p(x=1) = p$ and $p(x=0) = 1 - p$, its entropy can be calculated using Equation 1.4.

$$H(X) = -p \log p - (1 - p) \log (1 - p) \quad (1.4)$$

A graph of this function is shown in Figure 1.3. As we can see in this figure, fairness ($p_{(x=1)} = p_{(x=0)} = 0.5$) yields the maximum amount of entropy ($H(X) = 1$), whereas in certain cases ($p_{(x=1)} = 0$ or 1), entropy equals zero.

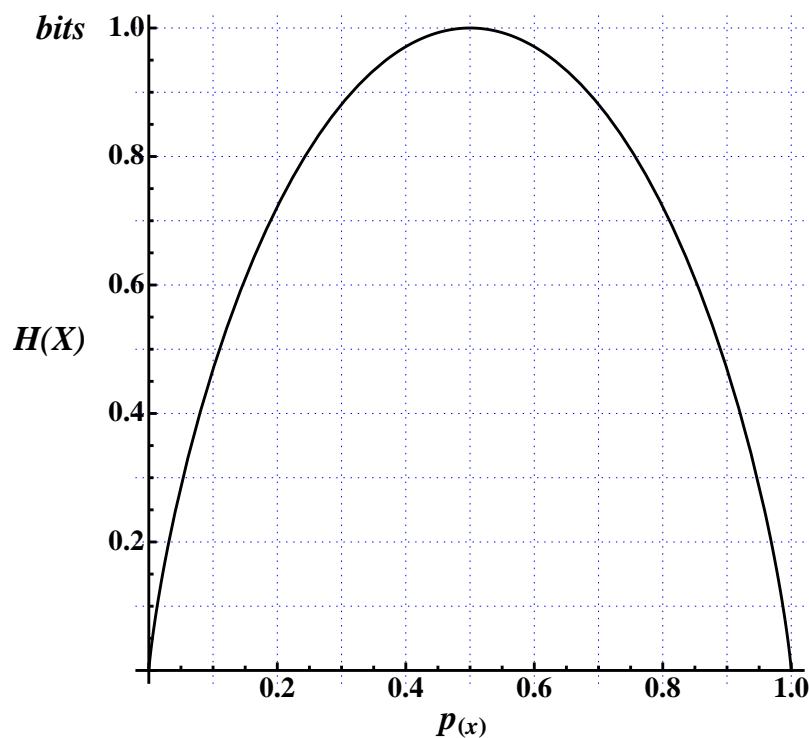


Figure 1.3: Entropy $H(X)$ as a function of $p_{(x)}$ for a binary variable

To extend the definition to a pair of discrete variables (X, Y) , the *joint entropy* $H(X, Y)$ is defined by Equation 1.5, in which $p_{(x,y)}$ is the joint probability mass function of the two variables, with $p_{(x_i, y_j)}$ denoting the probability of $x = x_i$ and $y = y_j$.

$$H(X, Y) = - \sum_i \sum_j p_{(x_i, y_j)} \log p_{(x_i, y_j)} \quad (1.5)$$

The *mutual information* between two random variables is a measure of the amount of information shared between these two variables, that is, the amount of information one random

variable contains about the other. Mutual information is the reduction in the uncertainty of one random variable due to the knowledge of the other. For a pair of discrete random variables X and Y , let $p_{(x,y)}$ be the joint probability mass function of the two variables, with $p_{(x_i,y_j)}$ denoting the probability of $x = x_i$ and $y = y_j$, and let $p_{(x)}$ and $p_{(y)}$ be the marginal probability mass functions with $p_{(x_{i+})}$ denoting the probability of $x = x_i$ and $p_{(y_{+j})}$ denoting the probability of $y = y_j$, the *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p_{(x_{i+})}p_{(y_{+j})}$, as defined in Equation 1.6.

$$I(X;Y) = \sum_i \sum_j p_{(x_i,y_j)} \log \frac{p_{(x_i,y_j)}}{p_{(x_{i+})}p_{(y_{+j})}} \quad (1.6)$$

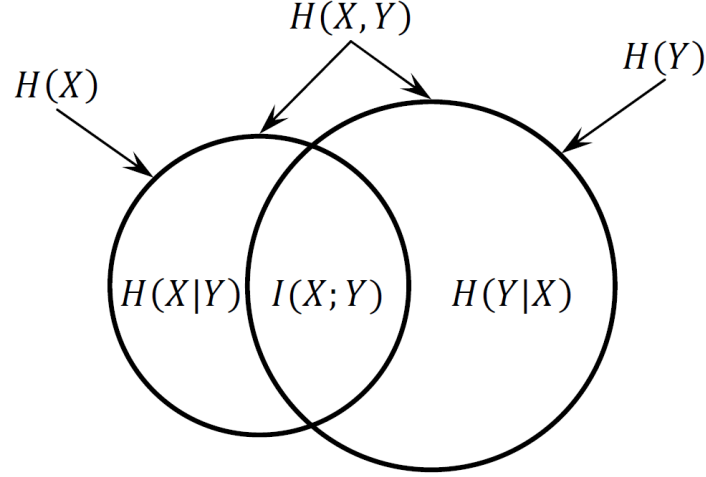
The relationship between entropy, joint entropy and mutual information can be illustrated by Equation 1.7 and the Venn diagram in Figure 1.4, in which we can clearly see that the mutual information $I(X;Y)$ corresponds to the intersection of the information in X with the information in Y .

$$I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X,Y) \quad (1.7)$$

Furthermore, Kvalseth (1987) promoted a method of normalizing the measure of mutual information to rescale it to the $[0,1]$ interval using the weighted average of the asymmetric entropy measures of the two nominal variables, as defined in Equation 1.8.

$$U^2 = \frac{2I}{H(X) + H(Y)} \quad (1.8)$$

Statistical inferences can then be performed on this normalized measure of association between two nominal variables.



Note: $H(X|Y)$ and $H(Y|X)$ are conditional entropies, which are not in the scope of this thesis

Figure 1.4: Relationship between entropy, joint entropy and mutual information.

On the other hand, a simple metric of mutual information-based *proximity measure* between two random variables X and Y can be calculated by taking the non-overlapping parts of information contained in each of the two variables using Equation 1.9, that is, the remained uncertainty that cannot be explained by the knowledge of the other variable (Cover & Thomas, 2006; Kraskov & Grassberger, 2009; Shannon, 1948).

$$d(X, Y) = H(X|Y) + H(Y|X) = H(X, Y) - I(X; Y) \quad (1.9)$$

Combining Equation 1.9 with Equation 1.7 and 1.8, Equation 1.10 can be easily derived.

$$d(X, Y) = H(X) + H(Y) - 2I(X; Y) = [H(X) + H(Y)] \times (1 - U^2) \quad (1.10)$$

This proximity measure can be normalized in the same way as how mutual information $I(X;Y)$ is normalized to U^2 , which simply gives us Equation 1.11.

$$D(X,Y) = \frac{d(X,Y)}{H(X) + H(Y)} = 1 - U^2 \quad (1.11)$$

1.4.2 Application of Information Theory in Network Analysis

Information theory has not been widely used in social network analysis. In fact, Eagle and Pentland's (2006) study on reality mining (see also Eagle, Pentland, & Lazer, 2008, 2009) is perhaps the only study that used information theory to construct measures of human behavior. The authors calculated the entropy of each individual's life based on their daily distribution of a number of locations (e.g., home, work, elsewhere, no signal) as registered by cell phones and then used this measure to represent the predictiveness of a person's life in terms of home/work transitions. They found that in MIT Media Lab, faculty and staff members lived a lower entropic life than graduate students. On the other hand, freshmen come to lab at the least regular basis, thus have the highest entropic life style. While this study is an application of information theory in understanding human behavior, although solely on the individual level, my literature search failed to find any study that used mutual information to represent tie strength between individuals, that is, on a relationship level, in social network analysis.

However, this method has been commonly used and proven to be effective in other areas of network analysis. For example, researchers in computational linguistics have developed algorithms to calculate word associations based on their occurrences in a large

corpus of text documents (e.g., Church & Hanks, 1990; P. Li & Church, 2007; Seretan & Wehrli, 2006). In bioinformatics, mutual information is also commonly used to estimate gene-gene associations based on the expression patterns as represented in sequential lists of nucleotides (e.g., Butte & Kohane, 2000; Dawy et al., 2006). Given the fact that networks from different knowledge domains share quite a number of similarities, and that researchers have started to analyze networks from different knowledge domains using similar techniques and to describe them using similar models (Newman, 2003; Watts & Strogatz, 1998), it is quite surprising that mutual information has not been applied in the area of social network analysis. Thus, the current study will be a first step exploration to apply information theory in social network analysis.

1.5 Hierarchical Cluster Analysis for Network Structure Inference

A crucial step to evaluate whether mutual information-based measures can be effectively used to represent strength of social ties in social network analysis is to examine the extent to which the network structures derived from mutual information-based measures resemble the true network structures. Hence, hierarchical cluster analysis is introduced in the current study for the purpose of network structure inference.

Hierarchical cluster analysis is one of the many strategies that have been used to visualize the relationship among elements of a network and to make inference on the overall structure of the network from proximity data among those elements (Aghagolzadeh, Soltanian-Zadeh, Araabi, & Aghagolzadeh, 2007; DeJordy, Borgatti, Roussin, & Halgin, 2007; Hubert, Arabie, & Meulman, 2006; Kraskov & Grassberger, 2009; Kraskov, Stogbauer,

Andrzejak, & Grassberger, 2005). Given a proximity matrix of n elements, the primary goal of hierarchical clustering analysis is to find a partition hierarchy. This analysis is usually performed as follows. Starting from a full partition in which each element forms a subgroup, elements are grouped together step by step. At each step, the joint of two subgroups is taken to form a larger group. New group formation at each step should ensure maximum preservation of relationships between elements as provided in the proximity matrix. The whole partition hierarchy can be formed at the n^{th} step and all clusters along with their substructures can then be detected.

In the current study, the algorithm developed by Hubert and colleagues (2006) will be used. This algorithm employs a method of combinatorial optimization that can be performed iteratively for linear unidimensional scaling with the goal of minimizing the least squares criterion. After acquisition of the partition hierarchy, a dendrogram (treeplot) representation can then be generated to visualize the underlying structure of how elements form subgroups inside the whole network (Aghagolzadeh et al., 2007; Hubert et al., 2006; Kraskov & Grassberger, 2009; Kraskov et al., 2005). This dendrogram representation will then be used for comparison and algorithm evaluation.

CHAPTER 2

THE CURRENT STUDY

The primary goal of the current study was to use mutual information to represent tie strength and proximity between nodes in bipartite social networks with non-metric associations. As a first step exploration of this method, I focused on networks with binary associations in this study. The method of mutual information were applied to two bipartite social networks, a small one with 18 nodes in the node group of interest and a medium sized one with 400 nodes in the node group of interest. Generalization of this method to bipartite networks with more than two nominal association types will then be discussed.

2.1 Study 1: The Southern Women Network

2.1.1 The Dataset

A small bipartite social network, the Southern Women Network, was used in the first study. This well-known social network data was collected by Davis, Gardner and Gardner (1941) in their anthropological study on caste and class in a small town in the southern rural area of the United States. They followed 18 women for nine months and recorded their participation in 14 events during that time. The original women-by-event matrix is shown in Figure 2.1. This dataset has been examined by a large number of researchers to illustrate applications of new techniques and algorithms they developed (e.g., Borgatti, 2009; Borgatti

NAMES OF PARTICIPANTS OF GROUP I	CODE NUMBERS AND DATES OF SOCIAL EVENTS REPORTED IN <i>Old City Herald</i>													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson.....	X	X	X	X	X	X	...	X	X
2. Miss Laura Mandeville.....	X	X	X	...	X	X	X	X
3. Miss Theresa Anderson.....	...	X	X	X	X	X	X	X	X
4. Miss Brenda Rogers.....	X	...	X	X	X	X	X	X
5. Miss Charlotte McDowd.....	X	X	X	...	X
6. Miss Frances Anderson.....	X	...	X	X	...	X
7. Miss Eleanor Nye.....	X	X	X	X
8. Miss Pearl Oglethorpe.....	X	...	X	X
9. Miss Ruth DeSand.....	X	...	X	X	X
10. Miss Verne Sanderson.....	X	X	X	X
11. Miss Myra Liddell.....	X	X	X	...	X
12. Miss Katherine Rogers.....	X	X	X	...	X	X	X
13. Mrs. Sylvia Avondale.....	X	X	X	X	...	X	X	X
14. Mrs. Nora Fayette.....	X	X	...	X	X	X	X	X	X
15. Mrs. Helen Lloyd.....	X	X	...	X	X	X
16. Mrs. Dorothy Murchison.....	X	X
17. Mrs. Olivia Carleton.....	X	...	X
18. Mrs. Flora Price.....	X	...	X

FIG. 3.—Frequency of interparticipation of a group of women in Old City, 1936—Group I.

Figure 2.1: Women-by-event matrix in the bipartite Southern Women Network (from A. Davis et al., 1941)

TYPE OF MEMBERSHIP	MEMBERS	EVENTS AND PARTICIPATIONS													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
<i>Clique I:</i> Core.....	{1	C	C	C	C	C	C	-	C	C					
	{2	C	C	C	-	C	C	C	C	-					
	{3	-	C	C	C	C	C	C	C	C					
	{4	C	-	C	C	C	C	C	C	-					
	{5			P	P	P	-	P	-	-					
Primary...	{6			P	-	P	P	-	P	-					
	{7					P	P	P	P	-					
	{8					-	S	-	S	S					
<i>Clique II:</i> Secondary...	{9					S	-	S	S	S					
	{10							S	S	S	-	-	S		
	{11							-	P	P	P	-	P		
	{12							-	P	P	P	-	P	P	
	{13								C	C	C	-	C	C	C
Core.....	{14							C	C	-	C	C	C	C	C
	{15							C	C	C	-	C	C	C	C
	{16								S	S	S	-	S		
Secondary...	{17									S	-	S			
	{18									S	-	S			

FIG. 5.—Types of members of, and relationships between, two overlapping cliques.

Figure 2.2: Clique membership of the Southern Women Network (from A. Davis et al., 1941).

& Everett, 1997; Borgatti & Halgin, 2011; Breiger, 1974; Doreian, 1979; Liu & Murata, 2009). The relatively small size of the dataset made it easy to illustrate key points in a new technique. More importantly, the underlying community structure (i.e., the clique membership of the 18 ladies) was a known fact (see Figure 2.2), which provided a ground truth for researchers to evaluate the effectiveness of their techniques and algorithms. The core and primary members in the two cliques (Clique I: Evelyn, Laura, Theresa, Brenda, Charlotte, Frances and Eleanor, Clique II: Myra, Katherine, Sylvia, Nora and Helen) were used as major criterion for algorithm evaluation in the current study.

2.1.2 Data Analysis

2.1.2.1 Mutual Information-Based Association Matrix

The first step of the analysis was to calculate the mutual information $I(X; Y)$ between every pair of individuals based on their event attendance patterns using Equation 1.6, and to calculate the entropy for each individual’s event attendance activity using Equation 1.3, which was equivalent to Equation 1.4 for this particular dataset. Take the first two women as an example, the mutual information between Evelyn and Laura was $I(Evelyn; Laura) \approx 0.179$. And their individual entropy were $H(Evelyn) \approx 0.683$ and $H(Laura) \approx 0.693$. According to Equation 1.8, the normalized measure of association between Evelyn and Laura would be $U^2 \approx 0.179/(0.683 + 0.693) \approx 0.260$. Following the same calculation, a matrix of normalized measure of association between each pair of ladies was generated in Table 2.1.

To help making inferences on the underlying community structure of the 18 women, a heat map visualization of the U^2 matrix in Table 2.1 was created (Figure 2.3a) using

	EV	LA	TH	BR	CH	FR	EL	PE	RU	VE	MY	KA	SY	NO	HE	DO	OL
Evelyn	0.260																
Laura	0.402	0.260															
Theresa	0.260	0.408	0.260														
Brenda	0.042	0.080	0.316	0.398													
Charlotte	0.316	0.398	0.316	0.398	0.072												
Frances	0.042	0.398	0.316	0.398	0.072	0.343											
Eleanor	0.235	0.025	0.235	0.025	0.149	0.160	0.160										
Pearl	0.042	0.080	0.316	0.080	0.072	0.072	0.343	0.160									
Ruth	0.006	0.000	0.042	0.000	0.002	0.002	0.072	0.160	0.343								
Verne	0.006	0.080	0.006	0.080	0.196	0.002	0.002	0.160	0.072	0.343							
Myrna	0.130	0.260	0.130	0.260	0.316	0.042	0.042	0.052	0.006	0.134	0.508						
Katherine	0.260	0.137	0.062	0.137	0.080	0.080	0.000	0.025	0.080	0.398	0.398	0.695					
Sylvia	0.529	0.260	0.164	0.260	0.134	0.134	0.006	0.009	0.006	0.042	0.042	0.164	0.260				
Nora	0.245	0.017	0.050	0.017	0.017	0.017	0.028	0.001	0.028	0.214	0.214	0.050	0.157	0.094			
Helen	0.162	0.000	0.162	0.000	0.104	0.034	0.034	0.589	0.421	0.421	0.421	0.251	0.201	0.003	0.013		
Dorothy	0.003	0.201	0.003	0.201	0.104	0.104	0.104	0.074	0.034	0.034	0.034	0.003	0.000	0.162	0.013	0.159	
Olivia	0.003	0.201	0.003	0.201	0.104	0.104	0.104	0.074	0.034	0.034	0.034	0.003	0.000	0.162	0.013	0.159	
Flora	0.003	0.201	0.003	0.201	0.104	0.104	0.104	0.074	0.034	0.034	0.034	0.003	0.000	0.162	0.013	0.159	1.000

Note: since the U^2 matrix is symmetric and with all 1's along the diagonal, only the lower triangle is presented in this table.

Table 2.1: Normalized measure of association (U^2) between each pair of ladies in Southern Women Network.

Mathematica 8 (2010). By paying special attention to the core and primary members of the two cliques, we can see that the mutual information was relatively higher among women from the same clique (Mean $U^2 = 0.274$ and 0.258 for Clique I and Clique II, respectively) than the overall average mutual information for all pairs of individuals (Mean $U^2 = 0.151$), as reflected in the darker orange color in the two blue rectangles in Figure 2.3a.

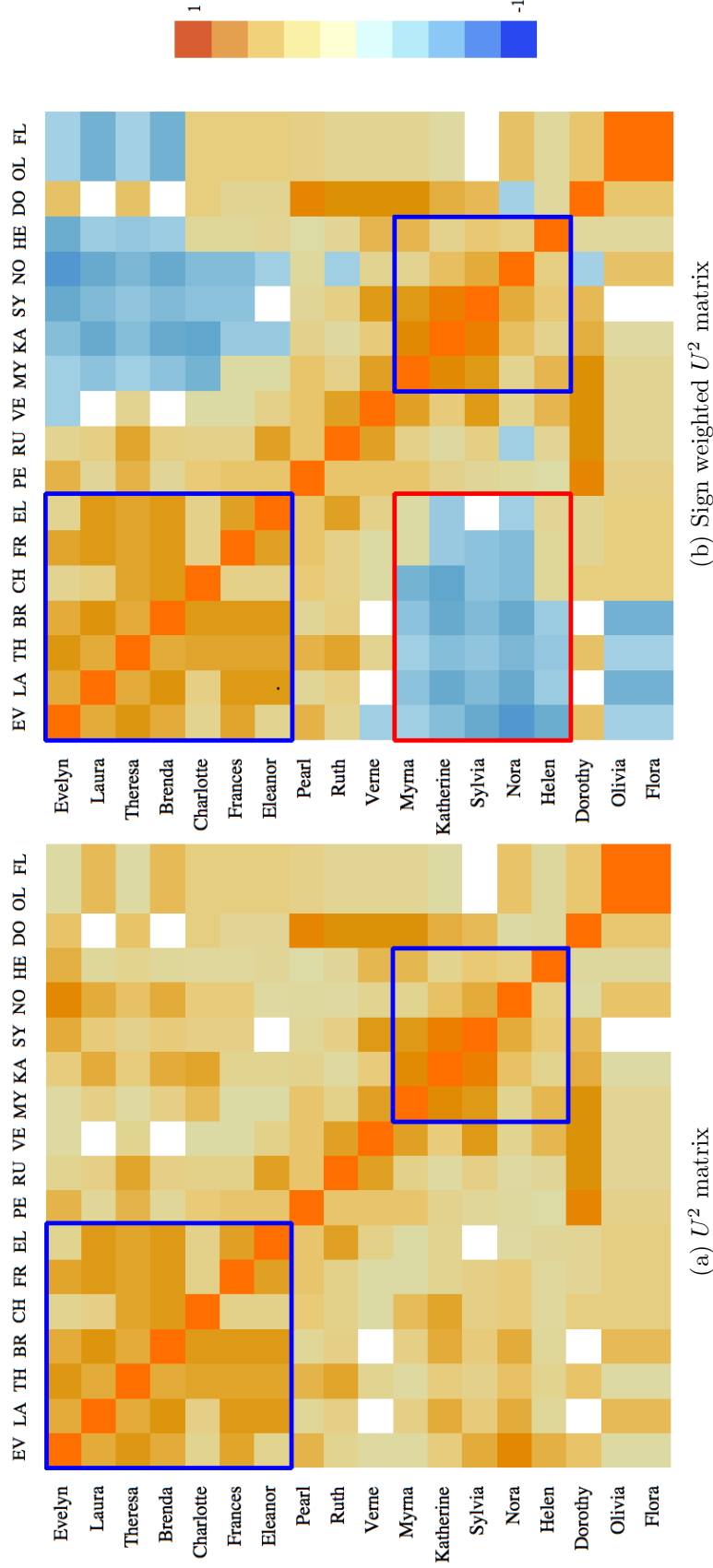
However, it is important to point out that since U^2 was an index of interdependence between two variables, this mutual information-based measure was different from simply counting the number of co-occurrences. For example, even though Laura and Verne attended the same event twice (Event 7 and 8), their mutual information was zero. For this particular case, it seemed that these two individuals attended the same events as a result of random choices, rather than homophily commonly assumed in social network analysis.

Moreover, it is also important to point out that any form of an imbalanced distribution of the joint probably mass function $p_{(x,y)}$ can result in a high measure of U^2 . Thus further investigation is needed to understand the causes of the highly interdependent behaviors between two individuals. Let's take the contingency tables of two pairs of women as an example (see Table 2.2). In Table 2.2a, the number of times both (or neither) Evelyn and Theresa attending the same event ($N_{(1,1)} + N_{(0,0)} = 12$) was much higher than the number of times one of them attending an event but the other not ($N_{(0,1)} + N_{(1,0)} = 2$). In

		Evelyn	
		1	0
Theresa	1	7	1
	0	1	5
(a) Evelyn and Theresa			

		Evelyn	
		1	0
Nora	1	2	6
	0	6	0
(b) Evelyn and Nora			

Table 2.2: Contingency table of the association patterns between two pairs of women



Note: the core and primary members of the two cliques are highlighted in the two blue boxes. Across-clique associations are highlighted in the red box in Figure 2.3b.

Figure 2.3: Visualization of association (U^2) matrix before and after weighting in Southern Women Network

contrast, the behavioral dissimilarity between Evelyn and Nora dominated the contingency table ($N_{(0,1)} + N_{(1,0)} = 12$, see Table 2.2b). Therefore, the high values of mutual information between the two pairs of women ($U_{Evelyn,Theresa}^2 = 0.402$, $U_{Evelyn,Nora}^2 = 0.529$) were due to different reasons.

In order to differentiate the two types of associations, a weighting system was introduced (see Equation 2.1).

$$U_{weighted}^2 = \begin{cases} U^2 & \text{if } (N_{(1,1)} + N_{(0,0)}) \geq (N_{(0,1)} + N_{(1,0)}) \\ -U^2 & \text{if } (N_{(1,1)} + N_{(0,0)}) < (N_{(0,1)} + N_{(1,0)}) \end{cases} \quad (2.1)$$

This weighting system simply compared the similar versus dissimilar cases for each pair of women and reversed the sign of U^2 if there were more dissimilar cases than similar ones. Hence, a positive U^2 indicated a tendency of two individuals attending the same events, whereas a negative U^2 indicated a tendency of two individuals avoiding each other by attending different events. We can clearly see the two types of relations in the heat map visualization of the sign weighted matrix ($U_{weighted}^2$) in Figure 2.3b, with orange indicating positive relations and blue indicating negative ones. The high concentration of negative relations between pairs of core and primary members from different cliques as indicated in the red box in Figure 2.3b (Mean $U^2 = -0.116$) provided further evidence that Clique I and Clique II were different from each other.

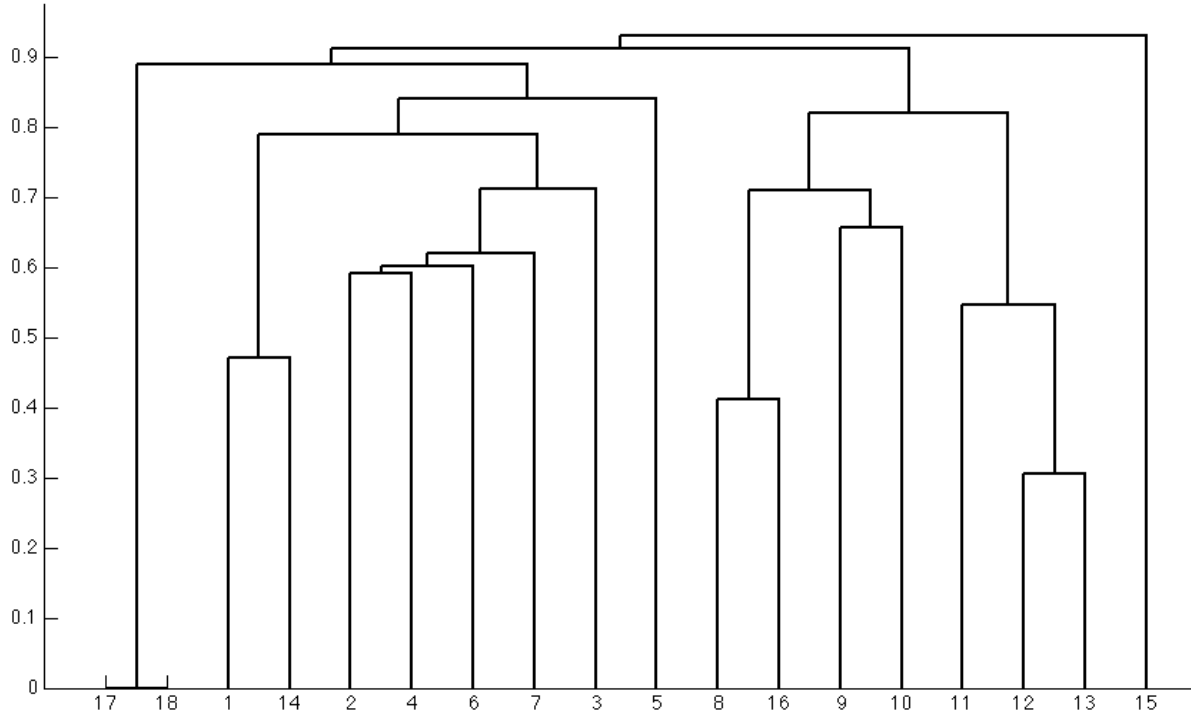
As we can see, the heat map visualization provided preliminary evidence on the effectiveness of the method of using normalized mutual information (U^2) to represent tie strength among the 18 women in this bipartite network. Although the average within-clique

U^2 was found to be higher than the overall average U^2 , the comparison was made post-hoc (i.e., given that we already knew the community structure of the network). Thus, the next question to ask was whether the network structure could be directly detected from the mutual information-based measures and whether the inferred network structure resemble the true community structure of the Southern Women Network.

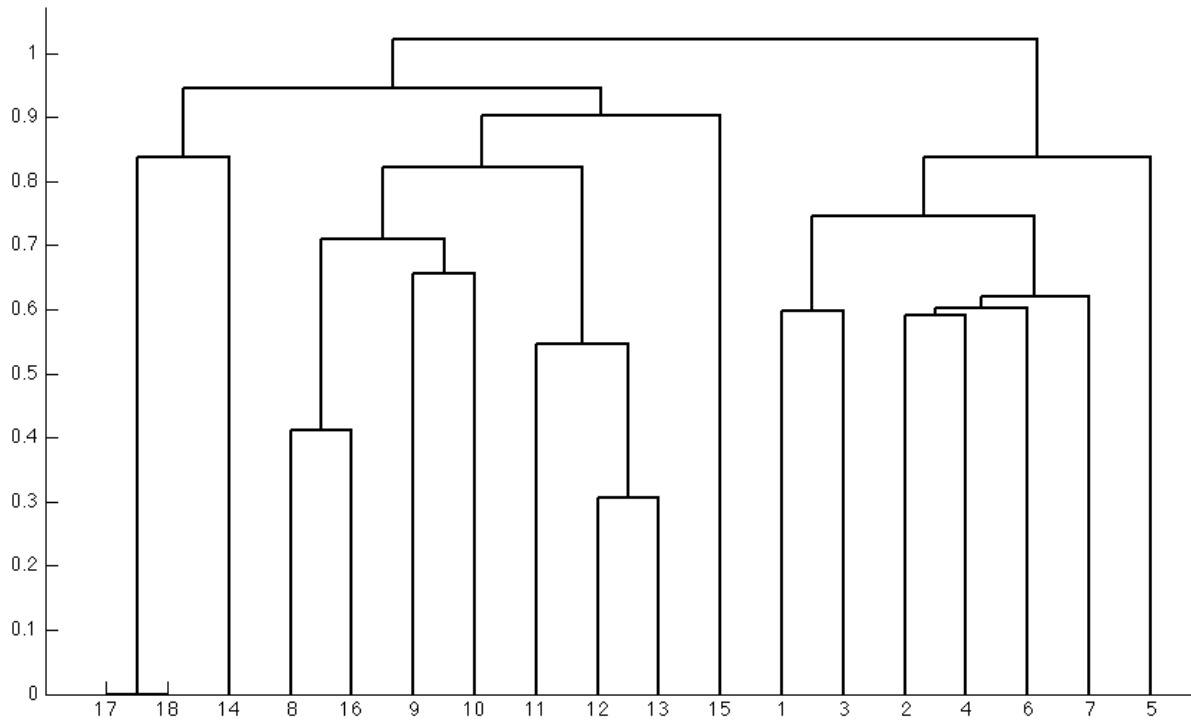
2.1.2.2 Network Structure Visualization

Two proximity matrices, $D = 1 - U^2$ and $D_{weighted} = 1 - U_{weighted}^2$, were constructed according to Equation 1.11. Hierarchical clustering analysis was then conducted on the two proximity matrices using Hubert and colleagues' (2006) algorithm in MATLAB (2011). A dendrogram representation was constructed for each proximity matrix (see Figure 2.4). In the dendrogram, the points along the x-axis represented the 18 women in the network. Vertical lines from each individual were connected by horizontal bars at different heights along the y-axis, representing the level at which the subgroup was formed. Level of subgroup formation was estimated from the proximity matrix.

In the dendrogram derived from the unweighted mutual information-based proximity matrix (Figure 2.4a), two large subgroups were identified, with women 15 not belonging to either group. Consistent with the community structure provided in the original study, women 1 through 7 belonged to the same subgroup. However, there were some discrepancies between the rest part of group identification and the results of the original study. In particular, the most salient difference was that women 14 (Nora) was grouped with women 1 (Evelyn) at an early step. Whereas in fact, Nora should belong to Clique II as indicated in the original study. As discussed earlier, this discrepancy was due to the fact that these two women did



(a) $D = 1 - U^2$



(b) $D_{weighted} = 1 - U_{weighted}^2$

Figure 2.4: Dendrogram representations of the Southern Women Network structure derived from hierarchical clustering analysis using proximity matrices

have high dependency in their event attendance, but in opposite directions. Therefore we would expect that by introducing the weighting system when calculating mutual information-based proximity between individuals, the resulted dendrogram representation should have a higher resemblance with the ground truth. This expectation was confirmed in Figure 2.4b, in which we could clearly see two subgroups identified, with women 1 through 7 belong to the first group and women 8 through 18 belonging to the second. The only discrepancy was that women 8 (Pearl) was grouped into Clique I in the original study. However, since this women was only a secondary member of Clique I, and the same discrepancy was also found in previous studies using different algorithms (Liu & Murata, 2009), we can conclude that the network structure derived from the weighted mutual information-based proximity matrix highly resembled the true structure of the Southern Women Network.

To summarize, the results in the first study showed that the method of using mutual information to represent tie strength and proximity between nodes in bipartite networks provided satisfying results in detecting the underlying community structure of the social network, especially after introducing the weighting system to help differentiate types of high interdependency that were due to different association patterns.

2.2 Study 2: Social Network from Cell Phone Call Records

In this study, the method of using mutual information to represent tie strength and proximity between individuals was applied to a larger social network dataset with 400 nodes.

2.2.1 The VAST 2008 Challenge Cell Phone Call Records Data

The original dataset was an artificial dataset from the third mini challenge of the Visual Analytics Science and Technology 2008 Challenge (VAST 2008), which was a contest part of the IEEE Symposium on Visual Analytics Science and Technology 2008. The goal of this mini challenge was to find the social network structure from a set of cell phone call records over a ten-day period and to characterize the changes in this social structure over this period. The original dataset was not in the form of a bipartite social network. It was a one-mode edgelist containing 9834 cell phone call records, for each of which five fields of information were provided: ID number of the calling phone, ID number of the receiving phone, date and time of the call, duration of the call and the location of the call origination cell tower. The name of the five core members of interest was given and the relationship among them was described as follows:

“We have medium confidence that Ferdinando Catalano is identifier 200. Close relatives and associates he would be calling would include David Vidro, Juan Vidro, Jorge Vidro, and Estaban Catalano. We believe Ferdinando would call brother Estaban most frequently. We also believe that David Vidro coordinates high-level Paraiso activities and communications.” (from IEEE VAST Challenge Descriptions, 2008)

Research on the award-winning contest submissions provided further information on the structure and dynamics of the social network structure (Chien, Tat, Proulx, Khamisa, & Wright, 2008; Correa et al., 2008; Farrugia & Quigley, 2008; Payne, Solomon, Sankar, & McGrew, 2008; Pellegrino, Pan, Robinson, Stryker, & Luo, 2008; Perer, 2008; Swing, 2008; Ye et al., 2008). First, the underlying structure of the Catalano/Vidro social network can be illustrated in Figure 2.5. The different coloring of the nodes in Figure 2.5 represented different roles of the each individual in the network. Estaban Catalano and David Vidro were critical coordinates of activities and communications in the organization. As the two largest “hubs” in the network who kept in contact with a large number of individuals in the 400 nodes, we would expect that these two orange nodes should have high ranks on their betweenness centrality measures. That means, the probability that these two nodes occurred on the geodesic paths between two other nodes should be particularly high. Jorge and Juan Vidro were two subordinates of David Vidro. These two individuals were also actively contacting a number of other individuals in the network, but not as much as Estaban Catalano and David Vidro. Therefore, the two blue nodes were “hubs” smaller than the orange nodes, as reflected in their betweenness centrality measures¹. Ferdinando Catalano was the core person in the network. He kept in contact with all the other 4 core members. But he did not contact the rest individuals in the 400-node network as often. Hence, this pink node was not a “hub” in the network and did not necessarily rank high in betweenness centrality measure. However, by keeping in contact with all the other four “hubs” in the

¹The status of these two individuals were relatively equal in the Catalano/Vidro social network in terms of centrality measures. Results in previous contest submissions were only able to identify two ID numbers that might correspond to these two individuals. But it was difficult to differential which ID correspond to which individual exactly. Therefore, these two individuals are not differentiated from each other and two interchangeable ID numbers will be reported when reporting analysis results in this thesis.

network, this node could easily reach all other nodes in the network within a few steps, which would be reflected in a high rank in the measure of closeness centrality. Second, according to the contest submissions, all five core members in the Catalano/Vidro network changed their cell phone numbers between day 7 and day 8. Therefore, although the five core members in the Catalano/Vidro social network and the relationship among them kept the same throughout the 10 days, their corresponding ID numbers were different from day 1 to 7 than those from day 8 to day 10.

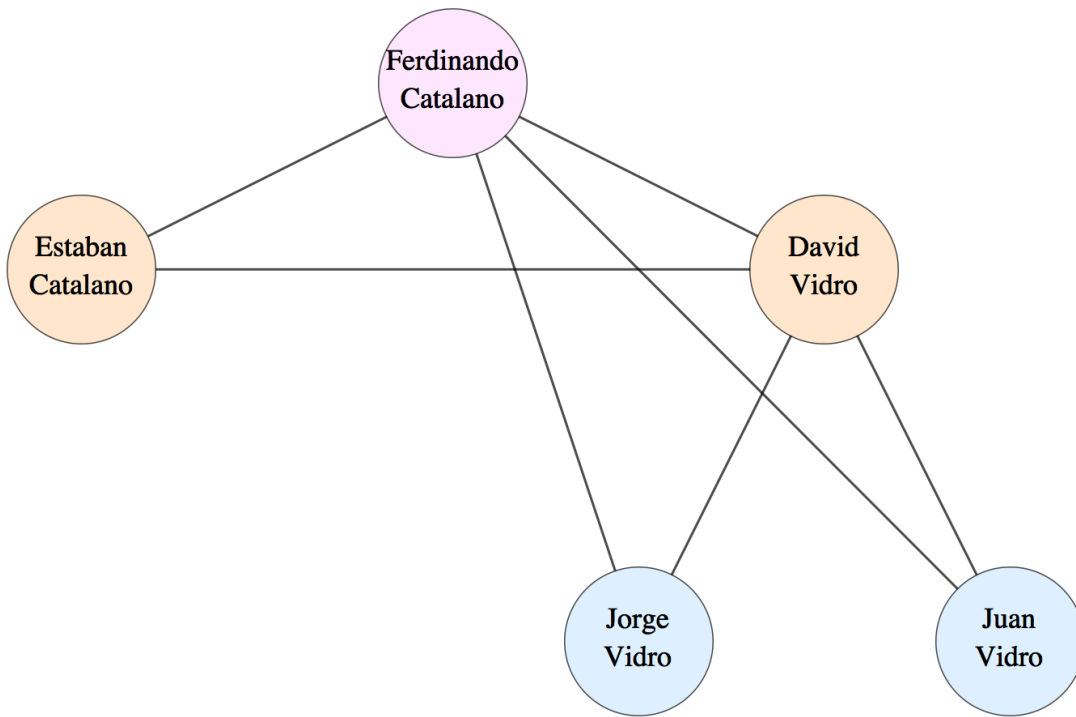


Figure 2.5: Structure of the 5 core members in the Catalano/Vidro social network summarized from award-winning contest submissions in VAST 2008 Challenge

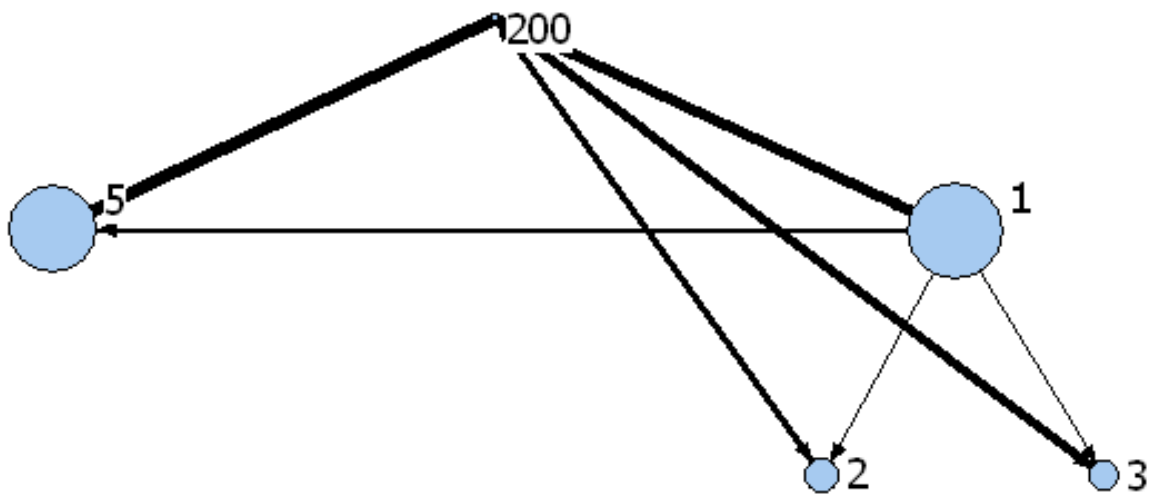
Study 2 was conducted in the following steps. First, the original dataset in the form of a one-mode edgelist was analyzed using a social network analysis software named UCINET (Borgatti, Everett, & Freeman, 2002) to see if the results converged with what was found in previous contest submissions. Second, since the original dataset was a one-mode edgelist,

not in the form of a bipartite network, it was reconstructed so that the method of mutual information calculation can be applied. Third, mutual information-based tie strength and proximity matrices was then created from the reconstructed data. Structural visualization techniques were then applied to these matrices to investigate whether the underlying social network structure could be effectively identified from mutual information-based measures. Lastly, a weighting system with different allocation of importance on different association patterns was introduced when calculating mutual information-based measures. Whether the effectiveness of identifying underlying social network structure increased after introducing the weighting system was then investigated.

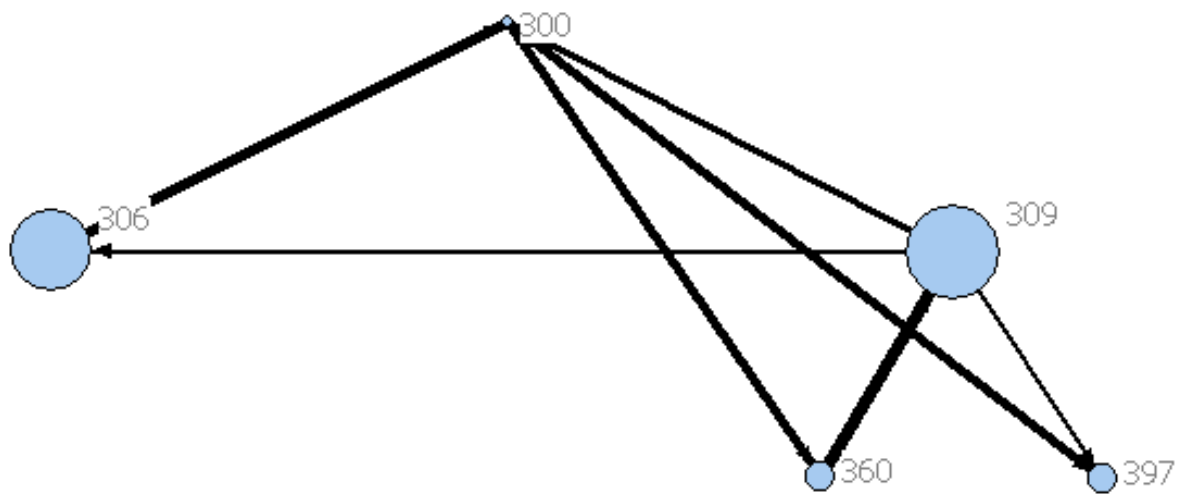
2.2.2 Analysis with Original Data

Based on the information provided in previous contest submissions, the original data was divided into two parts (phone call records in the first 7 days and the last 3 days) and each part was analyzed separately in UCINET. The length of each phone call was entered as an edge weight between the caller and the receiver. The resulted network structure of the five core members from the phone records in the two time periods is shown in Figure 2.6. Despite the fact that there were some small discrepancies among previous contest submissions themselves, a comparison of Figure 2.6 to the results in previous contest submissions suggested a high agreement that the ID numbers in the two figures corresponded to the 5 core members during each time period.

The betweenness and closeness centrality measures of the five core nodes were calculated for the two time periods, along with their rank among all 400 nodes (see Table 2.3).



(a) Day 1 to 7



(b) Day 8 to 10

Note: the five nodes in each subfigure have a one-to-one mapping with the individuals in Figure 2.5. The size of the circles represents the betweenness centrality of the node in the entire network of 400 nodes. The thickness of the links represents tie strength.

Figure 2.6: Social network structure of the 5 core members from original data

The results were consistent with the descriptions of the features of the 5 nodes in the previous section. Node 1 and 5 during the first 7 days and node 309 and 306 during the last 3 days ranked highest in both betweenness and closeness centrality, suggesting that they should correspond to the two orange nodes in Figure 2.5. Node 2 and 3 during the first 7 days and node 360 and 397 during the last 3 days also ranked high in both betweenness and closeness centrality, but lower than the previous group of nodes. Therefore, they should correspond to the two blue nodes in Figure 2.5. Most interestingly, node 200 during the first 7 days and node 300 during the last 3 days had relatively low measure on betweenness centrality, but ranked particularly high on closeness centrality. The unique characteristics of the two nodes along with their link patterns with the other four nodes suggested that they should correspond to the pink nodes in Figure 2.5, who was in fact the leader of the Catalano/Vidro social network.

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	200	343.49	112	1518	8
Estaban Catalano	5	21915.07	2	1320	2
David Vidro	1	23855.47	1	1306	1
Juan/Jorge Vidro	2	7109.90	4	1467	4
	3	5621.06	5	1512	7

(a) Day 1 to 7

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	300	1688.87	24	1563	4
Estaban Catalano	306	22205.35	2	1398	2
David Vidro	309	26442.41	1	1370	1
Juan/Jorge Vidro	360	6815.05	4	1580	6
	397	6350.75	5	1576	5

(b) Day 8 to 10

Table 2.3: Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from original data

2.2.3 Data Reconstruction

Since the original cell phone call records dataset was a one-mode edge list, it should be reconstructed to bipartite form before mutual information calculation can be applied. The beginning date and time as well as the duration of each phone call record was taken to calculate whether a particular individual was on phone talking with someone or off phone at a particular minute during the 10 days. A 400-node by 14416-minute 2-mode adjacency matrix was then constructed with a 1 or a 0 in the $(i, j)^{th}$ cell representing the on phone or off phone status of the i^{th} caller at the j^{th} minute. From the adjacency matrix obtained after data reconstruction, we were only able to tell if a pair of callers were on (or off) phone simultaneously or not. But we were not able to make definitive inferences on whether the two individuals who were on phone at the same time were actually contacting each other, since it was possible that they were each contacting someone else who were also on phone during that time. Therefore, mutual information-based measures could be used as a probabilistic measure to infer the tie strength between two nodes in this network. Moreover, it is also important to point out that this stage of data reconstruction resulted in information loss that was not recoverable since all explicit edges were removed. As a result, it would be normal to expect a relatively lower performance in social network structure inference from mutual information-based association measures.

The whole adjacency matrix was visualized in Figure 2.7a. The orange dots representing on phone status showed a clear pattern in which the density of dots were higher during day time and lower during night time. Thus it was quite easy to differentiate one day from another in the visualization. More importantly, from this visualization, we can

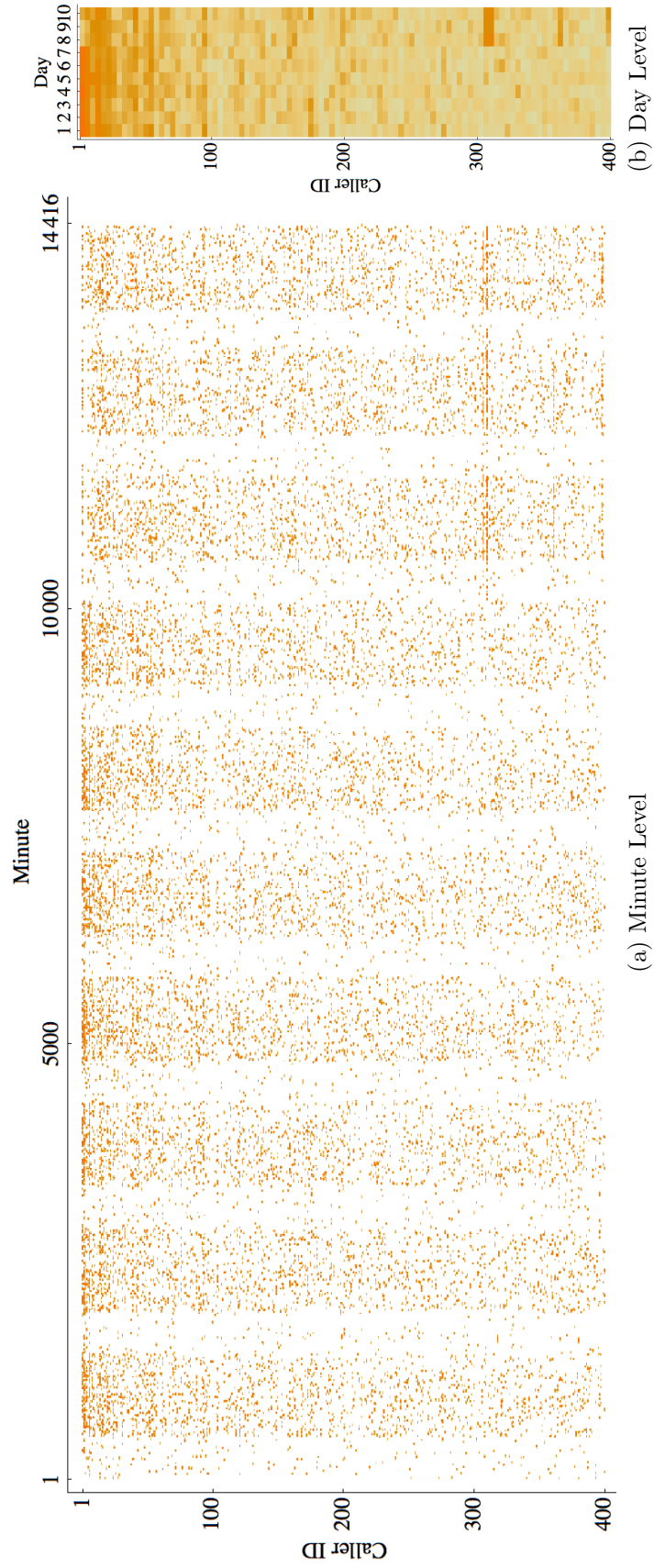


Figure 2.7: Visualization of minute-by-minute and day-by-day cell phone call activities for all 400 nodes

clearly see that the pattern of dots changed between the seventh and the eighth day. By simply looking at the density of the dots, we can see that the first several caller IDs made a lot of phone calls from day 1 to day 7 but stopped making as many calls in the last three days. In contrast, the caller IDs around 300 started to make quite a number of phone calls in the last three days. This pattern change can also be clearly seen when the callers' daily activity level were aggregated and shown in Figure 2.7b. Therefore, the visualization of the reconstructed cell phone call records adjacency matrix converged with the dynamics of the Catalano/Vidro social network found in previous contest submissions, that was, all five core members changed their cell phone numbers between day 7 and day 8.

2.2.4 Inferring Network Structure from Mutual Information-Based Measures

2.2.4.1 Mutual Information-Based Association Measures

After data transformation, the normalized measure of mutual information-based association between each pair of callers was then calculated and a 400 by 400 U^2 matrix was generated. The next step was then to investigate if the same network structure could be found from the U^2 matrix.

Due to the computational limitation of UCINET, a subset of 80 nodes were selected from the original U^2 matrix for each of the two time periods. The both on phone time for each pair of nodes were calculated and then ranks from highest to lowest. Pairs of nodes were added to selection one by one starting from the node pair with the highest overlapping on phone time, until the total number of nodes reached 80. The 80 by 80 U^2 matrix was then

entered into UCINET, with the measure of U^2 entered as edge weights. Extremely weak links ($U^2 < 0.01$) was filtered out. Then the betweenness and closeness centrality measures of each node as well as their rank among the 80 nodes were calculated (see Table 2.4).

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	200	24.28	54	257	27
Estaban Catalano	5	226.64	2	229	2
David Vidro	1	380.87	1	222	1
Juan/Jorge Vidro	2	117.29	8	247	9
	3	63.45	23	256	24

(a) Day 1 to 7

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	300	16.20	61	227	64
Estaban Catalano	306	90.48	1	205	1
David Vidro	309	53.72	7	211	3
Juan/Jorge Vidro	360	70.80	4	211	3
	397	71.50	3	208	2

(b) Day 8 to 10

Table 2.4: Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from mutual information-based tie strength estimation (U^2)

As we can see in Table 2.4a, node 1 and node 5 still ranked highest in both betweenness and closeness centrality measures, indicating that the two biggest hubs in the Catalano/Vidro social network can still be easily identified after removal of explicit edges and using mutual information-based association measures to estimate tie strength between nodes. The betweenness and closeness centrality measures were still relatively high for node 2, but not for node 3. But the unique characteristic of ranking low in betweenness but high in closeness disappeared for node 200, indicating that it probably would be difficult to find this important node when using mutual information-based measure to represent tie strength in a bipartite network. Similar result patterns were found with the U^2 matrix for the last

three days (Table 2.4b), except that all four big hubs had high ranking in both betweenness and closeness centrality measures. But the order of the ranking was not the same as what was found from the original data.

Similar as what was found in Study 1, the different on/off phone association patterns might be of different importance in contributing to the estimation of tie strength. For the particular case of making phone calls, usually people are on phone talking with someone for only a very small proportion of time. Thus, not surprisingly, the reconstructed adjacency matrix was a very sparse one that on average, these callers only spent 6.17% time on phone. And when we look at the on/off phone association patterns between two callers, both were off phone most of the time (88.14%, see Table 2.5a). Despite the large proportion of off-off associations, they were much less informative than the other association patterns in determining tie strength.

		Caller 2	
		on	off
Caller 1	on	0.0049	0.0568
	off	0.0568	0.8814

(a) Frequency of association patterns

		Caller 2	
		on	off
Caller 1	on	3.394	0.293
	off	0.293	0.019

(b) Weighting matrix

Table 2.5: Contingency table of association patterns and weighting matrix

Thus, a weighting system that placed more importance on less frequent cases was introduced (Table 2.5b). The weighting matrix was generated by taking the inverse of the frequency of each association pattern and then rescaling the four weights so the numbers added up to 4. The weighting matrix was then combined with Equation 1.6 to calculate the weighted mutual information between two callers as well as the weighted entropy of each individual caller (see Equation 2.2). The weighted mutual information and weighted

entropy were then entered in Equation 1.8 to calculate the normalized weighted mutual information-based association ($U_{weighted}^2$) between two callers.

$$I_{weighted}(X; Y) = \sum_i \sum_j w_{i,j} \times p_{(x_i, y_j)} \log \frac{p_{(x_i, y_j)}}{p_{(x_i+, y_{+j})}} \quad (2.2)$$

Following the same procedure, 80 nodes were selected and the $U_{weighted}^2$ matrix of those 80 nodes were entered into UCINET. After filtering out weak links ($U_{weighted}^2 < 0.04$), betweenness and closeness centrality measures of each node as well as their rank in the 80 nodes were calculated (see Table 2.6). Comparing to Table 2.4, improvements can be observed in that all four hubs received high rank in both betweenness and closeness centrality measures in both time periods, especially for node 3 in the first 7 days. However, node 200/300 still could not be easily identified since they still ranked low in both centrality measures in the two time periods. Therefore, the weighting system helped detecting hub

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	200	11.43	58	245	24
Estaban Catalano	5	432.61	2	210	2
David Vidro	1	536.40	1	204	1
Juan/Jorge Vidro	2	156.84	4	228	4
	3	93.61	9	228	4

(a) Day 1 to 7

Name	ID	Betweenness	Rank	Closeness	Rank
Ferdinando Catalano	300	2.84	78	237	68
Estaban Catalano	306	170.03	1	196	1
David Vidro	309	137.11	3	201	4
Juan/Jorge Vidro	360	157.68	2	200	2
	397	136.32	4	200	2

(b) Day 8 to 10

Table 2.6: Betweenness and closeness centrality of the 5 core members in the Catalano/Vidro social network from weighted mutual information-based tie strength estimation ($U_{weighted}^2$)

nodes with high betweenness centrality measures, but did not help with detecting the node with high closeness centrality but low betweenness centrality in this bipartite social network of cell phone call records.

2.2.4.2 Mutual Information-Based Proximity Measures

Hierarchical cluster analysis was also applied to the proximity matrices, (D and $D_{weighted}$) derived from Equation 1.11. Similarly, due to the computational limitation of the hierarchical cluster analysis algorithm developed by Hubert and colleagues (2006), the same 80 nodes were selected and dendrogram representations of the four proximity matrices were generated (see Figure 2.8 to Figure 2.11).

The results of hierarchical cluster analysis were not as ideal. As we can see in the first three dendrogram representations, the five important nodes were scattered and no consistent pattern of node clusters could be found. The only exception was Figure 2.11, in which all five important nodes were closely grouped together. However, how these five nodes were different from other ones in the social network still could not be directly observed in the dendrogram representation.

To summarize, the results of Study 2 was mixed. The mutual information-based estimation of tie strength between nodes combined with UCINET analysis based on graph theory helped with identification of 4 of the 5 important nodes that were high on betweenness centrality measures. Introducing the weighting system that placed different importance on different association patterns slightly improved the sensitivity of this method in detecting nodes of interest. The one node with a low rank in betweenness centrality but a high rank in closeness centrality could not be identified. Therefore, the effectiveness of this method

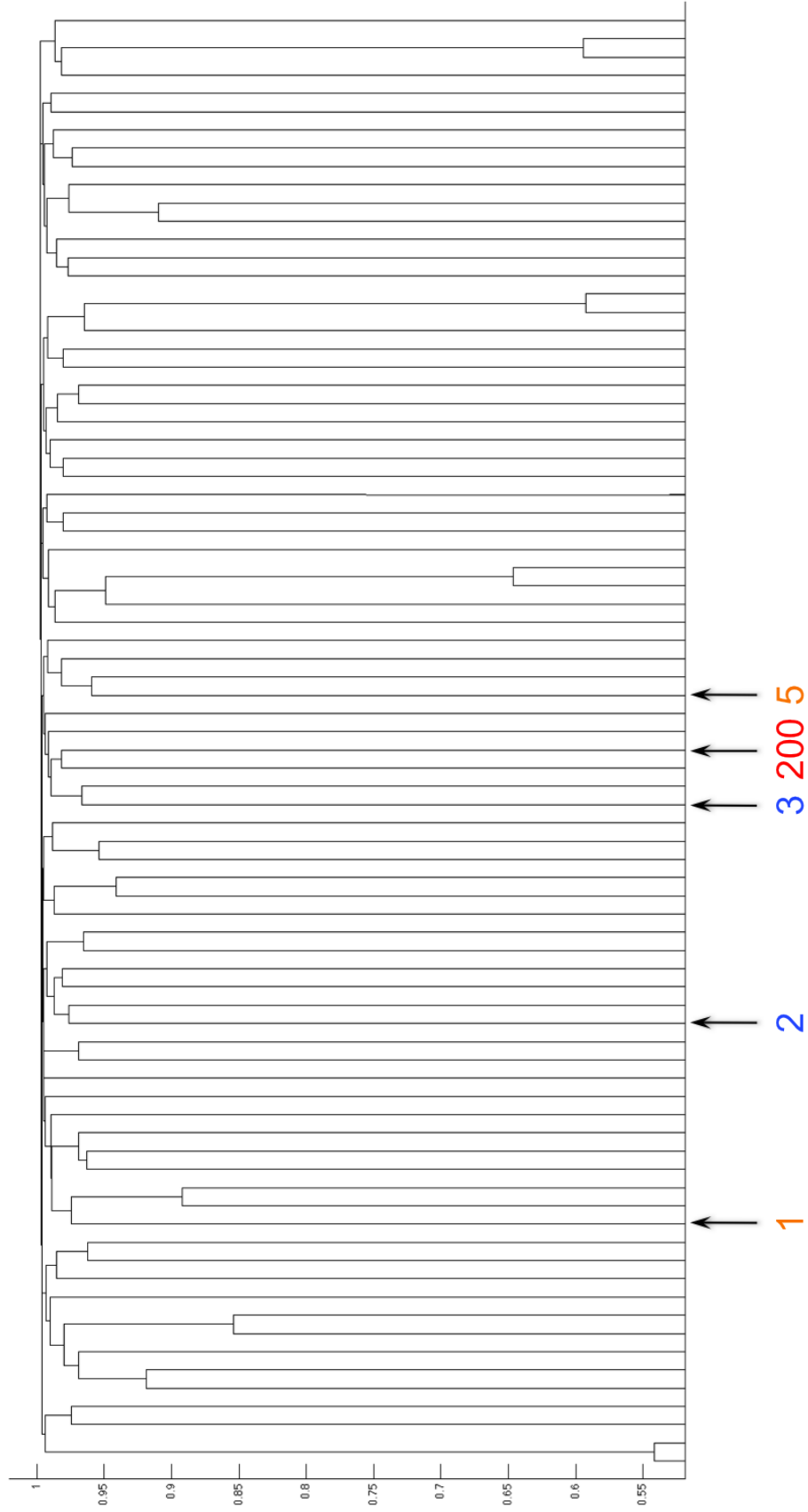


Figure 2.8: Dendrogram representations of the Catalanano/Vidro social network structure derived from hierarchical clustering analysis using proximity matrix ($D = 1 - U^2$) from day 1 to day 7

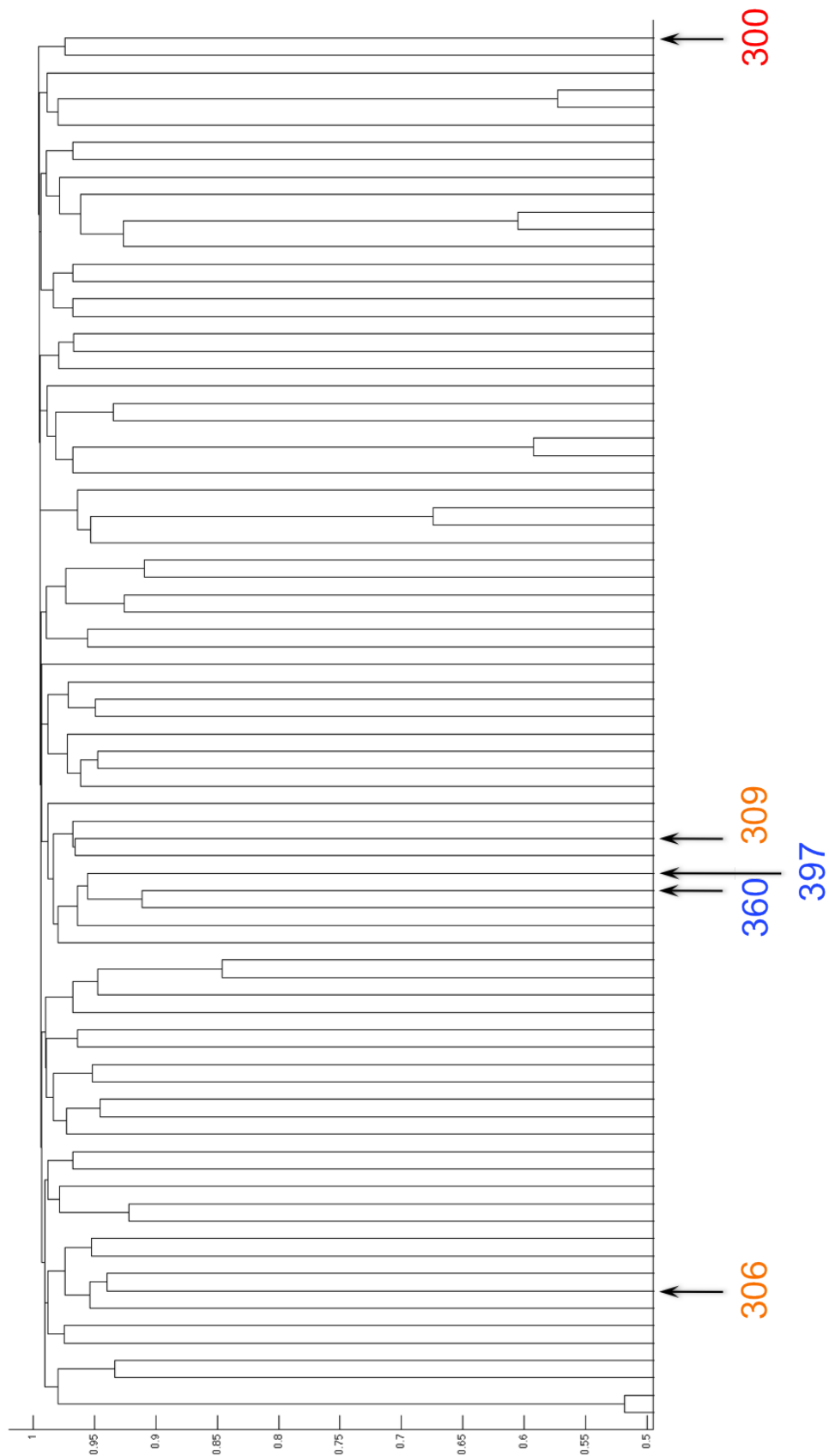


Figure 2.9: Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using proximity matrix ($D = 1 - U^2$) from day 8 to day 10

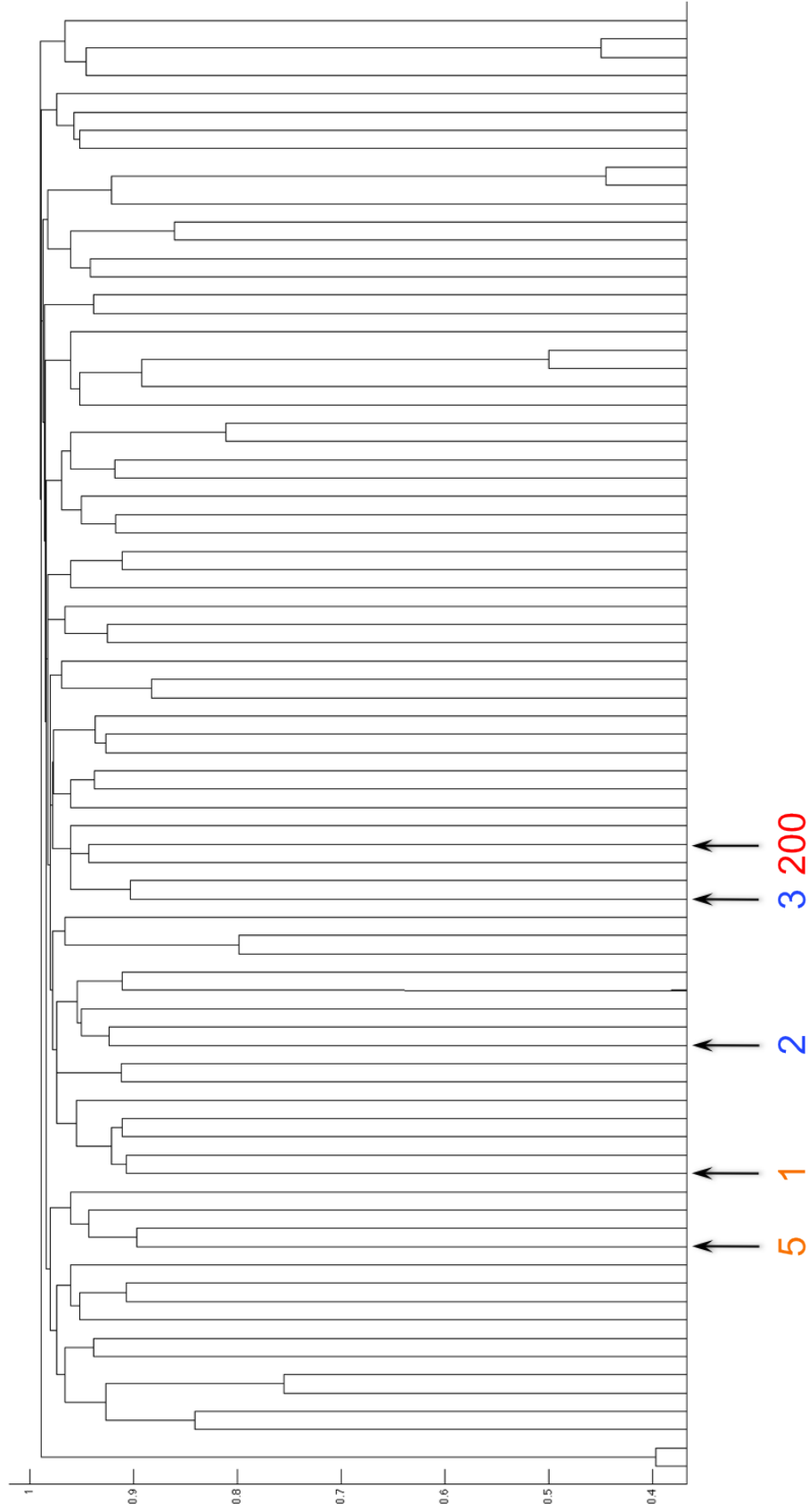


Figure 2.10: Dendrogram representations of the Catalanano/Vidro social network structure derived from hierarchical clustering analysis using weighted proximity matrix ($D_{weighted} = 1 - U_{weighted}^2$) from day 1 to day 7

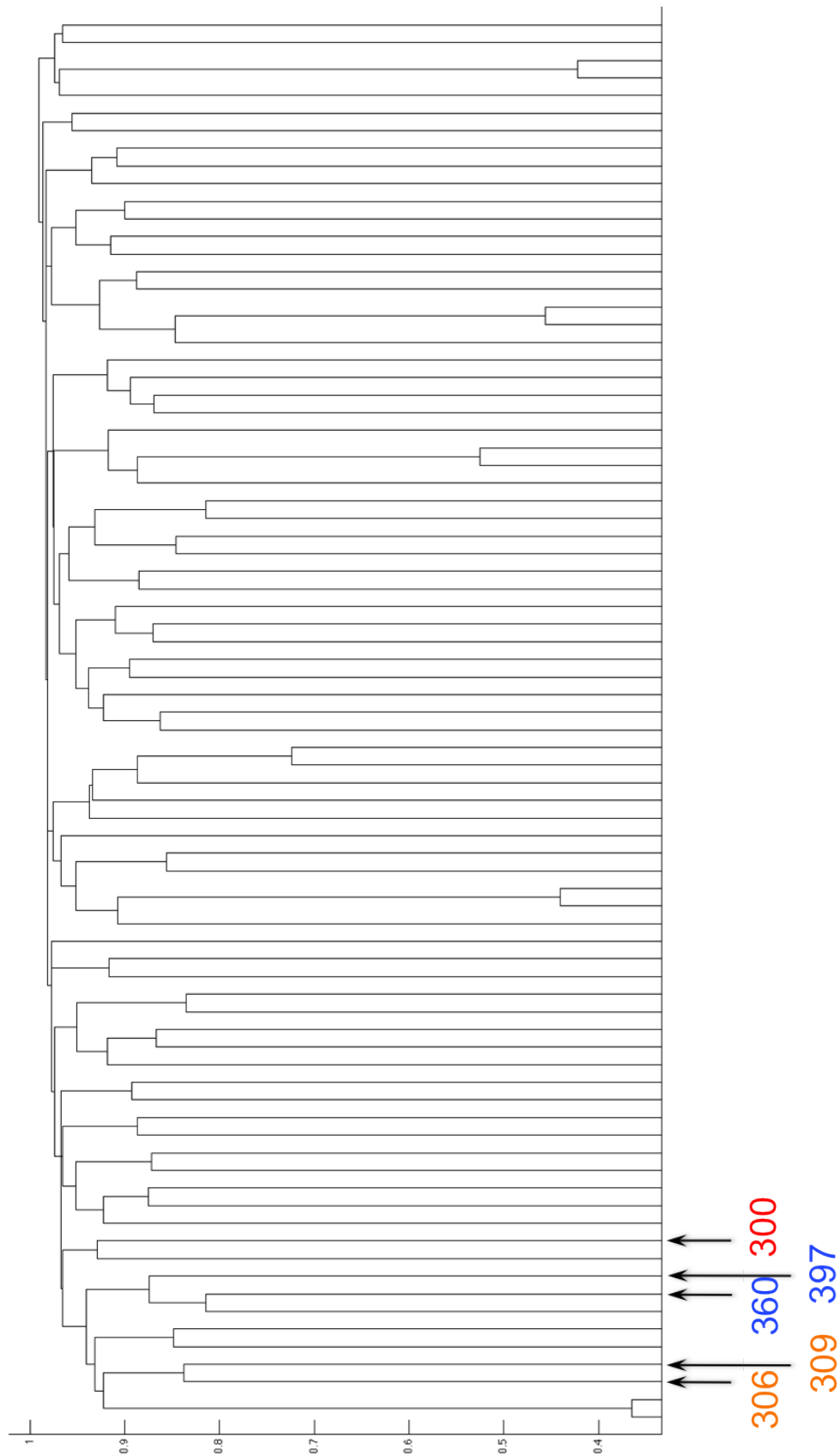


Figure 2.11: Dendrogram representations of the Catalano/Vidro social network structure derived from hierarchical clustering analysis using weighted proximity matrix ($D_{weighted} = 1 - U_{weighted}^2$) from day 8 to day 10

might vary depending on which centrality measure was used. In contrast, although the weighting system might have helped in analyzing the data of the last three days to a small extent, mutual information-based proximity estimation combined with hierarchical cluster analysis was in general not as effective in inferring underlying social network structure from this bipartite cell phone calls social network. However, given the fact that a significant amount of information was lost during data reconstruction, it was quite impressive that all of the 4 large hubs could still be detected from the reconstructed bipartite network using mutual information-based tie strength measures.

CHAPTER 3

DISCUSSION AND FUTURE DIRECTIONS

The current study was a first attempt to use mutual information-based measure to estimate tie strength and proximity between nodes in bipartite social networks with non-metric associations. Combined with other social network analysis techniques such as graph theory-based centrality measures and hierarchical cluster analysis, this method was proven to be at least moderately effective for bipartite social networks with binary associations. In Study 1, major members of the two subgroups in the Southern Woman Network were successfully detected from the mutual information-based proximity measures calculated from event attendance records of the 18 women. In Study 2, important hubs in the Catalano/Vidro social network with high betweenness centrality were successfully detected from the mutual information-based tie strength measures calculated from on/off phone status of the 400 individuals. Moreover, in both studies, the effectiveness of mutual information-based measures in detecting underlying social network structure improved after introducing weighting systems that placed different emphasis on different association patterns. To summarize, the two studies provided evidence that mutual information can be used as a useful measure in social network analysis. This brings back the question of why this method is not widely used in social network analysis but in other areas of network analysis.

3.1 Mutual Information in Social Network Analysis

As discussed in the introduction, surprisingly, mutual information is not used in social network analysis as often as in other areas of network analysis, such as computational linguistic studies on word association and bioinformatic studies on gene mapping and genomic clustering. This discrepancy might be due to several reasons.

First, unlike other types of networks, social network analysis has a high reliance on the homophily assumption and relatively low interest in other types of associations. This assumption is indeed adaptive in social network analysis since in a lot of situations, homophily is a prerequisite of the formation of a social tie. For example, two individuals need to be at the same location at the same time to be able to physically encounter and interact with each other, which is one of the major factors that would result in a social tie such as acquaintance and friendship. Similarly, two individuals need to have similar language and knowledge background to be able to communicate with each other effectively so that a social tie can form. In contrast, in other areas of network analysis, homophily is not assumed to be the major reason of tie formation. For example, in bioinformatics, two sequential lists of nucleotides that are complementary to each other also have strong associations. Hence, given the nature of how social ties form, it is indeed not surprising that homophily plays such an important role in social network analysis. However, although homophily is important in social tie formation, the relation between two individuals might evolve after a tie formation, which might be due to reasons other than homophily. For example, the “strategic allocation of limited resources” approach as discussed in the introduction might result in two individuals demonstrating behavioral patterns that are low in similarity, but

high in covariance. For instance, two co-workers might take different shifts in working time. In this case, they are not at the same location at the same time often, but the association between them is still high if we look into the covarying patterns in their locations. Another examples of non-homophily induced social ties is the policy of new graduate students taking lab rotations to get familiar with different types of research in some departments. Therefore, mutual information can be an important tool in social network analysis in both providing a full coverage of all association types and detecting non-homophily induced social ties in the bipartite social network data. This also brings the importance of introducing a good weighting system that differentiates different types of covariances and places great emphasis on the types of covariance of interest.

Second, unlike other types of network data, social network datasets are often incomplete and sparse, which might be one of the largest barriers that prevent the application of mutual information-based measures in social network analysis. For example, the cell phone call data in Study 2 is a sparse one since the proportion of time someone is on phone is usually very low, which would result in an even lower proportion of time when two individuals are on phone simultaneously. Similarly, the probability of someone traveling outside of home and work location is also low, which would also result in a sparse co-location data, such as the one in Crandall and colleagues' (2010) study. The sparsity in bipartite social network datasets might introduce bias in mutual information calculations. In addition, bipartite social network datasets are often incomplete too, such as the self report data in Crandall and colleagues' (2010) study that people only post images for interesting locations they visit, but not every location they have been to. Hence, the fact that completeness is

one of the prerequisite of mutual information calculation also post a barrier for using mutual information-based measures in analyzing bipartite social network data, unless systematic data collection was enforced to ensure data completeness.

Third, unlike the known total number, which is not a large one, of nucleotides in gene expressions, the number of possible non-metric association types in a bipartite social network can vary a lot. The simplest and probably most common case is bipartite networks with binary associations such as the two datasets analyzed in this thesis. Correlations among nodes are often calculated to represent tie strength in this type of network data (Borgatti, 2009), which might explain the low utilization of mutual information in social network analysis. Although correlation and mutual information can be used approximately interchangeably in binary sequences, this equivalence no longer exist for sequences with more than 2 status (W. Li, 1990). Thus, mutual information becomes a better choice when there are more than two non-metric association types in bipartite networks. However, the number of non-metric association types can be extremely large in some bipartite social network data, such as the number of possible locations of an individual at a particular time, the number of possible roles of an individual in a large organization, the number of tags a person can possibly use to describe a document, and so on. Given the fact that as the number of possible association types increase, the number of association type combinations between two individuals increase quadratically, the computational complexity for mutual information calculation can be particularly high for bipartite social networks with large number of non-metric associations. The increased complexity in computation might bring another barrier to the utilization of mutual information in social network analysis.

3.2 Limitations and Future Directions

Despite the difficulty discussed above, the current study is still a successful first step exploration of the application of this new method in social network analysis. The discussions of why mutual information is still not commonly used in the area of social network analysis also shed lights on several limitations of current study and future directions in improving the feasibility of using mutual information-based measures in bipartite social network data with non-metric associations.

First, both studies in this thesis examined a bipartite social network with binary associations. A critical step in future studies is to investigate the effectiveness of this method in datasets with more than two non-metric associations. Unlike correlation measures that cannot be generalized to more than two non-metric association types, mutual information can be simply generalized. Moreover, unlike the existing methods that only consider the co-occurrence of same association types, mutual information takes all possible combinations of association types into account. As a result, mutual information-based measures should be easily generalized and might be one of the most general methods to model relationship between nodes in all existing methods.

Second, as pointed out in the discussion above, computation complexity might be a significant barrier in applying mutual information-based measure in social network analysis. Moreover, the current study only examined social networks of small and moderate sizes. How the computation complexity would increase when applying this method to networks with a large number of nodes still needs to be investigated. Thus, future studies should be conducted to develop more efficient algorithms to overcome the computation complexity

when there are a large number of discrete association types and when there are a large number of nodes in the bipartite social network.

Third, one significant shortcoming of using mutual information-based measures is that the measure itself does not differentiate the reasons that causes the high interdependence between two nodes. That is, all combinations of association types between node pairs are treated equally in mutual information calculations. Thus, future studies should focus on how to differentiate between-node interdependence that are due to different reasons. As suggested in the two studies in the current thesis, introducing weighting systems can be helpful since different weights can be placed on different association combinations. The different goals of social network analysis should require different weighting systems since the association combinations of interest would vary. Thus, another important future direction is to investigate efficient ways for developing weighting systems that best support the goal of analysis.

Fourth, in the current study, other social network analysis techniques were introduced and combined with mutual information measures when inferring underlying social network structures. Although the techniques used in the current study were able to identify meaningful information from the mutual information-based measures, further studies are still needed to systematically investigate the compatibility of other social network analysis techniques with mutual information-based measures. In addition, the appropriateness of each technique should also vary depending on the goal of the analysis.

Fifth, as discussed in the previous section, mutual information-based measures have little tolerance on missing data. Since it is inevitable that social network data collected in

natural settings will have a significant amount of missing data, future studies should also focus on developing ways to make mutual information calculations applicable to datasets with missing values.

3.3 Conclusion

The major contribution of the current study is its exploratory nature of first applying a method into the area of social network analysis. Although the results were mixed and not all attempted analyses were successful, the current study pointed to several important directions of future studies to further refine this new method in social network analysis. The current study also brings the importance of learning and adopting useful methods across different research areas, which provides great insights in advancing the analysis techniques in social network analysis.

REFERENCES

- Aghagolzadeh, M., Soltanian-Zadeh, H., Araabi, B., & Aghagolzadeh, A. (2007). A hierarchical clustering based on mutual information maximization. In *Image processing, 2007. ICIP 2007. IEEE international conference on* (Vol. 1, pp. I-277–I-280).
- Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4), 590-614.
- Batagelj, V., & Mrvar, A. (2000). Some analyses of erdos collaboration graph. *Social Networks*, 22(2), 173-186.
- Borgatti, S. P. (2009). 2-mode concepts in social network analysis. In R. A. Meyers (Ed.), *Encyclopedia of complexity and systems science*. New York, NY: Springer.
- Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks*, 19, 243-369.
- Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet 6 for windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.
- Borgatti, S. P., & Halgin, D. S. (2011). Analyzing affiliation networks. In J. S. J. Scott & P. Carrington (Eds.), *The sage handbook of social network analysis* (chap. 28). Thousand Oaks, CA: Sage Publications.
- Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, 53(2), 181-190.

- Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5, 415-426.
- Chien, L., Tat, A., Proulx, P., Khamisa, A., & Wright, W. (2008). nSpace2 and geotime visual analytics. In *IEEE symposium on visual analytics science and technology* (p. 199-200). Columbus, OH: IEEE.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22-29.
- Correa, C. D., Crnovrsanin, T., Muelder, C., Shen, Z., Armstrong, R., Shearer, J., et al. (2008). Visual analytics of cell phone data using MobiVis and OntoVis. In *IEEE symposium on visual analytics science and technology* (p. 211-212). Columbus, OH: IEEE.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52), 22436-22441.
- Davis, A., Gardner, B. B., & Gardner, M. R. (1941). *Deep south: A social anthropological study of caste and class*. Chicago, IL: University of Chicago Press.
- Davis, G. F., & Greve, H. R. (1997). Corporate elite networks and governance changes in the 1980s. *American Journal of Sociology*, 103(1), 1-37.
- Davis, G. F., Yoo, M., & Baker, W. E. (2003). The small world of the american corporate elite, 1982-2001. *Strategic Organization*, 1(3), 301-326.

- Dawy, Z., Geobel, B., Hagenauer, J., Andreoli, C., Meitinger, T., & Mueller, J. C. (2006). Gene mapping and marker clustering using shannon's mutual information. *Transactions on Computational Biology and Bioinformatics*, 3(1), 47-56.
- DeJordy, R., Borgatti, S. P., Roussin, C., & Halgin, D. S. (2007). Visualizing proximity data. *Field Methods*, 19(3), 239-263.
- Doreian, P. (1979). On the evolution of group and network structure. *Social Networks*, 2(3), 235-252.
- Eagle, N., & Pentland, A. S. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), 255-268.
- Eagle, N., Pentland, A. S., & Lazer, D. (2008). Mobil phone data for inferring social network structure. In H. Liu, J. J. Salerno, & M. J. Young (Eds.), *Social computing, behavioral modeling, and prediction* (p. 79-88). Phoenix, AZ: Springer.
- Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobil phone data. *Proceedings of the National Academy of Sciences*, 106(36), 15274-15278.
- Farrugia, M., & Quigley, A. (2008). Animating multivariate dynamic social networks. In *IEEE symposium on visual analytics science and technology* (p. 215-216). Columbus, OH: IEEE.
- Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1, 215-239.
- Ghosh, S., Kane, P., & Ganguly, N. (2011). Identifying overlapping communities in folksonomies or tripartite hypergraphs. In *Proceedings of the 20th international conference companion on world wide web* (p. 39-40). New York, NY, USA: ACM.

- Huang, Y., Pei, J., & Xiong, H. (2006). Mining co-location patterns with rare events from spatial data sets. *GeoInformatica*, 10, 239-260.
- Hubert, L., Arabie, P., & Meulman, J. (2006). The structural representation of proximity matrices with matlab. In M. T. Wells (Ed.), *ASA series on statistics and applied probability*. SIAM, Philadelphia, PA, ASA, Alexandria, VA.
- IEEE. (2008). *IEEE VAST 2008 challenge detailed task descriptions for all challenges*. Available from <http://www.cs.umd.edu/hcil/VASTchallenge08/tasks.html>
- Kraskov, A., & Grassberger, P. (2009). MIC: Mutual information based hierarchical clustering. In F. Emmert-Streib & M. Dehmer (Eds.), *Information theory and statistical learning* (p. 101-123). Springer US.
- Kraskov, A., Stogbauer, H., Andrzejak, R. G., & Grassberger, P. (2005). Hierarchical clustering using mutual information. *Europhysics Letters*, 70(2), 278.
- Kvalseth, T. O. (1987). Entropy and correlation: Some comments. *IEEE Transactions on Systems, Man and Cybernetics*, 17(3), 217-219.
- Li, P., & Church, K. W. (2007). A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics*, 33(3), 305-354.
- Li, W. (1990). Mutual information function versus correlation functions. *Journal of Statistical Physics*, 60(5), 823-837.
- Lin, Z., & Lim, S. (2008). Fast spatial co-location mining without cliqueness checking. In *Proceeding of the 17th acm conference on information and knowledge management* (p. 1461-1462). New York, NY, USA: ACM.
- Liu, X., & Murata, T. (2009). Community detection in large-scale bipartite networks. In *International conference on web intelligence and intelligent agent technology - workshops* (p. 50-57). Milano, Italy: IEEE/WIC/ACM.

- Mardenfeld, S., Boston, D., Pan, S. J., Jones, Q., Iamntichi, A., & Borcea, C. (2010). Gdc: Group discovery using co-location traces. In *Social computing (SocialCom), 2010 IEEE second international conference on* (p. 641-648). Minneapolis, MN: IEEE Computer Society.
- Mariolis, P. (1975). Interlocking directorates and control of corporations: The theory of bank control. *Social Science Quarterly*, 56(3), 425-439.
- Mathematica 8.0.1.* (2010). Champaign, IL: Wolfram Research, Inc.
- Matlab 7.12.0.635 (R2011a).* (2011). Natick, MA: The Mathworks, Inc.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415-444.
- Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36, 363-377.
- Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213-238.
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404-409.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167-256.
- Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5200-5205.
- Payne, J., Solomon, J., Sankar, R., & McGrew, B. (2008). Palantir: The future of analysis. In *IEEE symposium on visual analytics science and technology* (p. 201-202). Columbus, OH: IEEE.

- Pellegrino, D., Pan, C.-C., Robinson, A., Stryker, M., & Luo, J. (2008). Visualization and collaboration in the VAST 2008 challenge. In *IEEE symposium on visual analytics science and technology* (p. 197-198). Columbus, OH: IEEE.
- Perer, A. (2008). Using SocialAction to uncover structure in social networks over time. In *IEEE symposium on visual analytics science and technology* (p. 213-214). Columbus, OH: IEEE.
- Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4), 581-603.
- Schifanella, R., Barrat, A., Cattuto, C., Markines, B., & Menczer, F. (2010). Folks in folksonomies: Social link prediction from shared metadata. In *International conference on web search and data mining* (p. 271-280). New York, NY: ACM.
- Seretan, V., & Wehrli, E. (2006). Multilingual collocation extration: Issues and solutions. In *Workshop on multilingual language resources and interoperability* (p. 40-49). Sydney, Australia: ACL.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423, 623-656.
- Swing, E. (2008). Solving the cell phone calls challenge with the Prajna project. In *IEEE symposium on visual analytics science and technology* (p. 221-222). Columbus, OH: IEEE.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York, NY: Cambridge University Press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442.
- Ye, Q., Zhu, T., Hu, D., Wu, B., Du, N., & Wang, B. (2008). Exploring temporal communication in mobil call graphs. In *IEEE symposium on visual analytics science and technology* (p. 207-208). Columbus, OH: IEEE.

Yoo, J. S., Shekhar, S., Smith, J., & Kumquat, J. P. (2004). A partial join approach for mining co-location patterns. In *Proceedings of the 12th annual acm international workshop on geographic information systems* (pp. 241–249). New York, NY, USA: ACM.