MULTI-MODE AND EVOLUTIONARY NETWORKS

by

Walid K. Sharabati

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Sciences and Informatics

Committee:

Edward J. Wegman	Dr. Edward J. Wegman, Dissertation Co-Director
Yasmin Said	Dr. Yasmin H. Said, Dissertation Co-Director
Robert Axtell	Dr. Robert Axtell, Committee Member
Igor Griva	Dr. Igor Griva, Committee Member
Tim Sauer	Dr. Tim D. Sauer, Committee Member
	Dr. Maxim Tsvetovat, Committee Member
D. Papaconstantopoulos	Dr. Dimitrios Papaconstantopoulos, Department Chairperson
P. Becker	Dr. Peter Becker, Associate Dean for Graduate Programs, College of Science
Vikas Chandhoke	Vikas Chandhoke, Dean, College of Science

Date: October 31, 2008

Fall Semester 2008
George Mason University
Fairfax, VA

Multi-Mode and Evolutionary Networks

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Walid K. Sharabati
Master of Science
Minnesota State University, 2003
Bachelor of Science
Bethlehem University, 1998

Co-Director: Dr. Edward J. Wegman, Professor
Co-Director: Dr. Yasmin H. Said
Department of Computational and Data Sciences

Fall Semester 2008
George Mason University
Fairfax, VA

Dedication

To The Memory Of My Mother

Acknowledgments

The assistance and support of many people, who in a way or another, helped make this possible. Above all, I am in debt to my advisor, mentor and teacher Professor Edward J. Wegman for his help, thoughtful ideas and guidance. I specially thank Dr. Yasmin H. Said for her constant support, critical feedback and encouragement throughout the different stages of this research. I would like to acknowledge my advisory committee, Dr. Tim D. Sauer, Dr. Robert L. Axtell, Dr. Igor Griva, Dr. Maxim Tsvetovat for the fruitful discussions, their time, critique and comments. I would like to express my sincere appreciation to my dear friend Dr. Abdallah Sayyed-Ahmad for his constant support and moral boosting. I would like to express my deepest gratitude to my uncles Ishaq Haddad, Dr. Maher Haddad and their families, and my brother-in-law Muhannad for welcoming me in their homes. A genuine and heartfelt appreciation is extended to my in-laws specially my mother-in-law, Khitam, who has been caring for me like a mother. Finally, I would like to thank my father and family who have been tremendously supportive and caring throughout the years.

Table of Contents

List of Tables

List of Figures

Abstract

MULTI-MODE AND EVOLUTIONARY NETWORKS
Walid K. Sharabati, PhD
George Mason University, 2008
Dissertation Co-Director: Dr. Edward J. Wegman
Dissertation Co-Director: Dr. Yasmin H. Said

In this dissertation, I present advanced mathematical methods underpinning networks, graphs and matrices. I develop a methodology to manipulate multi-mode high-dimensional networks and operate a mechanism for storing and performing matrix arithmetics on such networks and graphs. Additionally, I introduce the concept of having infinite networks and matrices and expand the literature involving traditional networks and matrices. Furthermore, I build up a model to estimate missing edges and vertices in a graph using covariate information and similarities among actors. The covariates are the exogenous attributes of entities, which could be numerical as well as categorical attributes. The model can be applied to social networks in addition to other networks. I then utilize the mathematical model to estimate missing vertices in a graph, a process that can be achieved through matrix transformation.

In the next stage, I present a method to predict the emergence of new actors in a network based on stochastic processes and suggest a model of preferential attachment. Finally, I apply quantitative methods to examine evolving networks.

Ultimately, I examine the structure of real networks and model their behavior. I perform a comprehensive analysis and simulation on applications in the social networks field, which includes coauthorship social networks (social networks of coauthors of scholarly publications), road fatal crashes networks in the United States, and news documents networks.

Chapter 1: Introduction

Social Network Analysis (SNA) or Network Theory is becoming important tools to analyze, model, and simulate the behavior of groups of people or entities both on the global level (how two or more groups interact with other group(s)) and on the local level (how individuals interact with each other within the same network.) In the past two decades, SNA has been used to analyze relations and ties among individuals of the same network and similarities between different networks to obtain a better understanding on how societies interact.

Social Network Simulation (SNS) is the branch of social networks that involves building, running and simulating artificial social networks. The fundamental component of network simulation is a set of homogeneous agents (actors) together with their individual properties and tasks [53]. Theses agents interact according to behaviorial rules set forth by the programmer, which resemble, in a simplified form, real world rules.

1.1 What is a Social Network?

A social network is an emerging tool frequently used in quantitative social science to understand how individuals or organizations are related. The basic mathematical structure for visualizing the social network is a graph. A graph is a pair V, E where V is a set of nodes or vertices and E is a set of edges or links. Social network analysis (also called network theory) has emerged as a key technique and a topic of study in modern sociology, anthropology, social psychology and organizational theory. The shape of the social network

helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More "open" networks, with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties.

Social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages. The social entities are referred to as actors that are represented by the vertices of the graph. Most social network applications consider a collection of actors that are all of the same type. These are known as one-mode networks. Social ties link actors to one another. The range and type of social ties can be quite extensive. A tie establishes a linkage between a pair of actors. Linkages are represented by edges of the graph. Examples of linkages include the evaluation of one person by another (such as expressed friendship, liking, respect), transfer of material resources (such as business transactions, lending or borrowing things), association or affiliation (such as jointly attending the same social event or belonging to the same social club), behavioral interaction (talking together, sending messages), movement between places or statues (migration, social or physical mobility), physical connection (a road, river, bridge connecting two points), formal relations such as authority and biological relationships such as kinship or descent. A linkage or relationship establishes a tie at the most basic level between a pair of actors. The tie is an inherent property of the pair. Many kinds of network analysis are concerned with understanding ties among pairs and are based on the dyad as the unit of analysis.

A network consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a significant feature of a social network. A partition of a network is a classification or clustering of the vertices in the network so that each vertex, sometimes called node, is assigned to exactly one class or cluster. Partitions may specify some property that depends on attributes of the vertices. Partitions divide

the vertices of a network into a number of mutually exclusive subsets. That is, a partition splits a network into parts. Partitions are also sometimes called blocks or block models. These are essentially a way to cluster actors together in groups that behave in a similar way.

In a network setting, actors or entities have several attributes to identify their role, behavior, background, and/or assets; some of which are unique to that actor and some are common among other actors. These attributes are the nodes' properties such as gender, age, political affiliation, ethnicity, race, nationality, religion, spoken languages, scientific field, income, education level, job class, and geographic location.

1.2 Preliminaries

Networks can be treated as directed graphs in which actors (individuals) are represented by vertices (nodes) while interactions between actors are represented by edges (ties), which may have weights. There are three basic representations of a network – the planar graph visualization, the adjacency matrix, and the sparse-graph representation.

There are several algorithms to study interactions within the network include centrality measures (vertex degree and closeness), network partitioning (cliques and clique overlapping), network connectivity (cut-points and bridges), structural equivalence, structural holes, brokerage roles and block-modeling, which will all be defined shortly.

Definition 1.1. A graph G, is a collection of vertices V and edges E; $G = \{V, E\}$, where $V = \{v_1,\ v_2,\ v_3,\ \cdots,\ v_i,\ \cdots,\ v_{|V|}\}$ and $E = \{e_1,\ e_2,\ e_3,\ \cdots,\ e_l,\ \cdots,\ e_{|E|}\}$.

Definition 1.2. An adjacency matrix A associated with a graph G, is a matrix of size

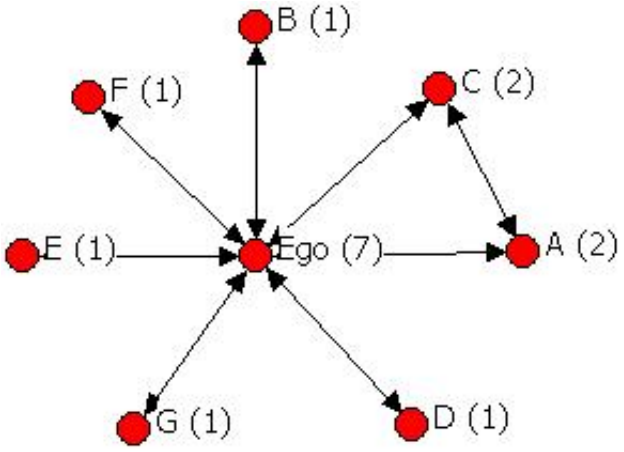


Figure 1.1: Example of a simple ego-network. Node Ego has degree = 7, vertices A and C have both degree = 2, vertices B, D, E, F and G all have degree = 1.

$|V| \times |V|$ and whose elements a_{ij} are

$$a_{ij} = \begin{cases} 1, & \text{if } \exists \text{ an edge } e_l \text{ connecting vertices } v_i \text{ and } v_j. \\ 0, & \text{if } \nexists \text{ an edge } e_l \text{ connecting vertices } v_i \text{ and } v_j. \end{cases}$$

$1 \leq i, j \leq |V|$ and $1 \leq l \leq |E|$.

1.2.1 Centrality Measures

There are three main centrality measures defined in [61]; namely, degree centrality, closeness centrality and betweenness centrality. To serve the purposes of this research, I will define degree and closeness centrality measures only. Degree of a vertex is the number of edges that connect it to other vertices, see Figure 1.1. Degree can be interpreted as measure of power or importance of a vertex, or measure of workload. The actor with most ties is the most important figure in a network. It has been shown that in a simple random graph, degree centrality has the Poisson distribution. Nodes with high degree are likely to be at the intuitive center. Deviations from a Poisson distribution suggest non-random processes, such processes form "scale-free" networks.

Definition 1.3. Degree of a vertex v_i; denoted $d(v_i)$, is defined by

$$d(v_i) = \sum_{j=1}^{|V|} a_{ij}, \quad a_{ij} \in A, \tag{1.1}$$

where d represents degree measure, and A is the adjacency matrix.

Closeness centrality measure is based on the inverse of the distance of each actor to every other actor in the network. Distance in this context is defined to be the number of steps a vertex v_i needs to reach a vertex v_j. If an actor is close to all other actors then this actor is considered important.

Definition 1.4. Closeness; denoted $c(v_i)$, is defined by

$$c(v_i) = \left[\sum_{j=1}^{|V|} d(v_i, v_j) \right]^{-1}, \tag{1.2}$$

where c represents closeness, $d(v_i, v_j)$ is the shortest distance between v_i and v_j.

Definition 1.5. The geodesic is the length of the shortest path between any two vertices.

1.2.2 Cohesive Sub-Groups: Cliques

Definition 1.6. A dyad is a pair of vertices and the edge connecting them.

Definition 1.7. A triad is a set of three vertices and the edges connecting them.

A triad is identified by a M-A-N number system of three digits and a letter; for more details refer to [14].

One of the interesting features in a network that caught structural analysts' attention is secondary sub-structures such as network cohesion. Researchers interested in cohesive subgroups gathered and studied sociometric data on affective ties in order to identify "cliquish" subgroups (face-to-face group). The clique is the foundational idea for studying and analyzing cohesive subgroups in social networks.

Definition 1.8. A clique in a graph is a maximal complete subgraph of three or more nodes, mutual dyads (2 nodes) are not considered to be cliques [61].

It consists of a subset of vertices all of which are adjacent to each other, and there are no other vertices that are also adjacent to all of the members of the clique. A clique is a very strict definition of cohesive subgroups. Cliques are a subset of the network in which the actors are more closely and intensely tied to one another than they are to other members of the network and if one actor disappears for any reason, the other two can still write/talk to each other. As an illustration, in Figure 1.1, the vertices {Ego, A, C} form a clique.

1.2.3 Structural Equivalence

Definition 1.9. Two actors are structurally equivalent if they have the same type of ties to the same people.

Identifying structurally equivalent actors can be done through the method of partitioning actors into subsets so that actors within each subset are closer to being equivalent than are actors in different subsets. One way to display the results of a series of partitions is to construct a dendrogram indicating the degree of structural equivalence among the positions and identifying their members. Each level of the diagram indicates the division resulting from a split of the previous subset [61]. A dendrogram thus represents a clustering of the

actors, those actors who are connected by branches low in the diagram are closer to being perfectly structurally equivalent, whereas subsets of actors who are joined only through paths high up the diagram are less structurally equivalent (or are not equivalent at all). In brief, the lowest position in the diagram indicates that every actor is different while the highest position indicates that all actors are the same; what is in between is more important in terms of structural equivalence.

1.2.4 Blockmodel

Definition 1.10. A blockmodel is the process of identifying positions in the network. A block is a section of the adjacency matrix "a group of people" structurally equivalent. It consists of two things according to Wasserman and Faust [61]:

- A partition of actors in the network into discrete subsets called positions.
- For each pair of positions a statement of the presence or absence of a tie within or between the positions on each of the relations.

A blockmodel is thus a hypothesis about a multirelational network. It presents general features of the network, such as the ties between positions, rather than information about individual actors.

A blockmodel is a simplified representation of multirelational network that captures some of the general features of a network's structure. Specifically, positions in a blockmodel contain actors who are approximately structurally equivalent. Actors in the same position have identical or similar ties to and from all actors in other positions. Thus, the blockmodel is stated at the level of the positions, not individual actors.

1.2.5 Structural Holes

Structural holes provide important information about the structure of a socio-network, they have low redundancy and cause stress because there are too many vertices connected to the brokerage. The basic form of structural holes is a triad with one edge missing, in which two actors communicate with the same person, but do not communicate with each other.

The purpose of equivalence analysis is to identify and visualize "classes" or clusters. In cluster analysis, it is implicitly assumed that the similarity or distance among actors reflects as single underlying dimension. It is possible, however, that there are multiple "aspects", "attributes" or "dimensions" underlying the observed similarities of cases. Components analysis could be applied to correlations among actors. Alternatively, MDS (Multi-Dimensional Scaling) could be used (metric for data that are inherently valued) to cluster the actors.

MDS represents the patterns of similarity or dissimilarity in the tie profiles among the actors (when applied to adjacency or distances) as a "map" in multi-dimensional space. This map lets us see how "close" actors are, whether they "cluster" in multi-dimensional space, and how much variation there is along each dimension. The aim of MDS is to minimize stress – distance between vertices. "Stress" is a measure of badness of fit; $0 \leq \text{stress} \leq 1$. In MDS, we look at a range of solutions with more dimensions, so we can assess the extent to which the distances are uni-dimensional. The "meaning" of the dimensions can sometimes be assessed by comparing agents that are at the extreme poles of each dimension.

1.3 Motivation

There are several applications to social network analysis in the areas of science and technology, this includes co-authorship social networks (social networks of coauthors of scholarly publications,) alcohol user social networks (or alcohol ecology networks,) covert and espionage networks, terrorist networks, disease social networks, and computer social networks.

1.3.1 Co-authorship Social Networks

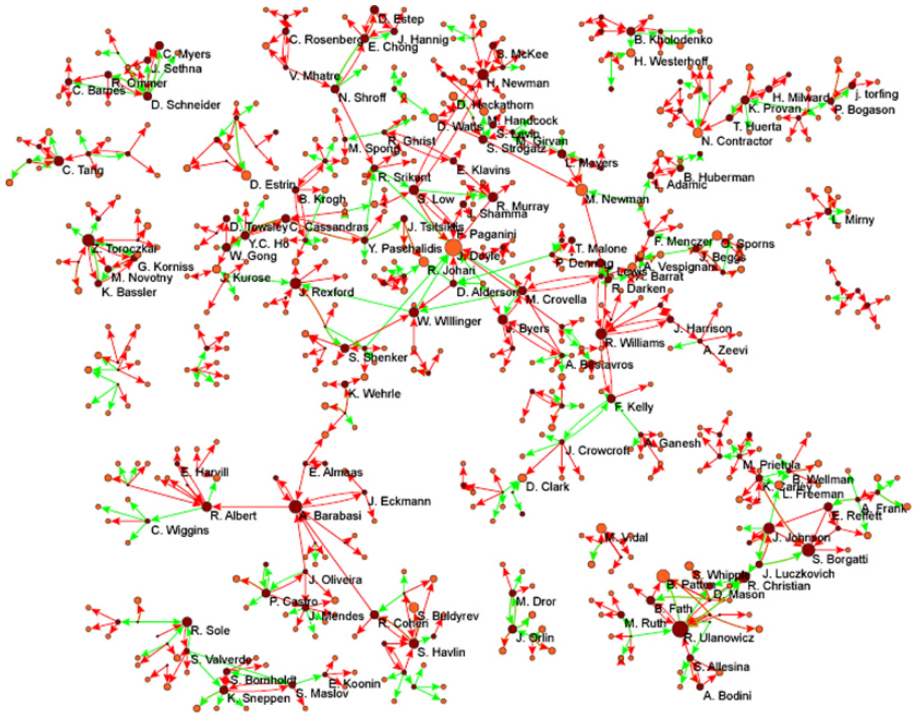


Figure 1.2: Sample author-coauthor social network [5].

Scholarly publication is considered a vital aspect in academia both for faculty members and researchers. Authors have many incentives to publish. For one reason it is prestigious. For another, authors of scholarly publications get financial compensation through research grants as well as promotions. Different disciplines and individuals have evolved distinguishable mechanisms for coping with the publication pressures [56]. Co-author social networks can reveal information on the style of co-authorship, which can be summarized as solo or no co-authors models, mentor models, entrepreneurial models, and team models. These styles are made evident clustering members of the network. I conjecture based on two papers published recently [55, 56] that certain styles of co-authorship lead to the possibility of group-think, reduced creativity, and the possibility of less rigorous reviewing processes. Of all the work that has been done on social networks, very few scientists had considered

coauthorship networks. The main goal of analyzing coauthorship networks is to be able to answer the question of "who-wrote-with-whom" and with what frequency. A sample social network of coauthors [5] is depicted in Figure 1.2.

The mathematical model for estimating missing edges in this scenario can reveal if two authors have worked together at some point even though they have not published based on their similarities in the field, geographic location, language, and school. It is worth mentioning that relationship among coauthors is generally symmetric, with the possible exception of the case when a distinction of leading coauthor and other coauthors is made. A symmetric relationship means that if author A published with author B, then this also implies that author B published with author A. This can be represented by a directed graph through the relations $A \overset{\text{published with}}{\longleftrightarrow} B$.

1.3.2 Covert and Espionage Social Networks

In covert social networks of both individuals and organizations such as gang networks, smuggling networks, alliances networks, analysts seek information about missing edges and actors in addition to key figures in the network. The method I propose in this dissertation can be used to estimate the probability of missing elements in a network taking into account edge dependency. Individuals and organizations strive to suppress their identities and interactions and try not to divulge any information of any kind to non-members. It is their interest to give the impression that there is a missing member or linkage/connection in the network, so that the network would look incomplete or dysfunctional to the outsider.

In such networks, analysts would want to predict the role of the members in the network who are intentionally hiding their roles and tasks. These networks can be divided into two subnetworks (the support network and active cell network). The study of the network's structure over time provides information on the formation of the network as well as the emergence of vertices (actors) and ties (linkages).

1.3.3 Alcohol Ecology Social Networks

The alcohol ecology social network is yet another important type to study and research because the outcome of heavy alcohol users has direct impact on the society. Alcohol abuse leads to acute outcomes that are violence related leading to injuries, assault, domestic violence, child abuse, sexual assault, murder, DWI, fatal crashes, violent crime, death due to trauma, juvenile violence, crime associated with drugs and deaths due to suicide as stated in [53]. Thus, the need for a social network model for ecological alcohol systems is needed.

The alcohol system involves the complex interactions among users, their family and peers, non-users, producers and distributors of alcohol products, law enforcement, courts, prevention activities, and treatment centers [53]. Figures 1.3 and 1.4 present a bipartite social network and matrix respectively of alcohol users and institutions. The interactions are asymmetric. For example, alcohol user may deal with one law enforcement officer, while an officer may deal with many alcohol users. Finally, Figure 1.5 shows the bipartite network of the ecological alcohol system.

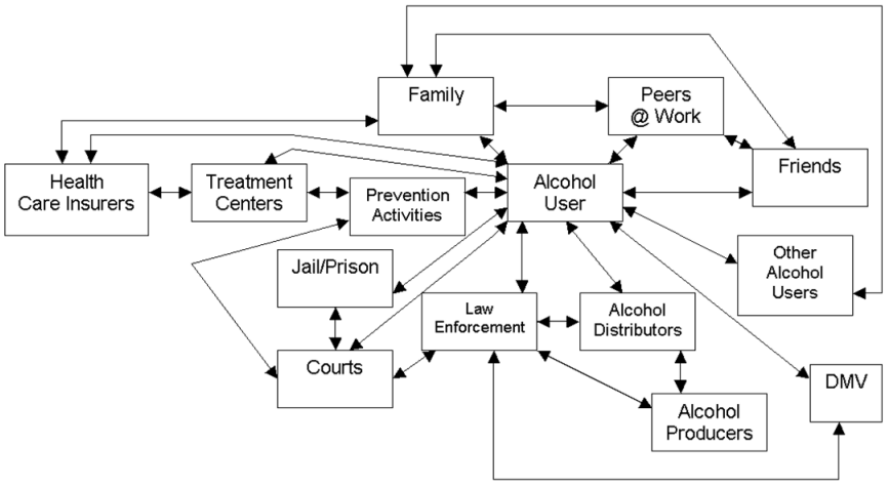


Figure 1.3: A two-mode graph for social network of alcohol users and institutions [62].

The analysis and modeling of alcohol user social networks may provide a policy tool for

	Alcohol User	Family	Peers @ Work	Friends	Other Users	Alcohol Distributors	Alcohol Producer	Law Enforcement	Courts	Prison	Rehab	Treatment	Health Insurance	DMV
Alcohol User		X	X	X	X	X		X	X	X	X	X	X	X
Family	X		X	X	X								X	
Peers @ Work	X	X		X										
Friends	X	X	X											
Other Users	X	X												
Alcohol Distribution	X						X	X						
Alcohol Producers						X		X						
Law Enforcement	X					X	X		X				X	
Courts	X							X		X	X			
Prison	X								X					
Rehab	X								X			X		
Treatment	X										X		X	
Health Insurance	X	X										X		
DMV	X							X						

Figure 1.4: A two-mode "bipartite" adjacency matrix for social network of alcohol users [62].

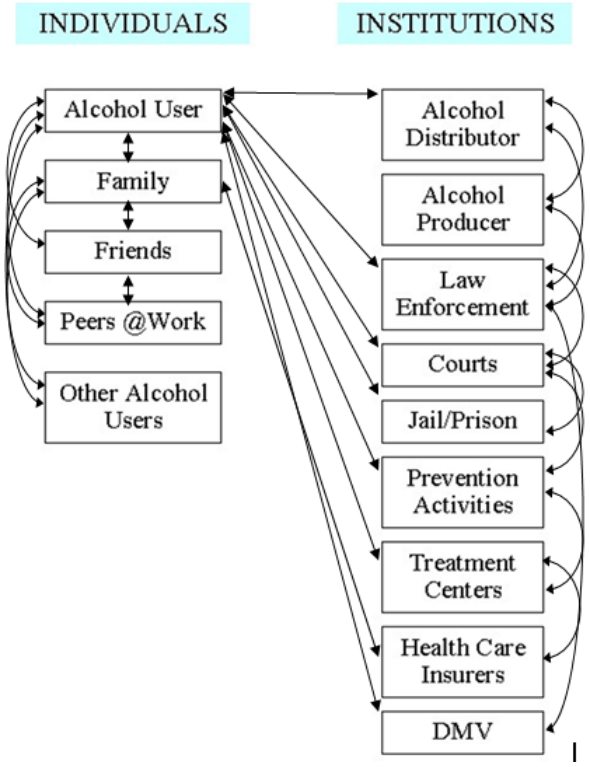


Figure 1.5: A two-mode social network of alcohol users [62].

examining the effects of interventions and encourage policy decision makers and law enforcement implement new rules for alcohol consumption for a more sophisticated and safer society. For instance, limiting alcohol usage to a certain individuals and places or even not allowing drinking at all in certain areas can reduce drastically the negative consequences of alcohol abuse. Policies should also involve courts, prevention activities and treatment facilities. It is of great importance to investigate the evolution of alcohol user social networks. Figure 1.6 shows a graph model for interventions. As suggested by [53], acute outcomes may occur any time in the day but the likelihood changes during the day. The social network probability changes during the day, week, and month depending on the circumstance.

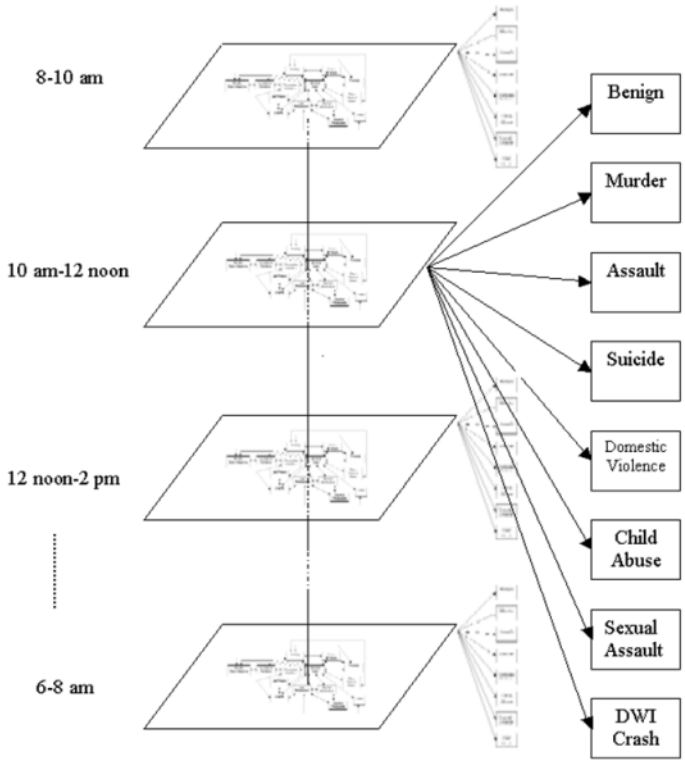


Figure 1.6: Graph model for interventions [62].

The terminal vertices in the one-mode alcohol user networks represent either acute or benign outcomes. An important concept in the work is to realize that suppression of one acute

outcome could increase the probability of another. Extra policing of off-license alcohol outlets may reduce assaults in the vicinity, but could lead to an increase in DWI and domestic violence. The goal is the simultaneous suppression of acute outcomes.

1.3.4 Computer Social Networks

There is an exponentially increasing interaction between the computers/Internet and people. Computer social networks link people to organizations in a bipartite asymmetric relationship. The acute outcomes related to computer intrusion include a clog in the Internet with junk email, viruses, worms, trojans, and major fraud risks. For example, if a computer attack happens, both private and public sectors are subject to disruption including E-commerce, critical military command and control functions, telecommunications, supply chains, and ordinary commerce.

1.3.5 Disease Social Networks

The evolving disease social network network of individuals and diseases may be represented by dynamic bipartite graphs. It is constantly changing and individuals and diseases continuously move and change locations. The contact network of individuals is a heavily connected graph. The major issue is the uncontrolled disease propagation, which raises an important question: How can several cases of disease cases be contained before becoming an epidemic? In such networks, the knowledge of degree distribution and clustering is important for local propagation. Yet, global propagation can be deduced from the general structure of the graph.

1.4 Problem Statement

1.4.1 Relational Networks

Some social network relationships can be treated as a two-mode "bipartite" networks, or

Table 1.1: The PCANS model.

	Person	Resource	Task
Person	N 1-mode	S 2-mode	A 2-mode
Resource			C 2-mode
Task			P 1-mode

three-mode "tripartite" networks. As an example, consider the author-paper networks, there are two types of vertices, one class of vertices represents authors while the other represents papers. There is one relationship type; "person A authored/coauthored paper P". This two-mode relational socio-network can be concluded from the PCANS model [34], [8]. Table 1.1 portrays the PCANS model.

I can perform matrix operations such as the product of matrices to obtain interesting results given that the two-mode matrix is binary. Let the two-mode "author-by-paper" binary social matrix AP be given, then

$$AP \times AP^T = AP \times PA = AA,$$

is the one-mode network of authors related through papers. Similarly,

$$AP^T \times AP = PA \times AP = PP,$$

is the one-mode network of papers related through authors.

The author-by-author social matrix AA is one of interest, it reveals relationships among authors, in other words, the author-by-author matrix resembles the "who-wrote-with-whom" relationship.

One of the issues I address in this dissertation is how to produce a one-mode relational network from a weighted two-mode matrix; the traditional matrix matrix-transpose multiplication fails to give meaningful values as the product generates perfect squares.

Consider a bipartite "coauthor-by-paper" social network. Let A be the adjacency matrix of size $m \times n$ representing the graph of the network, with $m =$ number of coauthors, and $n =$ number of papers. Then,

$$C_{m\times m} = A_{m\times n} \cdot A^T_{n\times m} = \text{ coauthorship proximity matrix, and}$$

$$P_{n\times n} = A^T_{n\times m} \cdot A_{m\times n} = \text{ paper-by-paper proximity matrix.}$$

where,

$$c_{ii} = \sum_{j=1}^{n} a_{ij} = \text{number of papers author } i \text{ published,}$$

$$p_{jj} = \sum_{i=1}^{m} a_{ij} = \text{number of coauthors coauthored paper } j, \text{ and}$$

$$c_{ij} = \text{tie-strength between coauthors } i \text{ and } j.$$

Finally, if $D_{m\times m} = C^2_{m\times m}$ then

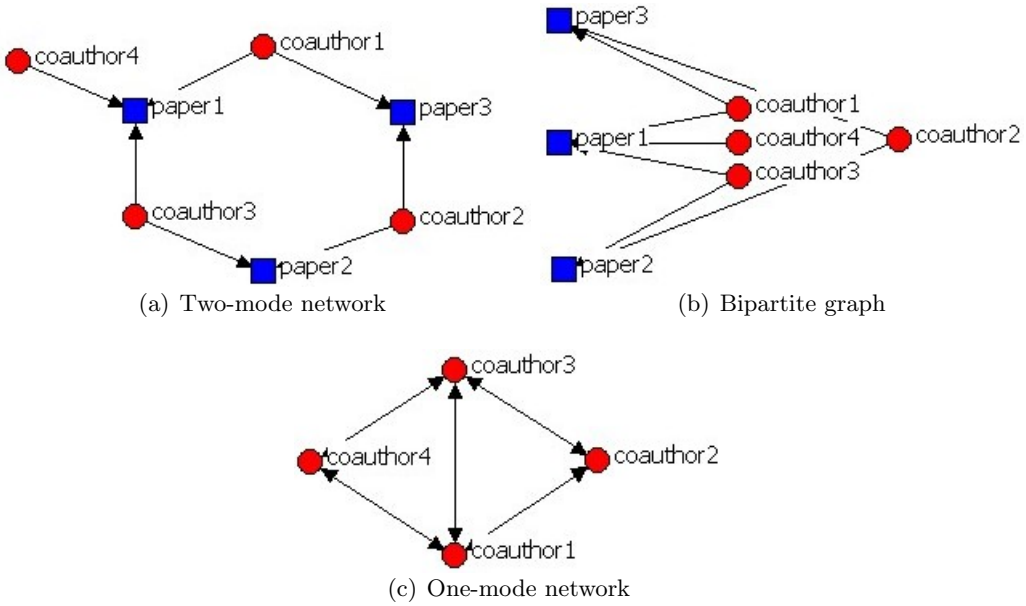
$$d_{ii} = \text{ vertex degree of coauthor } i.$$

Example 1. Suppose the coauthor-by-paper adjacency matrix A is given by

	paper1	paper2	paper3
coauthor1	1	0	1
coauthor2	0	1	1
coauthor3	1	1	0
coauthor4	1	0	0

$$\Rightarrow A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}_{4\times 3}$$

There is a one-to-one correspondence between the matrix representation of a social network and directed graphs. Graphical representations of the above matrices are depicted in Figures 1.7 and 1.8

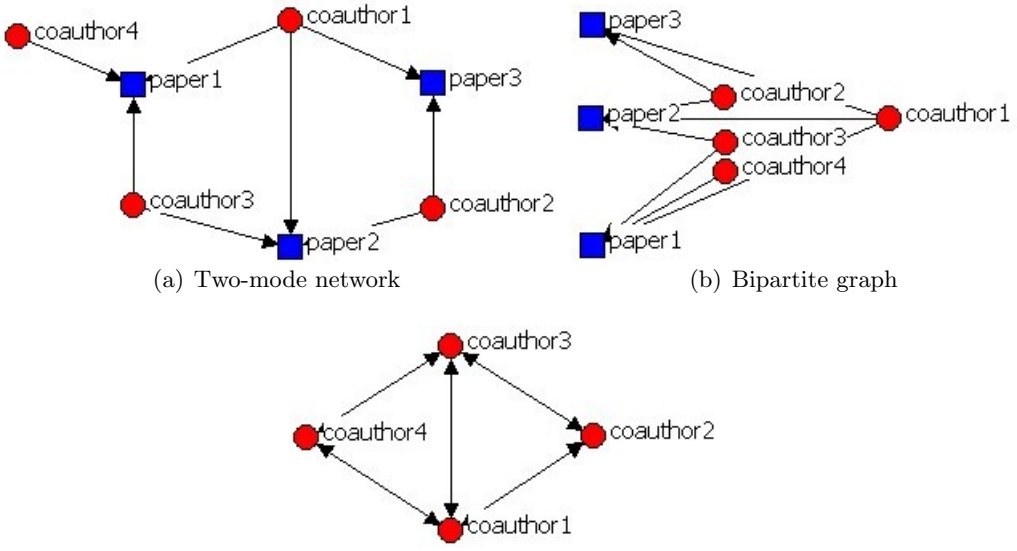


(a) Two-mode network (b) Bipartite graph

(c) One-mode network

Figure 1.7: Example of a co-authorship network

If we carefully examine the networks in Figures 1.7(a) and 1.8(a) we observe that these different 2-mode networks in fact have the same 1-mode graphical network representation, see Figures 1.7(c) and 1.8(c). This is due to the fact that when converting to 1-mode some network features are lost; the same effect when someone projects from 3-D to 2-D. This is an example of how the 1-mode network does not provide sufficient answer of how peer-ties



(a) Two-mode network (b) Bipartite graph

(c) One-mode network

Figure 1.8: Example of a co-authorship network

are formed. As a result, the one-mode network should be constructed from the two-mode and the two-mode network should the primary source to preserve any network features. This is important when cliques are present and one needs to determine which members formed which clique. The entrepreneurial and laboratory styles of coauthorship networks are different styles [55, 56], yet the blockmodel of the 1-mode network identifies both as one style. The blockmodel does not show how cliques were formed. The ultimate solution to this problem is to consider the weighted adjacency matrix as opposed to the binary adjacency and then construct the distributions of dyads, triads, tetrads, pentads, hexads, heptads, and octads. In section 2.3, I adopt a matrix representation that helps keep track of cliques at each time step.

Throughout this research I have encountered multidimensional networks some of which are weighted (not dichotomous) networks in which the traditional product of the matrix and its transpose fails to provide a meaningful outcome. I have invented in Chapter 2 a mechanism for manipulating multi-mode networks and developed a method to generate the one-mode network from a weighted two-mode network. The PCANS model presented above

is only a special case of a multidimensional network.

1.4.2 Covariate Information and Missing Edges

This research study provides a tool to model the interaction and co-relations between actors in social and other networks using covariate information defined on actors or entities. In this regard, I adopt a theoretical approach to analyze networks quantitatively and use advanced mathematical concepts to address certain issues related to networks and then present results and visualization of sample real networks.

Generally speaking, edges (ties) and vertices (nodes) in real networks may be imperfectly observed. This could be due to bad sampling, undercoverage, or more importantly, actors intentionally attempt to hide their roles or linkages to serve different purposes, which makes the network incompletely observed and difficult to monitor the behavior of its members and analyze its structure.

In some social networks such as covert and alliances networks, knowing whether there is a missing edge may be of interest. Actors in such networks strive to hide their identities giving the illusion there is a broken link. The method of estimating missing edges addresses this issue; it is based on using nodes' attributes to measure the level of similarity among vertices. If two vertices are very similar then it is more likely that they are connected or there is a strong potential for a link in the future; however, if they are very dissimilar then most likely these two vertices are not connected.

Thus, the objective is to predict the unobserved vertices and edges in incompletely observed social network using actors' attributes. Each entity has a vector of covariate information based on the external structure of the network; which can be used to derive pairwise similarity measures among actors. Conceptually, this is done through applying the inner product on the vector of attributes.

The probability of an edge between vertex v_i and vertex v_j is then the estimate $\mathcal{P}(\widehat{E(v_i, v_j)}) = \mathcal{S}(v_i, v_j)$, where $\mathcal{S}$ is the similarity measure. The basic idea of similarity, is to have a quantitative measure of those attributes that intersect, actors with more shared attributes are more likely to be similar. Thus, an edge with high probability implies the two actors are very similar and that the link may in fact exist, and if otherwise not, then this is an indication of a missing potential edge or there is a high potential to form a link in the future, which is in both cases a useful information. This helps predicting potential edges in the imperfect graph more accurately and offers valuable information to analysts to disambiguate any unknown relations among actors.

It is possible for the covariates to carry categorical attributes including nominal and ordinal. Ordinal values may be treated as discrete interval scale with no problem; however, nominal data may be quantified by introducing a dummy variable or an indicator variable. For example the variables "age" and "income" are continuous that take numerical values, "gender", "discipline", "spoken languages", "ethnicity" are all nominal, and "political affiliation", "degree" are considered ordinal. For these types of variables, I use contingency tables and the χ^2-test to obtain the similarities. The level of similarity between two given vertices is proportional to all common attributes these vertices have.

1.4.3 Evolutionary Networks

The next topic that needs to be addressed is the continuously changing networks. The purpose of studying the evolution of networks over time is, based on the current and previous structures of the network, to predict and simulate the future behavior of the network as a whole (macro/global level analysis) and evaluate the performance of each individual (micro/local level analysis.) One great advantage of observing and analyzing time series networks is the fact that analysts are able to monitor the process of introducing new actors,

edges, and roles as well as keep track of the changes, bookkeep and watch the status of the current actors and their roles. My approach helps identifying similar networks and similar figures in a network as well as recognizing "elite" subgroups within a network.

The mathematical model I propose to analyze evolutionary networks makes use of the similarity measures obtained on actors using covariate information. The measures are used to predict the behavior of evolving networks as well as predicting emerging groups within networks. The model essentially utilizes transition probabilities in a finite state stochastic process in discrete time. The status of the network at current state depends upon previous network settings in accordance to a Markov process. The sequence of states is time dependent and recursive. Thus, the transition adjacency matrix at time $t+1$ can be expressed as a function of the previous transition adjacency matrix at time t.

1.4.4 Simulated Social Networks

To conclude, I present a mathematical model of preferential attachment in coauthorship socio-networks. The process of one actor (new or existing) attaching to another actor and strengthening ties over time is a stochastic random process. The distributions of tie-strength and clique-size are derived from empirical data and utilized to determine the low level processes. The mathematical algorithm is then implemented using Agent Based Simulation model to simulate a coauthorship network. An agent is a computer representation of a human or other entity together with rules of behavior [53]. An agent follows these rules in interaction with other agents. These rules are usually probabilistic in nature. If many agents are introduced into the system, the general behavior of the system may be simulated.

1.5 Literature Review

1.5.1 Predicting Unobserved Edges

Marchette and Priebe [41] suggested a model to predict unobserved edges in incompletely observed networks. They used the constrained random dot product graph (CRDPG) model and covariates measured on actors to rank potential edges according to a probability that they are in fact present based on the internal network structure. It is the assumption sometimes that vertices and edges in an observed social networks are fully known and accurate; i.e. the network is well-observed.

Unless there is a complete database on the researched social network, it is almost impossible to have a clear picture of the entire network (actors, edges, and members' roles) regardless of the type of the network, either due to sampling problems (either because of missing data or undercoverage) or; more importantly, because actors attempt to suppress their existence and role in the network.

The model is based on the Erdös-Renyí random graph and runs by fitting an CRDPG to an observed network, and then rank potential edges according to the probabilities induced by the estimated model.

If $x \in \mathbb{R}^d$ is a vector assigned to vertex v. The conditional probability of an edge from v_i to v_j is a function of the dot product of the vectors:

$$P[v_i v_j \in E | x_i, x_j] := p_{ij} = f(x_i \cdot x_j),$$

where f is a simple threshold:

The vectors x_i correspond to latent variables. The iterative algorithm computes the maximum likelihood estimate and if there are missing values (edges) in the adjacency matrix, they are replaced with the estimated probability of the edge. For more details about the CRDPG algorithm, refer to [41]. It is illustrated on a dataset of alliances between 173 nations collected from 1816 to 2000.

The difference between this model and what I suggest in this dissertation though is the fact that; first of all, my method utilizes quantitative and qualitative covariates to estimate the probability of missing edges and vertices (using line graphs) as well. However, most importantly their method relies on the internal structure of the network, whereas, my method uses covariate information associated with vertices – an approach that depends on the external structure of the network.

1.5.2 The MDS For Clustering Similar Actors

The Multi-Dimensional Scaling (MDS) technique uses the feature matrix to cluster similar actors. The feature matrix F, is formalized as a set of observations, in which each observation consists of a set of variables "covariates".

$$F = \{\{a_{1,1},\ a_{2,1},\ \cdots,\ a_{n,1}\},\ \cdots,\ \{a_{1,m},\ a_{2,m},\ \cdots,\ a_{n,m}\}\}$$

ID	Name	Gender	Age	Income	City	Education
1	Peter	M	22	\$35,000	Fairfax	BS
2	Sarah	F	18	\$21,000	Baltimore	HS
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Zach	M	34	\$65,000	Washington DC	PhD

MDS clusters vertices by partitioning them into exhaustive non-overlapping subsets. It provides a spatial representation of similarity patterns among actors, in which similar vertices

appear closer together and dissimilar ones appear far apart. The input may be a set of attributes, Euclidean distances or graphical distances, while the output is a set of coordinates in 2-D, 3-D or higher dimensional space.

1.5.3 Mantel's Test for Association Between Transition Matrices

Because there is no mathematical model predicts any observed network perfectly, uncertainties and errors are generated. I define the residual matrix to be

$$\text{Residual Matrix} = \text{Observed Matrix - Predicted Matrix.}$$

$$R(t+1) = A(t+1) - \hat{A}(t+1).$$

In this section, I use Mantel's test to assess the model. Mantel's test measures the goodness of fit, which is done through applying a multivariate matrix regression analysis to measure the goodness of fit [44], [49]. But, since all the entries of the co-occurrence matrix are dependent upon each other (inter-correlated) and upon the status of the network at previous states, the traditional regression analysis methods would not provide the best association measure given such conditions.

In 1967, Nathan Mantel a biostatistician at the National Institute of Health suggested an approach that overcomes dependency issues. His test is widely used to assess species-environmental relationships. I believe the test can also be applied to networks as well. Mantel's test is essentially a regression in which the variables are themselves distance or similarity matrices summarizing pairwise similarities among sample location (in the social networks context, it summarizes pairwise similarities among actors.) The power and versatility of Mantel's test stems from the various ways that the distance matrices or the regression itself can be framed.

I proceed by applying Mantel's procedure to networks. Suppose there are two matrices

of interrelations at time $t+1$, an observed and predictor matrices. Without loss of generality assume the matrices are symmetric with zero elements along the main diagonal, i.e. no self-ties, the matrices are given below.

$$A(t+1) = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{12} & 0 & a_{23} & \cdots & a_{2n} \\ a_{13} & a_{23} & 0 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & a_{3n} & \cdots & 0 \end{bmatrix}_{n \times n}, \quad \hat{A}(t+1) = \begin{bmatrix} 0 & \hat{a}_{12} & \hat{a}_{13} & \cdots & \hat{a}_{1n} \\ \hat{a}_{12} & 0 & \hat{a}_{23} & \cdots & \hat{a}_{2n} \\ \hat{a}_{13} & \hat{a}_{23} & 0 & \cdots & \hat{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{a}_{1n} & \hat{a}_{2n} & \hat{a}_{3n} & \cdots & 0 \end{bmatrix}_{n \times n}$$

Note that the matrices must be of the same rank and because the matrix is symmetrical, the correlation between the upper triangular parts and the lower triangular parts is the same. And if there are n actors, the matrix contains $m = \frac{n(n-1)}{2}$ edges or adjacencies. They are not independent of each other: changing the "position" of one object would change $n-1$ of these adjacencies. As a result, the relationship between the two matrices cannot be assessed by evaluating the correlation coefficient between the two sets of adjacencies and testing its statistical significance.

Mantel's procedure is considered a randomization test. The correlation between the two sets of $\frac{n(n-1)}{2}$ distances is calculated. Because the elements of the adjacency matrix are dependent, the test of significance is evaluated via permutation procedures; the rows and columns of the distance matrices are randomly rearranged.

For the randomization test the elements of the either of the two matrices say the observed matrix are randomly permuted, while the elements of the second matrix say the predictor matrix are left in the same order. For example if a random permutation gives the order

$5, 3, \cdots, 1$, then the randomly permuted matrix is

$$A_n^{rand}(t+1) = \begin{bmatrix} 0 & a_{35} & a_{n5} & \cdots & a_{15} \\ a_{35} & 0 & a_{n3} & \cdots & a_{13} \\ a_{n5} & a_{n3} & 0 & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{15} & a_{13} & a_{1n} & \cdots & 0 \end{bmatrix}_{n\times n} = B_n(t+1) = \begin{bmatrix} 0 & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{12} & 0 & b_{23} & \cdots & b_{2n} \\ b_{13} & b_{23} & 0 & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & b_{3n} & \cdots & 0 \end{bmatrix}_{n\times n}$$

Mantel test is based on linear correlation and hence is subject to the same assumptions that beset the correlation. The Mantel's statistic is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{a_{ij} - \bar{a}}{s_a} \right) \cdot \left(\frac{b_{ij} - \bar{b}}{s_b} \right) \tag{1.3}$$

where r is the conventional Pearson correlation coefficient bounded on $[-1, 1]$.
$b_{ij} \in B(t+1)$ is the permuted matrix.

A randomized value of the correlation between the two matrices (one of which is the permuted matrix) is then calculated, and the distribution of values for the statistic is generated via many iterations ≈ 5000 for $\alpha = 0.05$, [30].

To assess significance, however, the rows and columns of one of the matrices are subjected to random permutations many times, with the correlation being recalculated after each permutation. The significance of the observed correlation is the proportion of such permutations that lead to a higher correlation coefficient.

If the null hypothesis of no correlation between the two matrices is true, then permuting the rows and columns of the matrix should be equally likely to produce a larger or a

smaller coefficient. The test can discover changes in the pattern of correlation at different scales.

The test considers explicitly the relationship between the observed adjacency matrix and the predictor adjacency matrix suggested by the model. As a formal hypothesis test, Mantel's test can be used to compare an observed covariate similarity matrix to the one posed by the numerical model. The test summarizes the strength of the correspondence between the two adjacency matrices. The model adjacency matrix can be provided as a simple binary matrix of 0's and 1's (dichotomous), or the transition probability matrix (weighted).

The samples of the test are sets of permuted matrices derived from the estimated adjacency matrix $\hat{A}(t+1)$. The question is, "Are samples taken from the predicted network also similar in terms of the observed network?" In this case two samples are similar having distance=0 if they both portray the same network feature, otherwise they are dissimilar having distance=1. The test is similar to an $F-$ratio test.

The test of significance can be determined by sampling the randomization distribution. Although, Mantel used normal approximation for the randomization distribution of Z to carry out the test of significance of an observed matrix, Mielke (1978) and Faust and Romney (1985) questioned the normal assumption and suggested that the significance can be determined by comparing the test statistic directly with the randomization distribution. There are $n!$ possible permutation for any matrix of order n which implies it is practical to determine the distribution of computed correlations.

1.6 Methodology

In the previous sections I have provided a brief introduction to networks and the current problems related to networks. In the introductory part, I covered basic concepts in network

theory such as centrality measures, cohesion, structural equivalence, structural holes, and blockmodeling. There are number of the challenges concerning social networks such as the manipulation of large-scale multi-dimensional networks, the continuously growing and changing networks (evolutionary networks), the prediction of edges and vertices, and the lack of advanced mathematical models. In this dissertation, I present solutions and vision to remedy these problems.

1.6.1 Road Map

In Chapter 2, I have focused on the mathematics underpinning networks both static and evolving. This part of the dissertation is largely theoretical and covers concepts in matrix theory, graph theory, estimation, geometry and fuzzy logic. For example, I have worked out a technique to undergo the storage and manipulation of a large-scale network that is continuously expanding using primitive network blocks represented with sub-matrices stored in a global matrix. This matrix representation has the advantage of keeping track of the changes the network endures and the magnitude of each change over time, but most importantly performing arithmetic operations on the matrix is simple due to the fact that the matrix is block-diagonal and the elements are sub-matrices whose entries are ones. To some extent the algorithm I presented is envisioned by the Finite Element Method (FEM) in the sense that small blocks (elements) contribute to the formation of the global network (matrix).

To address the issue of having a continuously growing network, I invented a tool to expand on vertices by introducing a matrix of infinite dimension resembling an infinite network in which vertices are categorized as active or inactive. The infinite matrix offers a mechanism of observing the development of a network over time.

Then I have developed an advanced approach which can be used to derive one-mode networks from weighted (valued) two-mode networks. For higher-mode dichotomous (binary)

networks the derivation is straightforward, more details are presented in the literature review section. I invented a mechanism to express the weighted network as a combination of binary networks that are used to obtain the one-mode weighted network. Perhaps one of the major contributions of my dissertation in the network theory field is the manipulation of higher dimensional relational networks to extract information and gain knowledge about networks on the different lower dimensions and modes. I have extended the differential of one-mode from two-mode networks on higher-mode networks and have defined new matrix multiplications accordingly.

The rest of Chapter 2 concerns mathematical tricks that efficiently compute graph and network measures such as the computation of edge count and graph density, an iterative algorithm to compute the network diameter, degree centrality of vertices and degree and distance matrices. Lastly, I have studied edge and vertex duality, in which edges transform to vertices and vertices transforms to edges through what is known as line graphs. A portion of this section has been utilized to discuss the importance and properties of some special line graphs.

In Chapter 3, I have focused my research on studying edges and vertices in a network. This part relates to the interchangeability and duality between vertices and edges in a graph. I have suggested a method that uses covariate information associated with vertices to estimate the probability of missing edges and covariate information associated with edges to estimate the probability of missing vertices. In order to predict missing vertices, I have utilized the line graph transformation to convert edges to vertices and vertices to edges and the problem now is to compute the probability of an edge in the line graph. Estimating the probability of an edge is obtained by taking the inner product of the vectors of covariates. Ultimately, I have extended the methodology of predicting edges (dyadic ties) to predict edges in a triad. The method incorporates covariate information as well; however, it is based on geometry and fuzzy logic rather than the inner product of two vectors. It is worth

mentioning, though, that my model assumes that if two entities share many common values and are close to each other distance-wise then these entities are similar and more likely to connect/communicate. I had a discussion on this issue with Dr. Tsvetovat in which he pointed out that in Social Sciences you can give a degenerate scenario. For instance, if a married person and his mother in-law are not close to each other geographically then they are more likely to be happy and content. However, if they live in a close distance from each other then they are more likely to argue and be dissatisfied.

In Chapter 4, I have integrated concepts from Chapters 2 and 3 to build models for evolutionary networks and preferential attachment. A common property of many large networks is the vertex connectivities (dyadic edges). This feature is a consequence of two generic mechanisms; the continuous expansion of networks by adding new vertices, which is called growth, and the preferential attachment of new vertices to sites that are already well connected. Network growth means that the number of vertices increases with time. Subsequently, I implement the notion of having an infinite matrix and the ideal edgeless graph, which are defined in Chapter 2. Preferential attachment means that the more connected a vertex is, the more likely it is to acquire new edges. Intuitively, preferential attachment can be understood if we think in terms of social networks connecting people. Here an edge from actor A to actor B means that actor A "knows" or "is acquainted with" actor B. Vertices with many edges represent well-known people with lots of relations. When a new actor enters the community, he or she is more likely to become acquainted with one of those more visible actors rather than with a relative unknown.

In Chapter 5, I have implemented the theory of networks on real-life social networks and other types to portray the different levels of interactivity, which includes the coauthorship social network of prominent statisticians, road fatalities network in the United States, news documents network, preferential attachment and the emergence of scientific subfields. In

coauthorship social networks, I identified special groups of coauthors that are high in degree and tie strength in which I called elite group. Coauthorship social networks data were collected from the online Current Index to Statistics (CIS) database [12] as well as personal curriculum vitae. The CIS database is jointly published by the American Statistical Association (ASA) and the Institute of Mathematical Statistics (IMS). Road fatalities network presents how states relate to other states through different crash factors and how states are similar with respect to crash factors. Road fatalities data were collected from the online Fatality Analysis Reporting System's (FARS) website [21], an affiliation of the National Center for Statistics and Analysis (NHTSA) on traffic safety facts. And finally, in the news documents example I performed an assessment of the documents network derived from the term-document and bigram-document networks. Text data were collected by the Linguistic Data Consortium in 1997. This is a superset of the data used in Martinez (2002). The data consisted of 15,863 news reports collected from Reuters and CNN from July 1, 1994 to June 30, 1995.

I have concluded the chapter with simulation of two evolutionary social networks to demonstrate preferential attachment; the first model simulates the evolution of a coauthorship social network, while the second simulates the evolution of a online music friendship social network.

The literature I present in the following chapters is solely my own work and represents my contribution to the field of network theory and analysis with the exception of few definitions in which it was difficult to separate them from other material. Consequently, I preferred to keep these definitions that were named after scholars or previous defined in the science in context. The dissertation offers solutions to many of the problems encountered by analysts and researchers and many of the ideas that are presented in the dissertation can be used successfully in the fields of network theory, graph theory and matrix theory.

Chapter 2: Network, Graph And Matrix Theory

The purpose of this dissertation as a whole is to elaborate mathematical thoughts in the field of networks. This chapter discusses advanced mathematical techniques and their applications to static and evolutionary networks. I start by introducing a new graph and matrix theory notation and terminology, followed by theory on multi-mode networks. One of the goals of this research is to integrate network theory, graph theory and matrix theory. Consequently, the approach is mostly theoretical and abstract.

2.1 Network Recipes

2.1.1 The Star Graph S_n

In this section I present a deeper analysis to matrices and their properties and how matrices are related to networks and graphs. I begin with studying some primitive building blocks of networks and their properties. Given a completely star shape graph $G(V, E) = S_n$ of size n with $n-1$ leaves as in Figure 2.1, where $|V| = n$ is the number of vertices. Let A be the

Figure 2.1: The Star Network.

square adjacency matrix of size n corresponding to S_n. A_n is of the form

$$A_{n\times n} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n\times n}$$

Then,

$$A_n^2 = \begin{bmatrix} n-1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & \cdots & 1 \end{bmatrix}_{n\times n}, \text{and} \quad A_n^3 = \begin{bmatrix} 0 & n-1 & n-1 & \cdots & n-1 \\ n-1 & 0 & 0 & \cdots & 0 \\ n-1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n-1 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n\times n}$$

We observe that,

$$A_n^2 = 1_n + (n-1)\cdot O_n - A_n, \text{ and}$$

$$A_n^3 = (n-1)\cdot A_n.$$

With 1_n being a square matrix whose elements are all 1's "complete graph" K_n, and O_n being the matrix whose elements are zeros except $o_{11} = 1$.

Corollary 1. In general, if A_n is the proximity matrix corresponding to a star-like (ego) network, then for $p = 1, 2, 3, \cdots$, we have

$$A_n^{2p+1} = (n-1)^p \cdot A_n$$

$$A_n^{2p} = 1_n + (n-1)^p \cdot O_n - A_n$$

In general, if $B_n = \frac{1}{n-1} \cdot A_n^2$, then $B_n^2 = B_n$, i.e B_n is idempotent. Furthermore, if $B_n = \frac{1}{\sqrt{n-1}} \cdot A_n$, then $B_n^{2p+1} = B_n$ and $B_n^{2p} = B_n^2$.

2.1.2 The Complete Graph K_n

Corollary 2. Consider the complete graph $G(V, E) = K_n$; a clique network with self-ties present. Let $B_n = \frac{1}{n} 1_n$, then $B_n^p = B_n$ is idempotent for $p = 1, 2 \cdots$. More specifically,

$$1_n^p = n^{p-1} \cdot 1_n,$$

where

$$1_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{n \times n}$$

In addition,

$$e^{1_n} = I_n + \frac{e^n - 1}{n} 1_n,$$

where I_n is the identity matrix of size n.

Suppose A_n is the adjacency matrix corresponding to a complete graph (clique) of size

n with no self-ties imposed,

$$A_n = \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}_{n \times n}$$

Then,

$$A_n^2 = \begin{bmatrix} n-1 & n-2 & n-2 & \cdots & n-2 \\ n-2 & n-1 & n-2 & \cdots & n-2 \\ n-2 & n-2 & n-1 & \cdots & n-2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n-2 & n-2 & n-2 & \cdots & n-1 \end{bmatrix}_{n \times n}$$

Thus,

$$A_n^2 = (n-2) \cdot A_n + (n-1) \cdot I_n = (n-2) \cdot 1_n + I_n$$

Moreover,

$$A_n^3 = (n-2)^2 A_n + (n-1)(n-2) I_n + (n-1) A_n$$

We can write A_n in terms of 1_n and I_n,

$$A_n = 1_n - I_n$$

Therefore,

$$A_n^2 = (n-2) \cdot 1_n + I_n$$

In general,

$$A_n^p = (1_n - I_n)^p = \sum_{i=0}^{p} \binom{p}{i} 1_n^i (-1 \cdot I_n)^{p-i} = \sum_{i=0}^{p} \binom{p}{i} \cdot n^{i-1} \cdot 1_n \cdot (-1)^{p-i} \cdot I_n$$

$$= \sum_{i=0}^{p} \binom{p}{i} \cdot n^{i-1} \cdot (-1)^{p+i} \cdot 1_n$$

2.1.3 The ℓ^p-norm and Networks

Because new actors can emerge at any time and become active members, the dimensions of the evolving adjacency matrices increase indefinitely and computations can be cumbersome. Consequently, the need for a tool to tackle the dimensionality issue is essential. The ℓ^p-norm offers a mechanism to expand on vertices (nodes) and allows having null vertices. The ℓ^p-norm is defined on an infinite dimensional vector space. In a sense, there will be a set of infinite vertices categorized as active or inactive, a zero will be assigned to all inactive null vertices. Thus, the adjacency matrix becomes very sparse and the infinite matrix functions as the operator. Actors can change status from inactive to active and vice versa if the network in evolutionary mode and the process of introducing new actors (vertices) to a network in a fixed period of time is a stochastic process modeled by the Poisson process.

Definition 2.1. The ℓ^p-norm is the vector norm $||\vec{x}||_p$ defined by

$$||\vec{x}||_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}},$$

where $\vec{x} = (x_1, x_2, \cdots, x_n)$.

The ℓ^p-norm is commonly encountered in vector algebra and operations such as the inner product, which I heavily rely on to compute similarities.

Definition 2.2. A real Hilbert space is a real inner product space that is a complete normed space (Banach space) under the norm defined by the inner product.

Definition 2.3. The inner product $< \cdot, \cdot >$ for a vector v is defined by

$$\langle x, y \rangle = \sum_i x_i y_i.$$

The ℓ^p-inner product of vectors produces a scalar. If $p = 2$, then the ℓ^2-inner product gives the cosine distance between the two vectors in Euclidean space.

2.2 The Theory of Infinite Networks

When the size of the network increases tremendously and new vertices and edges are constantly being added, the traditional matrix operations become computationally expensive. In this part of the dissertation, I present a tool to perform matrix multiplication on the micro level, in which the contribution of each element (a clique in this case) at a given time step is embedded into the global matrix. The mathematical model involves new entities (edges and vertices) introduced to the network in addition to existing edges and vertices. It applies to static and dynamic networks as well. In this regard, I adopt a new notation and terminology to deal with these issues.

Definition 2.4. Let $G(V, E)$ be a graph. The order of $G(V, E)$ is the number of its vertices, denoted $o(G(V, E)) = |V|$; or simply $o(G)$.

Definition 2.5. Let $G(V, E)$ be a graph. The size of $G(V, E)$ is the number of its edges, denoted $s(G(V, E)) = |E|$; or simply $s(G)$.

Definition 2.6. The edgeless graph; sometimes call the empty graph, is the graph with no edges, denoted $G^{\circ}(V, E = \phi)$, where $o(G^{\circ}) = |V| = n,\ 0 \leq n < \infty$, and $s(G^{\circ}) = |E| = 0$.

The edgeless graph is the initial object in the category of graphs.

Definition 2.7. The infinite edgeless graph, denoted $N(V, E = \phi)$, is a graph in which the number of vertices is infinite with no edges, where $V = \{v_1, v_2, \cdots\}$, $o(N) = \infty$ and $s(N) = 0$.

Definition 2.8. An infinite matrix is a matrix with infinite number of rows and columns, denoted A_{∞}.

Technically, the infinite edgeless graph is a graph with infinite vertices all of which are disconnected (isolated). It can be represented with the infinite zero-matrix ${}^{0}A_{\infty}$, whose elements ${}^{0}a_{ij}$ are all zeros.

Definition 2.9. The infinite complete graph; or simply the infinite graph, is the graph $I(V, E) = K_{\infty}$ with $o(I) = \infty$ and $s(I) = \infty$.

Definition 2.10. The $n-$component infinite subgraph is a graph with n connected subgroups of vertices excluding isolates, in which each component must contain at least one edge.

Any network with a finite number of entities can be treated as an infinite network, so that inactive entities are idle, they are represented with isolated vertices. Thus, the infinite graph representation of the network consists of infinitely many disconnected vertices in addition to a finite number of connected vertices. The vertices are labeled according to the natural

numbers system.

To mathematically formulate these concepts using matrices, I introduce a modified matrix notation. I define ${}^{b}\overset{t}{A}_{\infty}$ to be binary infinite dimensional adjacency matrix at time step t whose elements are zeros and ones. ${}^{w}\overset{t}{A}_{\infty}$ refers to the weighted infinite dimensional proximity matrix at time step t whose elements are real numbers. Finally, ${}^{0}\overset{t}{A}_{\infty}$ and ${}^{1}\overset{t}{A}_{\infty}$ refer to the infinite dimensional matrices at times step t of zeros and ones respectively. Hence, the symbol on top of the matrix variable indicates the time step and the symbol on the upper left side of the matrix variable is reserved for the type of the matrix. Possible types are "binary", "weighted", "zero", "one". The dimension of the matrix is placed in the lower right side of the matrix variable, while the power (exponent) or transpose of a matrix is placed in the upper right side.

Now, let ${}^{0}\overset{0}{A}_{\infty}$ be the square infinite zero matrix at time 0. ${}^{0}\overset{0}{A}_{\infty}$ is the network of infinite vertices – all disconnected.

Let $\left\{ {}^{b}\overset{i}{A}_{\infty} \right\}_{i=1}^{t}$ be the sequence of binary adjacency one-mode matrices of clique interactions. Table (2.1) portrays an example of the contribution of each matrix element in the proximity weighted matrix ${}^{w}\overset{t}{A}_{\infty}$ at time t, the derivation is presented below.

$$ {}^{w}\overset{t}{A}_{\infty} = {}^{0}\overset{0}{A}_{\infty} + \sum_{i=1}^{t} {}^{b}\overset{i}{A}_{\infty} = \sum_{i=1}^{t} {}^{b}\overset{i}{A}_{\infty} = \sum_{i=1}^{t-1} {}^{b}\overset{i}{A}_{\infty} + {}^{b}\overset{t}{A}_{\infty} = {}^{w}\overset{t-1}{A}_{\infty} + {}^{b}\overset{t}{A}_{\infty} \, . \tag{2.1} $$

Assuming each vertex is allowed only to make at most one clique interaction at each time

Table 2.1: The Derivation of the Weighted Proximity Matrix From Binary Matrices.

$^{1}_{b}A_{\infty}$

	1	2	3	4	$\cdots$	∞
1	1	1	1	0	$\cdots$	0
2	1	1	1	0	$\cdots$	0
3	1	1	1	0	$\cdots$	0
4	0	0	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
∞	0	0	0	0	$\cdots$	0

$+$

$^{2}_{b}A_{\infty}$

	1	2	3	4	$\cdots$	∞
1	1	1	0	1	$\cdots$	0
2	1	1	0	1	$\cdots$	0
3	0	0	0	0	$\cdots$	0
4	1	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
∞	0	0	0	0	$\cdots$	0

$=$

$^{2}_{w}A_{\infty}$

	1	2	3	4	$\cdots$	∞
1	2	2	1	1	$\cdots$	0
2	2	2	1	1	$\cdots$	0
3	1	1	1	0	$\cdots$	0
4	1	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
∞	0	0	0	0	$\cdots$	0

step, the sequence of complete subgraphs (cliques) $\left\{ {}^{i}_{b}A_{\infty} \right\}_{i=t_0}^{t}$ composes the global proximity matrix through additions of binary matrices. Therefore, ${}^{t}_{w}A_{\infty}$ resembles the evolution of the network over time. The infinite matrix clarifies any ambiguity resulted from introducing new vertices at each time step; in the sense that the number of vertices is fixed although that quantity is unbounded. The infinite matrix approach and the construction of the proximity matrix scheme offer a mechanism to track down structural formation pertaining to the network. It is of great importance to identify cohesive subnetworks within the network and how they are formed. The method is consistent in terms of building the weighted matrix from primitive blocks such as dyads–triads–tetrads–pentads–hexads–heptads–octads– and higher level. The statistical distribution of the primitive blocks can be used to identify cohesive subgroups (cliques). For example, in the author-coauthor social network application, it may not be that simple to distinguish the laboratory style from the entrepreneurial style of coauthorship if the matrix is in weighted format. Statistical inference for the clique-size is a way to separate the two styles through hypothesis testing. This is done by computing the probability of observing a given extreme clique-size in such networks assuming a certain style exists to determine wether or not the claim of that style is present.

Table 2.2: The Two-Mode Infinite Matrix.

	t_1	t_2	$\cdots$	t_m	t_{m+1}	$\cdots$	t_∞
a_1	at_{11}	at_{12}	$\cdots$	at_{1m}	0	$\cdots$	0
a_2	at_{21}	at_{22}	$\cdots$	at_{2m}	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_n	at_{n1}	at_{n2}	$\cdots$	at_{nm}	0	$\cdots$	0
a_{n+1}	0	0	0	$\cdots$	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_∞	0	0	0	$\cdots$	0	$\cdots$	0

$=$

	t_1	t_2	t_3	$\cdots$	t_∞
a_1	1	1	0	$\cdots$	0
a_2	1	1	0	$\cdots$	0
a_3	1	0	0	$\cdots$	0
a_4	0	1	0	$\cdots$	0
a_5	0	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_∞	0	0	0	$\cdots$	0

Now, suppose that $|V| = n$ is the order of the graph at time t. The finite version of 2.1 is

$$ {}^{w}\overset{t}{A}_{n\times n} = \sum_{i=1}^{t} {}^{b}\overset{i}{A}_{n\times n} = \sum_{i=1}^{t-1} {}^{b}\overset{i}{A}_{n\times n} + {}^{b}\overset{t}{A}_{n\times n} = {}^{w}\overset{t-1}{A}_{n\times n} + {}^{b}\overset{t}{A}_{n\times n} \,. \tag{2.2} $$

Furthermore, suppose at time $t+1$ another clique is generated and a new vertex/vertices are introduced. Let m be the number of the new unique vertices, then equation 2.1 becomes

$$ {}^{w}\overset{t+1}{A}_{(n+m)\times(n+m)} = {}^{w}\overset{t}{A}_{(n+m)\times(n+m)} + {}^{b}\overset{t+1}{A}_{(n+m)\times(n+m)} \,. \tag{2.3} $$

If no new vertices are introduced, then 2.1 simplifies into

$$ {}^{w}\overset{t+1}{A}_{n\times n} = {}^{w}\overset{t}{A}_{n\times n} + {}^{b}\overset{t+1}{A}_{n\times n} \,. \tag{2.4} $$

For infinite two-mode networks, the procedure of formulating the bi-partite infinite matrix and the corresponding one-mode infinite matrix is very similar and presented below.

Consider the two-mode binary infinite matrix ${}^{b}AT_\infty$, where the types A and T represent rows and columns respectively.

If t_i; $i \in \mathbb{N}$ represents a time-series or sequential network feature, then the two-mode matrix

is the evolutionary matrix corresponding to entities of type A; thus, ${}^{b}AT_{\infty}$ is a representation of the evolutionary network. The sequential one-mode binary matrices $\left\{ {}^{i}_{b}A_{\infty} \right\}_{i=1}^{t}$ are computed in the following manner. Let t_i be the column vector of time-step i, then

$$ {}^{i}_{b}A_{\infty} = t_i \cdot t_i^T, \quad 1 \leq i \leq t, \tag{2.5} $$

and the weighted ${}^{w}A_{\infty}$ matrix can be obtained from ${}^{i}_{b}A_{\infty}$ as previously described.

On the other hand, if ${}^{b}AT_{\infty}$ does not symbolize sequential network feature, a direct calculation of the infinite one-mode proximity matrix ${}^{w}A_{\infty}$ is obtained as follows

$$ {}^{w}A_{\infty} = {}^{b}AT_{\infty} \cdot {}^{b}AT_{\infty}^{T} = {}^{b}AT_{\infty} \cdot {}^{b}TA_{\infty}. \tag{2.6} $$

The infinite matrix in 2.2 is square and very sparse; all elements are zeros except for a fixed number of rows n and a fixed number of columns m. As a result, matrix multiplication can be performed by setting up a square sub-matrix of size $k = \max(m, n)$ and perform the product on the square matrix of size k as we would normally do. Theoretically, computation in 2.6 can be done with no problem because ${}^{b}AT_{\infty}$ is very sparse and $at_{ij} = 0, \ \forall \ i > n, \ j > m$. The infinite product is presented in Equation 2.7.

$$ {}^{w}A_{\infty} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} at_{ij} \cdot at_{ji} = \sum_{i=1}^{k} \sum_{j=1}^{k} at_{ij} \cdot at_{ji}. \tag{2.7} $$

Corollary 3. If ${}^{t_1}_{b}A_{\infty}$ and ${}^{t_2}_{b}A_{\infty}$ are infinite binary matrices corresponding to infinite graphs

G_1 and G_2 respectively. Then

$$ {}^{t_1}_{b}A_\infty + {}^{t_2}_{b}A_\infty \text{ is equivalent to } G_1 \cup G_2. $$

This can be thought of as the convoluted network, and ${}^{t_1}_{b}A_\infty$ and ${}^{t_2}_{b}A_\infty$ may represent bipartite graphs.

Corollary 4. Suppose ${}^{t_1}_{b}A_\infty$ and ${}^{t_2}_{b}A_\infty$ are infinite binary matrices corresponding to the infinite graphs G_1 and G_2 respectively for time steps t_1 and t_2. Then ${}^{t_2}_{b}A_\infty - {}^{t_1}_{b}A_\infty$ is a matrix representation of the network of all new ties at time step t_2.

Definition 2.11. For matrices A and B of the same size, the Hadamard product also known as entrywise or Schur product is defined by

$$ A \bullet B = a_{ij} \cdot b_{ij}, \quad a_{ij} \in A \text{ and } b_{ij} \in B. $$

Note: The Hadamard product is commutative.

Corollary 5. Let ${}^{b}A$ be a square one-mode adjacency matrix. The Hadamard product ${}^{b}A \bullet {}^{b}A^T$ gives the symmetric relations of ${}^{b}A$; a nonzero element indicates a symmetric relation. If ${}^{b}A \bullet {}^{b}A^T = {}^{b}A$ then ${}^{b}A$ is said to be symmetric; i.e. ${}^{b}A = {}^{b}A^T$.

Corollary 6. If ${}^{1}_{b}A_\infty$ and ${}^{2}_{b}A_\infty$ are infinite binary matrices corresponding to infinite graphs

G_1 and G_2 respectively. Then

$$ {}^{b}\overset{1}{A}_{\infty} \bullet {}^{b}\overset{2}{A}_{\infty} \text{ is equivalent to } {}^{b}\overset{1}{A}_{\infty} \text{ and } {}^{b}\overset{2}{A}_{\infty}, $$

which is analogous to the intersection $G_1 \cap G_2$. ${}^{b}\overset{1}{A}_{\infty}$ and ${}^{b}\overset{2}{A}_{\infty}$ may represent bipartite graphs.

This can be thought of as the network of solely maintained ties (strongest perfect interactions).

2.3 Block-Diagonal Matrix Representation

Next, I present a way to represent the one-mode matrix as a collection of blocks positioned along the main diagonal of a larger matrix. Given a sequence of discrete binary interactions in time $\left\{ {}^{b}\overset{i}{A}_{n} \right\}_{i=1}^{t}$. Define the block diagonal matrix ${}^{b}B_{\infty}$, as follows.

$$ b_{ii} = {}^{b}\overset{i}{A}_{n_i}, $$

where the diagonal elements of ${}^{b}B_{\infty}$ are the finite matrices, and $b_{ij} = 0$ for $i \neq j$. The block diagonal matrix representation of cliques in Table 2.3 has many advantages. First of all, it suggests a way to keep track of continuously changing networks, so that the order and size of interactions are easily determined. Moreover, it allows identifying unique cliques and their sizes; the size n_i of each block varies according to the clique size. The global matrix is built up from small primitive blocks or cohesive subnetworks; where the contribution of each block or element on the local level is added to form the global network and each block indicates and interaction at time step $t_i, i \in \mathbb{N}$. Finally, because the matrix is block

Table 2.3: Block Diagonal Matrix Representation of Cliques.

${}^{b}B_{\infty}=$

	$\overset{1}{a}_1$	$\cdots$	$\overset{1}{a}_{n_1}$	$\overset{2}{a}_1$	$\cdots$	$\overset{2}{a}_{n_2}$	$\overset{3}{a}_1$	$\cdots$	$\overset{3}{a}_{n_3}$	$\cdots$
$\overset{1}{a}_1$	1	$\cdots$	1							
$\vdots$	$\vdots$	$\ddots$	$\vdots$							
$\overset{1}{a}_{n_1}$	1	$\cdots$	1							
$\overset{2}{a}_1$				1	$\cdots$	1				
$\vdots$				$\vdots$	$\ddots$	$\vdots$				
$\overset{2}{a}_{n_2}$				1	$\cdots$	1				
$\overset{3}{a}_1$							1	$\cdots$	1	
$\vdots$							$\vdots$	$\ddots$	$\vdots$	
$\overset{3}{a}_{n_3}$							1	$\cdots$	1	
$\vdots$										$\ddots$

Table 2.4: The Square of The Block Diagonal Matrix of Cliques.

${}^{w}B^{2}{}_{\infty}=$

	$\overset{1}{a}_1$	$\cdots$	$\overset{1}{a}_{n_1}$	$\overset{2}{a}_1$	$\cdots$	$\overset{2}{a}_{n_2}$	$\overset{3}{a}_1$	$\cdots$	$\overset{3}{a}_{n_3}$	$\cdots$
$\overset{1}{a}_1$	n_1	$\cdots$	n_1							
$\vdots$	$\vdots$	$\ddots$	$\vdots$							
$\overset{1}{a}_{n_1}$	n_1	$\cdots$	n_1							
$\overset{2}{a}_1$				n_2	$\cdots$	n_2				
$\vdots$				$\vdots$	$\ddots$	$\vdots$				
$\overset{2}{a}_{n_2}$				n_2	$\cdots$	n_2				
$\overset{3}{a}_1$							n_3	$\cdots$	n_3	
$\vdots$							$\vdots$	$\ddots$	$\vdots$	
$\overset{3}{a}_{n_3}$							n_3	$\cdots$	n_3	
$\vdots$										$\ddots$

diagonal matrix operations are implemented and performed on the micro-level. For example, squaring ${}^{b}B_{\infty}$ is performed by simply squaring the diagonal elements, namely,

$$b_{ii}^2 = {}^{b}\overset{i}{A}_{n_i}^2, \quad \text{and} \quad {}^{b}\overset{i}{A}_{n_i}^2 = n_i \cdot {}^{b}\overset{i}{A}_{n_i} .$$

${}^{w}B^2{}_{\infty}$ will be a weighted matrix, the square of ${}^{b}B_{\infty}$ is shown in Table 2.4.

Furthermore, ${}^{w}A_n$, the weighted matrix is found by integrating the contribution of each block b_{ii} into the global matrix ${}^{w}\overset{t}{A}_n$.

Likewise, there are some drawbacks associated with this representation. For instance, there might be inconsistency associated with identities; labels may be duplicated. Moreover, the matrix as a whole is very sparse and the matrix dimension may grow up rapidly.

2.4 One-Mode Matrices From Two-Mode Weighted Matrices

The product of a matrix and its transpose method presented earlier to obtain the weighted one-mode proximity matrix from binary two-mode adjacency matrices fails to give meaningful implication if the bi-partite proximity matrix has weighted values; in this regard the values are magnified greatly yielding a matrix with the sum of squares along the main diagonal and sum of growing products off the main diagonal. In this section, I provide a modified version of this technique to tackle this problem.

Definition 2.12. For a bipartite graph $G(V^a, V^b, E)$ with sets of vertices $V^a = \left\{v_1^a, v_2^a, \cdots, v_i^a, \cdots v_{|V^a|}^a\right\}$ of type A; $|V^a| = n$, $V^b = \left\{v_1^b, v_2^b, \cdots, v_j^b, \cdots v_{|V^b|}^b\right\}$ of type B; $|V^b| = m$, and a set of edges $E = \left\{e_1, e_2, \cdots, e_{|E|}\right\}$ connecting types A and B vertices,

The weighted adjacency matrix ${}^{w}AB_{n\times m}$, also known as Edmonds matrix, is defined by

$$ {}^{w}AB_{n\times m} = \begin{cases} {}^{w}ab_{ij}, & \left(v_i^a, v_j^b\right) \in E \\ 0, & (v_i^a, v_j^b) \notin E. \end{cases} $$

$1 \leq i \leq n$, $1 \leq j \leq m$, and the indeterminate ${}^{w}ab_{ij} \in \mathbb{R}$.

Definition 2.13. A complete bipartite graph or biclique is a bipartite graph where every vertex of the first set is connected to every vertex of the second set. $G(V^a, V^b, E)$ is a bipartite graph such that for any two vertices $v_i^a \in V^a$ and $v_j^b \in V^b$, we have $\left(v_i^a, v_j^b\right)$ is an edge in G. The complete bipartite graph with partitions of size $|V^a| = m$, $\left|V^b\right| = n$ and $|E| = m \cdot n$ is denoted $K_{m,n}$.

Definition 2.14. The multiplicity of an edge, denoted $m\left(E(v_i^a, v_j^b)\right) = {}^{w}ab_{ij}$, is the number of multiple edges sharing the same end-vertices; i.e. the edge weight.

Definition 2.15. The multiplicity of a graph is the maximum multiplicity of its edges; the maximum of all weights.

Let ${}^{w}AB_{\infty}$ be the two-mode weighted proximity matrix of vertices of types A and B and let ${}^{b}AB_{\infty}$ be the binary version of ${}^{w}AB_{\infty}$. Assume $n =$ |type A active vertices|, so that ${}^{b}ab_{ij} = 0 \ \forall \ i, j > n$. Let $m =$ |type B active vertices|. Then, we have

$$ {}^{w}AB_{\infty} \cdot {}^{b}AB^{T}{}_{\infty} = {}^{w}AB_{\infty} \cdot {}^{b}BA_{\infty} = {}^{w}AA_{\infty}, \tag{2.8} $$

The elements off the main diagonal are

$$^{w}aa_{ij} = \sum_{k=1}^{\infty} {}^{w}ab_{ik} \cdot^{b} ab_{jk} = \sum_{k=1}^{m} {}^{w}ab_{ik} \cdot^{b} ab_{jk},$$

and along the main diagonal when $i = j$ we have,

$$aa_{ii} = \sum_{k=1}^{\infty} {}^{w}ab_{ik} \cdot^{b} ab_{ik} = \sum_{k=1}^{m} {}^{w}ab_{ik} \cdot^{b} ab_{ik} = \sum_{k=1}^{m} {}^{w}ab_{ik} = \text{vertex } i \text{ count of interactions},$$

Note that

$$^{b}ab_{ik} = \begin{cases} 1, & v_i \leftrightarrow v_k. \\ 0, & v_i \nleftrightarrow v_k. \end{cases}$$

The element aa_{ii} along the main diagonal is the total frequency of interactions for vertex i of type A, and the elements aa_{ii}, $1 \leq i \leq m$, are the marginal distribution of total interactions for vertices of type A. The off-diagonal elements $aa_{ij}, i \neq j$ of $^{w}AA_{\infty}$ represent the sum of edge weights for a determining (influential) vertex i. Vertex i influences the overall relationship; in other words, vertex i determines a_{ij}. Note that $^{w}AA_{\infty}$ is diagonally dominant; i.e. for a given i,$^{w}\, aa_{ij} \leq^{w} aa_{ii}$.

In 2.8, if we set $^{w}aa_{ij} = \max(^{w}aa_{ij},^{w}\, aa_{ji})$ for $i \neq j$, then the off-diagonal elements $^{w}aa_{ij}$ represent the sum of edge unions (max weights) excluding zero weight edge interactions between a pair of vertices related through the vertex type B. It is the sum of all determining active edges and represents the overall relationship between vertices i and j. Alternatively,

Table 2.5: Ordered Two-Mode Matrix.

$^{w}AB_{\infty}=$

	b_1	b_2	$\cdots$	b_j	b_{j+1}	$\cdots$
a_1	ab_{11}	ab_{12}	$\cdots$	ab_{1j}	0	$\cdots$
a_2	ab_{21}	ab_{22}	$\cdots$	ab_{2j}	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	ab_{i1}	ab_{i2}	$\cdots$	ab_{ij}	0	$\cdots$
a_{i+1}	0	0	$\cdots$	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

we can obtain $^{w}aa_{ij}$ as follows

$$^{w}aa_{ij} \quad = \quad \sum_{k=1}^{\infty} \max(^{w}ab_{ik},^{w}ab_{jk}) = \sum_{k=1}^{m} \max(^{w}ab_{ik},^{w}ab_{jk}). \tag{2.9}$$

If $ab_{ik} \vee ab_{jk} = 0$ in 2.9, then $\max(ab_{ik}, ab_{jk}) = 0$.

Special Case: Suppose the columns of $^{w}AB_{\infty}$ are all in descending order, see Table 2.5. $ab_{(i-1)j} \leq ab_{ij} \leq ab_{(i+1)j}$, $ab_{ij} = 0$ is excluded from the inequality; zero entries are allowed. Then, the lower triangle of $^{w}AA_{\infty}$ in 2.8 contains all minima count (sum of all intersections), while the upper triangle contains all maxima count not including edges with zero weights for a dominating vertex v_i.

$$^{w}aa_{ij} \quad = \quad \sum_{k=1}^{\infty} \min(^{w}ab_{ik},^{w}ab_{jk}) = \sum_{k=1}^{m} \min(^{w}ab_{ik},^{w}ab_{jk}), \quad \text{for } i > j. \tag{2.10}$$

$$^{w}aa_{ij} \quad = \quad \sum_{k=1}^{\infty} \max(^{w}ab_{ik},^{w}ab_{jk}) = \sum_{k=1}^{m} \max(^{w}ab_{ik},^{w}ab_{jk}), \quad \text{for } i < j. \tag{2.11}$$

Again, if $^{w}ab_{ik} \vee^{w} ab_{jk} = 0$ in 2.11, then $\max(^{w}ab_{ik},^{w}ab_{jk}) = 0$.

In the same manner, I proceed to obtain the one-mode matrix for vertices of type B

$$ {}^{w}AB^{T}{}_{\infty} \cdot {}^{b}AB_{\infty} = {}^{w}BA_{\infty} \cdot {}^{b}AB_{\infty} = {}^{w}BB_{\infty} \, . \tag{2.12} $$

2.12 can also be derived as follows.

Define,

$$ {}^{w}bb_{ii} = \sum_{i=1}^{\infty} {}^{w}ab_{ij} = \sum_{i=1}^{n} {}^{w}ab_{ij}, \quad {}^{w}ab_{ij} \in {}^{w}AB_{\infty} \tag{2.13} $$

$$ {}^{w}bb_{ij} = \sum_{k=1}^{\infty} \min({}^{w}ab_{ki}, {}^{w}ab_{kj}) = \sum_{k=1}^{n} \min({}^{w}ab_{ki}, {}^{w}ab_{kj}), \quad {}^{w}ab \in {}^{w}AB_{\infty} \tag{2.14} $$

The mechanism of converting a two-mode matrix to one-mode is similar to the effect of projecting from 2-D to 1-D; in the sense that one detailed feature about the network is being lost and the one-mode setup provides only one dimensionality encompassing the one-type marginal relationship. Besides, this process is irreversible; once the one-mode network is obtained it is impossible to retrieve the original two-mode network layout.

2.4.1 Multi-Layering Binary Decomposition

In the remainder of this section, I present a tool to express weighted matrices two-mode or one-mode as a linear combination of optimal binary matrices. These matrices represent the primitive subnetworks generating the global network. It is a process of degenerating the weighted graph into a maximally connected binary subgraphs.

Given the weighted proximity matrix ${}^{w}\overset{1}{A}_{n\times m}$ at time step 1. Let ${}^{b}\overset{1}{A}_{n\times m}$ be the binary

version of ${}^{w}\overset{1}{A}_{n\times m}$. Let,

$$\mathcal{R} = \left\{ {}^{w}\overset{1}{a}_{ij} \right\} = \{r_i\}.$$

Define r to be the number of unique nonzero elements of ${}^{w}\overset{1}{A}_{n\times m}$.

$$r = \left| \left\{ {}^{w}\ \overset{1}{a}_{ij} \right\} \right|,\ {}^{w}\ \overset{1}{a}_{ij} \neq 0,\ 1 \leq r \leq |E| \leq n \cdot m.$$

Theorem 2.1. ${}^{w}\overset{1}{A}_{n\times m}$ can be written as a linear combination of at most r distinct binary matrices ${}^{b}\overset{i}{A}_{n\times m}, 1 \leq i \leq r$ with ${}^{b}\overset{1}{A}_{n\times m} \sqsupset {}^{b}\overset{2}{A}_{n\times m} \sqsupset \cdots \sqsupset {}^{b}\overset{r}{A}_{n\times m}$. The notation $\sqsupset$ means that the binary matrix ${}^{b}\overset{1}{A}_{n\times m}$ has more ones and less zeros than ${}^{b}\overset{2}{A}_{n\times m}$. It also means that ones in ${}^{b}\overset{2}{A}_{n\times m}$ are located at the same position of ones in ${}^{b}\overset{1}{A}_{n\times m}$.

This is equivalent to saying a weighted graph can be expressed in terms of finitely stacked subgraphs of equally weighted (binary) edges.

The method is presented below.

Consider the set $\mathcal{A} = \left\{ {}^{w}\overset{1}{a}_{ij} : {}^{w}\overset{1}{a}_{ij} \in {}^{w}\overset{1}{A} \right\} = \{a_1, a_2, \cdots, a_r\}$.

Let $\alpha_1, \alpha_2, \cdots, \alpha_r \in \mathbb{R}$.

Define, $\alpha_1 = a_1 = \min \left({}^{w}\overset{1}{a}_{ij} \in {}^{w}\overset{1}{A} :\ {}^{w}\overset{1}{a}_{ij} \neq 0 \right)$.

$$\sum_{i=1}^{r} \alpha_i = \max \left({}^{w}\overset{1}{A}_{n\times m} \right) = \left\| {}^{w}\overset{1}{A} \right\|_{\infty}.$$

Set ${}^{w}{}^{2}A = {}^{w}{}^{1}A - \alpha_1 \cdot {}^{b}{}^{1}A$.

Define, $\alpha_2 = a_2 - a_1 = \min\left({}^{b}{}^{2}a_{ij} \in {}^{w}{}^{2}A \colon {}^{w}{}^{2}a_{ij} \neq 0\right)$.

Set ${}^{w}{}^{3}A = {}^{w}{}^{2}A - \alpha_2 \cdot {}^{b}{}^{2}A$.

$\vdots$

Define, $\alpha_r = a_r - a_{r-1} = \min\left({}^{b}{}^{r}a_{ij} \in {}^{w}{}^{r}A \colon {}^{w}{}^{r}a_{ij} \neq 0\right)$.

Set ${}^{w}{}^{r+1}A = {}^{w}{}^{r}A - \alpha_r \cdot {}^{b}{}^{r}A$.

Proof. Need to show that ${}^{w}{}^{r+1}A = {}^{0}A$; the zero matrix.

Suppose $r = 1$. This implies that ${}^{w}{}^{1}A$ and ${}^{b}{}^{1}A$ both have only one non-zero element or $n \cdot m - 1$ zero elements. Let $\alpha_1 = \min\left({}^{w}{}^{1}A\right)$. Then,

$$ {}^{w}{}^{1}A - \alpha_1 \, {}^{b}{}^{1}A = {}^{w}{}^{2}A = {}^{w}{}^{r+1}A, $$

has no non-zero elements or $n \cdot m$ zero elements. This means that all elements of ${}^{w}{}^{r+1}A$ are zeros. Thus, ${}^{w}{}^{r+1}A = {}^{0}A$.

Now, suppose $r = |E| = n \cdot m$. This means that ${}^{w}{}^{1}A$ has no zero elements, and ${}^{b}{}^{1}A$; the matrix of ones in this case, represents a complete graph. If the network is bipartite then by complete graph I mean every vertex of the first type is connected to every vertex of the second type.

Let $\alpha_1 = \min\left({}^{1}_{w}A\colon {}^{1}_{w}a_{ij} \neq 0\right)$. Then,

$$ {}^{2}_{w}A = {}^{1}_{w}A - \alpha_1 {}^{1}_{b}A $$

has one zero element or $n \cdot m - 1$ non-zero elements.

Let $\alpha_2 = \min\left({}^{1}_{w}A - \alpha_1 {}^{1}_{b}A\right) = \min\left({}^{2}_{w}A\colon {}^{2}_{w}a_{ij} \neq 0\right)$. Then,

$$ {}^{3}_{w}A = {}^{1}_{w}A - \alpha_1 {}^{1}_{b}A - \alpha_2 {}^{2}_{b}A $$

has two zero elements or $n \cdot m - 2$ non-zero elements.

Similarly, let $\alpha_3 = \min\left({}^{1}_{w}A - \alpha_1 {}^{1}_{b}A - \alpha_2 {}^{2}_{b}A\right) = \min\left({}^{2}_{w}A - \alpha_2 {}^{2}_{b}A\right) = \min\left({}^{3}_{w}A\colon {}^{3}_{w}a_{ij} \neq 0\right)$.

Then,

$$ {}^{4}_{w}A = {}^{1}_{w}A - \alpha_1 {}^{1}_{b}A - \alpha_2 {}^{2}_{b}A - \alpha_3 {}^{3}_{b}A $$

has three zero elements or $n \cdot m - 3$ non-zero elements.

Finally, let $\alpha_r = \min\left({}^{1}_{w}A - \alpha_1 {}^{1}_{b}A - \alpha_2 {}^{2}_{b}A - \cdots - \alpha_{r-1} {}^{r-1}_{b}A\right) = \min\left({}^{r-1}_{w}A - \alpha_{r-1} {}^{r-1}_{b}A\right) = \min\left({}^{r}_{w}A\colon {}^{r}_{w}a_{ij} \neq 0\right)$. Then,

$$ {}^{r+1}_{w}A = {}^{1}_{w}A - \alpha_1 {}^{1}_{b}A - \alpha_2 {}^{2}_{b}A - \alpha_3 {}^{3}_{b}A - \cdots - \alpha_r {}^{r}_{b}A $$

has r zero elements or $n \cdot m - r = r - r = 0$ non-zero elements.

This implies that the elements of ${}^{r+1}_{w}A$ are all zeros. Thus, ${}^{r+1}_{w}A_{n\times m} = {}^{0}A_{n\times m}$. □

Next, consider the weighted matrix ${}^{w}\overset{1}{A}B$ corresponding to a bipartite graph. Let $\alpha_1, \alpha_2, \cdots, \alpha_r$ be scalars and ${}^{b}\overset{1}{A}B, {}^{b}\overset{2}{A}B, \cdots, {}^{b}\overset{r}{A}B$ be given such that

$$ {}^{w}\overset{1}{A}B = \alpha_1 \cdot {}^{b}\overset{1}{A}B + \alpha_2 \cdot {}^{b}\overset{2}{A}B + \cdots + \alpha_r \cdot {}^{b}\overset{r}{A}B, $$

then

$$ \alpha_1 \cdot {}^{b}\overset{1}{A}B \cdot {}^{b}\overset{1}{A}B^T + \alpha_2 \cdot {}^{b}\overset{2}{A}B \cdot {}^{b}\overset{2}{A}B^T + \cdots + \alpha_r \cdot {}^{b}\overset{r}{A}B {}^{b}\overset{r}{A}B^T = {}^{w}AA \tag{2.15} $$

is the one-mode matrix of vertices of type A related through vertices of type B corresponding to the weighted two-mode graph ${}^{w}\overset{1}{A}B$.

The diagonal element aa_{ii} once again is the count of interactions for vertex i of type A. The elements aa_{ii}, $1 \leq i \leq m$, are the marginal distribution for vertices of type A. While, the off-diagonal elements $a_{ij}, i \neq j$ of ${}^{w}\overset{1}{A}A_{\infty}$ represent the sum of edge intersections (edge overlaps) between a pair of vertices related through the vertex of type B; a_{ij} resembles the overall common relationship between vertex i and j. This is a generalization of the product of ${}^{b}AB$ by its transpose.

Alternatively, the elements of ${}^{w}AA$ in 2.15 can also be computed as follows.

Define,

$$\begin{aligned} a_{ii} &= \sum_{j=1}^{\infty} ab_{ij} = \sum_{j=1}^{m} ab_{ij}, \quad ab_{ij} \in {}^{w}AB_{\infty} \\ a_{ij} &= \sum_{k=1}^{\infty} \min\left(ab_{ik}, ab_{jk}\right) = \sum_{k=1}^{m} \min(ab_{ik}, ab_{jk}), \quad ab \in {}^{w}AB_{\infty}, 1 \leq i, j \leq n. \end{aligned} \tag{2.16}$$

This is essentially the sum of the common interactions between vertices i and j of type A.

2.5 Three-Mode Matrices

Continuing with the same analogy; a third network feature may be introduced to add another dimensionality to the problem resulting in a 3-D cuboid (rectangular parallelepiped) matrix, in which 3-mode matrix manipulations can be explored. The cuboid matrix resembles a tripartite network. A cuboid matrix is in fact a tensor of rank 3; however, for the purposes of this research I will use the term cuboid instead. An example of a three-mode network might be author-by-papers-by-universities. Throughout this section, I make the assumption that the graph is finite to understand how the mathematics work for three-mode matrices, which implies that the matrices have finite dimensions. Finally, the graphs are assumed to represent dichotomous relations, i.e. the multiplicity of an edge is one.

Definition 2.16. Let $V^a = \left\{v_1^a, v_2^a, \cdots, v_i^a, \cdots, v_{|V^a|}^a\right\}$ be the set of vertices of type a, $V^b = \left\{v_1^b, v_2^b, \cdots, v_j^b, \cdots, v_{|V^b|}^b\right\}$ be the set of vertices of type b, and $V^c = \left\{v_1^c, v_2^c, \cdots, v_k^c, \cdots, v_{|V^c|}^c\right\}$ be the set of vertices of type c. Furthermore, let $E = \left\{e_1, e_2, \cdots, e_{|E|}\right\}$ be the set of edges connecting types a, b, c vertices; in this sense, e_i is a hyperedge. Assume $|V^a| = n$, $|V^b| = m$, and $|V^c| = p$. The binary adjacency matrix ${}^{b}ABC_{n\times m\times p}$ corresponding to the finite graph

$G(V^a, V^b, V^c, E)$ is defined by

$$ {}^{b}abc_{ijk} = \begin{cases} 1, & \left(v_i^a, v_j^b, v_k^c\right) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2.17} $$

$1 \leq i \leq n$, $1 \leq j \leq m$, $1 \leq k \leq p$.

I start by computing the two-mode weighted matrix ${}^{w}AB$, the method is presented below

$$ {}^{w}ab_{ij} = \sum_{k=1}^{p} abc_{ijk}, \quad 1 \leq k \leq p. $$

This is equivalent to projecting from 3D cuboid matrix onto the 2D planar matrix giving the marginal bipartite distribution for types A and B.

Because the cuboid is a three-dimensional object, there are several matrix arithmetic operations to perform on the cuboid matrix, some of which result in a rectangular matrix, while others result in a cuboid matrix. Here, I explore few meaningful operators related to networks, but before I explain how these operations are performed, I would like to discuss how a cuboid is being transposed in 3D. Unlike the 2D rectangular matrix, which only has two faces, the 3D cuboid has six faces leading to six different ways to view the block in terms of size, namely, $n \times m \times p$, $n \times p \times m$, $m \times n \times p$, $m \times p \times n$, $p \times n \times m$, and $p \times m \times n$. As a result, the transpose can be done in six different ways.

Suppose ${}^{b}ABC_{n \times m \times p}$ a dichotomous tripartite matrix. Then,

1. $$ {}^{b}ABC^{T_{cba}}_{n \times m \times p} = {}^{b} CBA_{p \times m \times n}, \text{ with } {}^{b}cba_{ijk} = {}^{b} abc_{kji}. $$

2.

$$ {}^{b}ABC^{T_{bac}}_{n\times m\times p} = {}^{b}\,BAC_{m\times n\times p}, \text{ with } {}^{b}bac_{ijk} = {}^{b}\,abc_{jik}. $$

The following cuboid matrix multiplication definition is the traditional 3D matrix product and results in another 3D matrix.

Definition 2.17. Given the matrix ${}^{b}ABC_{n\times m\times p}$. The weighted 3D matrix

$$ {}^{w}AAC_{n\times n\times p}={}^{b}ABC_{n\times m\times p}\cdot {}^{b}ABC^{T_{bac}}_{n\times m\times p}={}^{b}ABC_{n\times m\times p}\cdot {}^{b}BAC_{m\times n\times p}, $$

so that the product of the sub-matrices $AB_{n\times m}\cdot BA_{m\times n}$ is well-defined, is computed as follows

$$ {}^{w}aac_{k} = {}^{b}ABC_{k}\cdot {}^{b}BAC_{k}, \quad 1\leq k\leq p. $$

${}^{w}AAC_{n\times n\times p}$ is a two-mode graph (network) represented with a 3D matrix.

Definition 2.18. Given the matrix ${}^{w}AAC_{n\times n\times p}$. The weighted 3D matrix

$$ {}^{w}AAA_{n\times n\times n}={}^{w}AAC_{n\times n\times p}\cdot {}^{w}AAC^{T_{caa}}_{n\times n\times p}={}^{w}AAC_{n\times n\times p}\cdot {}^{b}CAA_{p\times n\times n}, $$

so that the product of the sub-matrices $AC_{n\times p}\cdot CA_{p\times n}$ is well-defined, is computed as follows

$$ {}^{w}aaa_{i} = {}^{w}AAC_{i}\cdot {}^{b}CAA_{i}\ . $$

${}^{w}AAA_{n\times n\times n}$ is the one-mode 3D matrix of triadic vertices (triplets) of type A related through vertices of types B and C. In graph terminology, a nonzero entry in ${}^{w}AAA_{n\times n\times n}$ indicates that the hyperedge is connecting three vertices altogether as opposed to two vertices in the traditional graph context. ${}^{w}AAA_{n\times n\times n}$ represents triadic relations, a nonzero value ${}^{w}aaa_{ijk}$ indicates that all three vertices are connected through a hyperedge.

The one-mode 2D matrix ${}^{w}AA_{n\times n}$ of pairwise vertices of type A related through vertices of types B and C is found by summing up over one of the dimensions. In this regard, the 3D matrix is assumed to be symmetric.

$$ {}^{w}aa_{ij} = \sum_{k=1}^{n} {}^{w}aaa_{ijk}. $$

Definition 2.19. Given a 3D binary cuboid matrix ${}^{b}ABC$ of size $n \times m \times p$ of relation types A, B and C respectively. The hyper product, denoted $A \circ B$, is define by

$$ {}^{b}ABC_{n\times m\times p} \circ {}^{b}ABC^{T_{cba}}_{n\times m\times p} = {}^{b}ABC_{n\times m\times p} \circ {}^{b}\, CBA_{p\times m\times n} = {}^{w}ABBA_{n\times m\times m\times n}, \quad (2.18) $$

where ${}^{w}ABBA_{n\times m\times m\times n}$ is the hyper-cuboid two-mode proximity matrix of pair of vertices (v_i, v_j) of types A and B related through the set of vertices v_k of type C. Let the product of the sub-matrices $BC_{m\times p} \cdot CB_{p\times m}$ be define. Then, the elements of ${}^{w}ABBA_{n\times m\times m\times n}$ are found as follows

$$ {}^{w}abba_{kl} = {}^{b}\, ABC_{kl} \cdot {}^{b}\, CBA_{kl}, \quad 1 \leq k, l \leq p. $$

The 4D hyper-matrix can be represented using 3D matrix by stacking n cuboid matrices each of size $m \times m \times n$, which results in a 3D matrix of size $m \times m \times n^2$. The 4D hyper-matrix is a tensor of rank 4.

The following 3D matrix multiplication definition results in a one-mode 2D matrix.

Definition 2.20. Given the matrix ${}^{b}ABC_{n\times m\times p}$. The weighted one-mode 2D matrix

$$ {}^{w}CC_{p\times p}={}^{b}ABC_{n\times m\times p}\odot {}^{b}ABC^{T_{bac}}_{n\times m\times p}={}^{b}ABC_{n\times m\times p}\odot {}^{b}BAC_{m\times n\times p},$$

so that the product of the sub-matrices $ABC_{n\times m}\cdot BAC_{m\times n}$ is well-defined, is computed as follows

$$ {}^{w}cc_{ij}=\sum_{q=1}^{n}\sum_{r=1}^{n}{}^{b}ABC_{(n\times m)_i}\cdot {}^{b}BAC_{(m\times n)_j}=\sum_{q=1}^{n}\sum_{r=1}^{n}{}^{w}AA_{(qr)_{ij}},\quad 1\leq i,j\leq p.$$

${}^{w}CC_{p\times p}$ is the one-mode 2D matrix of pairwise vertices of type C related through vertices of types A and B. ${}^{w}CC_{p\times p}$ represents diadic relations, a nonzero value ${}^{w}cc_{ij}$ indicates that the two vertices are connected through an edge.

2.6 Generalized $N-$Mode Matrices

As the network modes increase, more network features are revealed. Suppose a network has N possible features, then the $N-$hyper cuboid matrix (tensor of rank N) and hyper graph are used to analyze the $N-$mode network. All lower mode relations may be retrieved from the multi-mode network in the same fashion we convert the three-mode to two-mode and the two-mode to one-mode. In this part, I extend the rules of one, two, three-mode networks to work for a multi-mode matrix and assume the network is finite and represents dichotomous relations, so that matrix operations are easily explored. The $N-$mode graph is equivalent to clique relations of size N, in which cliques resemble hyperedges.

Definition 2.21. Let $V^1=\left\{v^1_1,v^1_2,\cdots,v^1_{i_1},\cdots,v^1_{|V^1|}\right\}$, $V^2=\left\{v^2_1,v^2_2,\cdots,v^2_{i_2},\cdots,v^2_{|V^2|}\right\}$,

$V^3 = \left\{v_1^3, v_2^3, \cdots, v_{i_3}^3, \cdots, v_{|V^3|}^3\right\}$, $\cdots$, $V^N = \left\{v_1^N, v_2^N, \cdots, v_{i_N}^N, \cdots, v_{|V^N|}^N\right\}$ be the sets of vertices of type $1, 2, 3, \cdots, N$ respectively. Furthermore, let $E = \left\{e_1, e_2, \cdots, e_{|E|}\right\}$ be the set of edges connecting types $1, 2, 3, \cdots, N$ vertices. Once again, e_i is a hyperedge. Assume $|V^i| = n_i$, $\forall\ 1 \leq i \leq N$. The binary adjacency matrix ${}^bA_{n_1 \times n_2 \times \cdots \times n_N}$ for the finite graph $G(V^1, V^2, \cdots, V^N, E)$ is defined by

$$ {}^ba_{i_1 i_2 \cdots i_N} = \begin{cases} 1, & \left(v_{i_1}^1, v_{i_2}^2, \cdots, v_{i_N}^N\right) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2.19} $$

$1 \leq i_j \leq n_j$, for $1 \leq j \leq N$.

The two-mode weighted matrix ${}^wA_{n_i \times n_j}$ corresponding to modes $1 \leq i, j \leq N$, $i \neq j$, is calculated below

$$ {}^wa_{ij} = \sum_{k_1=1}^{n_{k_1}} \cdots \sum_{k_{N-2}=1}^{n_{k_{N-2}}} a_{i_1 i_2 \cdots i_N}, \quad 1 \leq k_l \leq n_l. $$

This is equivalent to projecting from $N-$dimensional hyper-cuboid matrix onto the 2D planar matrix giving the marginal $N-$mode distribution for types i and j.

Arithmetic operations on a multi-dimensional matrix result in several meaningful network features related to the original network, ranging from $N-$mode network of relationships to one-mode relationships.

The transpose of an $N-$mode matrix is done in the same manner. Suppose ${}^bA_{n_1 \times n_2 \times \cdots \times n_N}$ is a dichotomous $N-$partite matrix. Then,

$$ {}^bA^{T_{m_1 \times m_2 \times \cdots \times m_N}}_{n_1 \times n_2 \times \cdots \times n_N} = {}^bA_{m_1 \times m_2 \times \cdots \times m_N}, \quad {}^ba_{i_1 i_2 \cdots i_N} = {}^ba_{j_1 j_2 \cdots j_N}, $$

where $j_1 j_2 \cdots j_N$ is a permutation of $i_1 i_2 \cdots i_N$, which is based on the transpose operator.

The following $N-$cuboid hyper matrix multiplication definition results in another $N-$cuboid hyper matrix.

Let ${}^{b}A_{n_1 \times n_2 \times \cdots \times n_N}$ be a dichotomous $N-$mode matrix. Assume $m_1 = n_i$, the weighted $(N-1)-$mode matrix for a mode i is calculated as follows

$$
\begin{aligned}
{}^{w}A_{m_1 \times m_1 \times m_3 \times \cdots \times m_N} &= {}^{b}A_{m_1 \times m_2 \times \cdots \times m_N} \cdot {}^{b}A^{T_{m_2 m_1 m_3 \cdots m_N}}_{m_1 \times m_2 \times \cdots \times m_N} \\
&= {}^{b}A_{m_1 \times m_2 \times m_3 \times \cdots \times m_N} \cdot {}^{b}A_{m_2 \times m_1 \times m_3 \times \cdots \times m_N},
\end{aligned}
$$

where the product of the sub-matrices $A_{m_1 \times m_2} \cdot A_{m_2 \times m_1}$, is well-defined.

Continuing in the same manner with m_2 being eliminated, the one-mode weighted matrix ${}^{w}A_{m_1 \times m_1 \times \cdots \times m_1}$ for a mode i is obtained in $N-1$ matrix multiplications and transpose.

${}^{w}A_{n_i \times n_i \times \cdots \times n_i} = {}^{w}A_{m_1 \times m_1 \times \cdots \times m_1}$ is the one-mode $N-$dimensional matrix of $N-$cliques of type i related through all other types. A nonzero entry in ${}^{w}A_{m_1 \times m_1 \times \cdots \times m_1}$ indicates that the hyperedge is connecting $N-$vertices.

The one-mode 2D matrix ${}^{w}A_{n_i \times n_i}$ of pairwise vertices of type i related through vertices of all other types may be found by summing up over $(N-1)$ dimensions.

$$
{}^{w}a_{qr} = \sum_{k=1}^{n_1} \cdots \sum_{k=1}^{n_{N-1}} {}^{w}a_{i_1 i_2 \cdots i_N}.
$$

2.7 Edge Count and Graph Density

Let ${}^bAB_\infty$ be the two-mode infinite matrix in binary format representing dichotomous relations; furthermore, let $x = (1\ 1\ 1\ \cdots)_{1\times\infty}$ be the infinite vector of ones.

The edge count (graph size) for a graph $G(V^a, V^b, E)$ having matrix ${}^bAB_\infty$, is defined by

$$\text{edge count} = s(G) = |E| = x\cdot\ {}^bAB_\infty \cdot x^T. \tag{2.20}$$

The finite version of 2.20 is

$$\text{edge count} = s(G) = x_{1\times n}\cdot\ {}^bA_{n\times m} \cdot x^T_{1\times m} = [|E|]_{1\times 1}, \tag{2.21}$$

where n = number of rows and m = number of columns.

Assume ${}^bA$ is one-mode excluding self-ties $a_{ii} = 0$, and let $m = n$. Then $x = (1\ 1\ 1\ \cdots\ 1)_{1\times n}$ and

$$\text{edge count} = s(G) = \frac{1}{2}\cdot x\cdot\ {}^bA \cdot x^T = [|E|]_{1\times 1}. \tag{2.22}$$

Definition 2.22. Graph density is defined as the ratio of number of edges in the graph to the total possible number of edges in a graph.

Given the complete graph K_n, let ${}^1A_n$ be the matrix of ones, and I_n be the identity matrix. Consider,

$$x \cdot \left({}^1A_n - I_n\right) \cdot x^T = x \cdot {}^1A_n \cdot x^T - x \cdot I_n \cdot x^T = n^2 - n = n(n-1).$$

Then,

$$\text{graph density} = \frac{x \cdot {}^bA_n \cdot x^T}{n(n-1)}. \tag{2.23}$$

Remark 2.1. $0 \leq$ graph density ≤ 1.

Definition 2.23. For matrices A and B, the Frobenius inner product is defined by

$$A : B = \sum_i \sum_j a_{ij} b_{ij} = \text{trace}\left(A^T B\right) = \text{trace}\left(AB^T\right).$$

In particular,

$$\text{trace}(A^T \cdot A) = \text{trace}\left(A \cdot A^T\right).$$

Given the Edmonds matrix ${}^wAB$ of a bipartite graph, then the edge count of ${}^wAB$ including multiplicities is

$$\text{edge count} = \text{trace}\left({}^wAB \cdot^b AB^T\right) = \text{trace}\left({}^wAB^T \cdot^b AB\right).$$

This essentially means that the number of edges of the relations $A - A$ graph equals the number of edges of the relations $B - B$ graph; for example, the number of edges of the one-mode author-by-author network is the same as the number of edges of the one-mode paper-by-paper network.

Given the weighted proximity matrix ${}^wAA$ of a one-mode graph, then the edge count of ${}^wAA$ is

$$\text{edge count} = \frac{1}{2} \cdot \text{trace}\left({}^bAB \cdot^b AB^T\right) = \frac{1}{2}\text{trace}\left({}^wAA\right) = \sum_{i=1}^{n} {}^waa_{ii} = \frac{1}{4} x \cdot^w AA_n \cdot x^T.$$

For a bipartite graph $G(V^a, V^b, E)$, suppose ${}^bAB$ is the two-mode binary matrix, then

$$\text{edge count} = \sum_i^n \sum_j^m {}^bab_{ij}.$$

2.8 Network Diameter and Degree

Network diameter and distance between vertices are important in network theory. Distance is related to farness and closeness. The following algorithm gives Shimbel's geodesic distance matrix D for a graph $G(V, E)$. Assume $|V| = n$, $|E| = m$. Start with the incidence matrix ${}^{b}\overset{1}{V}E_{n\times m}$. Note that the incidence matrix may be treated as two-mode matrix.

First, compute the first order vertex-vertex relation matrix from the incidence matrix,

$$ {}^{w}\overset{1}{V}V_n = {}^{b}\overset{1}{V}E_{n\times m} \cdot {}^{b}\overset{1}{V}E^{T}_{n\times m} . $$

Define $VV_n^1 = {}^{b}\overset{1}{V}V_n - I_n$.

Then, compute the second order vertex-vertex matrix

$$ {}^{w}\overset{2}{V}V_n = {}^{b}\overset{1}{V}V_n \cdot {}^{w}\overset{1}{V}V_n . $$

Define $VV_n^2 = {}^{b}\overset{2}{V}V_n - I_n$.

Similarity, compute the third order vertex-vertex matrix

$$ {}^{w}\overset{3}{V}V_n = {}^{b}\overset{2}{V}V_n \cdot {}^{w}\overset{1}{V}V_n . $$

Define $VV_n^3 = {}^{b}\overset{3}{V}V_n - I_n$.

$\vdots$

In the same manner, compute the $d-$th order vertex-vertex matrix

$$ {}^{w}\overset{d}{V}V_n = {}^{b}\overset{d-1}{V}V_n \cdot {}^{w}\overset{1}{V}V_n . $$

Define $VV_n^d =^b \overset{d}{VV_n} - I_n$, where d is the minimum integer so that VV^d is a complete graph. The process is repeated until $VV_n^d =^1 A_n$.

Note that,

$$1 \leq d \leq n-1, \quad \text{provided } n \geq 2.$$

$$D_n = d \cdot VV^d - VV^{d-1} - \cdots - VV^3 - VV^2 - VV^1.$$

D_n is the matrix of geodesic paths for a graph $G(V, E)$.

d is the exponent (number of steps) needed to transform $G(V, E)$ to a complete graph (clique) of size n.

Definition 2.24. The diameter of a graph $G(V, E)$ is $diam(G) = \max(d_{ij} : d_{ij} \in D_n) = d$; the largest shortest path or longest geodesic between any two vertices. The radius of a graph $G(V, E)$ denoted by $rad(G) =$ shortest geodesic.

If the above routine fails to transforms the $G(V, E)$ into K_n in $d = n$ steps, then G contains components. A zero row or column vector in the distance matrix D indicates that the vertex is not reachable. In such a cases, $diam(G) = \infty$ and $rad(G) = \infty$. However, the method applies separately to find the largest geodesic in each component by repeating the same steps to each component. The largest shortest path of component j is the number d_j needed to make component j a complete subgraph.

The following are some special graphs and their diameter:

Let $G(V, E)$ is K_n; the complete graph then $diam(K_n) = 1$.

Definition 2.25. A path graph P_n of size $n-$vertices is a graph that contains vertices of

degree two and one. All vertices have degree 2 except the end vertices, which have degree 1. A path graph is a broken cycle graph.

If $G(V, E) = P_n$, then $diam(P_n) = n - 1$.

The sum of shortest paths (geodesics) for each vertex is the row sum or column sum of D.

Definition 2.26. Peripheral vertices are vertices having the largest geodesic, while vertices forming the center are vertices having the shortest geodesic.

Definition 2.27. Let $G(V, E)$ be a graph. The accessibility of $v_i \in V$ is defined by

$$A(v_i) = \sum_{i=1}^{n} d_{ij} = \sum_{j=1}^{n} d_{ij}, \quad d_{ij} \in D_n.$$

If $G(V, E)$ is directed graph, D_n may not be symmetric; i.e. $d(v_i, v_j)$ does not necessarily equal $d(v_j, v_i)$. Yet, if D_n is symmetric, the following holds true.

The vertex with the lowest summation value is considered the most accessible,

$$\min(A(v_i)) = \min\left(\sum_{i=1}^{n} d_{ij}\right) = \min\left(\sum_{j=1}^{n} d_{ij}\right),$$

However, the vertex the highest summation value is considered the least accessible,

$$\max(A(v_i)) = \max\left(\sum_{i=1}^{n} d_{ij}\right) = \max\left(\sum_{j=1}^{n} d_{ij}\right).$$

If $G(V, E)$ is a directed graph and $v_i \rightarrow v_j$ does not imply $v_j \rightarrow v_i$, then D_n still holds shortest paths between vertices i and j.

Definition 2.28. The degree of vertex v_i is defined by $deg(v_i) = {}^{w}\overset{1}{v}v_{ii} \in {}^{w}\overset{1}{V}V_n$, while the degree vector is defined by

$$diag\left({}^{w}\overset{1}{V}V_n\right) = \text{diag}\left({}^{b}\overset{1}{V}E_{n\times m} \cdot {}^{b}\overset{1}{V}E^{T}_{n\times m}\right) = diag\left({}^{b}\overset{1}{V}V_n \cdot {}^{b}\overset{1}{V}V_n\right)$$

${}^{w}\overset{1}{V}V_n$ is weak diagonally dominant matrix because ${}^{w}vv_{ii} = \sum_i vv_{ij}$ for $i \neq j$.

Definition 2.29. The maximum degree of a graph $G(V, E)$ is $\max({}^{w}vv_{ii}) = \left\| {}^{w}\overset{1}{V}V \right\|_{\infty}$, where ${}^{w}vv_{ii}$ is the degree of vertex i.

Definition 2.30. A pendant vertex v_i is a vertex satisfying the criterion $d(v_i) = 1$.

Removing pendant vertices from a network reduces the graph diameter, as a result, vertices with high degree centrality dominate. When pendant vertices become isolates the core of the network and cohesive subgroups stand out.

2.9 Line Graphs

Definition 2.31. The line graph of $G(V, E)$ also known the edge graph, denoted G^l, is a graph satisfying the following criteria:

1. Each vertex of $G^l\left(V^l, E^l\right)$ represents an edge of $G(V, E)$.

2. Two vertices v_i^l and v_j^l of $G^l\left(V^l, E^l\right)$ are adjacent; i.e. $\left(v_i^l, v_j^l\right) \in E^l$ if and only if their corresponding edges are adjacent in $G(V, E)$.

The line graph is intersection graph of the edges of $G(V, E)$, it represents the adjacencies between edges of $G(V, E)$.

van Rooij and Wilf (1965) showed that if G is connected the sequence $G, G^l, \left(G^l\right)^l, \left(\left(G^l\right)^l\right)^l, \cdots$ of line graphs have four possible behaviors:

1. If G is a cycle graph C_n then G^l and each subsequent line graph is isomorphic to G itself. Cyclic graphs are the only connected graphs for which $G^l \cong G$.

2. If G is a claw $K_{1,3}$, then G^l and all subsequent line graphs are C_3.

3. If G is a path graph P_n, then each subsequent line graph is a shorter path P_{n-1} until eventually P_0 terminates with an empty graph.

4. In all remaining cases, the sizes of the line graphs increase without bound.

Suppose $G(V, E) = K_n$; the complete graph with $|V| = n$ and $|E| = \frac{(n-1)n}{2}$, $n \geq 4$. let $G^l\left(V^l, E^l\right) = K_n^l$ be the line graph or edge graph of G with

$$\left|V^l\right| = |E| = \frac{(n-1)n}{2} \quad \text{and} \quad \left|E^l\right| = \sum_{i=1}^{\frac{(n-1)n}{2}} n - 2 = \frac{(n-2)(n-1)n}{2},$$

then $diam(K_n) = 1$ and $diam(K_n^l) = 2$.

Assume $G(V, E) = P_n$; the path graph with $|V| = n$ and $|E| = n - 1$. let $G^l\left(V^l, E^l\right) =$

$P_n^l = P_{n-1}$ be the line graph of G with

$$\left|V^l\right| = n-1 \text{ and } \left|E^l\right| = n-2,$$

then $diam(P_n) = n-1$ and $diam\left(P_n^l\right) = n-2$.

Corollary 7. If $G(V,E) = P_n$, then

$$P_n^{l^{n-1}} = \overbrace{l \circ l \circ \cdots \circ l(P_n)}^{(n-1)-times} = \overbrace{\left(\cdots\left(P_n^l\right)^l\cdots\right)^l}^{(n-1)-times} = P_2.$$

Assume $G(V,E) = C_n$; the cycle graph with $|V| = |E| = n$, $n \geq 3$. let $G^l\left(V^l, E^l\right) = C_n^l = C_n$ be the line graph corresponding to C_n with

$$\left|V^l\right| = \left|E^l\right| = n,$$

then $diam(C_n) = diam\left(C_n^l\right) = \frac{n}{2}$ if $n-$even and $diam(C_n) = diam\left(C_n^l\right) = \frac{n-1}{2}$ if $n-$odd. C_n is the self line-graph.

Assume $G(V,E) = S_n$; the star graph with $|V| = n$ and $|E| = n-1$. let $G^l\left(V^l, E^l\right) = S_n^l = K_{n-1}$ be the line graph corresponding to S_n with

$$\left|V^l\right| = |E| = n-1 \text{ and } \left|E^l\right| = \frac{(n-2)(n-1)}{2},$$

then $diam(S_n) = 2$ and $diam\left(S_n^l\right) = diam\left(K_{n-1}\right) = 1$.

Definition 2.32. A wheel graph W_n is a graph with $|V| = n$ vertices, formed by connecting a single vertex to all vertices of an $(n-1)-$cycle. The smallest wheel graph is $W_4 = K_4$.

Assume $G(V, E) = W_n$; the wheel graph with $|V| = n$ and $|E| = 2(n-1)$, $n \geq 4$. let $G^l\left(V^l, E^l\right) = W_n^l$ be the line graph corresponding to W_n with

$$\left|V^l\right| = |E| = 2(n-1) \text{ and } \left|E^l\right| = \frac{(n-1)(n+4)}{2},$$

then

$$diam(W_n) = 2, \quad \text{and} \quad diam\left(W_n^l\right) = \begin{cases} \frac{n-1}{2}, & n - \text{odd} \\ \frac{n}{2}, & n - \text{even}. \end{cases}$$

Remark 2.2. Let $G(V, E)$ be a graph. $d(v_i, v_j) = 2 \Leftrightarrow e_{ik}$ and e_{kj} are connected in G^L; the line graph of G, where $d(v_i, v_k) = d(v_j, v_k) = 1$. In a sense, all vertices having distance $d = 2$ in G become adjacent (will have distance $d = 1$) in G^l.

2.10 Summary

To sum up what I did in this chapter, I suggested a tool to store and manipulate constantly growing large scale evolutionary networks using smaller subnetworks called cliques. The great benefit of this matrix representation resides in the efficiency in performing matrix operations. I then invented a mechanism to expand on vertices by introducing the notion of infinitely dimensional network and matrix to tackle the problem of having vertices continuously introduced to the network, it can also be used to monitor the development of the network as a time series. Furthermore, I have developed an algorithm to obtain one-mode networks from weighted (valued) two-mode networks, then I have generalized this method

to work multi-mode networks, in which I have defined new matrix multiplications of high dimensional matrices. Finally, I have presented several mathematical tools that efficiently compute graph and network measures such as edge count, graph density, network diameter, degree centrality of vertices and degree and distance matrices. Finally, I have studied the duality between edges and vertices, in which edges become vertices and vertices become edges.

In the next chapter, I discuss methods to predict missing edges and vertices in a network based on information about vertices and edges. The method concerns vector product to derive a similarity measure between pair of edges or vertices. The method will then be extended to computed the similarity of hyperedges, in which the similarity is obtained for groups of vertices or edges rather than pairs.

Chapter 3: Estimating Missing Edges And Vertices

Because edges determine connectivity between vertices, they are crucial to the structure of networks and knowing whether or not there is a missing edge in an incompletely observed network is of great importance. In many sampled networks, edges are imperfectly observed because of under-coverage or because actors are intentionally suppressing their roles and linkages to serve different purposes.

In this chapter, I present a mathematical model to predict unobserved edges and vertices in a network based on covariate information on vertices and edges. The covariates are the exogenous attributes of entities. There are two types of attributes a set of vertices or edges can have, quantitative attributes, which are numerical summaries associated with entities and qualitative attributes, which are categorical summaries associated with entities. The model consists of two similarity measures calculated simultaneously using both the quantitative and the qualitative attributes derived externally as opposed to endogenous approach. In the process of computing the similarity measure between vertices using the quantitative information I use the inner (dot) product technique to obtain an estimate. On the other hand, I use contingency tables and the χ^2-test to obtain another estimate to compute the similarity using qualitative information. The probability of having an edge between two given vertices is then a weighted sum of the two estimates. If two pairwise vertices wind up having a high similarity measure then there is a high probability the vertices have edge connecting them or there is a high potential for forming an edge.

Vertices and edges do not necessarily have the same set of attributes. Depending on the network setup and properties of the entities, different networks may have completely different set of vertex attributes. Therefore, before applying the method of estimating missing

edges, covariate information need to be carefully defined. For example, in the author-coauthor social networks, possible attributes defined on authors and coauthors include *age, education, gender, spoken languages, discipline, number of publications.* However, possible attributes related to papers include *field, topic, keywords, year of publication, publisher, single/multiple author(s)* are the main attributes. In the alcohol-consumer settings, *age, ethnicity, smoker, drug-user, alcoholic, income, job-class* are possible consumers attributes, whereas *zip-code, location, hours-of-day, days-of-week* are some attributes associated with ABC stores.

When estimating missing edges (dyads), I assume that the network is stationary, i.e. given a time slice, the order of the graph $|V|$ is fixed.

3.1 The Inner Product Method For Estimating Missing Edges Using Quantitative Covariates

Given a vertex v_i. Let $\mathcal{A}^q$ be the set of all quantitative attributes associated with v_i,

$$\mathcal{A}^q = \left\{ A_1^q,\ A_2^q,\ \cdots,\ A_k^q,\ \cdots,\ A_{|\mathcal{A}^q|}^q \right\}.$$

For instance, the quantitative set may be

$$\mathcal{A}^q = \{age, income\}.$$

Each of the variables A_k^q, $1 \leq k \leq |\mathcal{A}^q|$, takes on numerical values a_k^q in $\mathbb{R}$. A_k^q in this sense is not necessarily a discrete variable, and the quantities a_k^q need to be normalized.

Let $V = \left\{ v_1,\ v_2,\ \cdots,\ v_i,\ \cdots,\ v_{|V|} \right\}$ be the set of vertices in a network setting.

Define, the vector of quantitative attributes associated with each vertex $v_i,\ 1 \leq i \leq |V|$ as follows

$$\vec{V}_{v_i}(A^q) = \left(a^q_{v_{i1}},\ a^q_{v_{i2}},\ \cdots,\ a^q_{v_{ik}},\ \cdots,\ a^q_{v_{i|A^q|}}\right), \quad \text{or}$$

$$\vec{V}_{v_i}(A^q) = a^q_{v_{i1}}\mathbf{e}_1 + a^q_{v_{i2}}\mathbf{e}_2 + \cdots + a^q_{v_{ik}}\mathbf{e}_i + \cdots + a^q_{v_{i|A^q|}}\mathbf{e}_{|A^q|}.$$

Assume v_i and v_j are two vertices with corresponding vectors of quantitative attributes $\vec{V}_{v_i}(A^q)$, $\vec{V}_{v_j}(A^q)$ respectively. Then the inner product of the two vectors is

$$\vec{V}_{v_i}(A^q) \cdot \vec{V}_{v_j}(A^q) = \left|\vec{V}_{v_i}(A^q)\right| \cdot \left|\vec{V}_{v_j}(A^q)\right| \cdot \cos(\theta_{ij}),$$

where θ is the angle between $\vec{V}_{v_i}(A^q)$ and $\vec{V}_{v_j}(A^q)$.

Therefore,

$$\mathcal{S}_q(v_i, v_j) = \cos(\theta_{ij}) = \frac{\vec{V}_{v_i}(A^q) \cdot \vec{V}_{v_j}(A^q)}{\left|\vec{V}_{v_i}(A^q)\right| \cdot \left|\vec{V}_{v_j}(A^q)\right|}$$

is defined to be the quantitative similarity measure between $\vec{V}_{v_i}(A^q)$ and $\vec{V}_{v_j}(A^q)$. Because the numerical values all fall in the first quadrant, the angle $0^\circ \leq \theta \leq 90^\circ$.

The only situation the similarity $\mathcal{S}_q = 1$ is when $\vec{V}_{v_j}$ is a constant multiple of $\vec{V}_{v_i}$; i.e. $\vec{V}_{v_j} = \alpha \cdot \vec{V}_{v_i}$ for some real number α. To avoid having similarity measure $\mathcal{S}_q = 1$ whenever $\alpha \neq 1$, I introduce another quantitative attribute unique to each vertex. This can be done by setting that value to be the normalized vertex identity; i.e. $\frac{i}{n}$, where $n = |V|$. The quantity can also be set by generating uniform non-repeating random numbers between 0 and 0.01, in which each vertex has another unique covariate. Thus, $\vec{V}_{v_j}$ can never be a constant multiple of $\vec{V}_{v_i}$.

θ can be used to measure the level of similarity between a pair of vertices (actors) in social networks. If $\theta \to 0^\circ$ the two vectors are close to parallel which implies that the two vertices are very similar with a high probability of having/forming an edge. On the other hand, if $\theta \to 90^\circ$ the two vectors are close to perpendicular which implies that the two vertices are very dissimilar with a high probability of not having/forming an edge.

Let $\mathcal{P}_q(E(v_i,\ v_j))$ be the estimate of the probability of an edge between vertex v_i and vertex v_j based on quantitative exogenous covariates, then

$$\mathcal{P}_q\left(E(v_i,\ v_j)\right) = \mathcal{S}_q(v_i, v_j) = \cos(\theta_{ij}) = \frac{\vec{V}_{v_i}(A^q) \cdot \vec{V}_{v_j}(A^q)}{\left|\vec{V}_{v_i}(A^q)\right| \cdot \left|\vec{V}_{v_j}(A^q)\right|}$$

The basic idea of similarity is to have a quantitative measure of those attributes that intersect, actors having close attribute values are more likely to be similar. Thus, an edge with high probability implies the two actors are very similar and that the link do in fact exist, and if otherwise not, this is an indication of a missing potential edge between vertices i and j. This could also mean that there is a high potential to form an edge in the future, which in both scenarios is useful information. Such information helps to predict potential edges in an imperfectly observed network and assists analysts to elucidate disambiguate relations among actors.

3.2 Contingency Tables For Qualitative Attributes

Consider a vertex v_i. Let $\mathcal{A}^c$ be the set of all categorical attributes associated with v_i,

$$\mathcal{A}^c = \left\{A_1^c,\ A_2^c,\ \cdots,\ A_k^c,\ \cdots,\ A_{|\mathcal{A}^c|}^c\right\},$$

where $1 \leq k \leq |\mathcal{A}^c|$. A_k^c in this sense is a qualitative variable.

As an illustration, the qualitative set may be

$$\mathcal{A}^c = \{job - class, eduction - level, spoken - language\}.$$

If A_k^c is a nominal variable such as *gender*, then A_k^c is coded 0, if vertex v_i is female, and 1 if vertex v_i is male. In this regard, the indicator function of v_i given an attribute set A_k^c is used to code the variable *gender*.

$$I_{v_i(A_k^c)} = \begin{cases} a_k^c = 1, & \text{if } v_i \text{ has attribute } A_k^c. \\ a_k^c = 0, & \text{if } v_i \text{ does not have attribute } A_k^c. \end{cases}$$

Let $V = \left\{v_1,\ v_2,\ \cdots,\ v_i,\ \cdots,\ v_{|V|}\right\}$ be the set of vertices in a network setting.

However, if A_k^c is ordinal, it is treated as if A_k^c is quantitative.

Define the list of categorical attributes associated with vertex v_i as follows

$$L_{v_i}(A^c) = \left(a_{v_{i1}}^c, a_{v_{i2}}^c, \cdots, a_{v_{ik}}^c, \cdots, a_{v_{i|\mathcal{A}^c|}}^c\right).$$

Construct the contingency table for each set of pairs of vertices in the following manner

Nodes	Attributes					
	A_1^c	A_2^c	$\cdots$	A_k^c	$\cdots$	$A_{\lvert\mathcal{A}\rvert}^c$
$v_i(\mathcal{A}^c)$	$a_{v_{i1}}^c$	$a_{v_{i2}}^c$	$\cdots$	$a_{v_{ik}}^c$	$\cdots$	$a_{v_{i\lvert\mathcal{A}^c\rvert}}^c$
$v_j(\mathcal{A}^c)$	$a_{v_{j1}}^c$	$a_{v_{j2}}^c$	$\cdots$	$a_{v_{jk}}^c$	$\cdots$	$a_{v_{j\lvert\mathcal{A}^c\rvert}}^c$

Let $\mathcal{P}_c(E(v_i,\ v_j))$ be the estimate of the probability of an edge between vertex v_i and vertex v_j, then

$$\mathcal{P}_c\left(E(v_i,\ v_j)\right) = p - \text{value},$$

where $p-$value is the probability value obtained from the χ^2-test. In general, $0 \leq p - \text{value} \leq 1$. If $p-$value $\to 0$ indicates a low similarity level between v_i and v_j, however, if $p-$value $\to 1$ indicates a high similarity level between v_i and v_j.

The combined similarity measure between vertices v_i and v_j is computed as a weighted sum of the two measures $\mathcal{P}_q$ and $\mathcal{P}_c$ as follows:

$$\mathcal{P}(E(v_i,\ v_j)) = \omega_q \cdot \mathcal{P}_q(E(v_i,\ v_j)) + \omega_c \cdot \mathcal{P}_c(E(v_i,\ v_j)),$$

where

$$\omega_q = \frac{|A^q|}{|A^q| + |A^c|} \quad \text{and} \quad \omega_c = \frac{|A^c|}{|A^q| + |A^c|}.$$

Note that $\omega_q + \omega_c = 1$.

3.3 Predicting Vertices

In the previous sections, I presented a mechanism to predict missing dyads and triads in an incompletely observed networks. Vertices are not less important than edges. In fact, actors are the main element of a network; without actors a network is meaningless. Actors play a significant role in determining the dynamics of a network. In this section, I will introduce a technique to estimate missing vertices (nodes) in a network. The method is again based on covariate information for vertices (actors) rather than edges, and utilizes the line space of edges which becomes the space of vertices as discussed in section (2.9).

In optimization theory, maximizing a problem in the dual space is equivalent to; and sometimes tends to be more feasible than, minimizing it in the original space.

In the line space of graphs, vertices become edges and edges become vertices. Consequently, to estimate a missing vertex in the space of graphs, it suffices to estimate the missing edge corresponding to that vertex in the line space of graphs. In this regards, I use a mapping to transform from the space of graphs to the line space. Because graphs and matrices are isomorphic (one-to-one and onto), there is a function (transformation) that takes the graph and transforms it from the original space onto the line space and vice versa using matrices. In this sense, the matrix is the operator.

Before I present the mapping, I would like to point out that first any one-mode graph can be expressed in terms of a two-mode graph using the incidence matrix as opposed to the adjacency matrix where the set of edges represent the second type set of vertices. In this context, a new edge is introduced connecting the new vertex (former edge) with the original set of vertices. Additionally, it is crucial to use the covariates of edges in the original space which are now the covariates of vertices in the line space when performing the estimation.

Let $V = \{v_1,\ v_2,\ \cdots,\ v_i,\ \cdots,\ v_{|V|}\}$ be a set of vertices. Furthermore, let

$E = \{e_1,\ e_2,\ \cdots,\ e_{|E|}\}$ be the set of edges associated with V.

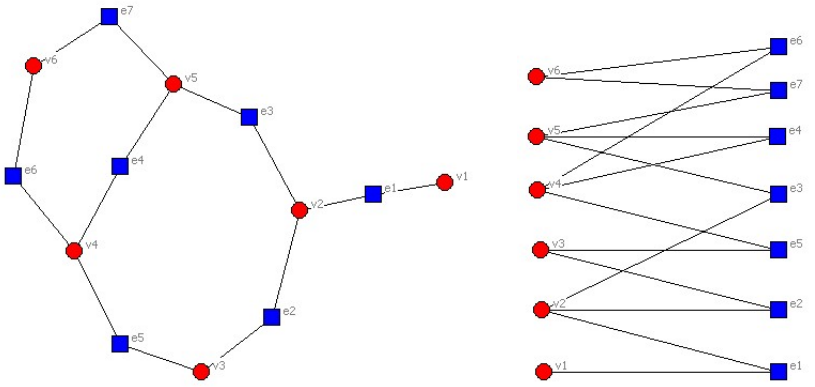
Given a graph $G(V, E)$. Let $f : G(V, E) \rightarrow G^l(V^l = E, E^l)$ be a mapping from the space of graphs onto the line space of graphs.

$$f = VE^T \cdot VE = EV \cdot VE = EE,$$

where VE is the vertex-edge incidence matrix corresponding to $G(V, E)$.

Note that $\left|V^l\right| = |E| \leq \left|E^l\right| \leq \frac{1}{2}|V|(|V| - 1)$. Moreover, in the line space every vertex (node) has at least degree = 2.

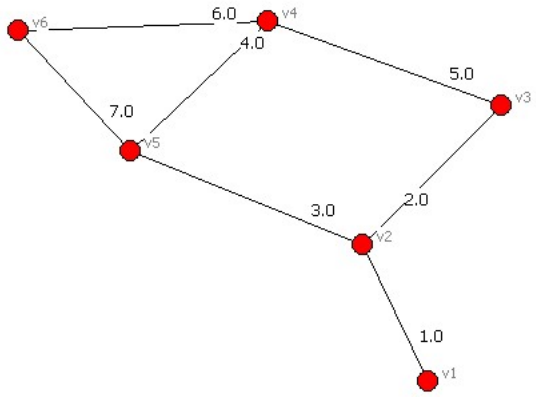
Example 2. To demonstrate this, consider the following graph $G(V, E)$, with $V = \{v_1,\ v_2,\ v_3,\ v_4,\ v_5,\ v_6\}$ and $E = \{e_1,\ e_2,\ e_3,\ e_4,\ e_5,\ e_6,\ e_7\}$. $|V| = 6$ and $|E| = 7$.



First, obtain the incidence matrix VE:

$VE =$

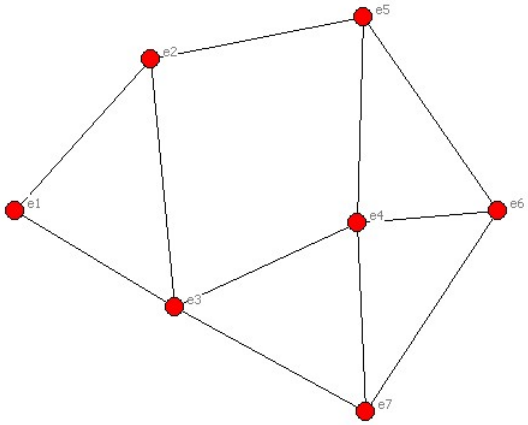
	e_1	e_2	e_3	e_4	e_5	e_6	e_7
v_1	1	0	0	0	0	0	0
v_2	1	1	1	0	0	0	0
v_3	0	1	0	0	1	0	0
v_4	0	0	0	1	1	1	0
v_5	0	0	1	1	0	0	1
v_6	0	0	0	0	0	1	1



The square matrix EE is the edge-edge representation of $G(V, E)$ in the line space, i.e. $G^l\left(E, E^l\right)$.

$EE =$

	e_1	e_2	e_3	e_4	e_5	e_6	e_7
e_1	0	1	1	0	0	0	0
e_2	1	0	1	0	1	0	0
e_3	1	1	0	1	0	0	1
e_4	0	0	1	0	1	1	1
e_5	0	1	0	1	0	1	0
e_6	0	0	0	1	1	0	1
e_7	0	0	1	1	0	1	0

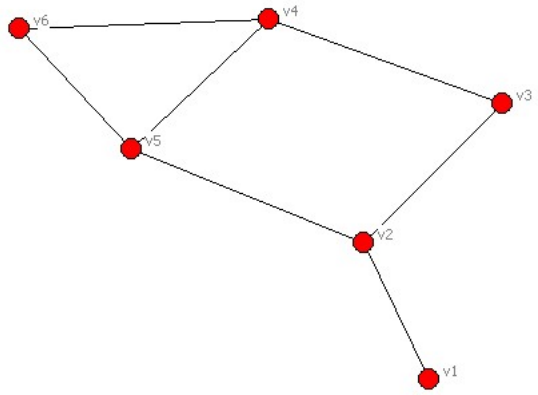


To map back to the original space of graphs we use the inverse transformation f^{-1}, where

$$f^{-1} = VE \cdot VE^T = VE \cdot EV = VV.$$

$VV =$

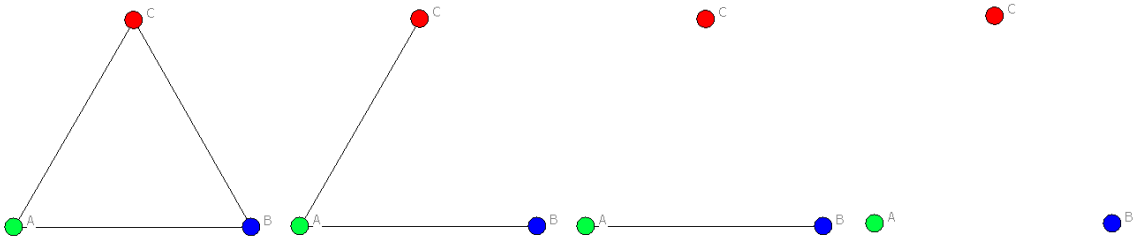
	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	0	0
v_2	1	0	1	0	1	0
v_3	0	1	0	1	0	0
v_4	0	0	1	0	1	1
v_5	0	1	0	1	0	1
v_6	0	0	0	1	1	0



VV is the vertex-vertex representation of $G(V, E)$.

3.4 Estimating Edges Of A Triad

Triadic relations are the basic foundation of edge dependency; the decision to form an edge may be dependent upon or related to the formation or absence of another edge(s) within a network. Because triads are the minimal representation of a cohesive subgroup in a network setting that involves more than one edge, they reveal how edges are affected by the external structure of the network. In this regard, I use exogenous covariate information on three entities to study the four possible scenarios of having edges in a triad, which are depicted in Figure 3.1. The first scenario is to have mutual agreement among all three actors; this is very interesting because this triad is a clique or a hyperedge of three vertices. It means that the three entities are very close where all ties are supposed to be strong. The second situation is to have a brokerage role. This is present in a path network or star network with three vertices. One vertex is maintaining connectivity with two disconnected vertices. In this sense, the brokerage communicates with both actors making sure the two actors are not interacting. Then, the third possibility is to have two mutually agreeing actors and a third isolated one. Finally, it is possible that all three actors have mutual disagreement and thus resulting in three isolated vertices. This indicates that all three actors are distant from each other and do not share any common interests or properties. Although, the term edge or

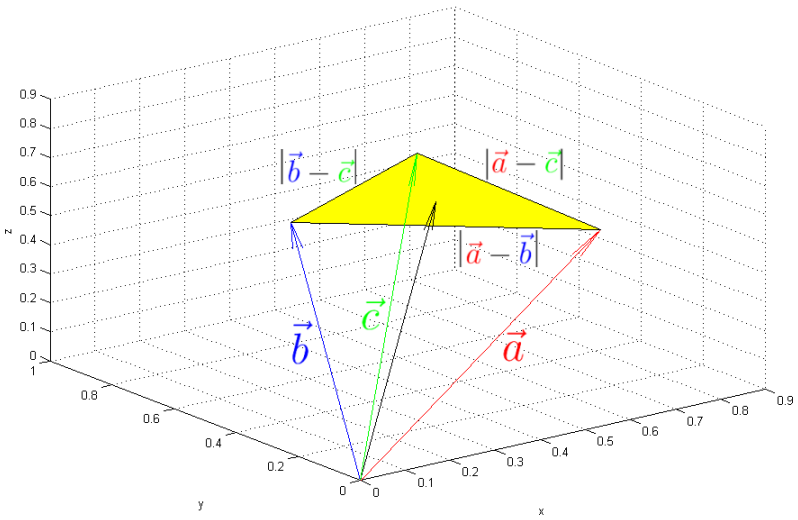


(a) Mutual agreement. (b) Brokerage. (c) Isolated dyad. (d) Mutual disagreement.

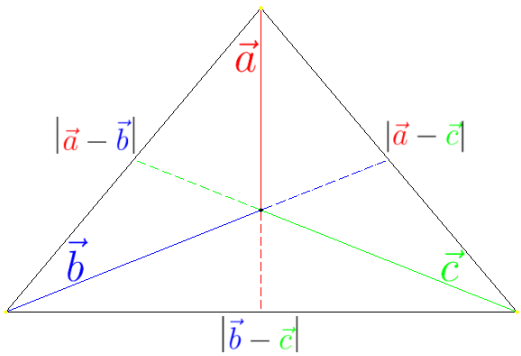
Figure 3.1: Possible Ties of a Triad.

diadic dependence may sound stochastic, which is modeled by exponential random graphs (P^* model), logistic model or Markov Chain Monte Carlo [24, 50, 57, 60], the approach I suggest is geometric and vector related.

Covariate information on three vertices define three unit vectors associated with each entity in the first quadrant forming a tetrahedron, thus the minimum space need to analyze such object is $\mathbb{R}^3$, see Figure 3.2. The geometric shape of the tetrahedron is the basis of connectivity in the triad, see Figure 3.4. The tetrahedron height represented by a vector always points in the direction of the centroid of the base triangle. Therefore, the distance between each vertex of the base triangle and the centroid determines edges among three vertices, see Figure 3.4. The more attributes the three vertices share among themselves the more close distance wise the three vectors are. The four situations described above are due to the fact that if the three vertices are within a epsilon distance from the base triangle centroid then they are more likely to be all connected, thus generating clique; see Figure 3.4(a). The extreme case of this situation is to have all three vectors coincide meaning the vertices are very similar; then $d(v_A, v_B) = d(v_A, v_C) = d(v_B, v_C) = 0$ and the angle between the vectors to $0°$. Yet, if only two vertices are below the threshold then the two vertices are more likely to be connected leaving the third vertex isolated; the two vertices are close to the centroid of the base triangle while the the third is distant; see Figure 3.4(b). The extreme case of this situation is to have two coincidental lines with the third being perpendicular. It is also possible for vertex to be smaller than the critical point while the other two vertices pass

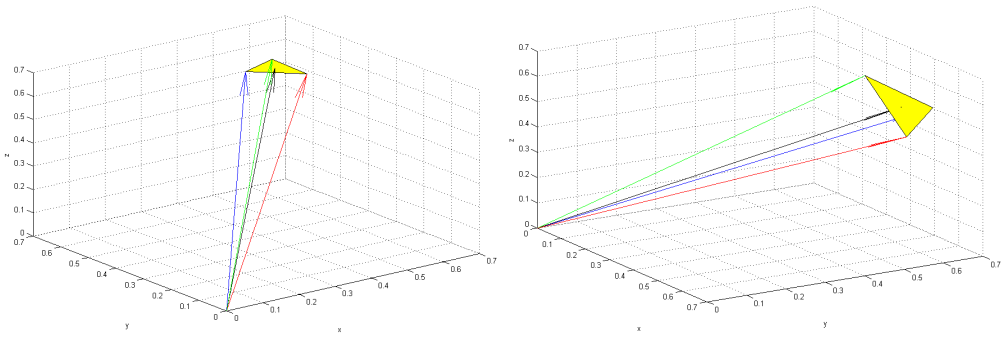


(a) Tetrahedron formed by three vectors.

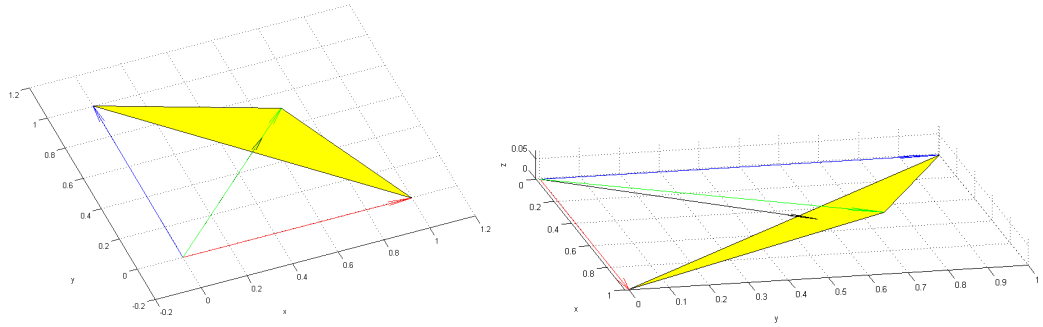


(b) Tetrahedron base triangle.

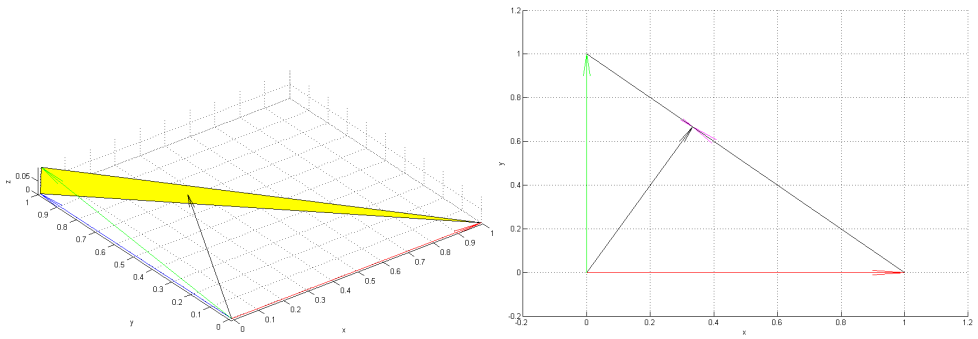
Figure 3.2: Tetrahedron and its base triangle.



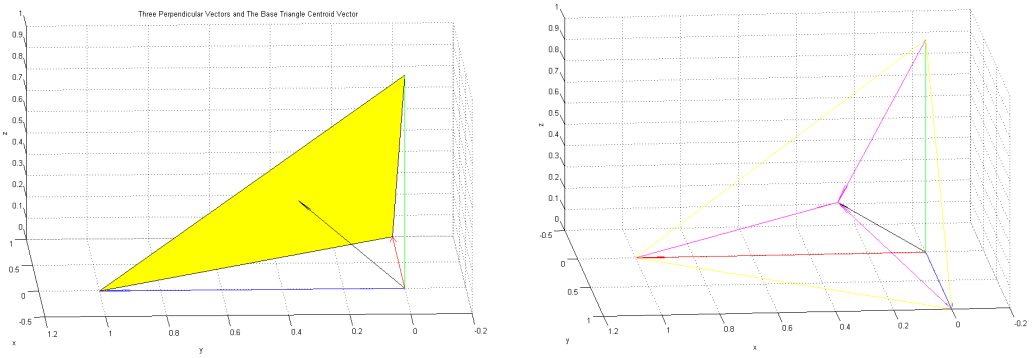
(a) Three almost parallel vectors results in a cohesive group.



(b) Two perpendicular vectors and a third lying half-way results in a structural hole.

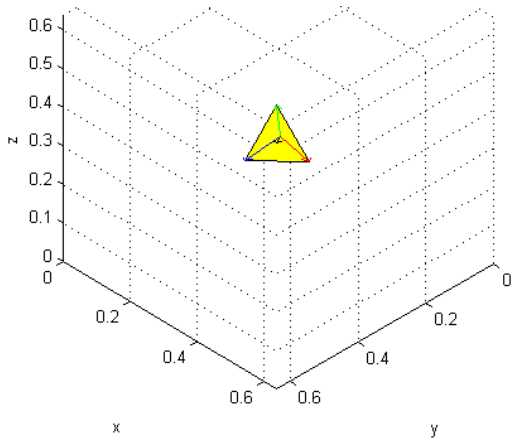


(c) Two perpendicular vectors and parallel third results in an isolated dyad.

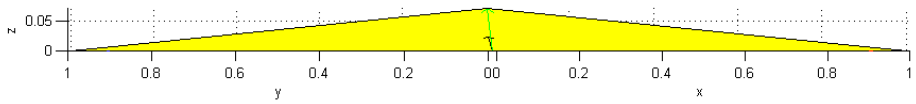


(d) Three perpendicular vectors results in isolated vertices.

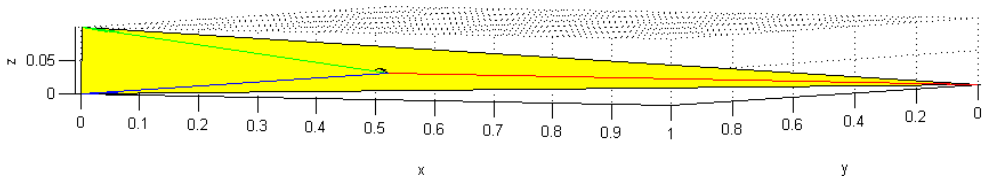
Figure 3.3: Scenarios of three vectors in space.



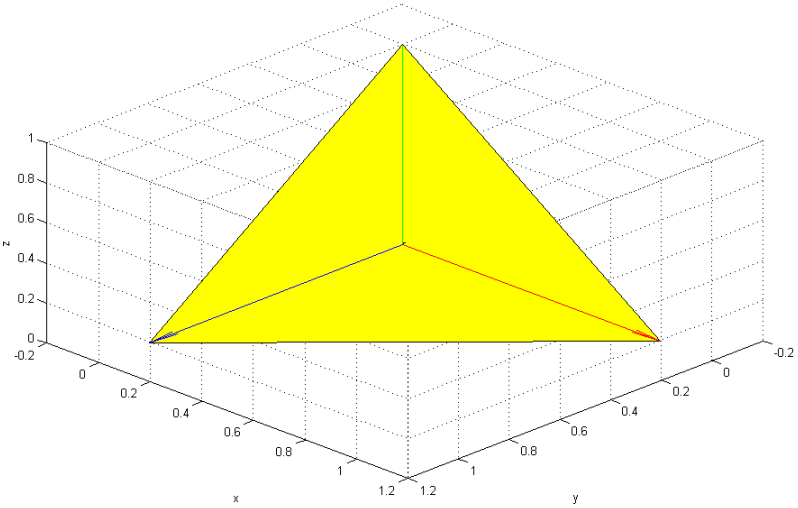
(a) All three vectors are within ϵ distance from the base triangle centroid.



(b) One vector is close to the base triangle centroid, the other two are distant.



(c) Two vectors are below the threshold while the third exceeds it.



(d) All three vectors are far from the base triangle centroid.

Figure 3.4: Distance of unit vectors from the base triangle centroid.

the critical value; this leads to one vertex being connected to two disconnected vertices in which their vectors are distant from the centroid of the base triangle, see Figure 3.4(c). The extreme case of this situation is to have two perpendicular vectors and a third vector lying on the plane joining the two vertices with a 45° from both vectors. All three vectors lie on a plane. Finally, if the distance between the three vertices and the centroid of the base triangle is more than epsilon then the three vertices are dissimilar and thus they are disconnected, see Figure 3.4(d). The extreme case of this situation is to have three perpendicular vectors; very dissimilar. The distance between any two vertices is then $\sqrt{2}$ and the perimeter of the base triangle is $3\sqrt{2}$.

In Figure 3.2(b), the centroid of the base triangle $\vec{C}$ is found by averaging the coordinates of the vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ respectively using the following formula

$$\vec{C} = \frac{1}{3}\left(\vec{a} + \vec{b} + \vec{c}\right).$$

The length of each side is

$$\left|\vec{a} - \vec{b}\right| = \sqrt{\vec{a}^2 + \vec{b}^2 - 2|\vec{a}| \cdot |\vec{b}| \cdot \cos\left(\theta_{\vec{a}\vec{b}}\right)} = \sqrt{2\left(1 - \cos\left(\theta_{\vec{a}\vec{b}}\right)\right)},$$

$$|\vec{a} - \vec{c}| = \sqrt{\vec{a}^2 + \vec{c}^2 - 2|\vec{a}| \cdot |\vec{c}| \cdot \cos\left(\theta_{\vec{a}\vec{c}}\right)} = \sqrt{2\left(1 - \cos\left(\theta_{\vec{a}\vec{c}}\right)\right)},$$

$$\left|\vec{b} - \vec{c}\right| = \sqrt{\vec{b}^2 + \vec{c}^2 - 2|\vec{b}| \cdot |\vec{c}| \cdot \cos\left(\theta_{\vec{b}\vec{c}}\right)} = \sqrt{2\left(1 - \cos\left(\theta_{\vec{b}\vec{c}}\right)\right)}.$$

$\vec{a}$, $\vec{b}$, and $\vec{c}$ are all unit vectors.

The triangle has a perimeter of

$$p = \sqrt{2}\left(\sqrt{1-\cos\left(\theta_{\vec{a}\vec{b}}\right)} + \sqrt{1-\cos\left(\theta_{\vec{a}\vec{c}}\right)} + \sqrt{1-\cos\left(\theta_{\vec{b}\vec{c}}\right)}\right).$$

This means that the maximum possible perimeter $3\sqrt{2}$; it is obtained when all three vectors are perpendicular to each other. Thus, they are all disconnected (no edges) as in Figure 3.1(d). A perimeter of zero indicates that all three vertices are parallel. Thus, they are all connected (hyperedge connecting three vertices) as in Figure 3.1(a). Therefore, $0 \leq p \leq 3\sqrt{2}$. Generally, if $2\sqrt{2} \leq p < 3\sqrt{2}$ then two vertices are connected while the third is isolated (one edge) as in Figure 3.1(c). However, and if $3\sqrt{2} - 2 \leq p < 2\sqrt{2}$ then one vertex is connected to two disconnected vertices (two edges) as in Figure 3.1(b); this is often known as a star or path graph of size 3.

The volume of the tetrahedron and the area of its base triangle are also useful, the formulas are derived below. Without loss of generality, suppose $\vec{ca} = \vec{a} - \vec{c}$ and $\vec{cb} = \vec{b} - \vec{c}$. Then the area of the base triangle is given by,

$$A = \frac{1}{2}\left|\vec{ca} \times \vec{cb}\right| = \frac{1}{2}\sqrt{3 + 4\cos\left(\theta_{\vec{a}\vec{c}}\right) \cdot \cos\left(\theta_{\vec{b}\vec{c}}\right) - 2\cos\left(\theta_{\vec{a}\vec{c}}\right) - 4\cos\left(\theta_{\vec{b}\vec{c}}\right) - \cos^2\left(\theta_{\vec{a}\vec{c}}\right)}.$$

The volume of the tetrahedron is given by,

$$V = \frac{1}{6}\left|\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)\right|$$

$$V = \frac{1}{6}\sqrt{1 + 2\cos\left(\theta_{\vec{a}\vec{b}}\right) \cdot \cos\left(\theta_{\vec{a}\vec{c}}\right) \cdot \cos\left(\theta_{\vec{b}\vec{c}}\right) - \cos^2\left(\theta_{\vec{a}\vec{b}}\right) - \cos^2\left(\theta_{\vec{a}\vec{c}}\right) - \cos^2\left(\theta_{\vec{b}\vec{c}}\right)}.$$

The four extreme scenarios described above are discussed more in detail below

- Scenario 1: $\theta_{\vec{ab}} = \theta_{\vec{ac}} = \theta_{\vec{bc}} = 90°$; the three vectors are perpendicular, which implies that the three vertices are all disconnected, see Figures 3.1(d) and 3.3(d). In such a case,

$$p = 3\sqrt{2}, \quad A = \frac{\sqrt{3}}{2}, \quad V = \frac{1}{6}.$$

- Scenario 2: $\theta_{\vec{ab}} = \theta_{\vec{ac}} = 90°$ and $\theta_{\vec{bc}} = 0°$; two of the vectors are within zeros distance while the third is perpendicular to both, which means that two vertices are connected and the third is disconnected, see Figures 3.1(c) and 3.3(c). In such a case,

$$p = 2\sqrt{2}, \quad A = 0, \quad V = 0.$$

- Scenario 3: $\theta_{\vec{ab}} = 90°$ and $\theta_{\vec{ac}} = \theta_{\vec{bc}} = 45°$; two of the vectors are perpendicular while the third is halfway between the two, which means that a vertex is connected to two disconnected vertices, see Figures 3.1(b) and 3.3(b). In such a case,

$$p = 3\sqrt{2} - 2, \quad A = \frac{1}{2}\sqrt{3 - 2\sqrt{2}}, \quad V = 0.$$

- Scenario 4: $\theta_{\vec{ab}} = \theta_{\vec{ac}} = \theta_{\vec{bc}} = 0°$; the three vectors are parallel, which implies that the three vertices are all connected, see Figures 3.1(a) and 3.3(a). In such a case,

$$p = 0, \quad A = 0, \quad V = 0.$$

The problem with this method of identifying triads is that the fact that it does not reveal which vertices are connected/disconnect. But, terribly enough it is possible to misclassify triads if the tetrahedron is a regular tetrahedron; for example, if each side of the base triangle has length $\left|\vec{a} - \vec{b}\right| = |\vec{a} - \vec{c}| = \left|\vec{b} - \vec{c}\right| = 0.8$, then $p = 2.4 > 3\sqrt{2} - 2$. Meanwhile, the

three vectors are equidistance from each other meaning they should be either all connected or all disconnect; yet, they are classified as the triad having two edges. The above method is good in predicting the triads in Figures 3.1(a) and 3.1(d), but not practical in predicting the triads in Figures 3.1(b) and 3.1(c). Consequently, the method is inconsistent in terms of predicting the correct triad and requires improvement. One solution to tackle this problem is to avoid having regular tetrahedron which may be achieved by adding a small random or unique perturbation to each vector.

However, the centroid of the tetrahedron base triangle can be utilized to address the issue of inconsistency through measuring the distance of each of the three vectors from from the centroid vector. The centroid acts like the average distance or center of mass of the three vectors. The question to ask then is how far each entity is away from that center of mass. If a vertex is at a distance from the centroid below a certain threshold then the vertex is more likely to be connected to at least one other vertex.

Predicting an edge between only two vertices is straightforward because the two vertices lie at an equidistance from the centroid and thus are both below the critical point; i.e. connected, or above it; i.e. disconnected. But, when a third vector is present that judgement may not be obvious. The new vector may be close to one vector and distant from the other; in such a case, two vertices of the base triangle are near the centroid while the third is far from it; thus, the triad contains only one edge. Or, the new vector is positioned halfway between the other two vectors; in such a case two vertices of the base triangle are far from the centroid while the third is close to it; thus, the triad contains two edges. Ultimately, if all three vertices of the base triangle are near the centroid the triad contains three edges (clique) and if all three vertices of the base triangle are distant from the centroid the triad contains no edges. This takes care of the problem of having vertices with the same distance from the centroid.

Having said that, the issue now is how to predict edges among three vertices mathematically besides visually. The solution is by using fuzzy logic. Assume the distance from the centroid is identified as close or far. Furthermore, define a boolean variable associated with vertices that has the value y if it is close to the centroid and n if it is far from the centroid. The tag y for "yes" means the vertex is available/willing to make an interaction, while the tag n for "no" is an indicator that the vertex is not available/willing to make an interaction. Then, I define the fuzzy operator $\oplus$ on three vertices (vectors) with two choices y and n as follows: $y \oplus y = 1$, $n \oplus n = 0$, and

$$y \oplus n = n \oplus y = \begin{cases} 1, & \text{if two vectors have tags } n \text{ and } n. \\ 0, & \text{if two vectors have tags } y \text{ and } y. \end{cases}$$

All four possible combinations of the boolean variable defined using fuzzy logic on three vertices are presented in Table 3.1 and resemble the triads in Figure 3.1. Essentially, if the first two vertices have tags n and the third has tag y then $y \oplus n = 1$. The third vertex must make interaction with at least one other vertex and because the other two have tags n, it is forced to connect to both of them allowing the brokerage triad with two edges to be present. Using the same logic, it also means that $y \oplus n = 0$, given the fact that the first two vertices have tags y and the third has tag n. The third vertex must make no interaction with at least one other vertex and because the other two have tags y, it is forced to connect to none because the other two vertices agree to form an edge leaving the third out; it allows for the triad with one edge to be present. Finally, if all vertices have tags y then they all mutually agree on communicating setting $y \oplus y = 1$, which allows the clique triad to be present. And, if all vertices have tags n then they all mutually disagree on communicating setting $n \oplus n = 0$, which allows for the triad with no edges to be present.

Remains to discuss is the choosing of critical distance from the centroid at which the tags y

Table 3.1: Two-level fuzzy operator defined on three vertices.

$\oplus$	y	y	y
y	$\star$	1	1
y	1	$\star$	1
y	1	1	$\star$

$\oplus$	y	n	n
y	$\star$	1	1
n	1	$\star$	0
n	1	0	$\star$

$\oplus$	y	y	n
y	$\star$	1	0
y	1	$\star$	0
n	0	0	$\star$

$\oplus$	n	n	n
n	$\star$	0	0
n	0	$\star$	0
n	0	0	$\star$

and n are set. For this purpose, I take the mean distance of all distances from the centroid for the four extreme scenarios. If all vectors are parallel then the distance from the centroid is 0 for each vertex. If two vectors are perpendicular and the third is at equidistance from both then the distance from the perpendicular vectors to the centroid is $\frac{\sqrt{6-\sqrt{2}}}{3}$ and from the third to the centroid is $\frac{\sqrt{6-4\sqrt{2}}}{3}$. If two vectors are parallel and the third is perpendicular to both then the distance from the centroid to the parallel vectors is $\frac{\sqrt{2}}{3}$ and from the centroid to the perpendicular vector is $\frac{2\sqrt{2}}{3}$. Finally, if the three vectors are all perpendicular then the distance from the centroid is $\sqrt{\frac{2}{3}}$. Hence, the average distance is 0.4965.

The boolean tags y and n may be extended to include a third option m for "maybe" for vertices that are not too close or too far from the centroid of the base triangle; yielding in six combinations; namely, $y \oplus y$, $y \oplus m$, $y \oplus n$, $m \oplus m$, $m \oplus n$, and $n \oplus n$. There are four clear cases and two fuzzy ones. The obvious cases are $y \oplus y = 1$, $y \oplus m = m \oplus y = 1$, $m \oplus n = n \oplus m = 0$, and $n \oplus n = 0$. Yet, the fuzzy situations are

$$y \oplus n = n \oplus y = \begin{cases} 1, & \text{if two vectors have tags } (n \text{ and } n) \text{ or } (n \text{ and } m). \\ 0, & \text{if two vectors have tags } (y \text{ and } y) \text{ or } (y \text{ and } m). \end{cases}$$

$$m \oplus m = \begin{cases} 1, & \text{if the third vector has tag } n \text{ or } m. \\ 0, & \text{if the third vector has tag } y \ . \end{cases}$$

The ten combinations of the three levels of closeness from the centroid defined by the fuzzy operator are depicted in Table 3.2 and resemble the triads in Figure 3.1. If two of the vertices have tags n and n or tags m and m and the third has tag y then $y \oplus n = 1$ and $m \oplus m = 0$. The vertex with the tag y must make interaction with at least one other vertex and because the other two capture the values $n \oplus n = 0$ or $y \oplus m = 0$, it is forced to connect to both vertices allowing for the brokerage triad with two edges to be present. Concurrently, if two vertices have tags y and y or tags m and m and the third has tag n then $y \oplus n = 0$ and $m \oplus m = 1$. The vertex with the tag n must make no interaction with at least one other vertex and because the other two capture the values $y \oplus y = 1$ or $m \oplus n = 0$, it is forced to connect to none because the other two vertices have an edge in common which leaves the third vertex isolated; this in turn allows for the triad with one edge to be present. Finally, if all vertices have tags y, y and y, or tags y, y and m or tags m, m and m, then they all mutually agree and thus $y \oplus y = 1$, $y \oplus m = 1$ and $m \oplus m = 1$, which allows the clique triad to be present. And, if all vertices have tags n, n and n or tags n, n and m, then they all mutually disagree and thus $n \oplus n = 0$, $n \oplus m = 0$, which allows for the triad with no edges to be present.

I have demonstrated a method to estimate triadic edges using a two-level fuzzy operator and a three-level fuzzy operator irrespective of edge dependency which is the core of the Markov Chain Monte Carlo model to predict edges in general. The method can be generalized to any clique size using the same analogy. The two-level fuzzy operator generated one mutually connected vertices (triadic edge), one isolated vertex (dyadic edge), one brokerage vertex (star or path graph), and one mutually disconnect vertices (isolates). However, the three-level fuzzy operator produced three mutually connected vertices (triadic edge), three isolated vertex (dyadic edge), two brokerage vertex (star or path graph), and two mutually disconnect vertices (isolates). Thus, the three-level fuzzy operator does not equally distribute edges.

Table 3.2: Three-level fuzzy operator defined on three vertices.

$\oplus$	y	y	y
y	$\star$	1	1
y	1	$\star$	1
y	1	1	$\star$

$\oplus$	y	y	m
y	$\star$	1	1
y	1	$\star$	1
m	1	1	$\star$

$\oplus$	m	m	m
m	$\star$	1	1
m	1	$\star$	1
m	1	1	$\star$

$\oplus$	y	n	n
y	$\star$	1	1
n	1	$\star$	0
n	1	0	$\star$

$\oplus$	y	m	m
y	$\star$	1	1
m	1	$\star$	0
m	1	0	$\star$

$\oplus$	y	y	n
y	$\star$	1	0
y	1	$\star$	0
n	0	0	$\star$

$\oplus$	y	m	n
y	$\star$	1	0
m	1	$\star$	0
n	0	0	$\star$

$\oplus$	m	m	n
m	$\star$	1	0
m	1	$\star$	0
n	0	0	$\star$

$\oplus$	n	n	n
n	$\star$	0	0
n	0	$\star$	0
n	0	0	$\star$

$\oplus$	m	n	n
m	$\star$	0	0
n	0	$\star$	0
n	0	0	$\star$

3.5 Summary

The main theme of this chapter is the study of edges and vertices and how they can be predicted (if they are missing in a network) using covariates associated with vertices and edges respectively. In Chapter 2 I addressed the interchangeability and duality between vertices and edges in a graph, which was the foundation to estimating the probability of vertices in an unobserved network. I applied the inner product method on the vector of covariates to estimate the probability of dyadic edges. Moreover, I have extended the method to predict triadic edges using covariate information as well; however, the method is based on geometry and fuzzy logic rather than the inner product of two vectors. To this end, I made the assumption that if two actors share many common values and attributes then they are considered close to each other which means that these actors are similar and thus the probability of interaction increases.

Chapter 4: Evolutionary Networks And Preferential Attachment

In this chapter, I develop a theory on evolutionary networks using concepts from calculus such as the difference and the average rate of change. The method helps detecting emerging "elite" actors within a network. Actors with high interacting rates are considered important actors. I then utilize tools and methods presented in Chapters 2, 3 to build a model of preferential attachment in which dyadic edges are generated based on the similarity measure computed on the vectors of covariates.

4.1 Evolving Networks And Emerging "Elite" Groups

Having data on a social network over time provides an insight on the formation and evolution of the network, which includes vertex-vertex interaction in the form of introducing and strengthening edges as well as introducing vertices. For each time slice, there is a network of actors and ties and a graph. Graphs may be dependent upon each other in which the formation of edges and vertices at time t is related to the status of edges and vertices at time $t-1$. They may also be treated independently, but are not assumed to be disjoint.

Time series analysis on evolving networks is useful in detecting emerging subgroups or subnetworks within the mother network. For example, the emergence of scientific subfields in author-coauthor networks, the emergence of alcoholic communities in an alcohol society, the emergence of elevated crime areas in an alcohol ecological system, and the emergence of a disease in a community.

Below I suggest a mathematical algorithm to identify emerging cliques in a network from

a series of matrices. Assume actors are allowed to make only one interaction with at least one other actor in the network at each time step, that is to say edges' weight is increased by one at each time and cliques may be present as well.

Definition 4.1. Let ${}^{w}\overset{0}{M}, {}^{w}\overset{1}{M}, {}^{w}\overset{2}{M}, \cdots, {}^{w}\overset{t}{M}$ be the weighted adjacency matrices of an evolving one-mode individual-by-individual social network of vertices over the time period $T = \{0, 1, 2, \cdots, t\}$ or a sequence of t matrices, then

$$ {}^{w}\overset{t}{D} = {}^{w}\overset{t}{M} - {}^{w}\overset{0}{M} $$

is the time t difference (change) adjacency matrix.

${}^{w}\overset{t}{D}$ can be thought of as the weighted change in the ties strengths since time $t_0 = 0$. A zero entry in the ${}^{w}\overset{t}{D}$ matrix, i.e. ${}^{w}\overset{t}{d}_{ij} = 0$, where ${}^{w}\overset{t}{d}_{ij} \in {}^{w}\overset{t}{D}$, indicates no change or interaction between actors i and j. If ${}^{w}\overset{t}{m}_{ij} = {}^{w}\overset{t}{d}_{ij}$, where ${}^{w}\overset{t}{m}_{ij} \in {}^{w}\overset{t}{M}$, then actor i and actor j have formed new edge with magnitude ${}^{w}\overset{t}{d}_{ij}$ since time $t_0 = 0$.

Now consider,

$$ {}^{w}\overset{t}{C} = \frac{1}{t} \cdot {}^{w}\overset{t}{D}, $$

${}^{w}\overset{t}{C}$ represents the subnetwork(s) of emerging group(s) within the original network, where ${}^{w}\overset{t}{c}_{ij}$ is a measure for the rate at which actors i and j are strengthening their ties at time t. Because edge weights are incremented by one and thus the maximum difference at time t gives edge weight difference of t, ${}^{w}\overset{t}{C}$ generates numbers that are between 0 and 1. ${}^{w}\overset{t}{C}$ may be used as the edge probability matrix.

A column or row vector ${}^{t_i}_{w}R$ of ${}^{t}_{w}D$, where

$$ {}^{t_i}_{w}r = {}^{t_{ij}}_{w}D $$

represents all the actors that actor i interacted with, including clique interactions.

Example 3. Assume ${}^{0}_{w}M, {}^{1}_{w}M, {}^{2}_{w}M, {}^{3}_{w}M$ are the one-mode weighted adjacency matrix of a network over four time periods with

${}^{0}_{w}M =$

	v_1	v_2	v_3	v_4	v_5
v_1	0	1	3	0	0
v_2	1	0	2	4	0
v_3	3	2	0	1	0
v_4	0	4	1	0	0
v_5	0	0	0	0	0

, ${}^{1}_{w}M =$

	v_1	v_2	v_3	v_4	v_5
v_1	0	1	3	0	0
v_2	1	0	2	4	0
v_3	3	2	0	2	0
v_4	0	4	2	0	0
v_5	0	0	0	0	0

${}^{2}_{w}M =$

	v_1	v_2	v_3	v_4	v_5
v_1	0	2	3	0	0
v_2	1	0	3	4	0
v_3	3	3	0	3	1
v_4	0	4	3	0	0
v_5	0	0	1	0	0

, ${}^{3}_{w}M =$

	v_1	v_2	v_3	v_4	v_5
v_1	0	2	4	0	0
v_2	2	0	3	4	0
v_3	4	3	0	4	2
v_4	0	4	4	0	0
v_5	0	0	2	0	0

The time 3 difference matrix is then

$$
{}^{w}\overset{3}{D} = {}^{w}\overset{3}{M} - {}^{w}\overset{0}{M} =
$$

	v_1	v_2	v_3	v_4	v_5
v_1	0	1	1	0	0
v_2	1	0	1	0	0
v_3	1	1	0	3	2
v_4	0	0	3	0	0
v_5	0	0	2	0	0

$$
, \quad {}^{w}\overset{3}{C} = \frac{1}{3} \cdot {}^{w}\overset{3}{D} =
$$

	v_1	v_2	v_3	v_4	v_5
v_1	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0
v_2	$\frac{1}{3}$	0	$\frac{1}{3}$	0	0
v_3	$\frac{1}{3}$	$\frac{1}{3}$	0	1	$\frac{2}{3}$
v_4	0	0	1	0	0
v_5	0	0	$\frac{2}{3}$	0	0

The matrices ${}^{w}\overset{3}{D}$ and ${}^{w}\overset{3}{C}$ reveal features about the formation of the network represented by the matrix M_3 that is otherwise hard to detect. For instance, the clique $v_1 - v_2 - v_3$ is an emerging subgroup as well as the star subnetwork. As a matter of fact, the central figure v_3 is making interaction with every other actor in the network. Not only v_3 and v_4 are maintaining high interaction, but also they capture the highest possible time 3 difference rate and thus making them important figures. ${}^{w}\overset{3}{C}$ suggests that $v_3 - v_4$ and $v_3 - v_5$ are more likely to continue strengthening their ties.

4.2 Preferential Attachment Using Covariate Information

In this section, I propose a mathematical model that utilizes similarity measures calculated on actors using covariate information. The measures are used to predict the behavior of evolving networks. The model is based on the transition probabilities in a finite state stochastic process in discrete time, in which the network status at the current state relies on the previous network status according to a Markov chain process. The sequence of states is time dependent and recursive. Thus, the transition adjacency matrix at time $t + 1$ can

be expressed as a function of the previous transition adjacency matrix at time t:

$$\overset{t+1}{{}_{w}A}_{\infty} = f_1\left(\overset{t+1}{{}_{w}B}_{\infty}\right),$$

where $\overset{t+1}{{}_{w}B}_{\infty} = f_2\left(\overset{t}{{}_{w}B}_{\infty}\right)$. This implies that

$$\overset{t+1}{{}_{w}A}_{\infty} = f_3\left(\overset{t}{{}_{w}B}_{\infty}\right).$$

But $\overset{t}{{}_{w}A}_{\infty} = f_2\left(\overset{t}{{}_{w}B}_{\infty}\right)$.

Hence,

$$\overset{t+1}{{}_{w}A}_{\infty} = f_1\left(\overset{t}{{}_{w}A}_{\infty}\right) = \cdots = f_t\left(\overset{0}{{}_{b}A}_{\infty}\right).$$

I decompose the adjacency matrix at time $t+1$ and express it as the sum of two matrices, one representing the observed matrix at time t, while the other matrix corresponds to the new change in the edges or the transition probabilities.

Let $\overset{t}{{}_{w}A}_{\infty}$ be the infinite square weighted adjacency matrix corresponding to an observed network (social or nonsocial) at time t. At time $t+1$, the network undergoes one of three possible types of change,

1. inactive actors become active by forming new edges,
2. already existing connected actors (active ones) strengthen their ties/form new ties, or
3. 1 and 2 simultaneously.

For large time intervals, the evolution of the network can be examined through the ${}^{w}\overset{t}{D}_{n}$ time t difference adjacency matrix. For small t, the change may not be noticeable. However, covariate information on actors may be helpful in estimating the change in the status among actors between states. Therefore, the covariate similarity matrix is used to estimate the observed matrix at time $t+1$.

Let ${}^{b}\overset{t+1}{T}_{\infty}$ be the weighted adjacency covariate matrix of actors at time $t+1$. ${}^{b}\overset{t+1}{T}_{\infty}$ is obtained in the following manner:

1. $t'_{ij} = \mathcal{S}(v_i, v_j) = P\left(E(v_i, v_j)\right)$.

2. If $t'_{ij} \geq 0.7$, then $t_{ij} = 1$ is an edge with multiplicity one.

The actual adjacency matrix, which resembles the status of the network at time $t+1$, can be approximated by

$$ {}^{w}\overset{t+1}{A}_{\infty} \sim {}^{w}\overset{t+1}{\hat{A}}_{\infty} = {}^{w}\overset{t}{\hat{A}}_{\infty} + {}^{b}\overset{t+1}{T}_{\infty} \, . $$

Or,

$$ {}^{w}\overset{t+1}{A}_{\infty} = {}^{w}\overset{t+1}{\hat{A}}_{\infty} + {}^{b}\overset{t+1}{e}_{\infty} = {}^{w}\overset{t}{\hat{A}}_{\infty} + {}^{b}\overset{t+1}{T}_{\infty} + {}^{b}\overset{t+1}{e}_{\infty}, $$

Now we can write the predictor matrix ${}^{w}\overset{t+1}{\hat{A}}_{\infty}$ in terms of the predictor matrix ${}^{w}\overset{t}{\hat{A}}_{\infty}$ and the covariates similarity matrix ${}^{b}\overset{t+1}{T}_{\infty}$ as follows

$$ {}^{w}\overset{t+1}{\hat{A}}_{\infty} = {}^{b}\overset{0}{A}_{\infty} + {}^{b}\overset{1}{T}_{\infty} + {}^{b}\overset{2}{T}_{\infty} + \cdots + {}^{b}\overset{t}{T}_{\infty} + {}^{b}\overset{t+1}{T}_{\infty}, $$

where ${}^{b}\overset{0}{A}_{\infty}$ is the initial observed network at time $t = 0$.

The similarity matrix $\mathcal{S}$ is updated at every time step before the covariate matrix ${}^{t+1}_{b}T_{\infty}$ is computed for the next step; thus, ${}^{t+1}_{b}T_{\infty}$ varies over time. However, if I make the assumption that ${}^{t+1}_{b}T_{\infty}$ undergos a constant change at the micro-level time, then similarities among actors are held constant on the short run and as a result we have,

$$ {}^{t+1}_{w}\hat{A}_{\infty} = {}^{0}_{b}A_{\infty} + (t+1) \cdot {}^{t+1}_{b}T_{\infty} \, . $$

This indicates that edge evolution between actors grows linearly, i.e. actors strengthen their ties following a linear function.

Suppose a new actor v_i has been introduced to the network. Then, based on the covariate similarity measure $S(v_i, v_j)$, actor v_i prefers to attach to actor v_j if the similarity is very high, namely, $S(v_i, v_j) \geq 0.7$.

4.3 Summary

I have suggested models for evolutionary networks and preferential attachment that are based on the theory I have put together in Chapters 2, 3. Vertex connectivity (dyadic edge) is a common property of many large-scale networks. As far as evolutionary networks, the model takes into account that the number of vertices is continuously growing, edges are constantly being introduced, or edges weights increase in value. As far as preferential attachment, the more connected a vertex is, the more likely it is to acquire new edges.

Chapter 5: Applications To Networks

There are many applications to this discipline. In this chapter, I case study the author-coauthor social networks and suggest a model of preferential attachment in emerging scientific fields. Researchers and faculty members benefit from publishing in terms of financial compensation and prestige. Distinguished scholars have developed, over time, certain styles of coauthorship depending on their career and field of study. The purpose of analyzing coauthorship networks is to be able to answer questions such as "who-wrote-with-whom", how often coauthors publish, who maintains strong relations leading to four basic styles of coauthorships.

Then, I explore the network of road fatal crashes across the United States and analyze relationship and similarities between states and crash factors as well. This network is an example of a non-social static two-mode network in which the number of states and crash factors are fixed.

Additionally, I present the network of news documents. For this non-social network, I construct the two-mode term-document network and bigram-document network. I then derive the one-mode document-document network with respect to terms and bigrams using the methods discussed in Chapter 2.

I close this chapter with two examples of simulated social networks; one demonstrating a coauthorship social network, while the other demonstrating an online music friendship network.

5.1 Edward Wegman Coauthorship Social Network

To begin, I would like to provide the reader some background on Dr. Wegman. Edward J. Wegman is a prominent professor of Statistics at George Mason University, Fairfax VA, USA. He received his Ph.D. degree in Mathematical Statistics from University of Iowa in 1968. Immediately after his Ph.D., he went to the Statistics Department at the University of North Carolina, Chapel Hill, which was one of the leading statistics departments in the world. His early career focused on the development of aspects of the theory of mathematical statistics. In 1978, he went to the Office of Naval Research (ONR) where he was the Head of the Mathematical Sciences Division. He has been in the research and academia field for some time and has published an array of work, which includes over 200 prestigious refereed journal, articles, books, and technical reports, authored individually and with a number of colleagues and Ph.D. students.

In this example, I examine the structure of the autho-coauthor social network and model its behavior on multiple levels. For this purpose, I perform a comprehensive social network analysis on the first-level and second-level network.

Data were collected from his personal website and his updated curriculum vitae. Initially, I built the one-mode adjacency/proximity matrix manually in *MS-Excel*, but at later stage I was able to obtain the one-mode matrix from the two-mode matrix through matrix multiplication. The first-level coauthorship network is of size 102×102. This matrix is symmetric because relationships among actors are reflexive; if author A published with author B then this also implies B published with A. It is worth mentioning though that co-authorship networks are not generally symmetric.

5.1.1 Network Visualization

I explore the network first then I present a detailed analysis of the network; Figure 5.1 shows the first-level layout of the network. The general structure of the network is a weighted digraph consisting of vertices (coauthors) and weighted edges (ties) representing frequency of coauthorship. The graph is an example of "ego" style network; in which all vertices are connected to one focal vertex, see Figure 2.1. In graph theory terms, this is referred to as a star graph with a network diameter of 2 and density of 0.0986.

The star-like network is a governing feature of any first-level coauthorship network due to

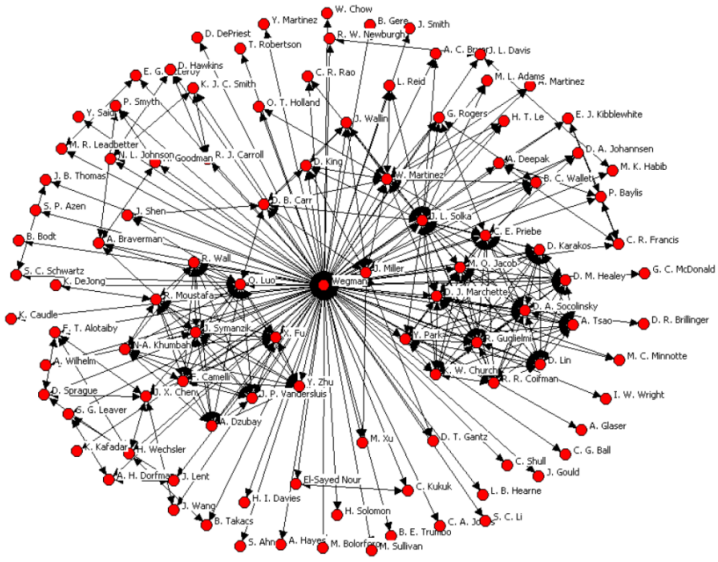


Figure 5.1: Wegman's author-coauthor social network.

the fact that all coauthors are explicitly connected to that one main author. Some coauthors share edges not only with Wegman, but also with other coauthors; more than one name may appear on a paper. There are two visible "clouds" (clump networks) fully connected – complete subgraphs, formally known as cliques.

Figure 5.2 shows the matrix representation of the network, each black square indicates a coauthorship relation. In the figure, block number 1 represent the principal author Wegman. The block diagonal structure indicates strong clusters (cliques), and the black horizontal

and vertical lines suggest that the principal author maintains interactions with all other co-authors.

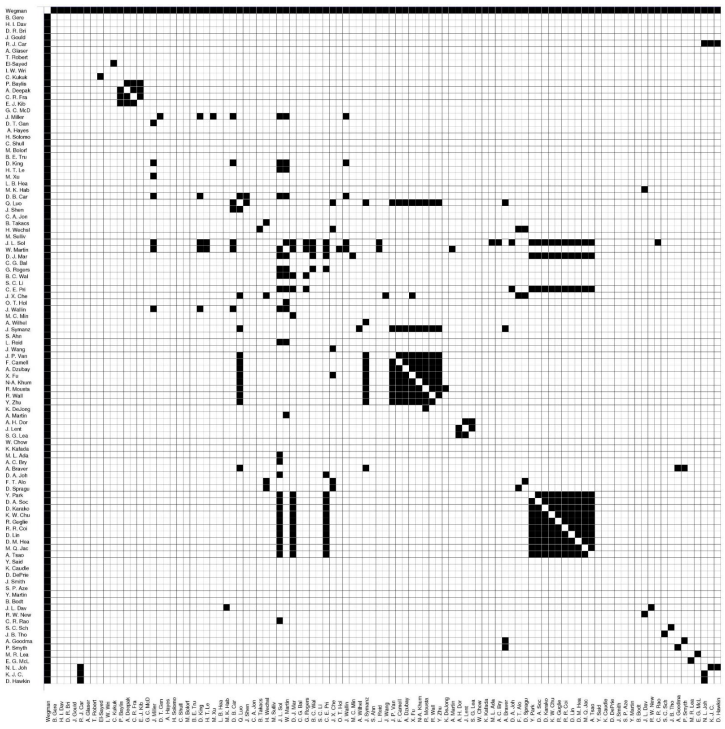


Figure 5.2: Adjacency matrix of Wegman's network.

Figure 5.3 portrays an optimized blockmodel partition with three clusters based on structural equivalence.

5.1.2 Centrality Measures

In social network concepts, centrality metrics are quantitative tools used to measure central figures or organizations in the network. Centrality provides information on position of actors on the individual level. Conceptually, centrality is about identifying actors residing in the center (core) of the network, such actors have access to information.

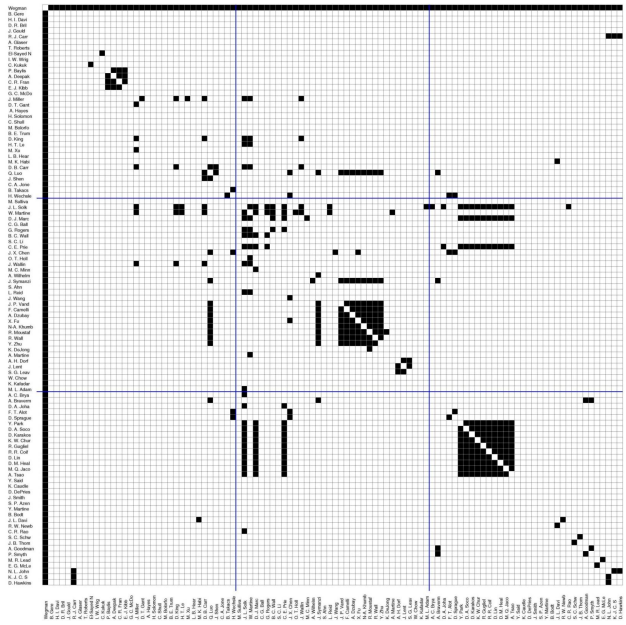


Figure 5.3: Random partition with three clusters.

I ran the centrality metrics vertex degree – a local measure, and closeness on Wegman's first-level coauthorship network; Figure 5.4 shows the results. Aside from Wegman, who is the most central person connected to all other actors in the network because he wrote with every member of the network; thus he captured the highest vertex degree, Solka with normalized vertex degree of 25.743 comes in second place. The third place is shared by Marchette and Priebe with normalized vertex degree of 15.842.

Using closeness centrality measure, Wegman captured the first place. Solka has the second highest closeness with a normalized value of 57.386, he is considered important because he is relatively close to all other actors. Marchette and Priebe have a closeness normalized value of 54.301, capturing the third place. W. Martinez with a normalized closeness value of 53.723 comes in fourth place. Notice that no other coauthor has a normalized closeness value less than 50.249. The reason is that the first-degree coauthorship network resembles a star network where most coauthors are close to the principal author. While the first-level network has limited structure, the second-level coauthorship network is expected to reveal

		Degree	Closeness				
1	Wegman	100.000	100.000	51	L. Reid	2.970	50.754
2	B. Gere	0.990	50.249	52	J. Wang	1.980	50.500
3	H. I. Davies	0.990	50.249	53	J. P. Vandersluis	9.901	52.604
4	D. R. Brillinger	0.990	50.249	54	F. Camelli	9.901	52.604
5	J. Gould	0.990	50.249	55	A. Dzubay	9.901	52.604
6	R. J. Carroll	3.960	51.010	56	X. Fu	10.891	52.880
7	A. Glaser	0.990	50.249	57	N-A. Khumbah	9.901	52.604
8	T. Robertson	0.990	50.249	58	R. Moustafa	10.891	52.880
9	El-Sayed Nour	1.980	50.500	59	R. Wall	9.901	52.604
10	I. W. Wright	0.990	50.249	60	Y. Zhu	9.901	52.604
11	C. Kukuk	1.980	50.500	61	K. DeJong	1.980	50.500
12	P. Baylis	3.960	51.010	62	A. Martinez	1.980	50.500
13	A. Deepak	3.960	51.010	63	A. H. Dorfman	2.970	50.754
14	C. R. Francis	3.960	51.010	64	J. Lent	2.970	50.754
15	E. J. Kibblewhite	3.960	51.010	65	S. G. Leaver	2.970	50.754
16	G. C. McDonald	0.990	50.249	66	W. Chow	0.990	50.249
17	J. Miller	7.921	52.062	67	K. Kafadar	0.990	50.249
18	D. T. Gantz	1.980	50.500	68	M. L. Adams	1.980	50.500
19	A. Hayes	0.990	50.249	69	A. C. Bryant	1.980	50.500
20	H. Solomon	0.990	50.249	70	A. Braverman	4.950	51.269
21	C. Shull	0.990	50.249	71	D. A. Johannsen	2.970	50.754
22	M. Bolorforoush	0.990	50.249	72	F. T. Alotaiby	3.960	51.010
23	B. E. Trumbo	0.990	50.249	73	D. Sprague	3.960	51.010
24	D. King	5.941	51.531	74	Y. Park	12.871	53.439
25	H. T. Le	2.970	50.754	75	D. A. Socolinsky	12.871	53.439
26	M. Xu	1.980	50.500	76	D. Karakos	12.871	53.439
27	L. B. Hearne	0.990	50.249	77	K. W. Church	12.871	53.439
28	M. K. Habib	1.980	50.500	78	R. Guglielmi	12.871	53.439
29	D. B. Carr	7.921	52.062	79	R. R. Coifman	12.871	53.439
30	Q. Luo	13.861	53.723	80	D. Lin	12.871	53.439
31	J. Shen	2.970	50.754	81	D. M. Healey	12.871	53.439
32	C. A. Jones	0.990	50.249	82	M. Q. Jacob	12.871	53.439
33	B. Takacs	1.980	50.500	83	A. Tsao	12.871	53.439
34	H. Wechsler	4.950	51.269	84	Y. Said	0.990	50.249
35	M. Sullivan	0.990	50.249	85	K. Caudle	0.990	50.249
36	J. L. Solka	25.743	57.386	86	D. DePriest	0.990	50.249
37	W. Martinez	13.861	53.723	87	J. Smith	0.990	50.249
38	D. J. Marchette	15.842	54.301	88	S. P. Azen	0.990	50.249
39	C. G. Ball	0.990	50.249	89	Y. Martinez	0.990	50.249
40	G. Rogers	4.950	51.269	90	B. Bodt	0.990	50.249
41	B. C. Wallet	4.950	51.269	91	J. L. Davis	2.970	50.754
42	S. C. Li	0.990	50.249	92	R. W. Newburgh	1.980	50.500
43	C. E. Priebe	15.842	54.301	93	C. R. Rao	1.980	50.500
44	J. X. Chen	6.931	51.795	94	S. C. Schwartz	1.980	50.500
45	O. T. Holland	1.980	50.500	95	J. B. Thomas	1.980	50.500
46	J. Wallin	5.941	51.531	96	A. Goodman	2.970	50.754
47	M. C. Minnotte	1.980	50.500	97	P. Smyth	2.970	50.754
48	A. Wilhelm	1.980	50.500	98	M. R. Leadbetter	1.980	50.500
49	J. Symanzik	11.881	53.158	99	E. G. McLeroy	1.980	50.500
50	S. Ahn	0.990	50.249	100	N. L. Johnson	3.960	51.010
				101	K. J. C. Smith	2.970	50.754
				102	D. Hawkins	2.970	50.754

Figure 5.4: Normalized vertices degree and closeness for all actors.

more features.

Figure 5.5, is a modified version of Figure 5.1, emphasizing nodes degree and tie strength. Nodes color and size are set by the attribute vertex degree, while edges color and thickness are set by tie strength – frequency of communication. Color palettes of nodes degree and ties strength are shown in Figure 5.5. Solka has the second largest vertex degree, and Priebe and Marchette have the same vertex color and size. Additionally, W. Martinez and Luo have relatively large nodes with the same vertex color. As far as considering ties strength instead degree centrality; Solka has the highest frequency of coauthorship then W. Martinez. Concurrently, another interesting hidden feature is revealed by the graph, Solka and W. Martinez have the strongest tie among all coauthors, they coauthored 17 times more than any other two coauthors did with the exception of the principal author. The edges (Solka, Priebe) and (Solka, Marchette) have the same color and thickness, which suggests that Priebe and Marchette coauthored with Solka the same number of times, in fact, they

coauthored four times.

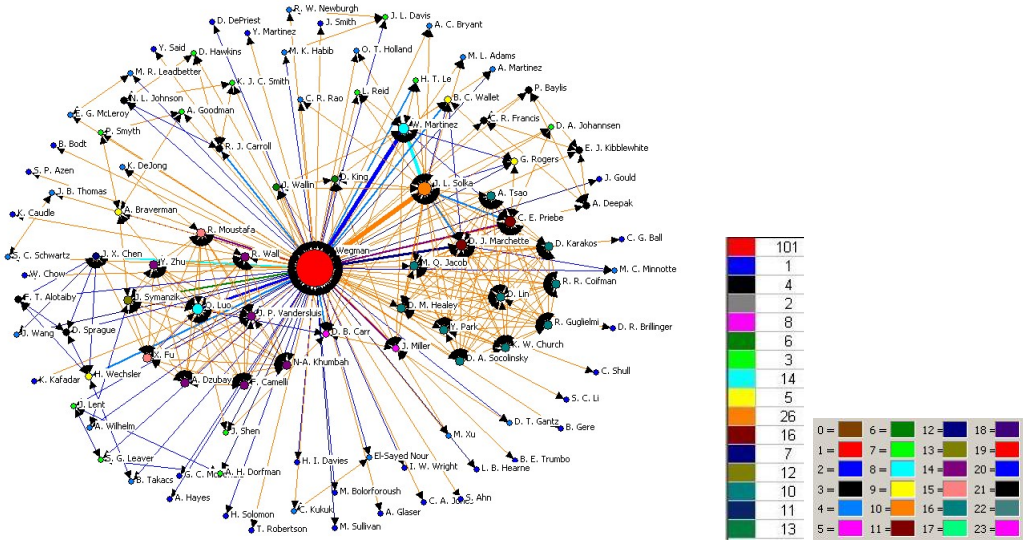


Figure 5.5: Wegman's coauthorship network. Color palettes. Left: nodes degree, right: tie strength.

To discover secondary structures and investigate connectivity I exclude the star vertex "Wegman", see Figure 5.6. The graph is partially disconnected with several isolated nodes; in this case, Wegman is considered a cutpoint and the edges {(Carr, Luo), (Carr, Shen)} are considered local bridges. Global bridges are less frequent in such networks. Both of Carr and Luo are cut-points; by removing either the subnetwork becomes disconnected. The figure also shows that several coauthors are still connected forming cohesive subnetworks such as the actors {Lent, Leaver, Dorfman}.

5.1.3 Cohesive Subgroups

Network cohesion can reveal highly connected subgroups – active networks acting in conjunction with the mother network, captured through cliques. In Wegman's network there are 36 cliques, Figure 5.7 shows all 36 clique sets. For example, clique number 11 contains the nodes (Wegman, Solka, Bryant), clique number 9 contains the actors (Wegman, Solka,

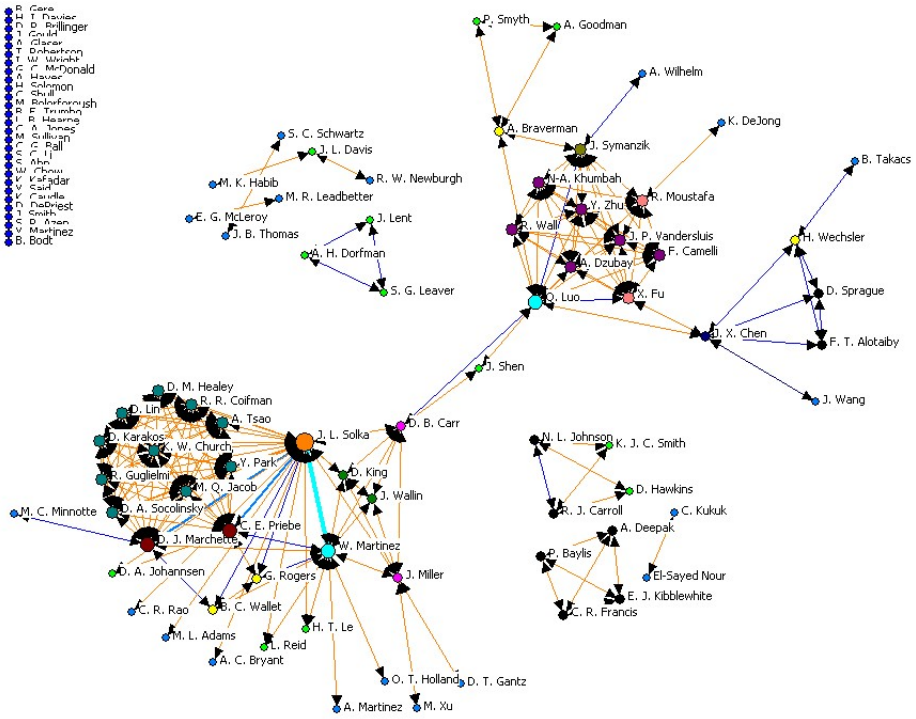


Figure 5.6: Wegman's network without Wegman.

W. Martinez, Reid), clique number 2 contains the actors (Wegman, Solka, W. Martinez, Marchette, Priebe). Actors of cohesive groups can be members of one or more cliques simultaneously such as Solka.

Cliques in a graph may overlap – the same vertex or set of nodes may belong to more than one clique (some cliques contain more than one member in common). It possible though that some nodes may not belong to any clique. However, no clique can be entirely contained within another clique, because if it were the smaller clique then it would not be maximal. Figure 5.8 shows the clique overlap. There is a considerable overlap among the cliques in the coauthorship relation, more than one coauthor belongs to one or more cliques.

Even more interesting, suppose that actors in one network form two non-overlapping cliques; and that the actors in another network also form two cliques, but that the memberships

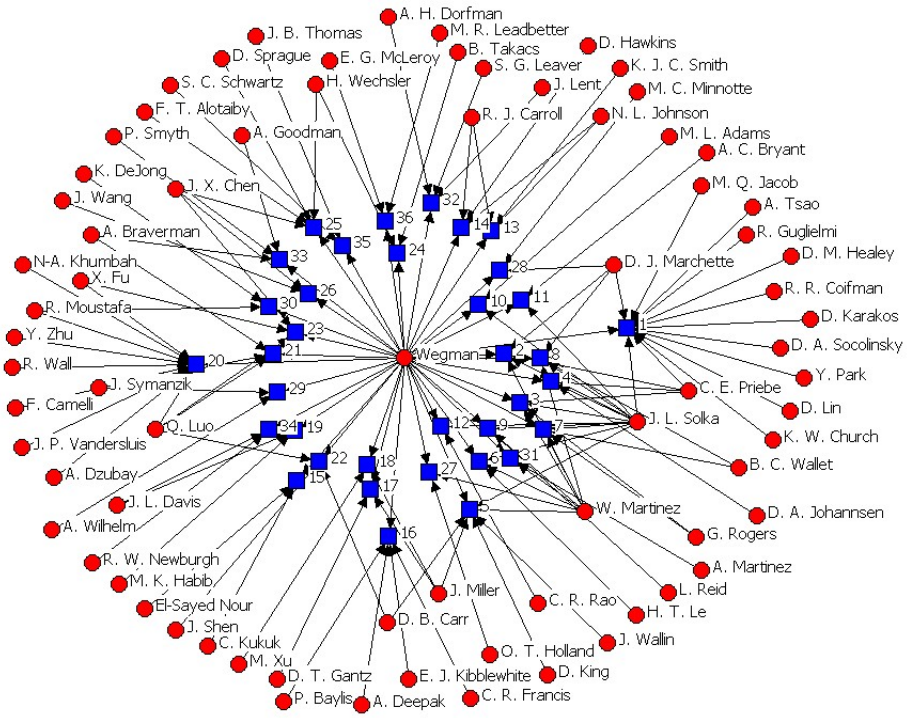


Figure 5.7: The 36 clique sets in Wegman's network.

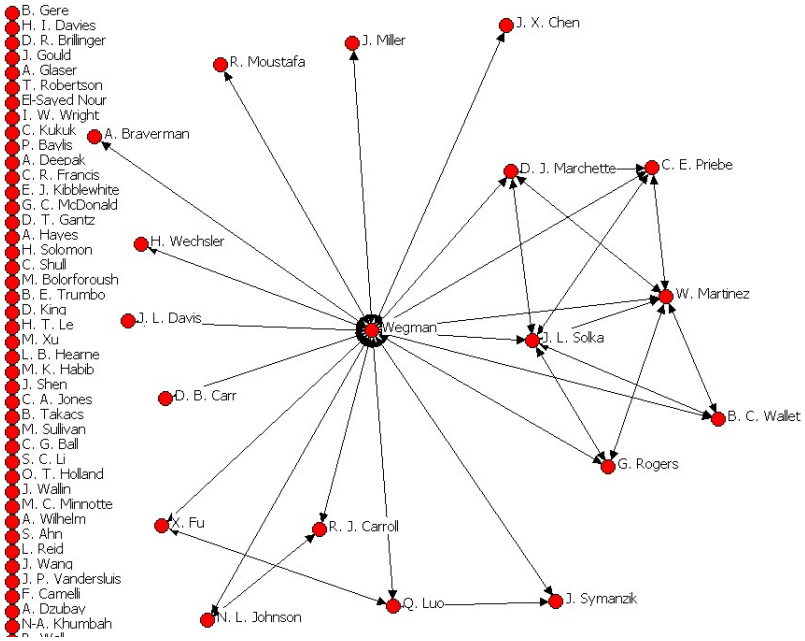


Figure 5.8: The clique overlap in Wegman's network.

overlap (some people are members of both cliques). Where the groups overlap, conflict between them is less likely than when the groups do not overlap [28]; Wegman, Solka, W. Martinez and Marchette demonstrates this feature. Where the groups overlap, mobilization and diffusion may spread rapidly across the entire network; where the groups do not overlap, traits may occur in one group and not diffuse to the other.

5.1.4 Block-Modeling

The method of partitioning actors into subsets so that actors within each subset are closer to being equivalent than are actors in different subsets is known in the network literature as blockmodel. Figure 5.9 shows the cluster diagram of Wegman's network. Actors at each level of the dendrogram are structurally equivalent. The highest level of the dendrogram indicates all actors are the same, however, the lowest level of the dendrogram indicates all actors are different. What is between theses levels is more informative in terms of structural equivalence.

A partition of a network is a classification or clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster [55]. Partitions divide the nodes of a network into a number of mutually exclusive (disjoint) subsets. Figure 5.10 shows these two clumps clustered in the upper left corner of the adjacency matrix, there is a total of four clusters in the graph. Actors of these clusters are structurally equivalent. The graph is based on random start block-modeling applied on the network using structural equivalence with four clusters.

5.1.5 Discarding Weak Ties

Some coauthors published only few times and maintained low interaction with the principal author. They may be graduated students or colleagues who are no longer maintaining strong ties with the principal author. Weak ties – edges with coauthorship frequency of 1

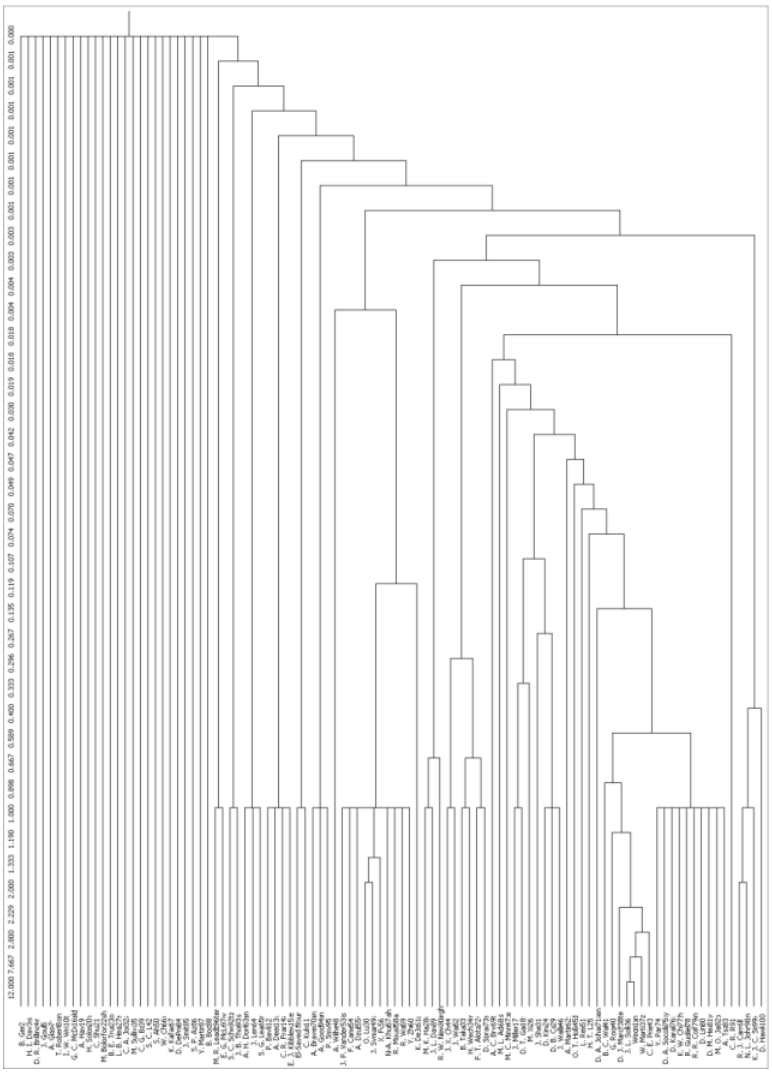


Figure 5.9: Dendrogram of Wegman's network.

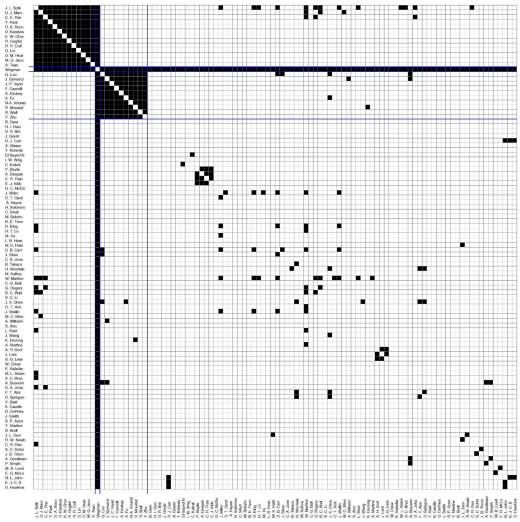


Figure 5.10: Random blockmodel using structural equivalence with four clusters.

have minimal impact on the structure of the network and as a result they are discarded. Coauthors with tie strength equal to one are assumed to not have coauthored with the principal author and hence are treated as isolated nodes. Figure 5.11 shows all edges with frequency ≥ 1 together with their corresponding edge weight. As before, nodes' color and size are set by the attribute "vertex degree" while edges' color and thickness are set by the attribute "tie strength". It is worth mentioning that weak ties in some types of networks are crucial to the structure of the network; however, in coauthorship networks they are not critical to the structure of the network.

Figure 5.12 presents the cliques of the network with weak ties being removed, the number of cliques is 14.

Figure 5.13 presents the network without Wegman, it is disconnected with fewer components. There are three subgroups that are still relatively strong, these subgroups are self-sustainable; they form a separate subnetwork independent of the principal author. In the absence of the principal author, these coauthors can still get together and publish. This type of networks is called support network, which has a flat structure with few holes and

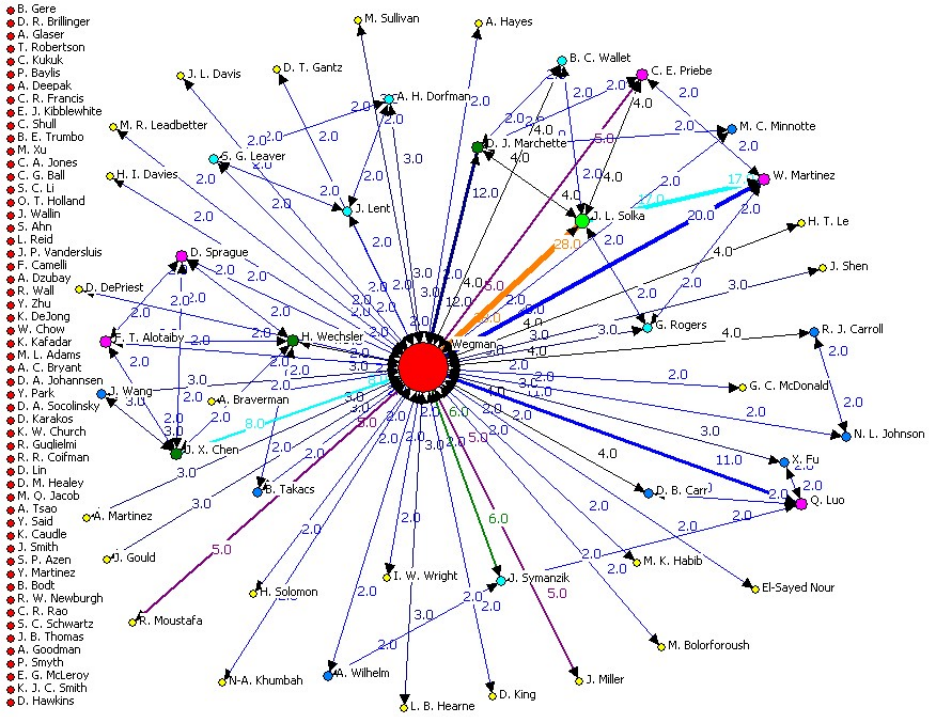


Figure 5.11: The network with coauthors having tie frequency=1 isolated.

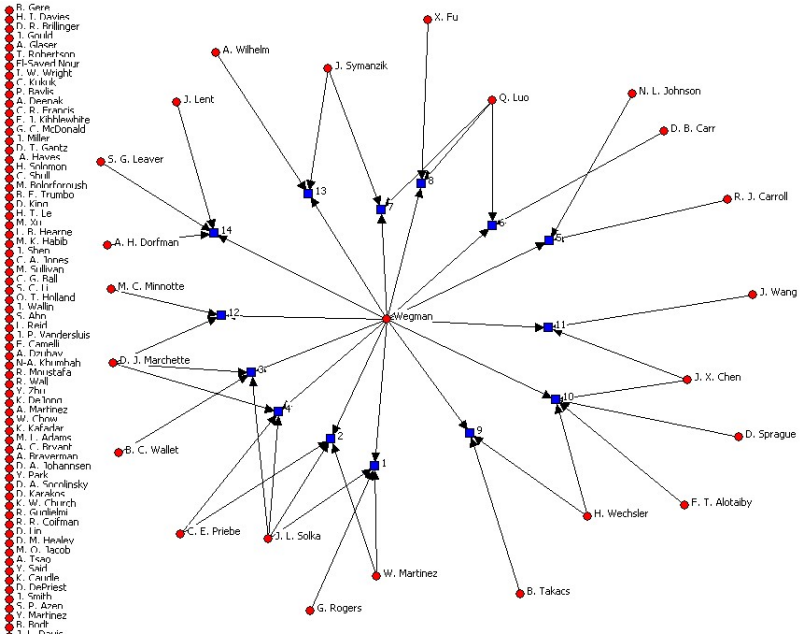


Figure 5.12: The clique set with coauthors having tie frequency=1 isolated.

high redundancy with an increased cost of coordination.

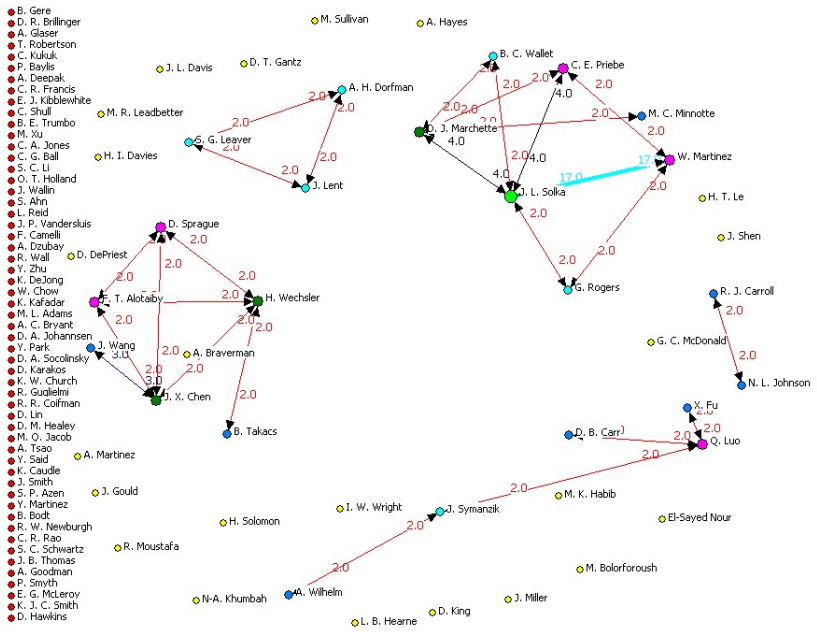


Figure 5.13: The network without Wegman.

5.1.6 Discarding Irrelevant Nodes

At this level of analysis, I remove two of the clump subnetworks (cliques). The first set of names to be deleted is {Y. Park, D. Socolinsky, D. Karakos, K. Church, R. Guglielmi, R. Coifman, D. Lin, D. Healey, M. Jacob, A. Tsao}. After discussing it with the principal author, it turned out that there was a project on Automated Serendipity, in which the main author of the paper decided to put these individuals, who contributed minimally, on the paper. The second set of names to be removed contains the names {R. Wall, Y. Zhu, J. Vandersluis, A. Dzubay, F. Camelli}; these individuals attended a course on Virtual Reality taught by principal author who decided to credit everyone on the publication even though they did not write anything in the paper. Figure 5.14 shows the network with the aforementioned set of nodes being removed.

Structural holes have low redundancy and cause stress because there are too many nodes

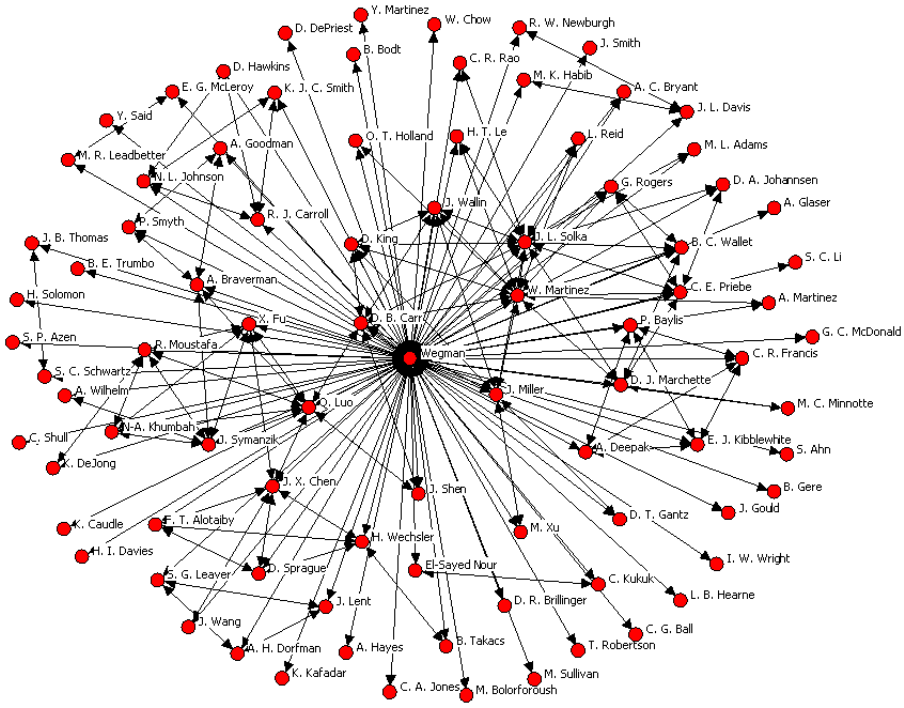


Figure 5.14: Wegman's network without the two clumps.

connected to the brokerage. The basic form of structural holes is a triad with one edge missing, in which two actors communicate with the same person, but do not communicate with each other. Wegman's first-level coauthorship network portrays this feature.

Figure 5.15 shows the adjacency matrix after removing the two sets of nodes. Figure 5.16 shows the network emphasizing vertex degree and tie strength. Figure 5.17 shows the cliques sets. Figure 5.18 shows the proximity (weighted adjacency) matrix in grey scale; the darker the color the higher the frequency of coauthorship.

[htbp]

The blockmodel for this modified network is presented in Figure 5.19, which shows four clusters of structurally equivalent coauthors. Wegman (the brokerage of the network) defines the first cluster because he coauthored every paper, coauthors (Miller, King, Carr, Solka, W. Martinez, Wallin) define the second cluster, coauthors (Luo, Symanzik, Fu, Khumb

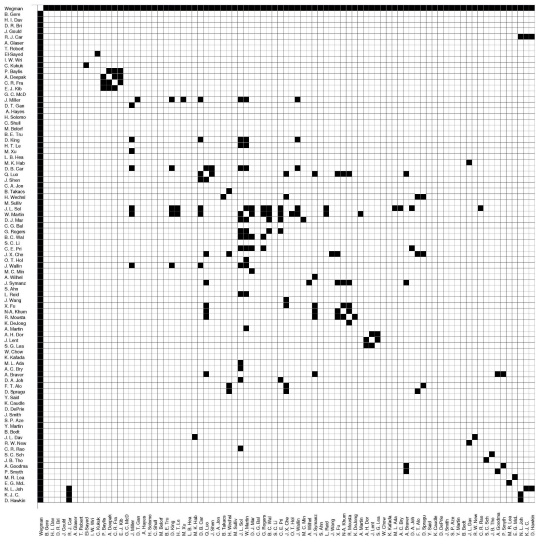


Figure 5.15: The adjacency matrix. Each black square indicates a coauthor relation.

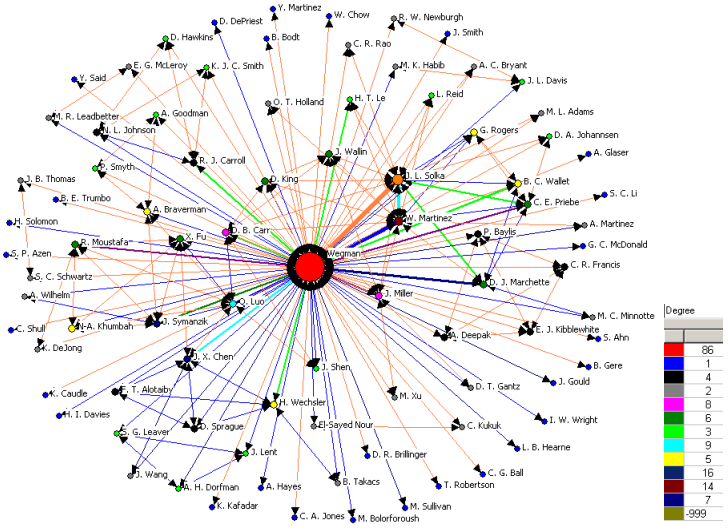


Figure 5.16: Wegman's network without the two clumps emphasizing vertex degree and tie strength.

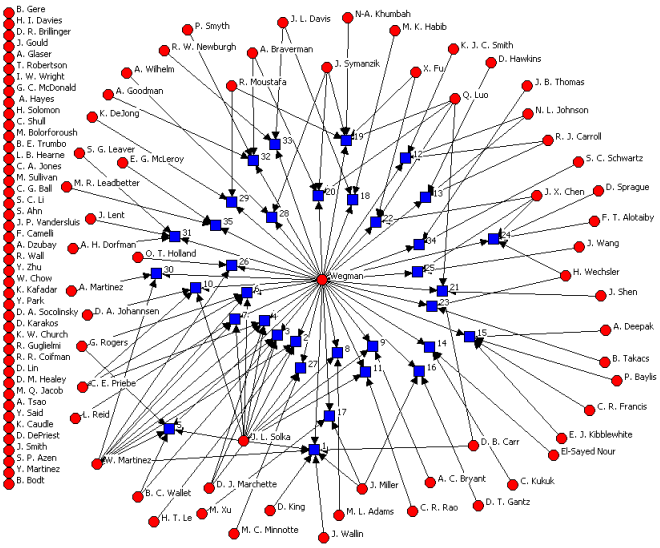


Figure 5.17: The 35 clique-sets.

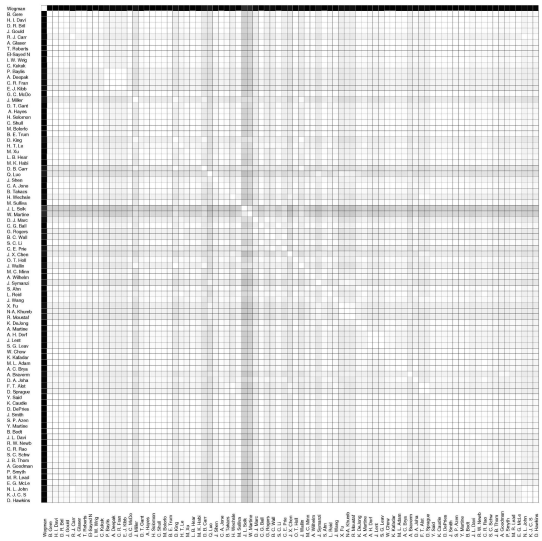


Figure 5.18: The proximity matrix in grey scale; the darker the color the higher the frequency.

Moustafa) define the third cluster and all other coauthors define the fourth cluster. Members of each cluster are structurally equivalent. The first cluster contains only one author; the principal author, this is of no surprise because the network resembles a first-level star-like coauthorship relations.

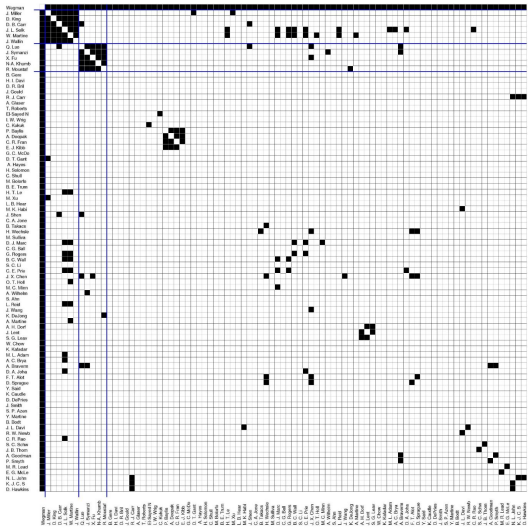


Figure 5.19: Random start blockmodel with four clusters using structural equivalence.

I finish this section by commenting on Wegman's first-degree coauthorship social network. The set of nodes (Wegman, Solka, W. Martinez, Marchette, and Priebe) are candidates for a potential "elite" group of coauthors; members of this group are high in degree, closeness and tie strength, removing these vertices results in a disconnected network and structural holes are also evident.

The analysis of the first-degree coauthorship network suggests that the principal author operates a "mentor" with most of the coauthors being younger than him. Most of these individuals worked with him to establish their future academic or industrial career and then left. The exception is the elite group, they were already established in the field and have maintained coauthorship relations.

Investigating the second-level coauthorship network by expanding on the elite group is the next part of this analysis, the goal is to get a clear picture of who-wrote-with-whom and which coauthors are critical to the network's structure. Furthermore, hidden features not captured by the first-level analysis are presumed to surface out and the network is expected to expand and fold into itself.

5.2 Second-Level Wegman's Coauthorship Network

5.2.1 Exploring the Network

The second-degree socio-network is expanded based on members of the potential elite group including Solka, W. Martinez, Marchette, Priebe, Chen, Carr, Symanzik. Figure 5.20 depicts the general structure of the second-degree network at the macro level. The network is growing with more structural holes as more coauthors are now on the periphery. There are 464 unique coauthors in total with density of 0.0121. The potential elite group is at the core of the network.

Figure 5.21 shows a bar graph of the top 15 coauthors in terms of vertex degree centrality.

Figure 5.22 shows the network in principal component layout. Figure 5.23 is a graphical representation of the adjacency proximity matrix, black squares indicate a coauthorship relation. There are several groups of highly connected coauthors (cliques) with strong ties.

I ran a measure of structural holes using the effective network size metric, a bar graph of the top 15 coauthors in terms of effective network size is shown in Figure 5.24.

Burt's effective network size measure gives the number of non-redundant ties in an *ego-network*, a measure of structural holes. Basically, how many actors is *Ego* connected to that are not connected to each other? Lack of ties among alters benefit the ego in terms of

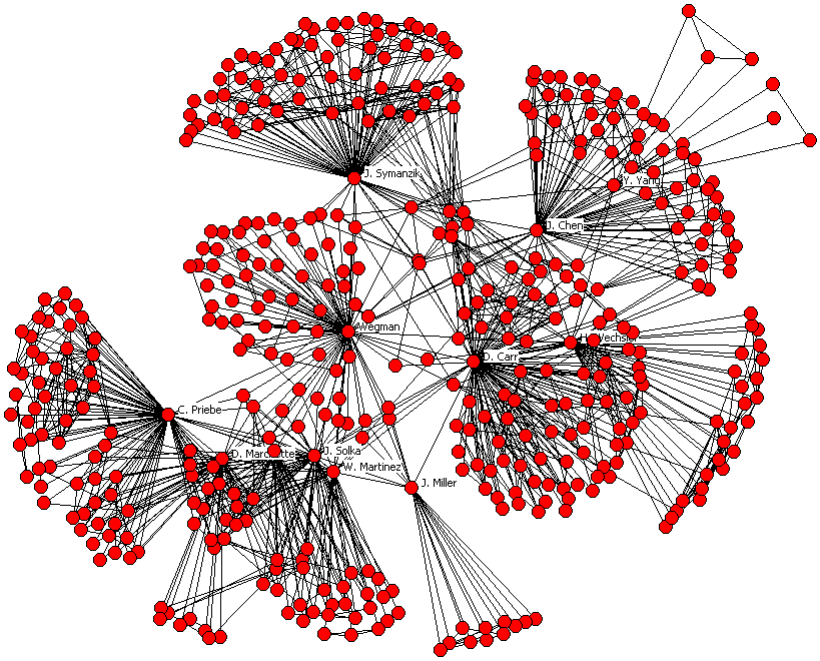


Figure 5.20: The second-level coauthorship social network.

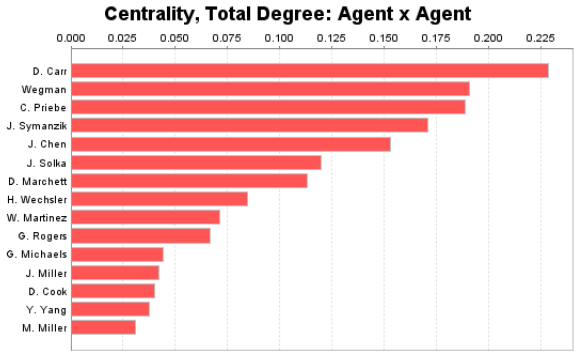


Figure 5.21: The top 15 actors in terms of vertex degree centrality.

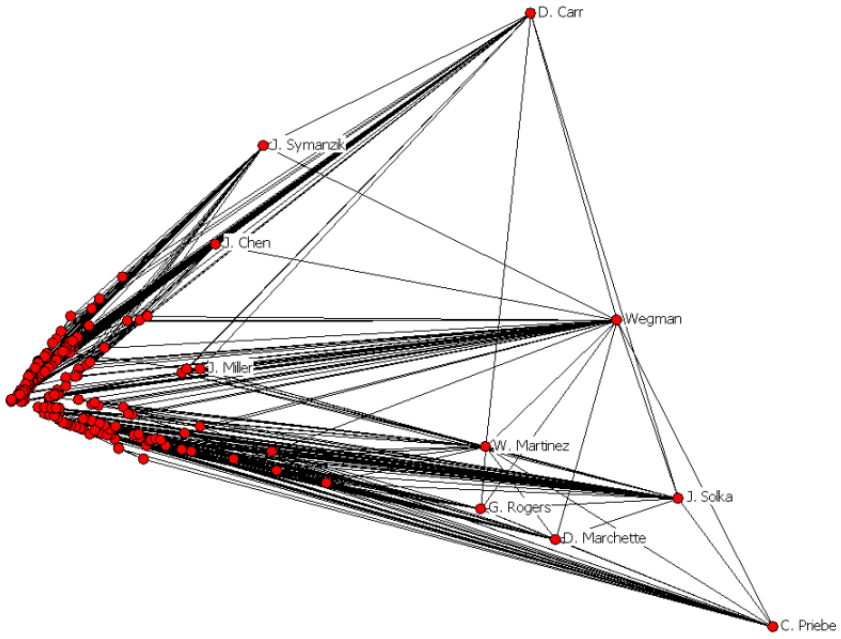


Figure 5.22: The network in principal component layout.

autonomy (independence, self-governing), control flow and information nexus. More structural holes mean power - information - freedom - low redundancy. Structural holes are fragile and require aggressive maintenance. In a structural hole situation, side nodes (the periphery) are marginalized. Structural holes are present mainly in ego-networks.

The top 15 coauthors of the second-level coauthorship network in terms of effective network size play the brokerage role (the core of the network). Carr captured the highest effective size value of 100.350. Wegman with an EffSize of 82.724 comes in second place. Priebe comes in third place with a value of 74.925. The metric indicates that these coauthors are connected to many other coauthors, but the opposite is not necessarily true. Structural holes also cause stress because there are too many nodes connected to the brokerage.

Figure 5.25 shows the 201 clique sets; few actors are members of many cliques, while most coauthors are members only of few cliques. Figure 5.26 is a bar graph of the top 15 coauthors in terms of clique count.

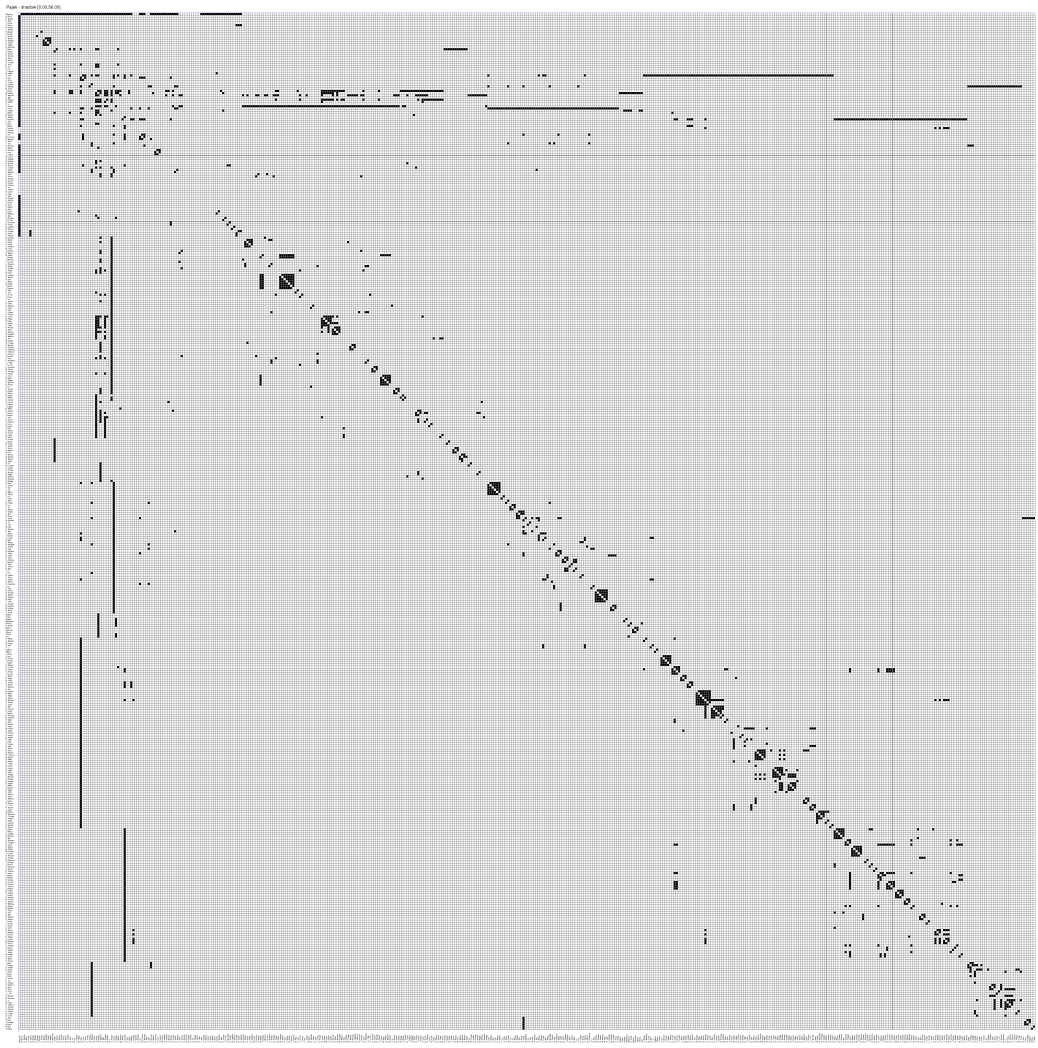


Figure 5.23: Structure matrix of Wegman's second-degree network.

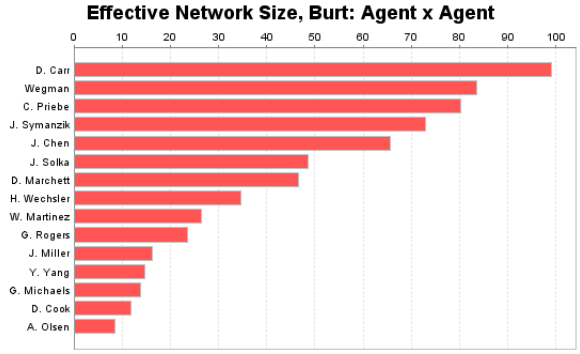


Figure 5.24: The top 15 actors in terms of effective network size.

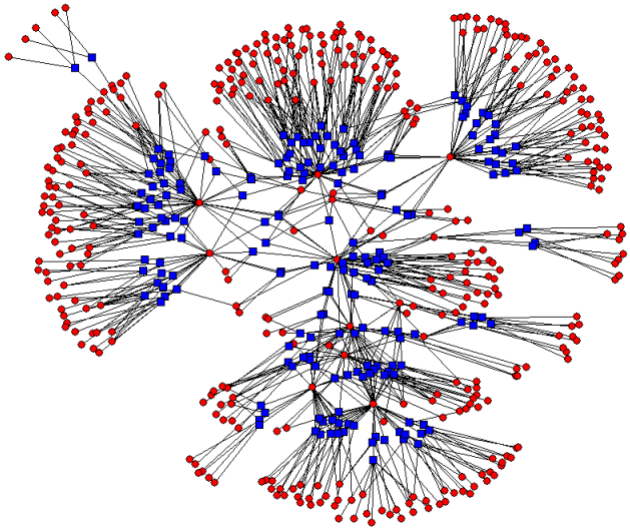


Figure 5.25: The 201 clique sets. A square indicates a clique set, while a circle indicates a coauthor.

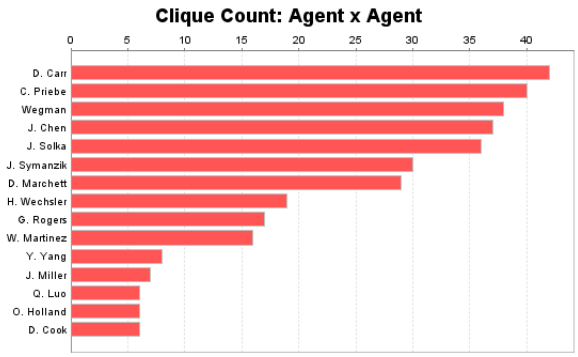


Figure 5.26: The top 15 actors in terms of clique counts.

Finally, I ran a metric of structural equivalence on the network. Because `Pajek` can't handle more than 256 nodes for blockmodeling, weak ties (ties with frequency=1) were removed. Figure 5.27 presents 3 clusters of blockmodeling based on structural equivalence, it suggests a distinct structural roles within the core; members of each cluster are structurally equivalent, i.e. these actors have the same ties to all other actors they are perfectly substitutable or exchangeable. In this example, structural equivalence generated three positions in the network.

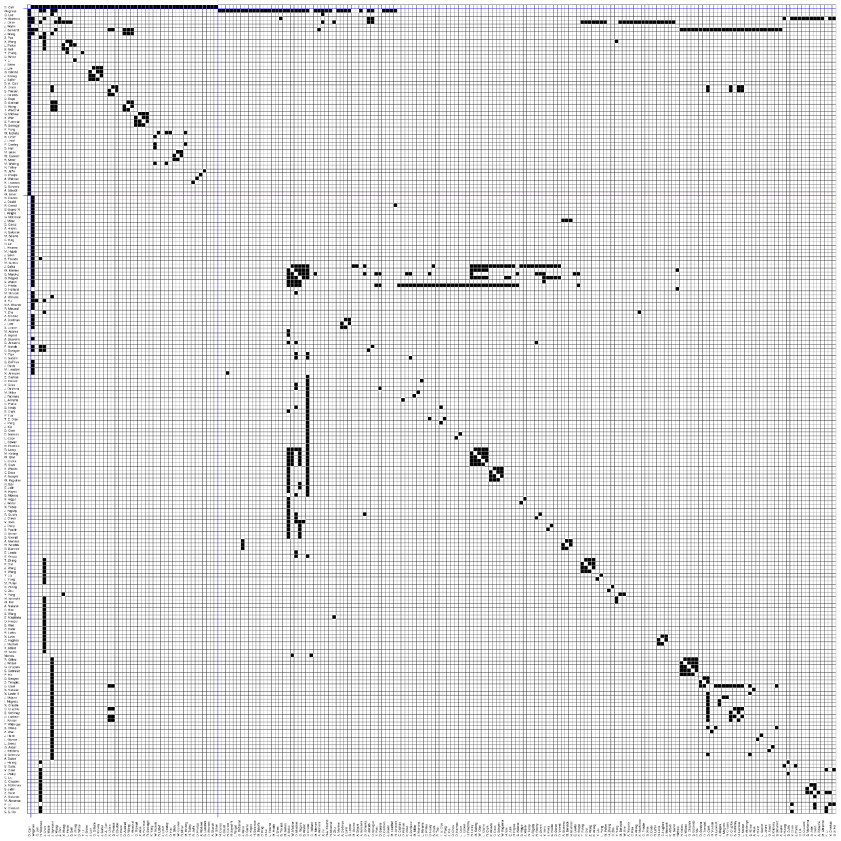


Figure 5.27: A metric of structural equivalence with 3 clusters.

5.2.2 Multi-Dimensional Scaling Clustering

The purpose of equivalence analysis is to identify "classes" or clusters based on similarity. I implicitly assume that distances among actors reflect as a two dimensional; although, it is possible that the data are multi-dimensional. MDS is used (metric for data that are inherently valued) to cluster actors based on distance.

MDS represents the patterns of similarity or dissimilarity in the tie profiles among actors (when applied to adjacency or distances) as a "map" in multi-dimensional space. The map lets us see how "close" actors are, whether they "cluster" in multi-dimensional space,

and how much variation there is along each dimension. The goal of MDS is to minimize stress – distance between nodes. "Stress" is a measure of badness of fit; $0 \leq$ stress ≤ 1. The range of solutions with more dimensions is sought, so that the analyst can assess the extent to which the distances are uni-dimensional. The meaning of the dimensions can sometimes be assessed by comparing agents that are at the extreme poles of each dimension.

Now, I attempt to cluster actors of the second-level coauthorship network using the MDS. Figure 5.28 shows the result of applying MDS-metric clustering on CONCOR (CONverging CORrelations) 1^{st} correlation data to the adjacency matrix of the coauthorship network, and selecting a two-dimensional solution. Nodes are plotted according to their coordinates. Close tight clusters of points identify actors that are highly similar on both dimensions. It appears, though, that some clusters are emerging, the closer the nodes the more similar they are. Coauthors residing on opposite poles (distant actors) are dissimilar. Notice, the transition among various disciplines from left to right. For example, clusters in the left-most side contain mainly coauthors from the computer science field, whereas clusters in the rightmost side are mainly coauthors from the statistics field.

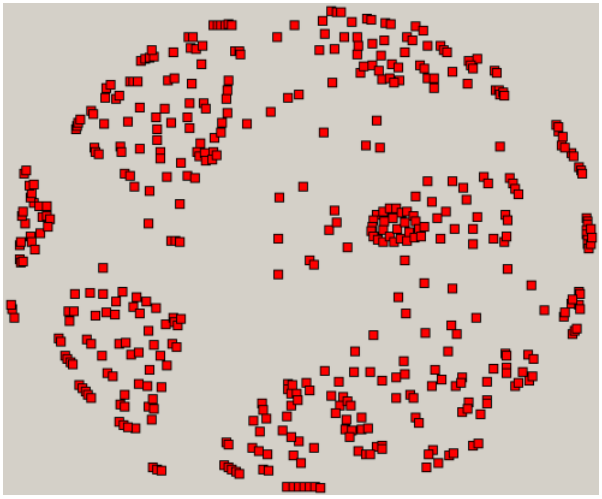


Figure 5.28: A 2-D MDS-metric clustering.

Stress for a two-dimensional solution is 0.371 and for a three-dimensional solution is 0.271.

5.2.3 Investigating the Elite Group

Coauthors who maintain strong ties are considered important and thus are member of the elite group; a coauthor is in an elite group(s) if the frequency of coauthorship exceeds a certain threshold. The process of determining elite groups is systematic. Figure 5.29 shows the second-level network excluding ties with frequency ≤ 2. Structural holes are present, the network is more centric with Wegman being at the core of the network.

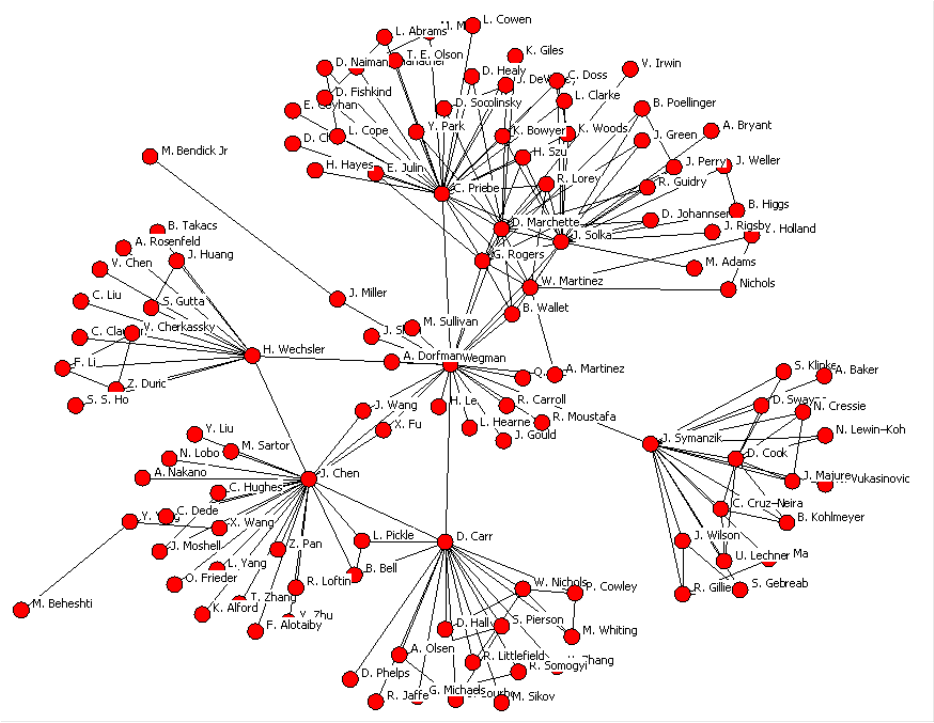


Figure 5.29: Edges with tie strength ≤ 2 are removed.

Figure 5.30 shows the network excluding ties with frequency ≤ 6, i.e. all the ties with strength $= 7$ or more. The network is disconnected with 3 immanent components. Again, notice the transition in positions; for example, actors on the rightmost side are statisticians while actors on the leftmost side are computer scientists.

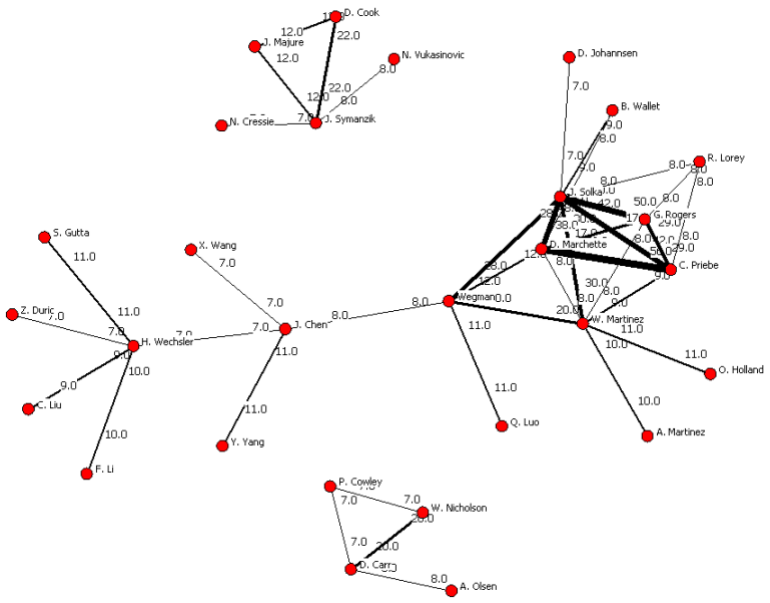


Figure 5.30: Edges with tie strength ≤ 6 are removed.

Figure 5.31 shows the coauthors with tie strength = 10 or more. Coauthors who published that many times with other coauthor(s) must be special and hold strong relations. Cliques are also present; for example, the set {Rogers, Marchette, Priebe, Solka } forms the strongest clique in the entire network. In addition, the sets {Cook, Symanzik, Majure}, {Wegman, W. Martinez, Solka}, {Wegman, Marchette, Solka} form another strong clique sets in the network. Each actor in this representation favors one or two coauthors in which he/she writes with the most. Style of coauthorship is discussed more in detail in the next section.

I finish the analysis of the second-level coauthorship network with some comments. First of all, this network has some interesting features on the macro and micro levels. On the macro level, the emergence of an "elite" group of coauthors high in vertex degree, tie strength, closeness, clique sets, and effective network size such as Wegman, Carr, Priebe, Solka, W. Martinez, Rogers, Symanzik, J. Chen, Wechsler. It seems like scholars favor few fellows to publish with more than anyone else in the entire network. Furthermore, MDS produced clusters of similar actors. Finally, the network revealed another exciting property

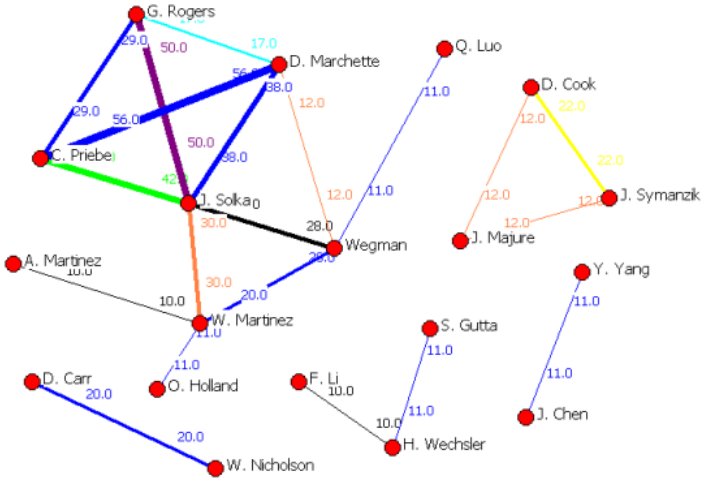


Figure 5.31: Edges with tie strength ≤ 9 are removed.

encompassing the transition in positions among different disciplines and coauthors clustered according to their relative scientific fields.

5.3 A Model of Preferential Attachment for Emerging Scientific Subfields

In this section, I focus on demonstrating scale-free author-coauthor social networks. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices (growth), and (ii) new vertices attach preferentially to sites that are already well connected (preferential attachment). A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems. Growth means that the number of vertices (actors) increases with time. Preferential attachment means that the more connected a vertex is, the more likely it is to acquire new edges. Intuitively, preferential attachment can be understood if we think in terms of social networks connecting people. Here an edge from actor A to actor B means that actor A

"knows" or "is acquainted with" actor B. Vertices with many edges represent well-known people with lots of relations. When a new actor enters the community, he or she is more likely to become acquainted with one of those more visible actors rather than with a relative unknown. Models that satisfy these two principles are known as Barabsi-Albert models [2]. In this section, I seek to demonstrate that author-coauthor networks in the statistical literature satisfy these two criteria. I first present the model that is based on the theory presented in Chapter 4 and then compare a randomly generated network with a real network. The process of one actor attaching to another actor (author) and strengthening the tie over time is a stochastic random process based on the distributions of tie-strength and clique size among actors, which are obtained from empirical data. I then utilize the model to predict emerging scientific subfields of the evolutionary coauthorship network. Followed by a discussion on style of coauthorship among prominent scholars that is using the distribution of tie-strength.

There has been work on author-coauthor networks and the emergence of global brain in [6], preferential attachment in [51] and implications for peer review in [55]. Coauthorship relationships can be treated as a 2-mode networks in which there are two types of nodes; the authors nodes and the papers nodes, and one relationship type; "person A authored/coauthored paper P".

Data on statisticians and statistics subfields were collected from the online Current Index to Statistics (CIS) database [12]. The CIS database is jointly published by the American Statistical Association (ASA) and the Institute of Mathematical Statistics (IMS). There are many analogous databases, for example, in Computer Science [13] and in Medicine [48]. I focused on the CIS database in this dissertation. The procedure used to harvest data involved two stages using names of well-established statisticians affiliated with prominent US universities. These data were used to build a social network of coauthors and to derive the distribution of tie-strength "frequency of coauthorship" among coauthors. A different dataset was used to derive the distribution of clique size. In the second stage, I used the

biopharmaceutical as keywords to query the database, the dataset was used to discover the emergence of scientific subfields by exploring the evolution of the coauthorship socio-networks over time as a time series.

5.3.1 Distribution of Tie Strength

In weighted coauthorship socio-networks, strength of a tie indicates the frequency of coauthored papers between two actors; in other words, it is a measure of how close two actors are and how much they trust each other. Therefore, studying tie-strength is a subject of interest in coauthorship social networks. We developed a MATLAB program to build the 1-mode proximity matrix of the data collected from the CIS database on contributing scientists in the field of statistics. This adjacency weighted matrix was later manipulated to construct the distribution of tie-strength. The statisticians dataset contained 1767 published papers that had 874 unique author(s)/coauthor(s), the 1-mode network of coauthors is shown in Figure 5.32, while the adjacency proximity matrix is shown in Figure 5.33, a black square indicates a coauthor relation..

Figure 5.34 shows the adjacency matrix of biopharmaceutical statisticians, a black square indicates a coauthor relations. The figure suggests that different independent isolated clusters (groups) of scholars working in the field. Note that the structure of this network is somewhat different from the statisticians network. Clusters and blocks are present along the main diagonal, cliques are more evident in the biopharmaceutical socio-network. There are no vertical/horizontal bars present as opposed to the statisticians socio-network because the subfield was the keyword used to query the database rather than names of authors. The distribution of tie-strength is shown in Figure 5.35(a)

Figure 5.35(a) suggests a power law distribution [11]. To investigate this, I first plotted the distribution in log-log scale, this is shown in Figure 5.35(b). Because the density curve is close to linear in log-log space, it is reasonable to conjecture that the distribution is power

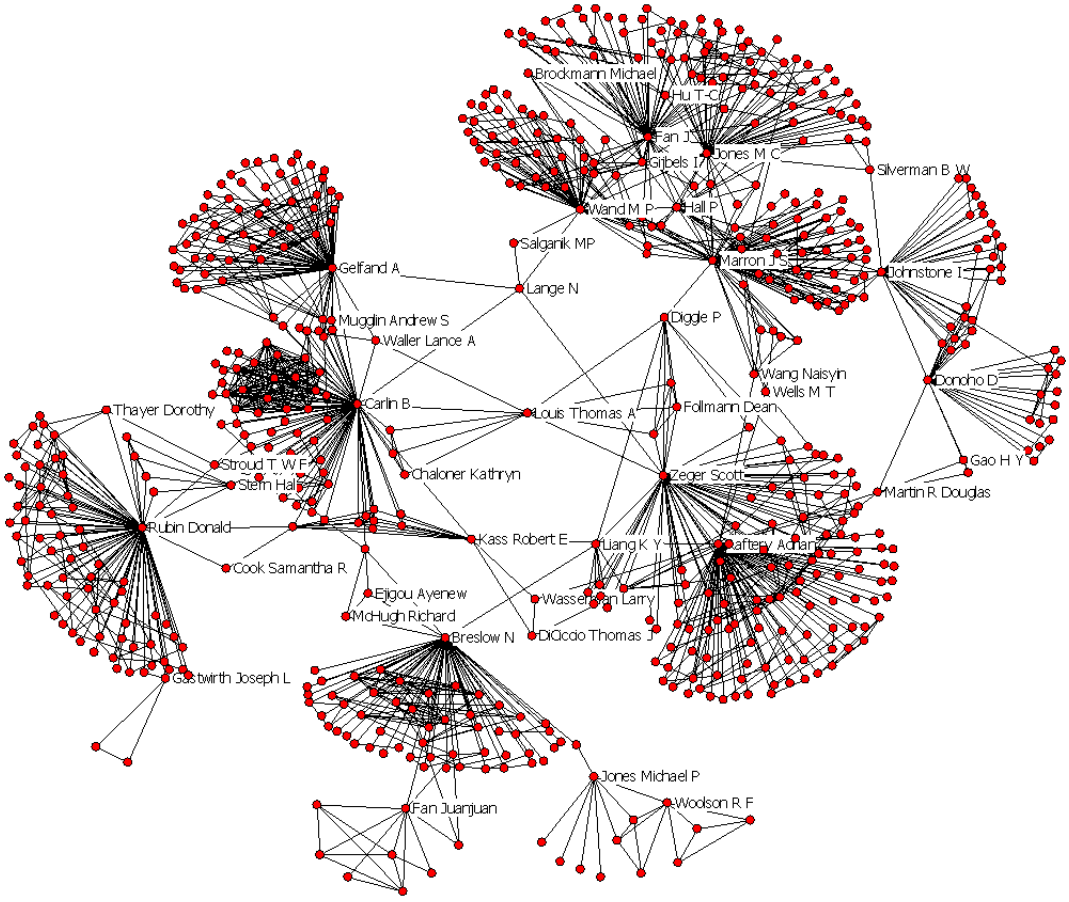


Figure 5.32: Coauthorship social network of prominent statisticians.

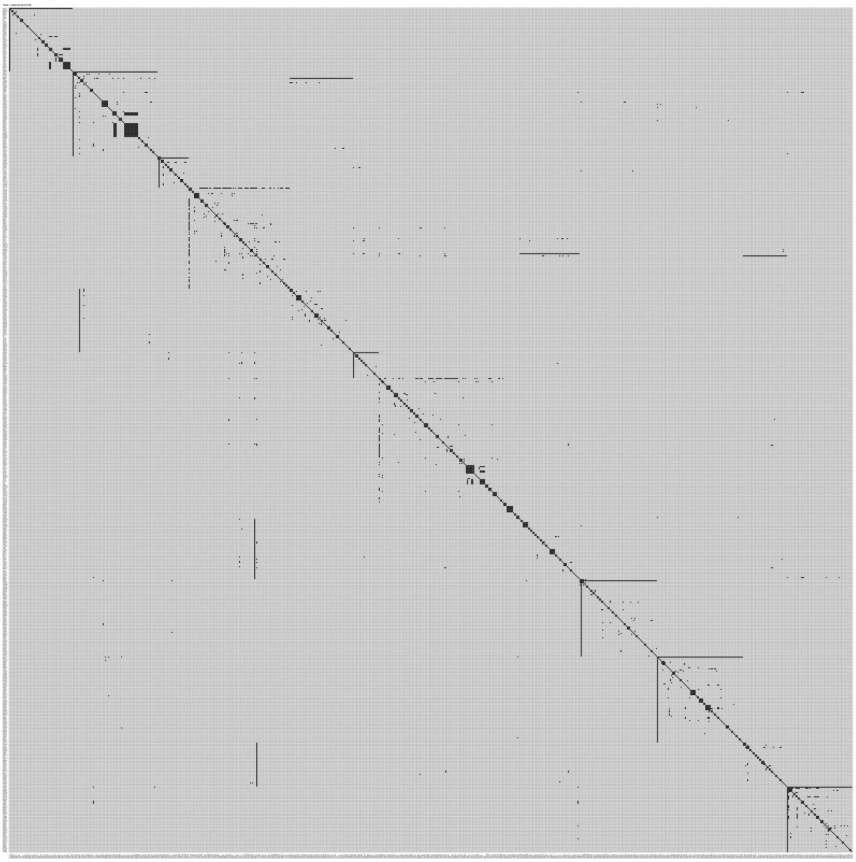


Figure 5.33: The adjacency proximity matrix of well-established statisticians.

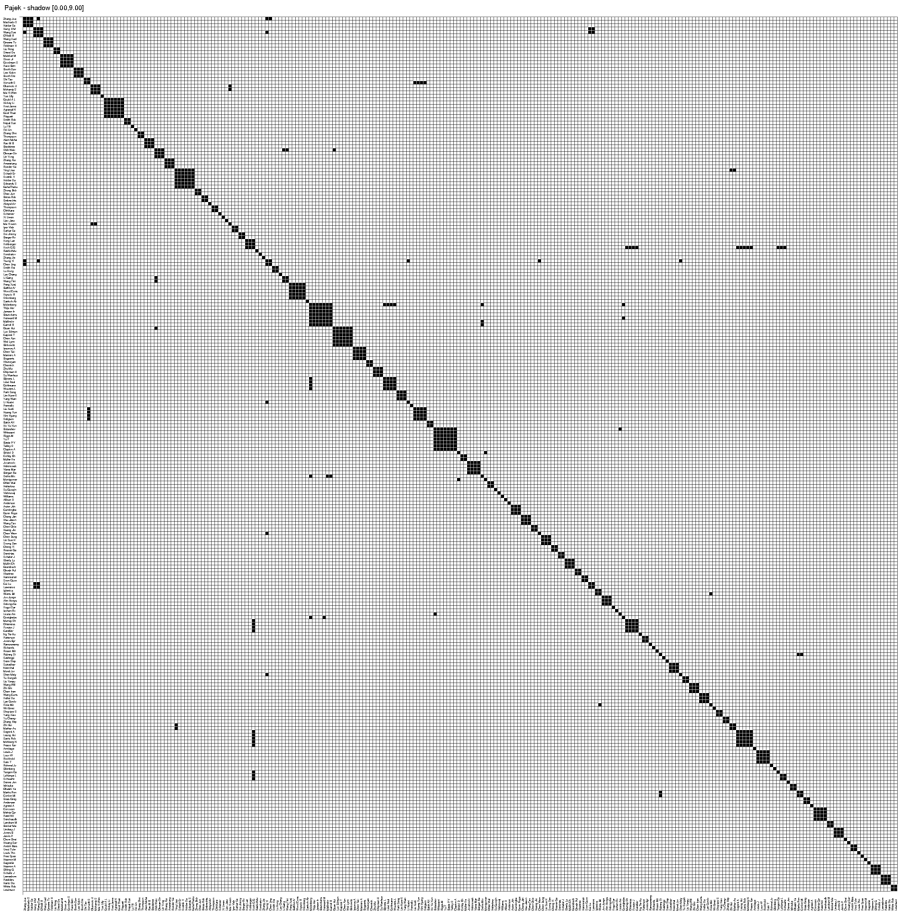


Figure 5.34: The adjacency matrix of Biopharmaceutical statisticians.

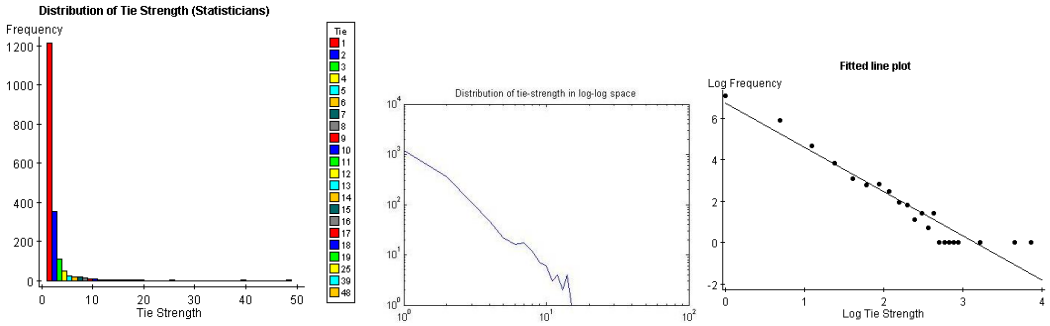
law. The next step would be computing the exponent α of the power law. This can be done by either finding the slope of the least-squares regression line in log-log space or by using the following aggregation method for calculating the exponent α.

$$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1},$$

where x_i is the observed tie-strength,

x_{min} is the minimum observed tie-strength; one in our problem,

n is the size of the vector.



(a) Distribution of tie-strength. (b) Distribution of tie-strength in log-log space. (c) Linear regression on tie-strength in log-log space.

Figure 5.35: Examining the attribute tie-strength.

An implementation of the aggregation method in MATLAB produced an α-value of 2.1716, the least squares regression model confirmed this result with an α-value of 2.13 and $r^2 = 0.915$. Therefore, I can observe that the distribution of tie-strength is power law with exponent value of approximately 2.1716.

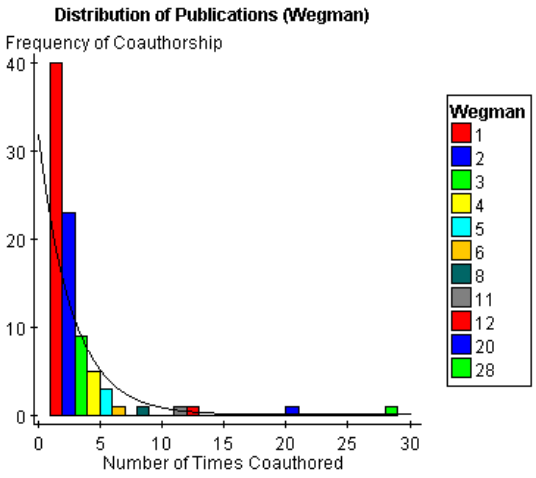
Looking into the low-level processes that produced the many-some-few power law pattern, I conjecture that this behavior is generated in view of the following reasons. Firstly, there

are higher chances to find two coauthors who simply published together few times or perhaps once. Many of these statistician are professors who may have a number of graduates working on projects or papers at a given time period. Upon graduation, many of these students prefer a career in the industry, therefore, they lose contact with their professors leaving behind one or two published papers with that professor. On the other hand, some scientists find themselves in the research area, as a result, the likelihood that two already coauthored individuals publish again rises. If you coauthored a good quality paper with someone and you liked him/her, chances are you are going to publish with him/her again if there is mutual agreement increase. And finally, there are those authors who favor only very few coauthors; a colleague or a fellow student who maintains good contacts and relations with that author, to publish with the most.

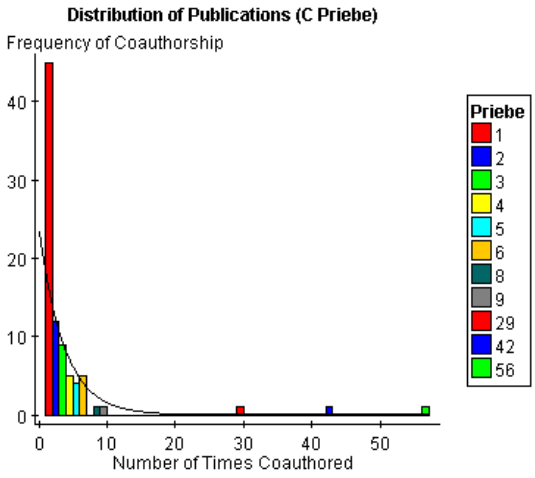
I further investigated the distribution of tie-strength of individual authors. Figure 5.36 shows the distribution of tie-strength of four different authors. Surprisingly, the distribution is again power-law with exponent α ranging $1.5 - 1.85$. Because $\alpha < 2$ both the mean and the variance of the distribution of the power-law are not defined and hence the power-law is said to be not well-behaved. In order for the mean and variance of a power-law to be well-behaved α has to be greater than 3, if $2 < \alpha < 3$ only the mean is finite. The distribution of tie-strength is a self-similar power-law distribution for coauthorship social networks.

5.3.2 Distribution of Clique Size

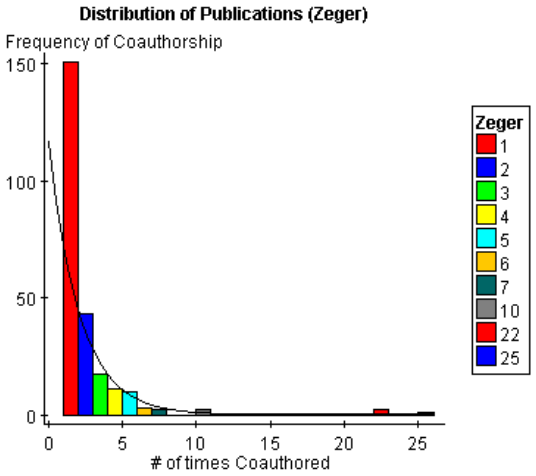
An important factor in preferential attachment is the clique size; the number of people coauthored a single paper. Note that a paper with sole author or two coauthors is technically not considered a clique. A clique in a graph must have at least three fully connected nodes "complete graph" [61]. I used the dataset of prominent statisticians to construct the distribution of clique size to obtain a better understanding of how coauthors interact.



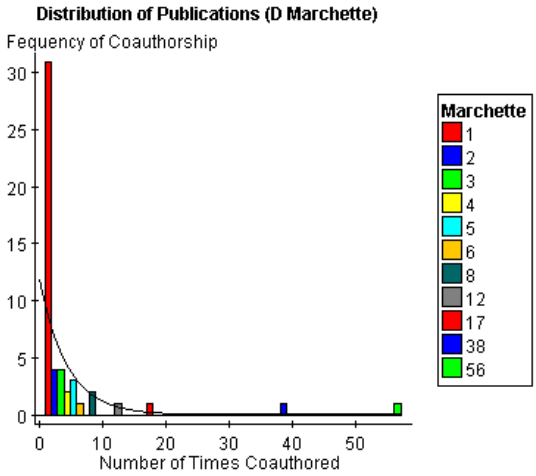
(a) E. Wegman



(b) C. Priebe



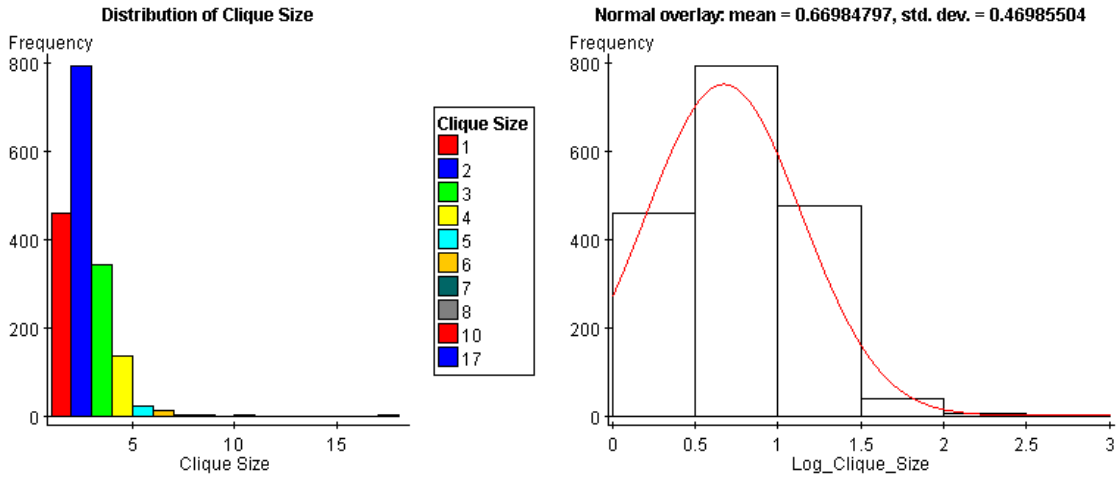
(c) S. Zeger



(d) D. Marchette

Figure 5.36: Distribution of Tie-Strength Among Authors.

Figure 5.37 shows the distribution of clique size. The distribution of clique size is approximately lognormal with mean $\mu = 1.954$ and standard deviation $\sigma = 1.6$.



(a) Clique size.

(b) Clique size in loglog scale.

Figure 5.37: Distribution of clique size.

5.3.3 The Emergence of Scientific Subfields

The biopharmaceutical subfield joins the fields biology and pharmacy. In this part, I explore the biopharmaceutical statisticians socio-network over time to inspect the emergence of this discipline. The data include papers published between the years 1977 and 2003. There are 157 published papers with 260 unique coauthor(s). Figure 5.38 shows the evolution of the network over time. In 2000, very few statisticians started writing about biopharmaceutical statistics, the graph in figure 5.38(a) shows isolated authors with two cliques of size three and dyadic relations. In figure 5.38(b), we start seeing more cliques, more groups are publishing in the biopharmaceutical subfield. In figure 5.38(c), the network is growing tremendously with more individuals publishing, it seems like H. James and W. Jane are leading coauthors in the new field. Finally, in 2003, the subfield is well-established with several independent and mutually exclusive groups working simultaneously, the leading figures are still H. James and W. Jane.

Two main factors controlled the evolution of this new field. Firstly, small groups and isolated scientists started researching the field, and then over time more scholars and larger groups are becoming more involved and interested in the subject. The second factor, resides with the fact that certain coauthors became the key figures in the field, this is evident from the high number of publications they coauthored in the subfield, see figure 5.39.

5.3.4 Random Graph Model

The model is based on stochastic "random" processes, in which vertices are generated randomly at each time step. At each time step, a new paper gets published and one of three things could happen.

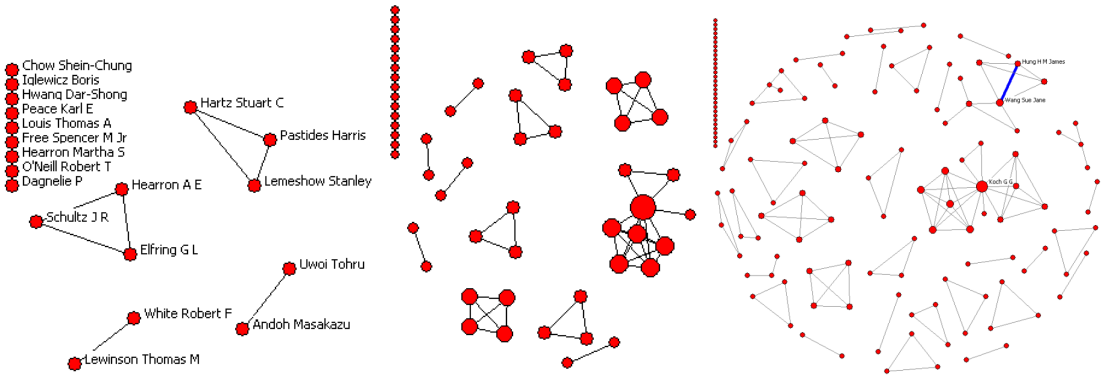
1. New actor(s) try to attach to existing actor(s).

2. Already existing non-attached actor(s) attempt to make an attachment(s).

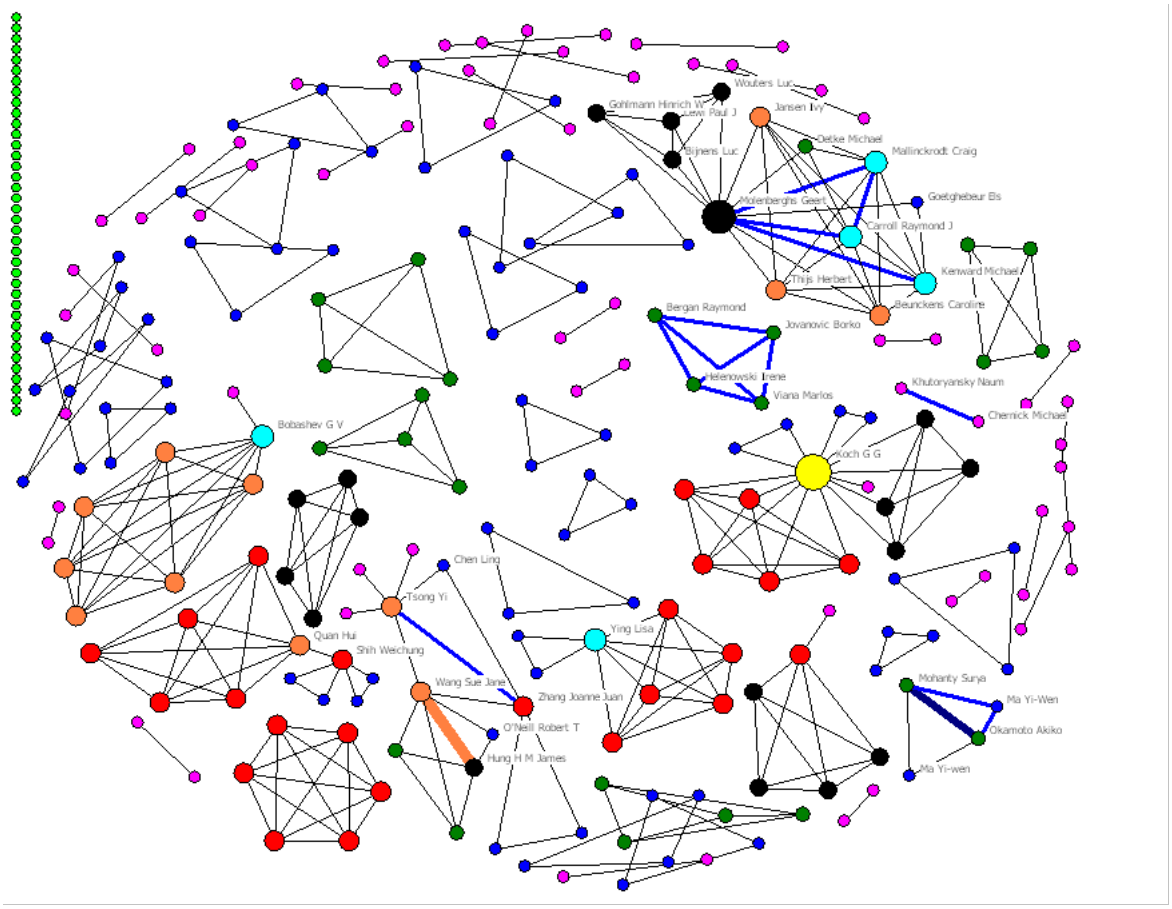
3. Already attached actor(s) strengthen their ties.

And each vertex has the attributes: name – age – weight – preference – status – field – active flag. These attributes uniquely identify actors, some of which change rapidly/slowly over time while other attributes remain the same over time. For example, the attributes "name" and "field" do not change. The evolution of "weight" and "status" attributes can be viewed as a time series because they change faster than any other attributes. "Age" changes linearly over time. Meanwhile, the "active" flag operates as a switch initially set to "on", but later could change to "off", once it is changed to "off" it remains in that state forever. Certain actors might change the attribute "preference".

The model was implemented in MatLab and consists of approximately 350 lines of code, it exploits the distributions of tie-strength and clique-size to build the coauthorship network. Figure 5.40(a) is a 2-mode author-by-paper simulated network. Note that a new publication surfaces at each time step. Figure 5.40(b) shows the 1-mode coauthorship network corresponding to the matrix in Figure 5.40(a). Figure 5.41 shows author's 1 attributes



(a) The network in 2000. (b) The network in 2001. (c) The network in 2002.



(d) The network in 2003.

Figure 5.38: The Evolution of the Biopharmaceutical Statistical Coauthorship Social Network.

Figure 5.39: 1-mode biopharmaceutical coauthorship social network.

after 10 iterations. Figure 5.42 shows another simulated coauthorship social network, the program ran for 100 iterations. The simulated network is similar to the network obtained from empirical data, see section 5.3.3.

	101	102	103	104	105	106	107	108	109	110
0	1	0	0	1	0	1	0	0	0	0
1	1	0	0	1	0	1	0	0	0	0
2	0	1	0	0	1	1	0	0	0	0
3	0	0	1	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0	1	0	0
5	0	0	0	0	0	0	1	0	0	0
6	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	0	1	0

(a) 2-mode author-by-paper socio-network.

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

(b) 1-mode author-by-author socio-network.

Figure 5.40: A simulated coauthorship social network.

```
      name: 1
       age: 23.8330
     field: [1 0 0 1 0 0]
    weight: 5
    status: 0.0053
preference: 2
    active: 1
```

Figure 5.41: Node attributes.

5.3.5 The Network of Well-Established Scholars

Figures 5.32, 5.33 present the social network of prominent statisticians affiliated with US universities. I will use the method of deleting weak ties and and pendants vertices (vertices with degree = 1) to expose the important coauthors in the network. In coauthor social networks, weak edges and hanging vertices do not impose great impact on the status of the network, however, in other types of networks weak ties could be crucial to the status and performance of the network. What is worth knowing in social networks is who maintains strong ties with who and who is connected to the most actors, such authors resemble the heart of the network and their strong ties is the blood that keeps it alive and active.

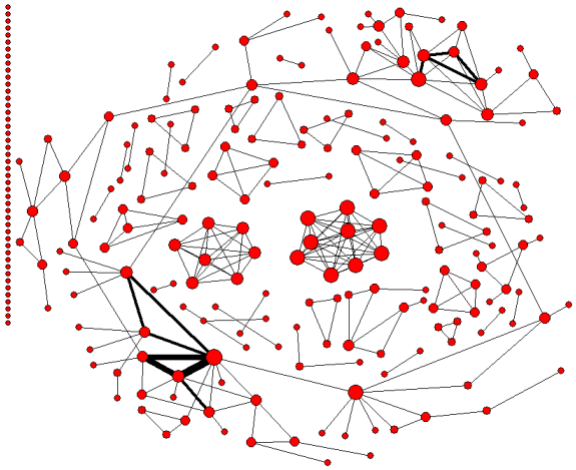


Figure 5.42: A simulated coauthorship network.

Brokerage roles are evident in this network. For example, the vertex "Lange N" in Figure 5.32 can be in the cut-point set, this author is connected to four key player scholars, namely, "Gelfand A", "Carlin B", "Wand M" and "Zeger S"). While maintaining good relations with prominent authors in the field of statistics, this author also connects structurally different parts of the network and styles of coauthorship. In addition, "Louis T" can also be considered in the cut-point author set, he is in contact with two mutually exclusive subgroups of authors in which none of the members of each subgroup publishes with member(s) of the other subgroup. "Hall P", "Diggle P" and "Gijbels I" are not cut-point authors but yet connected to key figures in the network, they are publishing with authors most of which are affiliated with different universities and geographically located in different continents. Further investigation reveals that some of these authors although they are not geographically in the same place, but they went to the same school, majored in the same field and spoke the same language and thus maintained good relations.

Continuing with the same spirit, I proceed by removing ties with weight = 1, Figure 5.43 depicts the altered network. Thick edges indicate higher weight value, the thicker the link is the higher the number of publications. Big nodes indicate higher degree, the bigger the vertex is the higher the number of coauthors that particular author has. The network is

not centric, in fact, it is more like a chain-network with network diameter = 12. It contains three separate components. In this layout, "Donoho" and "Gelfand" are far away from each other. However, "Zeger" and "Breslow" form two independent subnetworks. Finally, the authors "Marron", "Hall", "Fan", "Gijbels", "Wand" and "Jones" are very close and similar authors, they form inbred subnetwork.

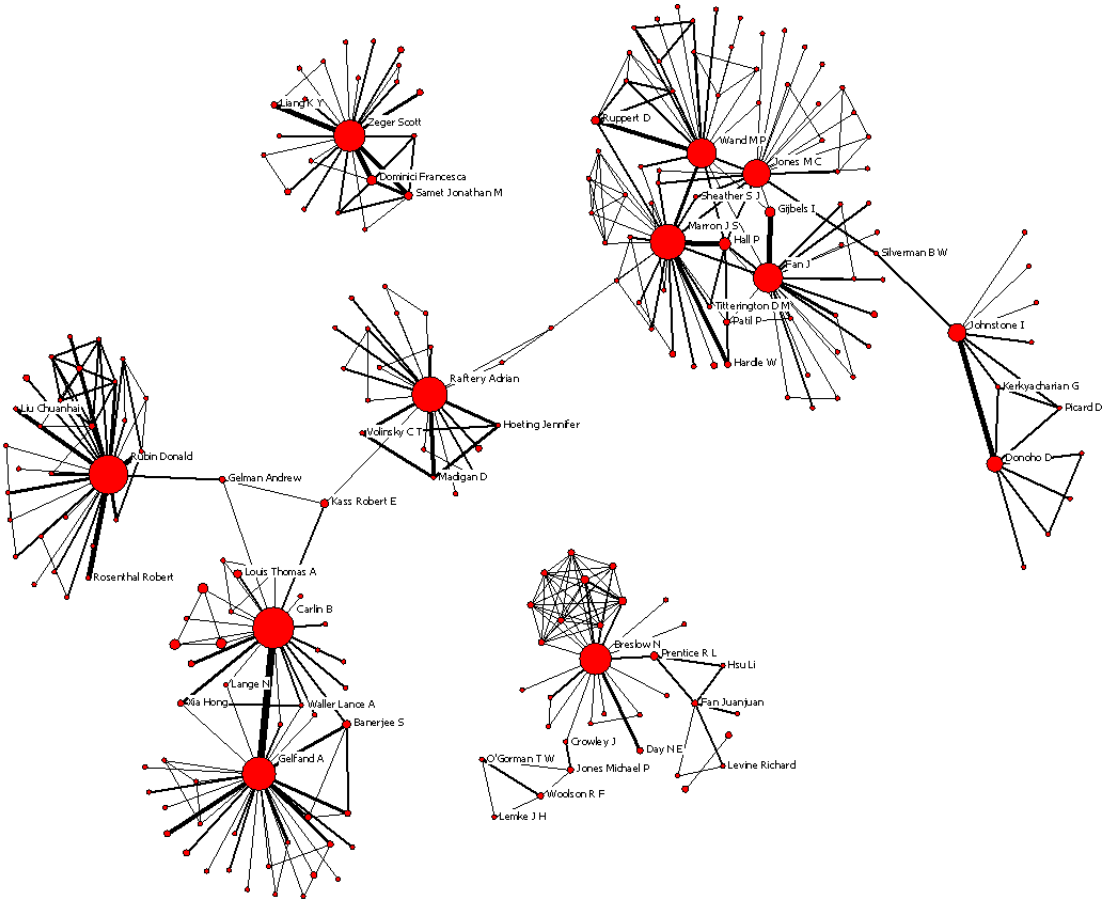


Figure 5.43: Social network of statisticians without nodes with degree and frequency = 1.

Figure 5.44 shows the network of authors with tie strength of seven or higher. Clearly, there are components of the original network consist of authors with high coauthored papers. Members of each component form an elite group of well-trusted authors and coauthors.

To sum up, this section addressed two issues, the first concerned empirical data to in-

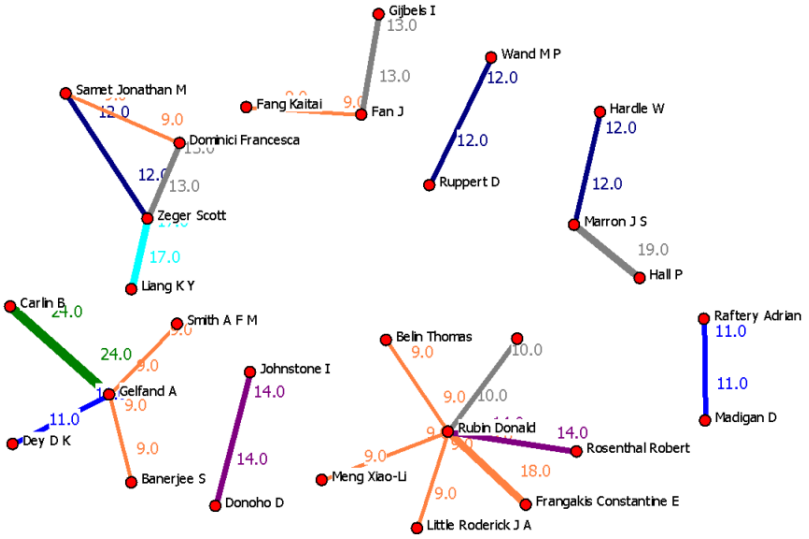


Figure 5.44: Social network of statisticians showing authors having tie strength 7 or higher.

vestigate the distributions of tie-strength and clique-size in coauthorship social networks. The distribution of tie-strength among authors is a well-behaved power law; however, the distribution of clique size is lognormal. While the second concerned the development of a program to generate random coauthorship network, the model takes into account the fact that authors' status and attributes change over time. The resulting artificial network looked similar to a real coauthorship social network of statisticians in the Biopharmaceutical subfield.

I close by showing the distribution of authors and papers over time for the prominent statisticians. Figure 5.45 suggests that there is a period in the early and late nineties captured the highest number of publications. This indicates that during that time slot the prominent statisticians were working with higher number of students and colleagues compared with other years, which is supported by the number of new unique authors during that period, see Figure 5.46. Surprisingly enough, frequency of coauthorship since 2005 is comparable with the mid eighties time period as far as number of papers, but higher with respect to new unique authors in recent years. It means that scholars were producing more literature back in the eighties. It appears, though, as if the interest in research is declining

since the year 2000.

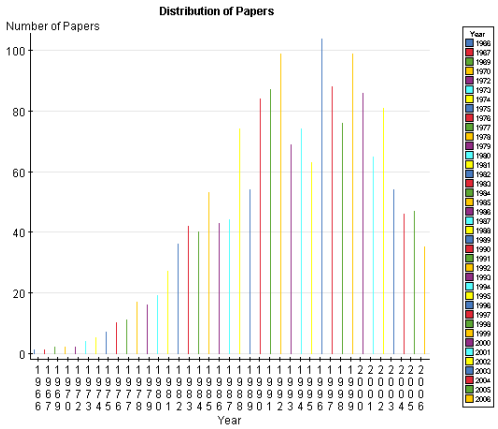


Figure 5.45: Distribution of papers.

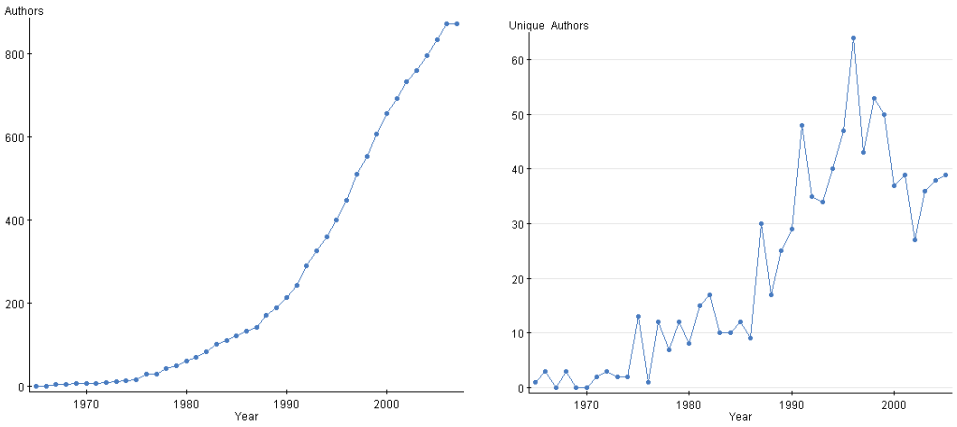


Figure 5.46: Distribution of authors.

5.4 Road Fatal Crashes In The United States

The road crashes network is an interesting example to study because accidents outcome may be devastating. It is the thought that many fatal crashes are due to over speed, high blood alcohol content level (BAC), high drug dosage, and/or bad weather conditions. In this section, I analyze road fatalities across the United States in 2006 both statistically

and from social networks perspectives. There are several significant factors and variables to examine among which are State, Age, BAC level, Travel Speed, Gender, Road Function Class (rural vs urban), Time of Day, Day of Week, Month of Year, Registered Vehicle, and Driver Related Factors.

It is the hope that the results presented in this example will provide an insight to policy makers and law enforcement authorities to take appropriate actions that may reduce road fatal crashes in states with high fatality rates.

Data on road fatalities were collected from the online Fatality Analysis Reporting System's (FARS) website [21], an affiliation of the National Center for Statistics and Analysis (NHTSA) on traffic safety facts. I queried the database on the variables listed above and obtained both univariate tabulation and cross tab datasets on the aforementioned variables.

Figure 5.47 shows the scatterplot matrix of seven factors related to road fatalities. There is association between age and time of day. For example, most crashes occurred late at night involve young adults. Older people are involved in the afternoon crashes; it may be because older people get off work late in the afternoon, which increases the odds of involving in a crash. Furthermore, outliers are present in the speed versus BAC level scatterplot, they correspond to speed in the range 40-60 mph and BAC level over 0.8%. Lastly, there is an outlier present in the BAC level and Day of Week scatterplot, it corresponds to Friday and BAC level over 0.9%. One reason related to this extreme value is the fact that people party and drink more on weekends as opposed to week days.

The cross tab of "state" and "alcohol" can be viewed as a two-mode network. Consider the bipartite "state-by-alcohol" network ${}^{w}SA$, then the one-mode network of states related through the variable "alcohol" is

$$ {}^{w}SA \cdot {}^{b}SA^{T} = {}^{w}SA \cdot {}^{b}AS = {}^{w}SS \, . $$

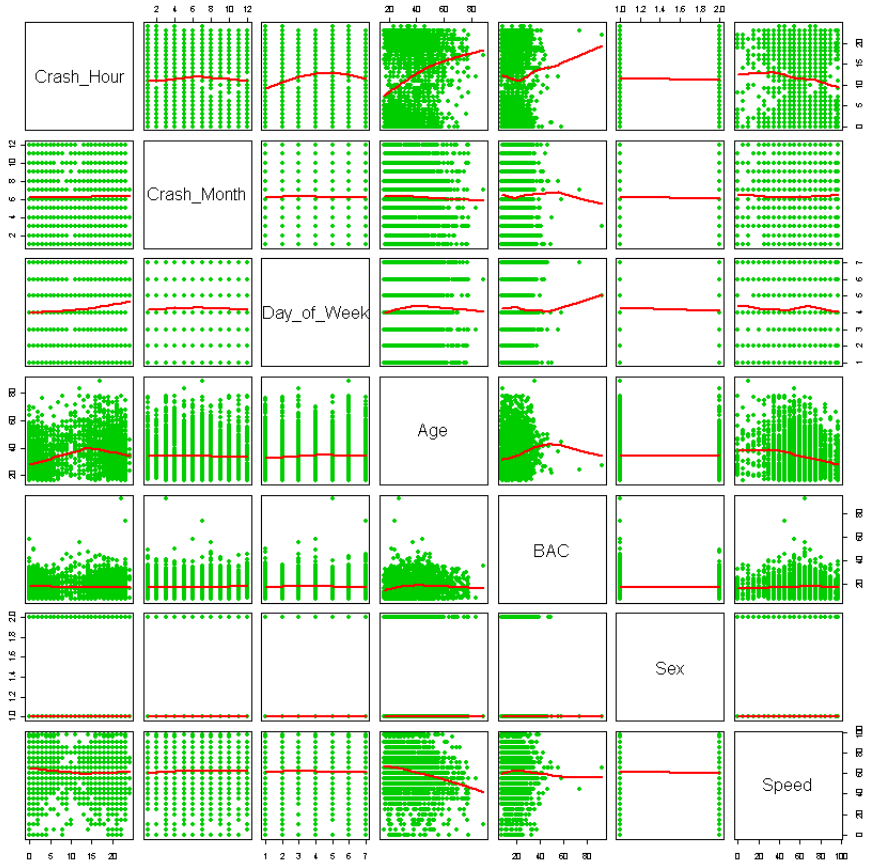


Figure 5.47: Scatterplot Matrix of Seven Road Fatal Crash Factors

The off-diagonal elements of ${}^{w}SS$ represent the sum of dominant interactions for a given "state" i and another "state" j related through the variable "alcohol". The diagonal elements are the marginal distribution of fatalities of "state".

Similarly,

$$ {}^{b}SA^{T} \cdot {}^{w}SA = {}^{b}AS \cdot {}^{w}SA = {}^{w}AA, $$

is the one-mode network of BAC levels related through the variable "states". The off-diagonal elements of ${}^{w}AA$ represent the sum of dominant interactions for a given "BAC level" i and another "BAC level" j related through the variable "state". The diagonal elements are the marginal distribution of fatalities of "BAC level".

While the bi-partite matrix ${}^{w}SA$ has its own implication, the 1-mode matrices ${}^{w}SS$ and ${}^{w}AA$ both have their unique properties.

Next, I apply the inner product technique on the vector of values between two *states* to obtain a similarity measure. This can be thought of as the Euclidean distance or ℓ_2-norm, correlates are special case of the cosine similarity measure. Because the number of fatalities ≥ 0, the vectors associated with the vertices (states) are in the first quadrant and thus $0^{\circ} \leq \theta \leq 90^{\circ}$. The cosine similarity is a quantitative measure of attributes that intersect; therefore, "states" with more shared values are considered very similar.

I start the analysis by exploring the driver related factors. The following is a list of the top six driver related factors for causing a fatal crash:

1. Under the influence of alcohol, drugs or medication;
2. Inattentive (talking, eating, etc);
3. Failure to keep in proper lane;

4. Failure to yield right of way;

5. Failure to obey actual traffic sign, traffic control devices or traffic officers; Failure to obey safety zone traffic laws;

6. Driving too fast for conditions or in excess of posted maximum.

The horizontal bar plot shown in Figure 5.48 presents the driver related factors that captured the highest crash rates. The 18 factors explain 90% of all crashes across the US. Failure to keep in proper lane is the first dominant factor, followed by Driving too fast; the third is the Influence of alcohol, drug or medication.

The top 6 Driver Factors Related to Fatal Crashes are:

1. Failure to Keep in Proper Lane.

2. Driving too Fast for Conditions or in Excess of Posted Maximum.

3. Under the Influence of Alcohol, Drugs Or Medication.

4. Failure to Yield Right of Way.

5. Inattentive (Talking, Eating).

6. Failure to Obey Actual Traffic Sign,Traffic Control Devices or Traffic Officers;Failure to Obey Safety Zone Traffic Laws.

The computed similarity matrix of the driver related factor is presented in Figure 5.49. There are several blocks along the main diagonal of highly similar states with respect to this measure.

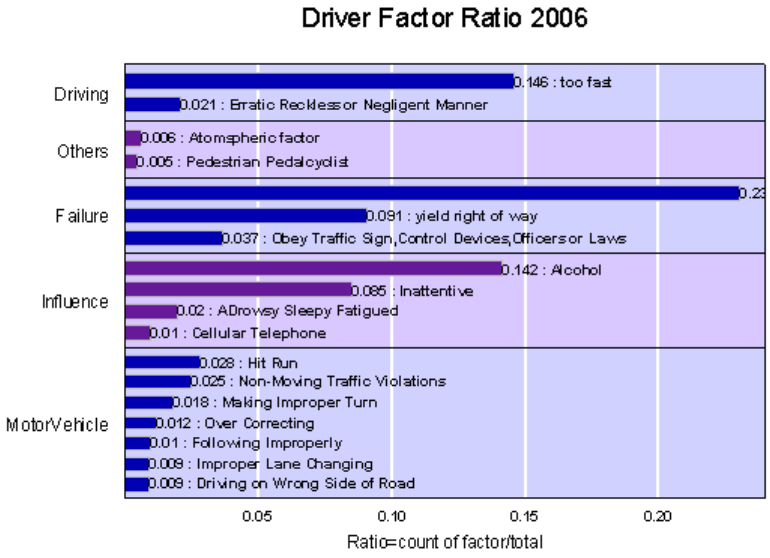


Figure 5.48: Horizontal Bar-plot of the Top 18 Main Driver Related Factors.

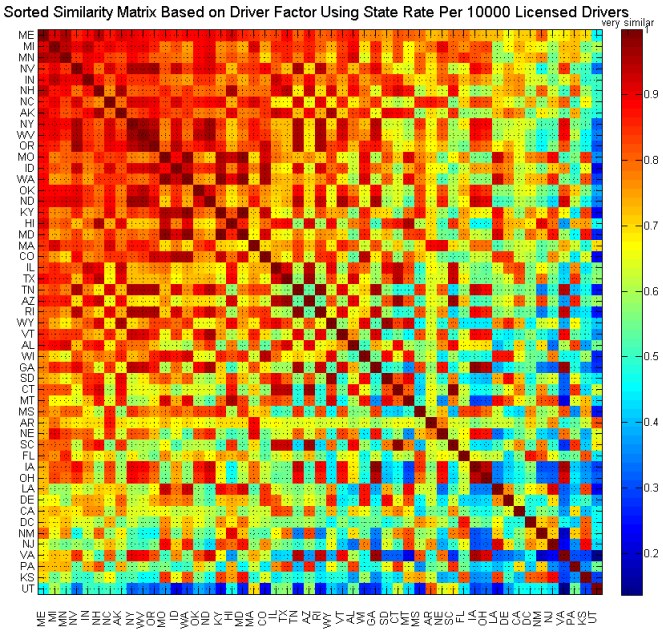


Figure 5.49: Similar States Based on Driver Factors Related Crashes.

5.4.1 Alcohol Factor

The next factor to study related to road fatalities is the Blood Alcohol Content (BAC) level, because alcohol abuse is more common among people than drug addiction or medication the latter will be out of focus at this stage. Figure 5.50 shows the distribution of alcohol related road fatalities. Fatalities with BAC level below 0.08%; the legal limit in all states to be alcohol-impaired, are excluded.

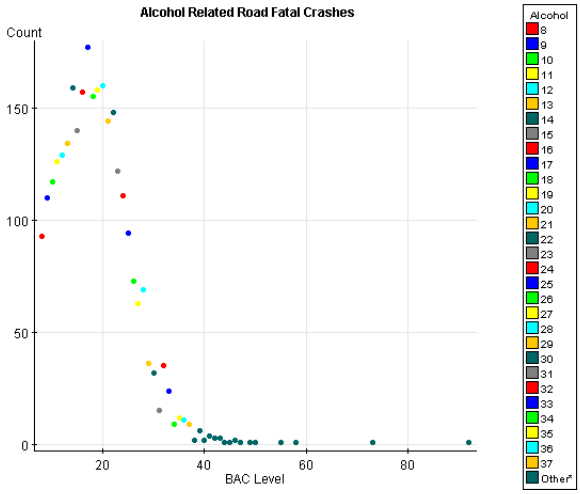


Figure 5.50: Distribution of Alcohol Related Road Fatalities.

Most fatal crashes are associated with BAC levels 0.01% through 0.23%. The concept that drivers having high levels of BAC are more prone to devastating accidents as opposed to low levels is counterintuitive, it is more likely to have someone with low BAC level than high BAC level and thus increasing the chances of a crash.

Let us visualize the two-mode state-by-alcohol network in all 51 states, the graph in MDS layout is shown in Figure 5.51. There is an obvious cluster of relatively low BAC level $(0.08\% - 0.33\%)$ associated with nearly all states residing at the core of the network.

Figure 5.52 shows the state-by-alcohol matrix and Figure 5.53 depicts the network with

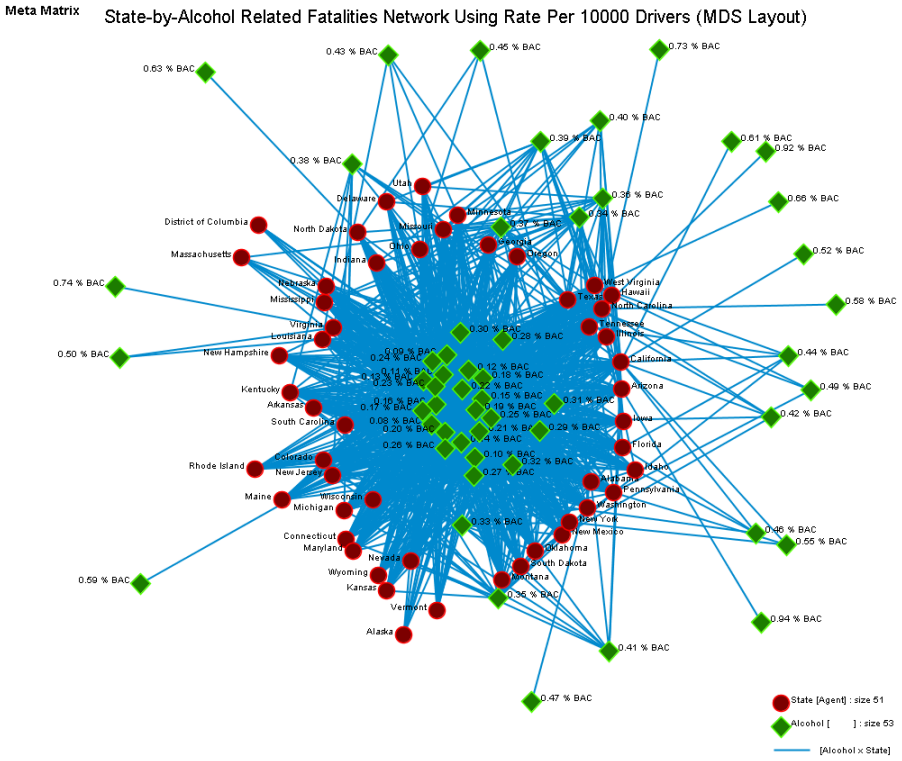


Figure 5.51: The Graph of The State-by-Alcohol Social Network.

an emphasis on edges with rates (weights) ≥ 0.1 per 10,000 registered drivers. Filtering out weak ties and focusing only on strong ties has the advantage of making the graph more readable. This is done by hiding or removing low frequency edges. In some networks, it may be more important to keep weak ties, so that edges surpassing the threshold are deleted. Identifying states with high rates is of interest in the road fatality network. Figure 5.53 suggests an elite group of states with high alcohol related road mortality rates consisting of (Wyoming, Montana, South Dakota, and Mississippi). Mississippi forms an isolate component with 0.16% & 0.18% BAC levels. Wyoming is associated with 7 different BAC levels one of which is the highest rate (0.1536 $\rightarrow$ 0.22% BAC).

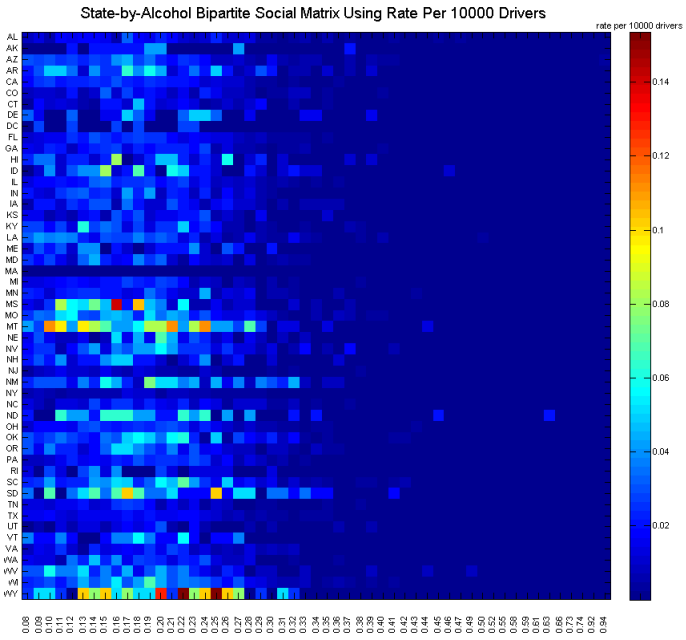


Figure 5.52: State-by-Alcohol Matrix.

The 1-mode states similarity matrix due to alcohol related fatal crashes is shown in Figure 5.54. The values are sorted based on the total similarity for each state to make it easy to distinguish the highly similar states (value$\rightarrow$1) from the highly dissimilar states (value$\rightarrow$0). Texas – California – Florida – Illinois and Ohio stand out as highly similar states, whereas Alaska – Washington DC and Massachusetts stand out as highly dissimilar states among

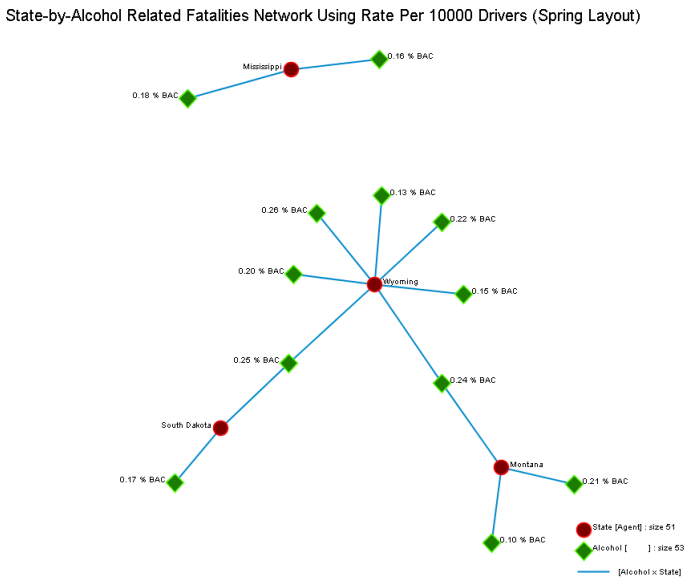


Figure 5.53: State-by-Alcohol.

themselves and somewhat dissimilar with all other states. Hawaii, Maine and Rhode Island is a group of relatively similar states.

Then I investigate how pair of states are related to each other through all the BAC levels, the results are shown in Figure 5.55. Using count (an abstract measure), the top 3 states are (California, Texas, Florida), they also form highly pairwise related states. Using rates, the top 4 states are (Montana, Mississippi, South Dakota, Wyoming) with Wyoming capturing the highest rank. These results are consistent with previous conclusions from the two-mode analysis of state-by-alcohol network both using counts and ratios.

In Figure 5.56(a) I am showing the highly unrelated pairwise states on a map, and in Figure 5.56(b) I am showing the highly related pairwise states on a map. Note that among the highly related states; Figure 5.56(b), the clique (MS-MT-WY-SD) is noticeable.

Figures 5.57 and 5.58 show the 1-mode network of pairwise BAC levels related through all states. The clique of heavily pairwise connected BAC levels is visible and a difference is

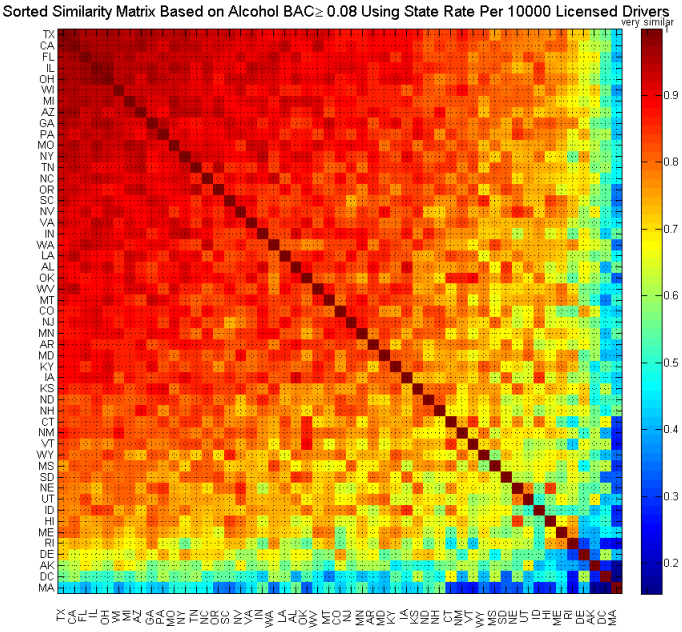


Figure 5.54: State-State Similarity Based on BAC Level.

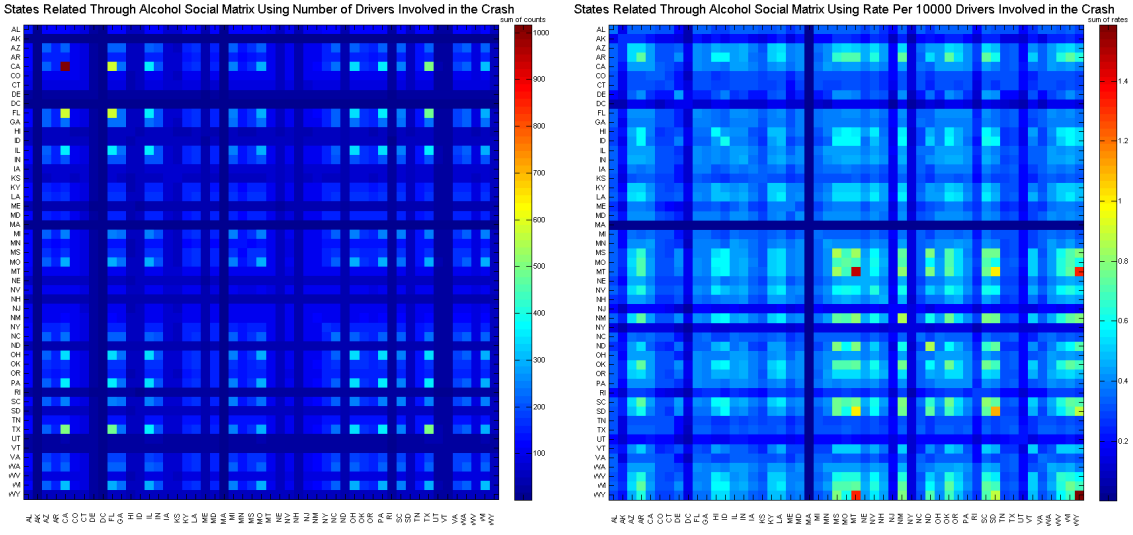
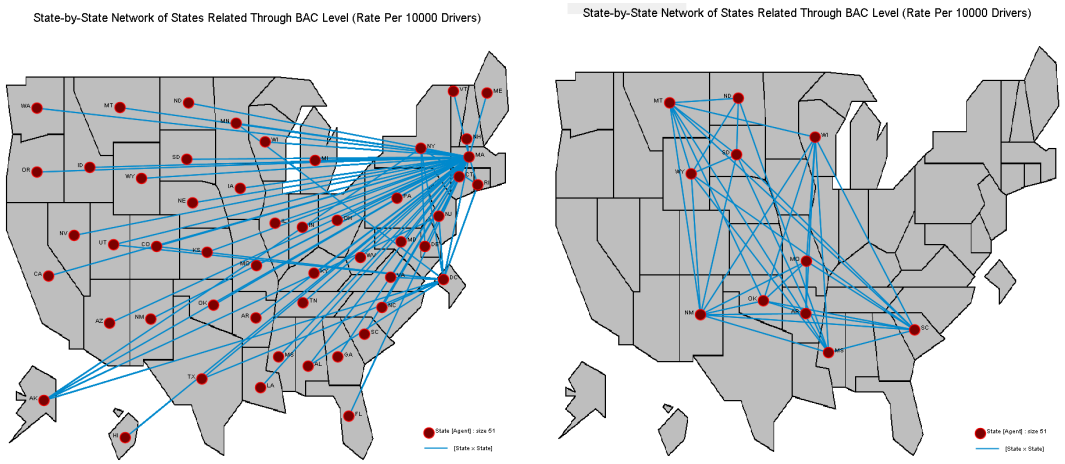


Figure 5.55: State-State Relationship Through BAC Levels.



(a) Low Ranking States (Weight ≤ 0.1). (b) High Ranking States (Weight ≥ 0.7).

Figure 5.56: State-State Relations Through BAC (Showing Strong and Weak Ties).

clear between using counts and ratios. Note that line crossings may be an issue sometimes.

5.4.2 Age Factor

In this part, I analyze states and age factor related to road fatalities. The question of interest is: which ages are more likely to have crashes and in which states? Figure 5.59 shows how age and states are related, the graph shows that teenagers are the main contributor for crashes across all states. The distribution of age of fatalities is heavily skewed right in all states because of the outliers present in some states. Additionally, states with red or dark orange squares resemble high fatality rate per 10000 registered drivers in the state. States such as Wyoming and Mississippi have the highest rates of 0.359 (age = 30) and 0.305 (age = 21) respectively.

Referring to Figure 5.60, there seem to be relatively large fatality rates associated with ages 15 through 62, and in some states counts reach older citizens such as 87 years old in California and Florida. As for Texas, there are more crashes involving children, California has a similar situation. In contrast, there are some states with low crash numbers for all

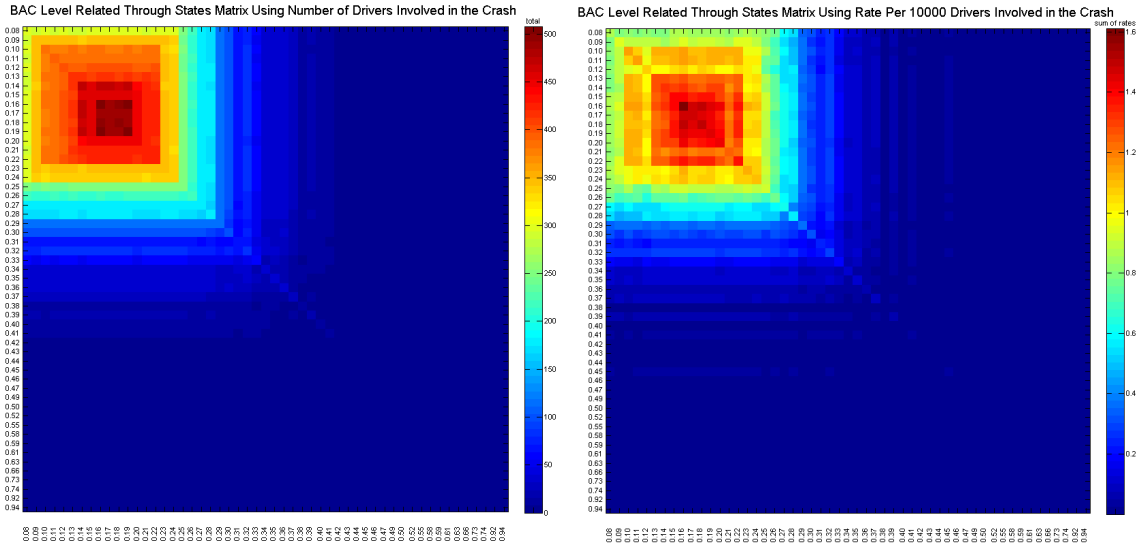


Figure 5.57: Alcohol-Alcohol Relationship Through States.

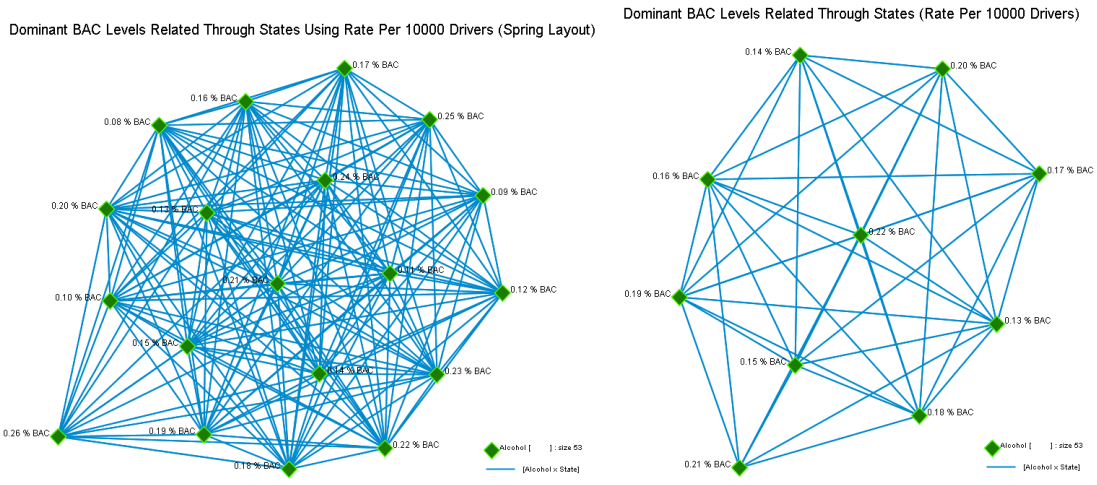


Figure 5.58: Alcohol-Alcohol Relationship Through States (Top BAC Levels).

ages such as Alaska, District of Columbia, Delaware, North Dakota and Vermont. It may relate to people's driving habit, the strictness of police and/or speed limit. Various reasons contribute to the crash, as various reasons lead to safer driving condition.

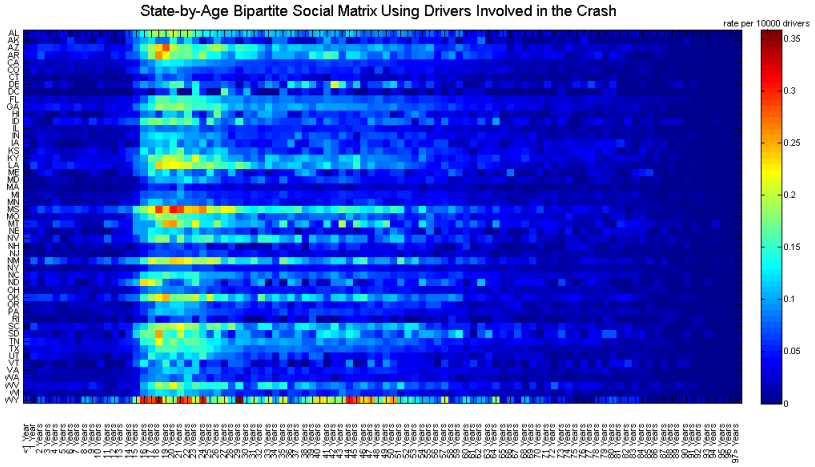


Figure 5.59: State-by-Age Bipartitie Social Matrix.

Figure 5.61 shows the ratio of males to females involved in fatal crashes using linked micromap, the original `R` code was provided by Dan Carr. This explains why insurance rates are much higher for teenagers and for males more than any other category. Males ratio is always about two times larger than of females'.

From the graph we conclude that older people have safer driving habits, which is also confirmed by the scatterplot in Figure 5.62. There appears to be a linear downward trend for the distribution of age, a simple linear regression model gave a coefficient of determination of $R^2 = 0.74$.

Figure 5.63 depicts the state-by-age two-mode network in MDS layout. Although there are two main clusters for age and another two for states, the graph is too busy and edges crossings are problematic. Consequently, I filter out the network by discarding weak ties to

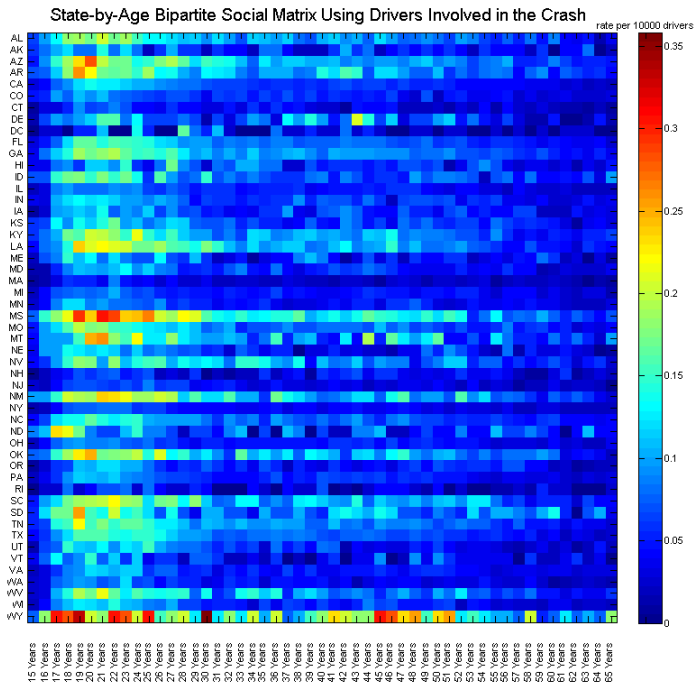


Figure 5.60: State-by-Age Two-Mode Social Matrix For Ages 15-65.

focus only on high rates, the graph with rates more than 0.1 is shown in Figure 5.64. The graph in Figure 5.65 depicts the state-by-age network in spring layout. In the graph, I am showing edges with frequency (rate per 10000 drivers) ≥ 0.2. Wyoming is connected to 21 different ages with a high degree centrality.

The similarity matrix of states based on age factor related crashes is shown in Figure 5.66. The majority of states are highly similar when it comes to age except for Washington DC with a similarity measure in the range (0.5-0.7).

Finally, I construct the one-mode network of states related through ages and the network of ages related through states, the results are presented in Figures 5.67 and 5.69. Using counts, California – Texas – Florida are the top three states. However, ratios put the states Mississippi and Wyoming up high in the list. Ages 16-27 are relatively high in rates, MDS

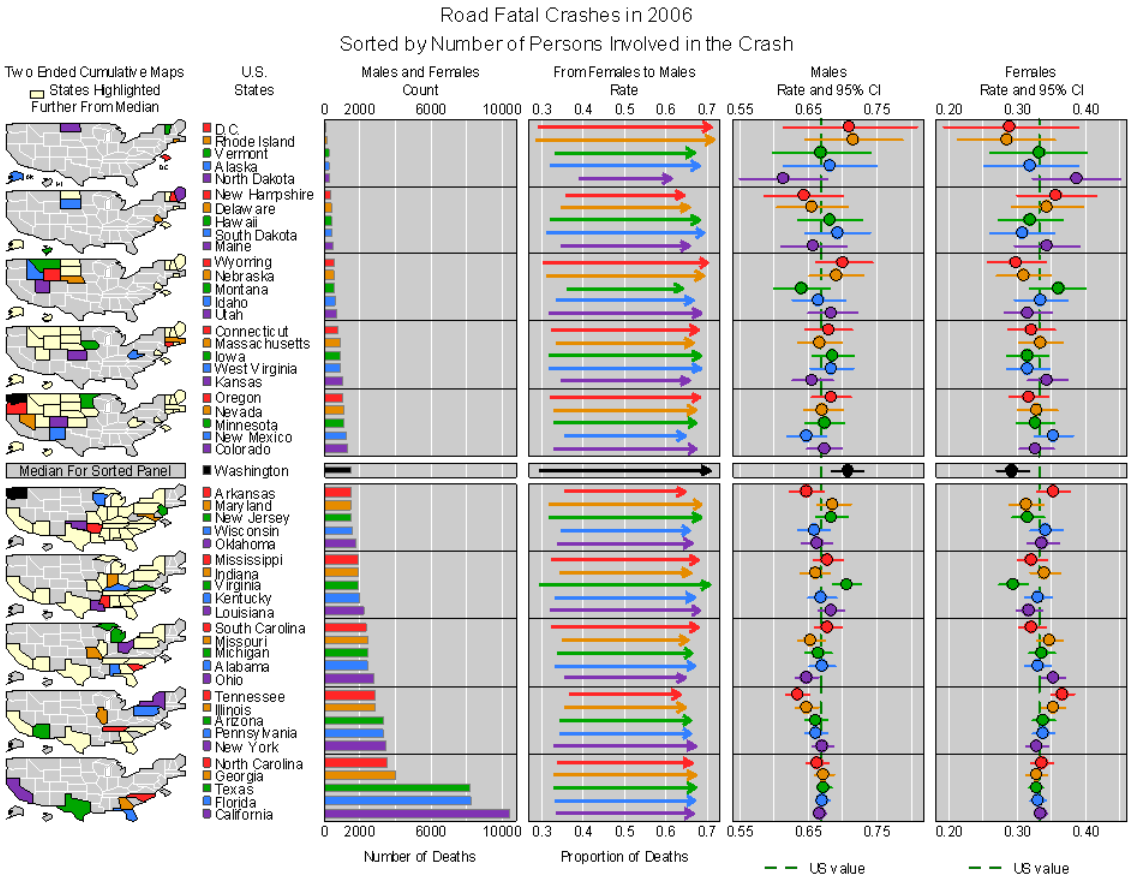


Figure 5.61: Gender State Micromap.

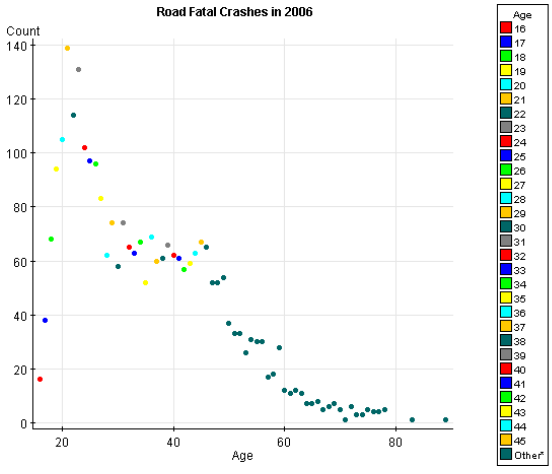


Figure 5.62: Marginal Distribution of Age.

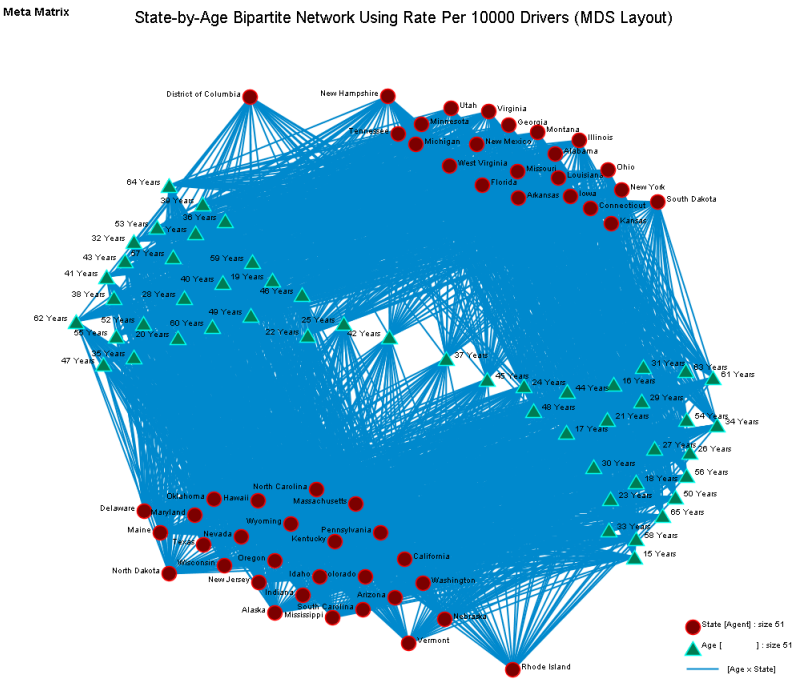


Figure 5.63: State-by-Age Two-Mode Graph For Ages 15-65 in MDS Layout.

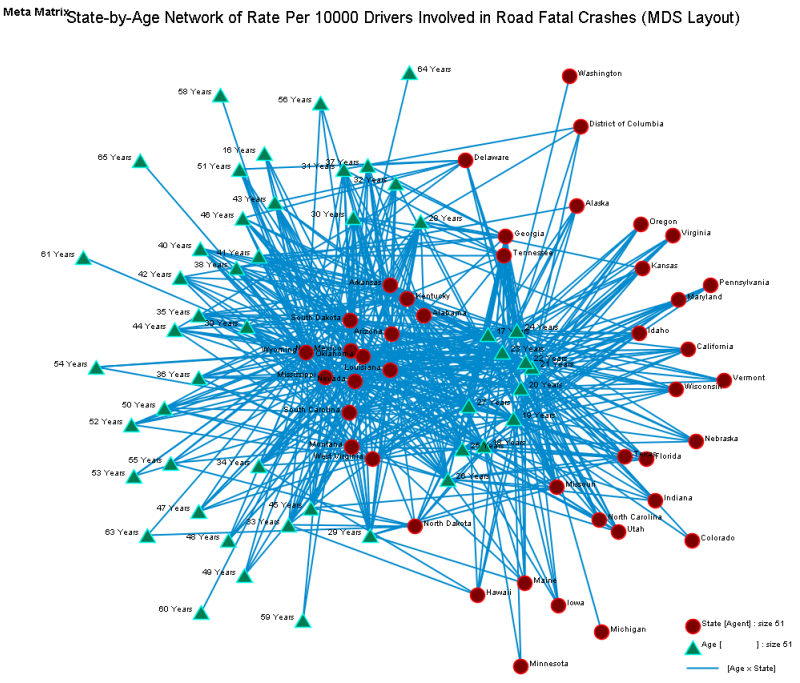


Figure 5.64: State-by-Age Graph For Ages 15-65 With Edge Weight ≥ 0.1.

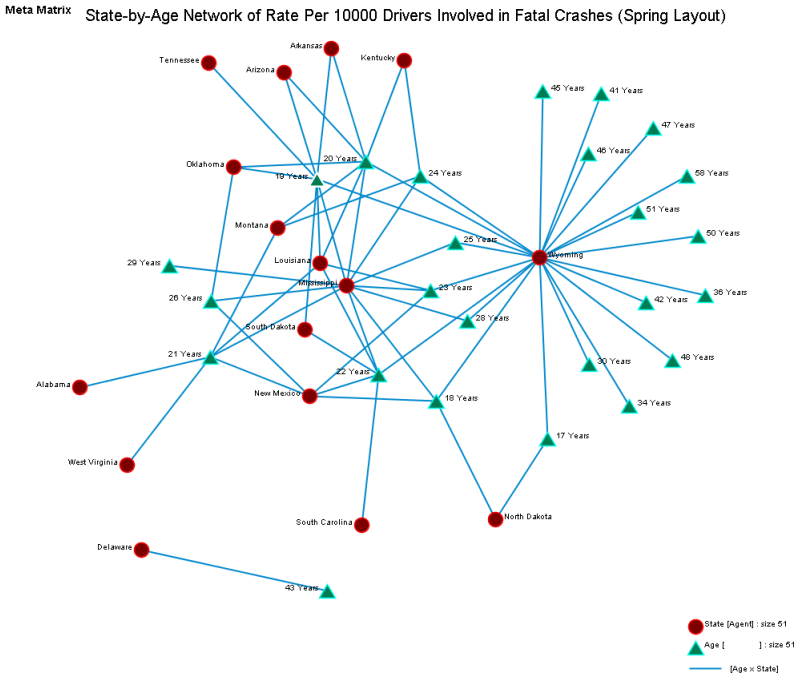


Figure 5.65: State-by-Age Two-Mode Social Graph For Ages 15-65.

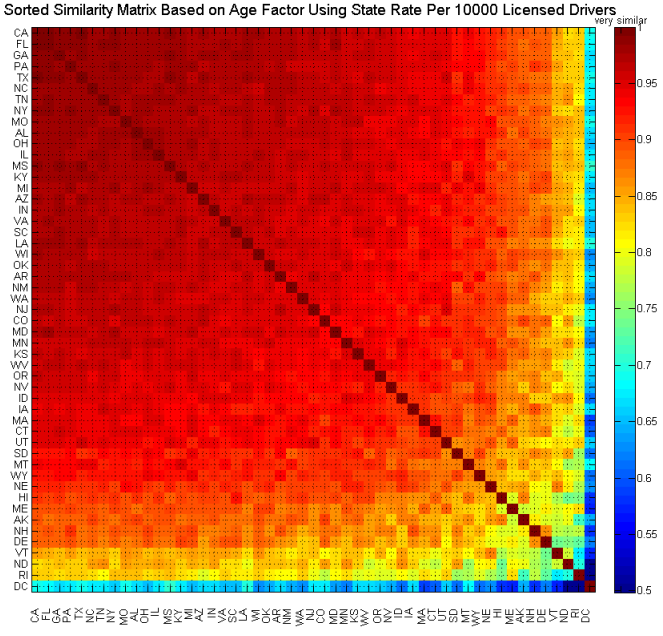


Figure 5.66: Sorted Similarity Matrix Based on Age Factor.

layout in Figure 5.64 suggested the same range. The diagonal values are the marginal distribution for age in all states, and the off-diagonal values represent the relationship between a dominant age i and another age j. And finally, an emphasis on weak ties and strong ties is shown in Figure 5.68.

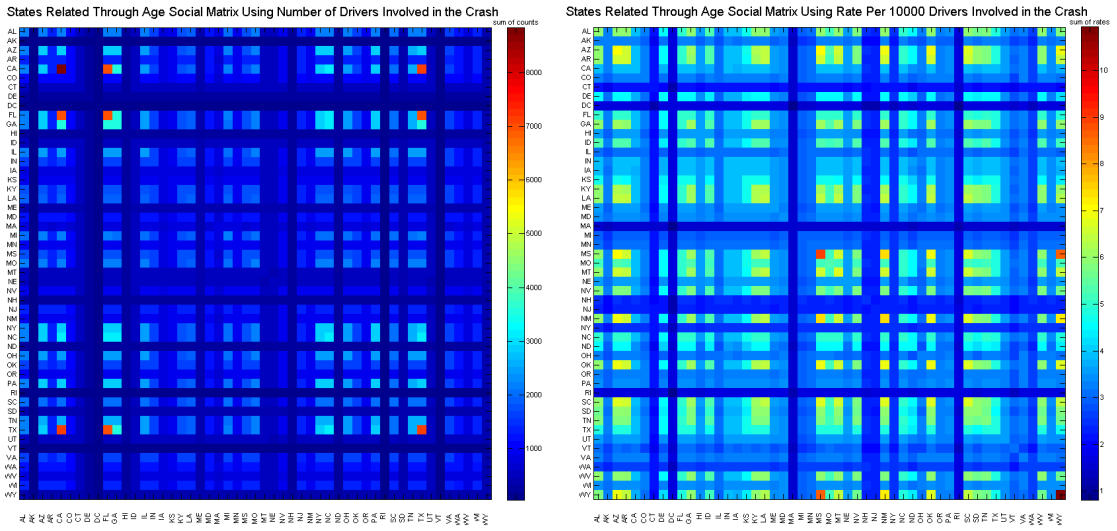


Figure 5.67: State-State Relationship Through Ages.

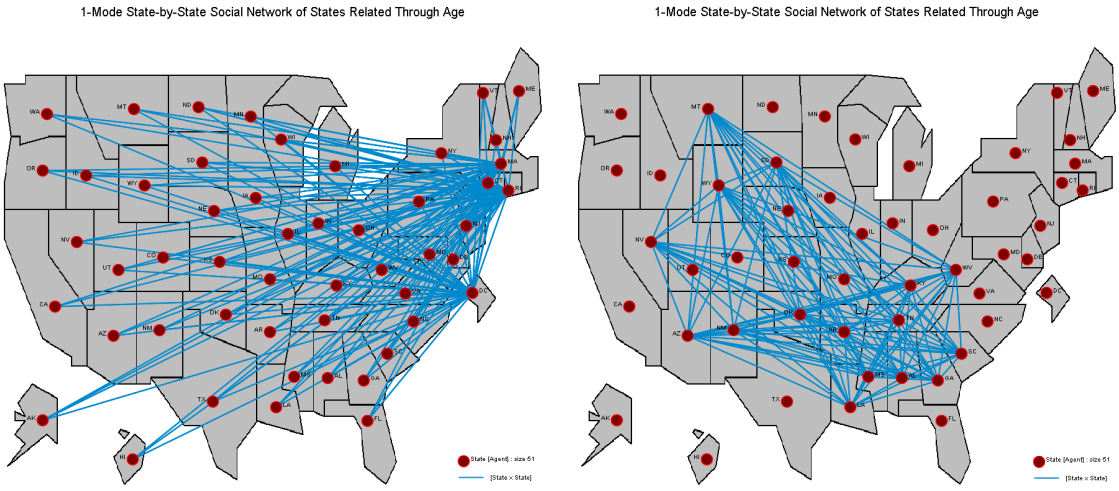


Figure 5.68: State-State Relationship Through Ages.

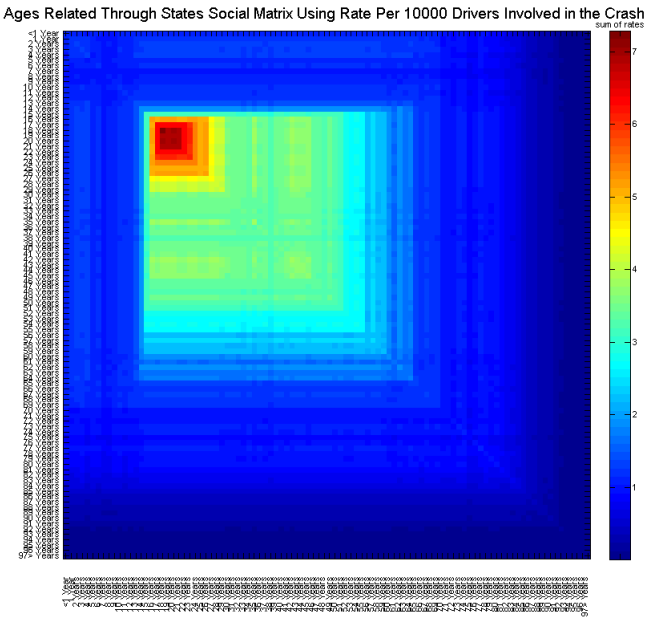


Figure 5.69: Age-Age Relationship Through States.

5.4.3 Travel Speed Factor

Travel speed factor is yet another important variable to examine, a scatterplot of the travel speed is presented in Figure 5.70. Most crashes happen at 45 mph and 55 mph, different from what people usually think that high speeds lead to more crashes. It may relate to a neighborhood where drivers drive much faster than the posted low speed limit.

Continuing with the same analogy, I start by exploring the state-by-speed matrix of fatalities, two plots are provided in Figure 5.71; one using counts and one using ratios. As far as counts, the top three states are (CA, FL, NC). However, ratios have totally different distribution; some high ranking states include (DE, LA, NC, OK, SC, WY). Figure 5.72 shows a graph version of the network. A big portion of highway 95 which goes from New York all the way to Florida crosses DE–VA–NC–SC–GA. A possible explanation as to why there are high road fatality rates due to speed in NC and SC is that by the time drivers enter these states they are tired and exhausted from the long drive, they want to arrive to Florida quickly and as a result a crash happens. As presented in Figure 5.79, there are two

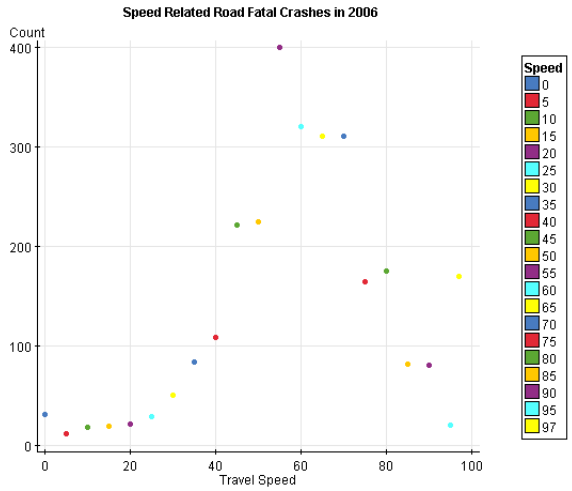


Figure 5.70: Marginal Distribution of Travel Speed.

edges with high weights connecting Florida and the vehicle is registered in New York and vice versa. The registered vehicle factor is discussed in the next section.

Once again, I discard weak ties to identify elite groups of states with high road fatalities related to travel speeds. The graph in Figure 5.73 depicts the filtered out network; ties thickness is set by rate per 10000 drivers. Wyoming has a strong tie associated with speed 65 mph (ratio = 1.101 per 10000 registered drivers in WY). This extreme value is a concern because Wyoming has also the highest alcohol related road fatality rate. Delaware has also a strong tie associated with speed 50 mph (ratio = 1.0002). And finally, Louisiana and North Carolina both have strong ties associated with speed 55 mph (ratio = 0.8027).

Figure 5.74 represents the sorted similarity matrix of states based on the travel speed factor. The plot shows pattern of similarity among states, there are groups of highly similar states among themselves, but at the same time highly dissimilar with other states. Blocks of similar states are also present along the main diagonal, these blocks resemble cliques. Interestingly enough, the states of Maine and Iowa are nearly dissimilar with almost all other states.

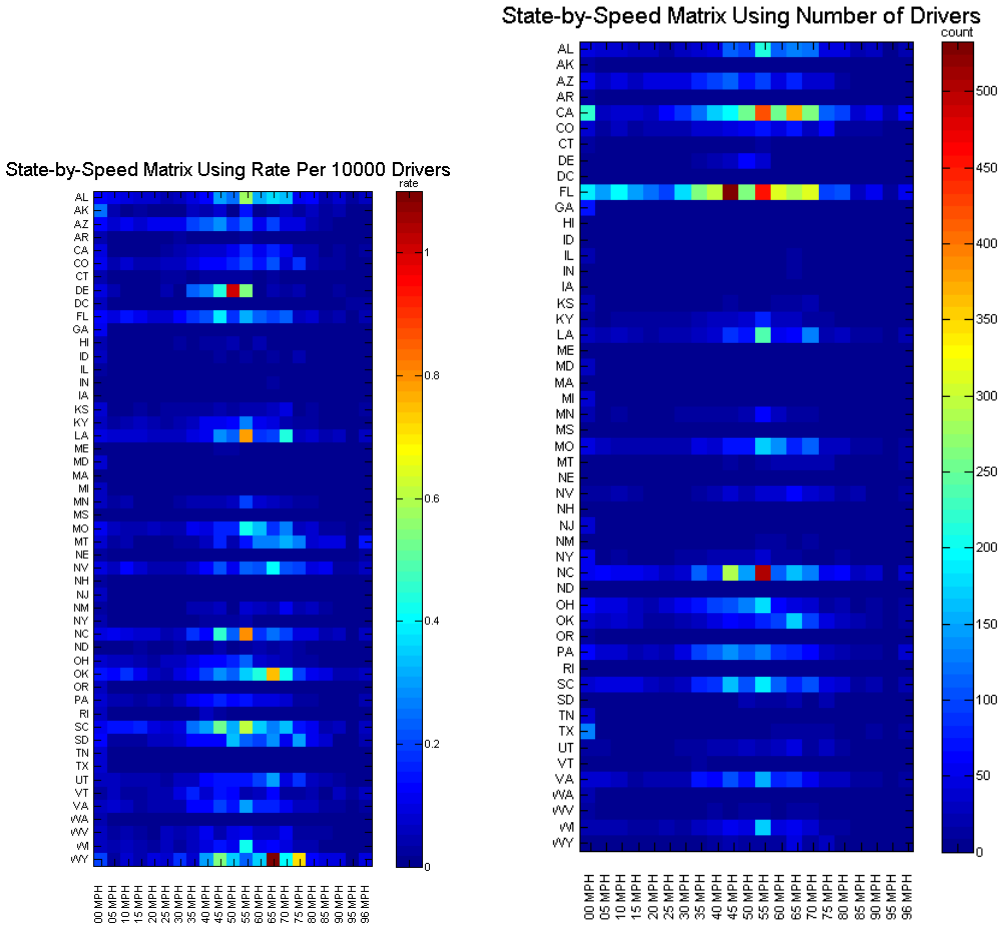


Figure 5.71: State-by-Speed Matrix.

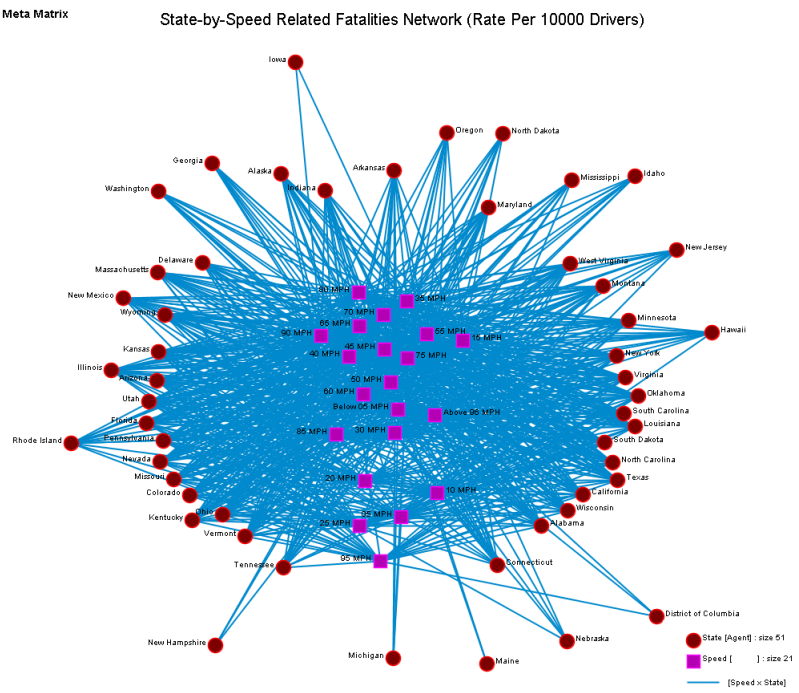


Figure 5.72: State-by-Speed Network.

The results of the matrix multiplication technique on the state-speed matrix show that (WY, OK, SC, NC, LA, AL) are the top states related through road fatalities due to high speeds, see Figure 5.75.

Figure 5.76 shows the top 20 states related through speed on the left and on the right the top 17.

Figure 5.77 is the speed-speed matrix resulted from multiplying the the speed-state matrix by the binary version of its transpose. The speeds 45–50–55–65 are highly related, they are the dominant speeds in all states. The plot on the right is showing the top five speeds.

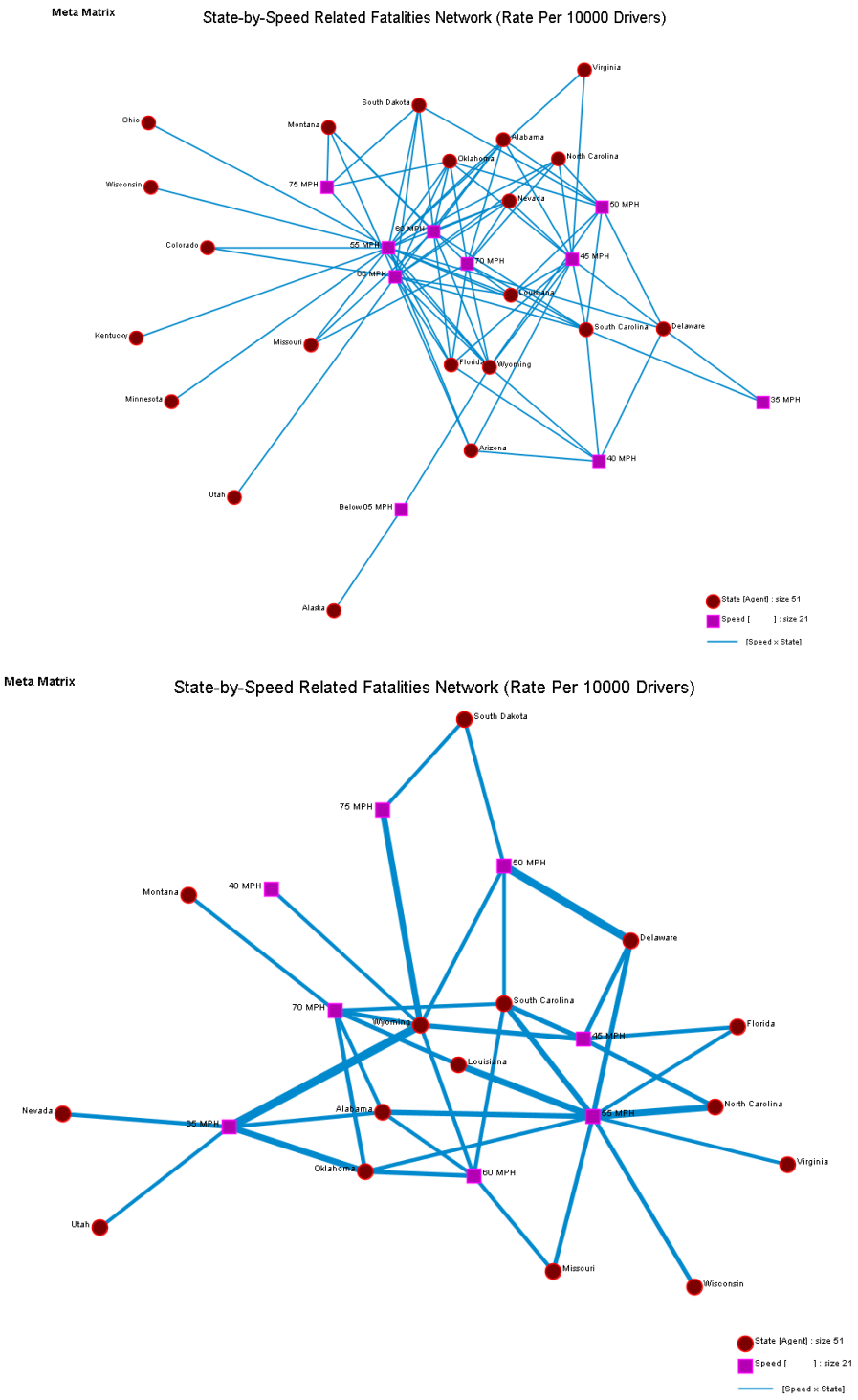


Figure 5.73: State-by-Speed Network.

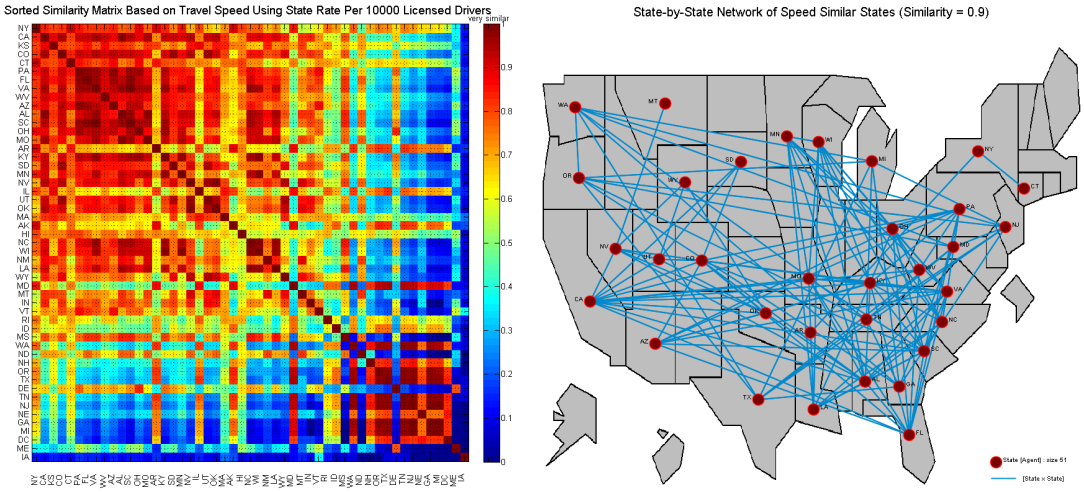


Figure 5.74: Similar States Based on Travel Speed.

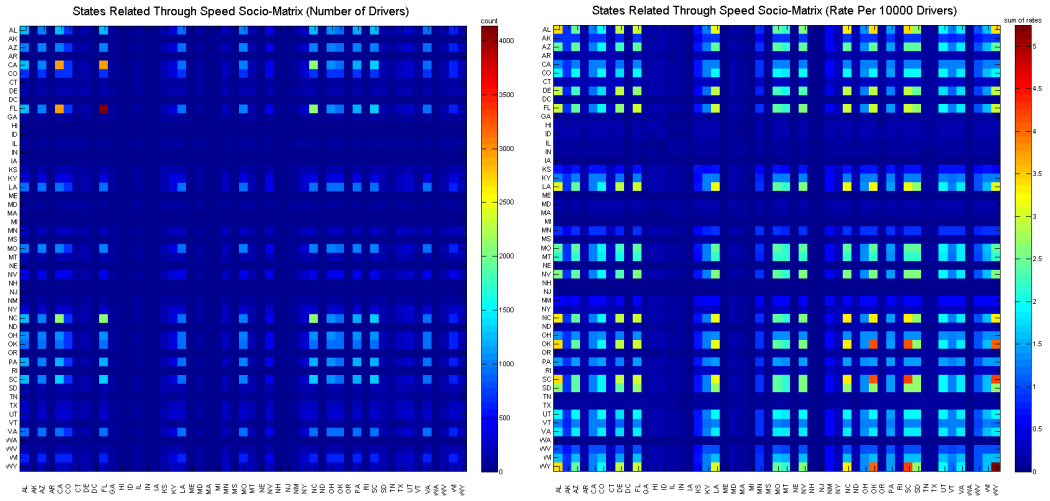


Figure 5.75: 1-Mode State-State Relationship Through Travel Speed.

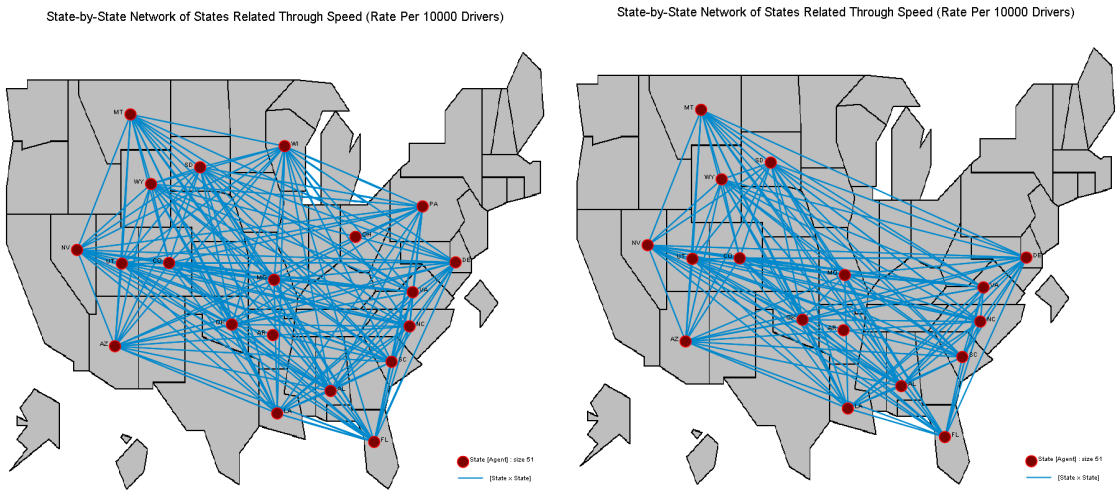


Figure 5.76: State-State Relationship Through Travel Speed.

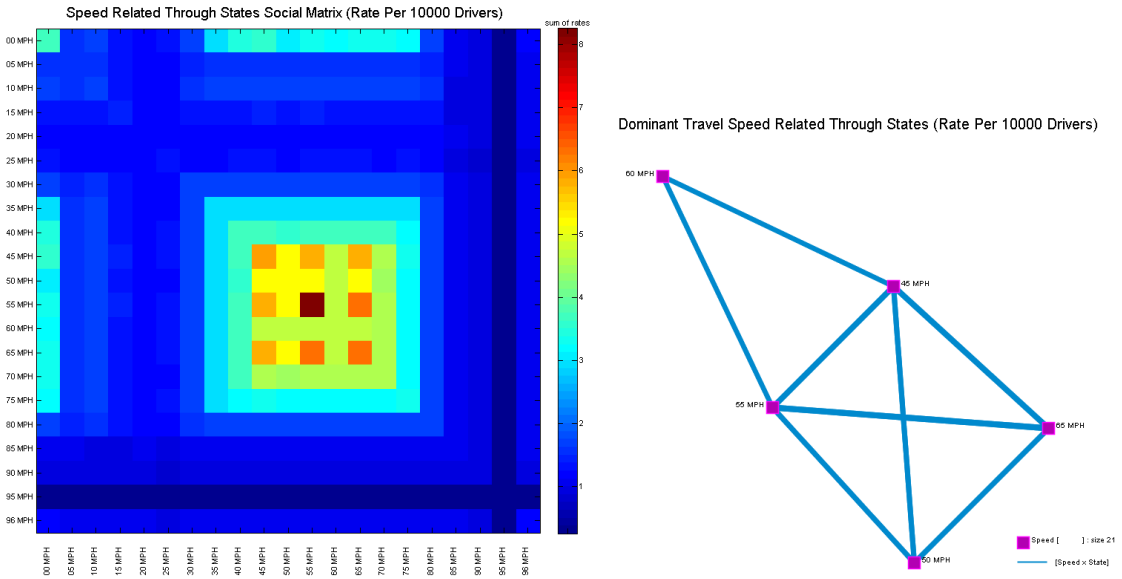


Figure 5.77: Speed-Speed Relationship Through States.

5.4.4 Registered Vehicle Factor

In this section, I examine the relationship between the state where accidents happened and the registered plate of the vehicle involved in the crash. The two-mode network is portrayed in Figure 5.78. The size of the nodes is set by degree centrality; in this case it is the number of other states/countries that state is connected to, while the thickness of edges is set by weight (number of fatal crashes). Arrows are pointing in the direction of the target state. Surprisingly enough, weak ties (edges) are connecting states such as (MA, NC, UT, AZ, OR) to Hawaii although there is no bridge between these states and the Caribbean island. What might be the reason is that there are military bases in Hawaii where military personnel can maintain their home-state license even if they are physically do not live in their home-state. Additionally, there are weak edges connecting the countries Canada and Mexico to Florida. The graph on the right emphasizes edges having weights $\geq$ 20 fatalities. Interestingly enough, are the strong edges between Mexico and Arizona and the two edges between Florida and New York. Florida and New York appear to have high interaction; Florida is a decent place for people to vacation or retire in. But even a more interesting conclusion is the high connectivity along the way between NY and FL and between FL – TX – CA – WA; it forms sort of a long chain starting in NY and ending in WA.

Finally, I computed the similarity matrix of the states and registered vehicles network, the result is presented in Figure 5.79. What can be concluded is that most states are highly dissimilar and only few states that are somewhat similar.

5.4.5 Road Function Class Factor

Road function class is yet another factor to explore, it is mainly categorized as urban and rural. The plot in Figure 5.80 shows the difference in the number of road fatalities between urban roads and rural roads, the difference is taken as (urban count - rural count). A pattern among states with low difference and states with high difference is visible. States

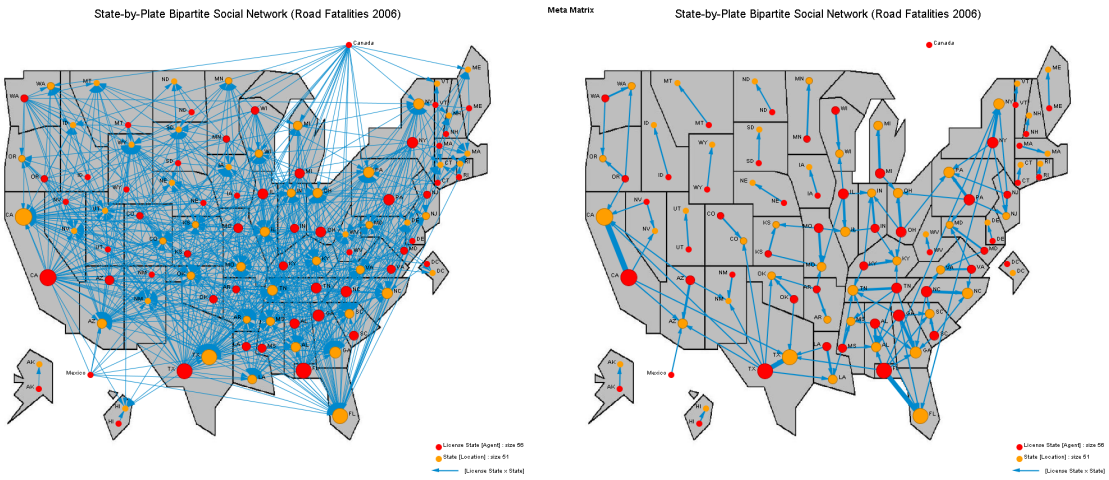


Figure 5.78: Bi-partite Network of States Related With Registered Vehicles.

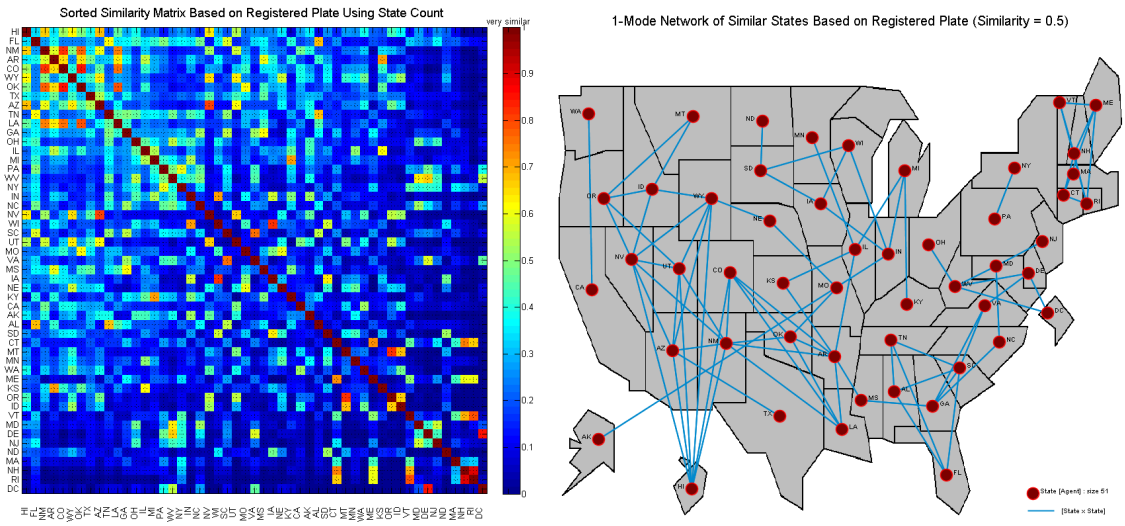


Figure 5.79: Similar States Based on Registered Vehicle Factor.

having above median differences are located on the east coast, west coast besides the states (Illinois, Florida, Louisiana, Georgia) with the exception of the states Oregon and Maine.

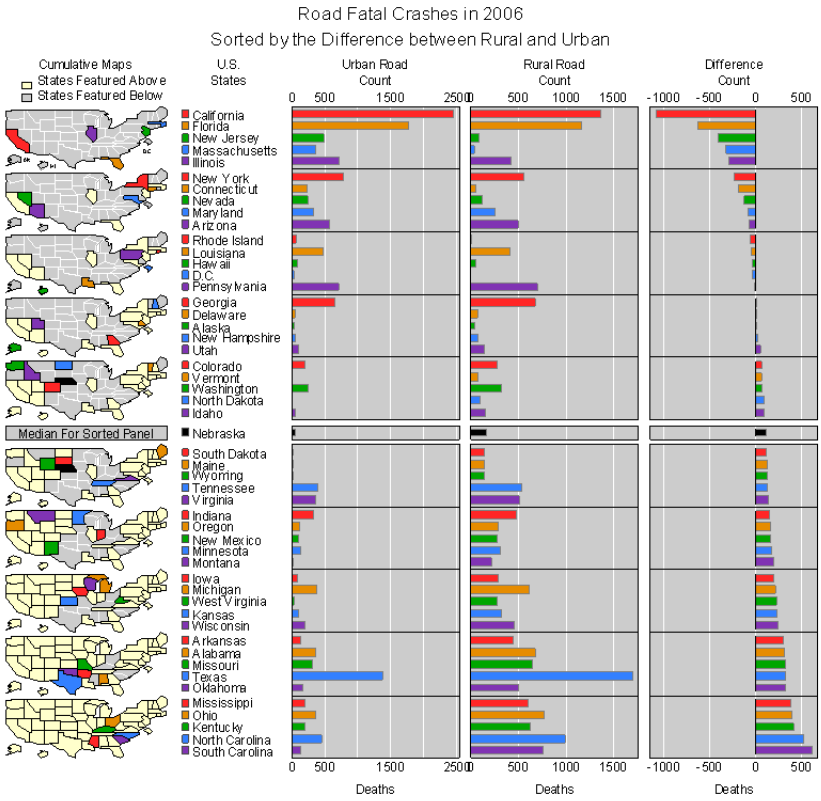


Figure 5.80: Road Function Class State Micromap.

The state-state similarity measure based on road function class data is the last computation to perform on this network, the results are shown in Figure 5.81. Patterns and cliques of highly similar states are clear, Washington DC has the lowest similarity measure with all other states (with the exception of MA) simply because DC is a city and not a state.

5.4.6 Conclusion

Using network theory techniques I identified states with high road fatality rates per 10,000 registered drivers that are due to alcohol, travel speed and age related. For example, the

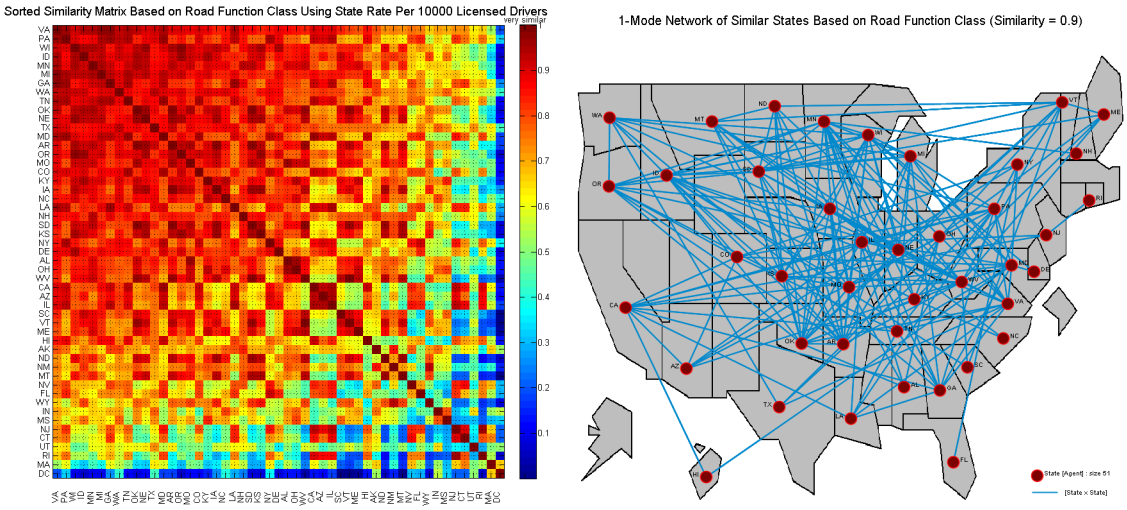


Figure 5.81: Sorted Similar States Based on Road Function Class Factor.

top four states with respect to alcohol related fatality rates are Montana, Mississippi, South Dakota and Wyoming. The top four states with respect to age related fatality rates are Wyoming, Mississippi, New Mexico and Arizona. The top four states with respect to travel speed related fatality rates are Delaware, Louisiana, North Carolina and Wyoming. The most noticeable observation is Wyoming, which stands out in all three factors.

Furthermore, I identified high ranking fatalities associated with alcohol, age and speed factors. For instance, the top BAC levels are between 0.08% and 0.33%. The top ages are between 16 and 27 years old. And the top travel speeds are 45–50–55–60–65 mph.

Additionally, I found out that in Florida there is a high number of fatal road crashes of vehicles holding New York registration and in Arizona of vehicles registered in Mexico.

Road crashes are generally not fatal but in some occasions when they happen they are catastrophic and therefore further attention is required from the statutory law-enforcement authority. I hope that this analysis provides an insight to authorities and policy makers to impose appropriate rules and laws in states with high road fatality rates that may reduce

chance of road fatalities.

5.5 Term-Document, Bigram-Document Networks

The following case study concerns news documents network. I start with a brief description of this example then I apply the techniques I developed in this dissertation to visualize and analyze the network. News text data were collected by the Linguistic Data Consortium in 1997 and were originally used in Martinez (2002). The data consisted of 15,863 news reports collected from Reuters and CNN from July 1, 1994 to June 30, 1995. The full lexicon for the text database included 68,354 distinct words. In all 313 stopper words are removed and after denoising and stemming, there remain 45,021 words in the lexicon. Dr. Martinez provided the MatLab code to stem and denoise text within a document. In my dissertation I only used 503 news documents. The documents were constantly being updated and in many cases the new document was a copy of the old one with some additions. The objective of this study is to categorize documents based on similarity and compute correlations between documents with respect to terms and with respect to bigrams. In this regard, I associate terms, bigrams and documents. Classically, the analysis of text data can be done through the use of text mining techniques; however, I use the network theory approach to serve this purpose.

Consider the two-mode term-document network and bigram-document network whose matrices are ${}^{w}TD$ and ${}^{w}BD$ respectively. Figures 5.82, 5.83 show the distribution of terms in the 503 documents and the top 51 terms respectively.

The distribution of terms is a skewed distribution and demonstrates the many-some-few pattern, many terms have relatively low frequency and few terms have relatively high frequency. The terms “sai” and “said” both correspond the root “say” when stemming the term. Among the highly ranked terms are the terms “peopl” - “on” - “north” - “go” - “here” - “two” - “out” - “just” - “cnn” and others.

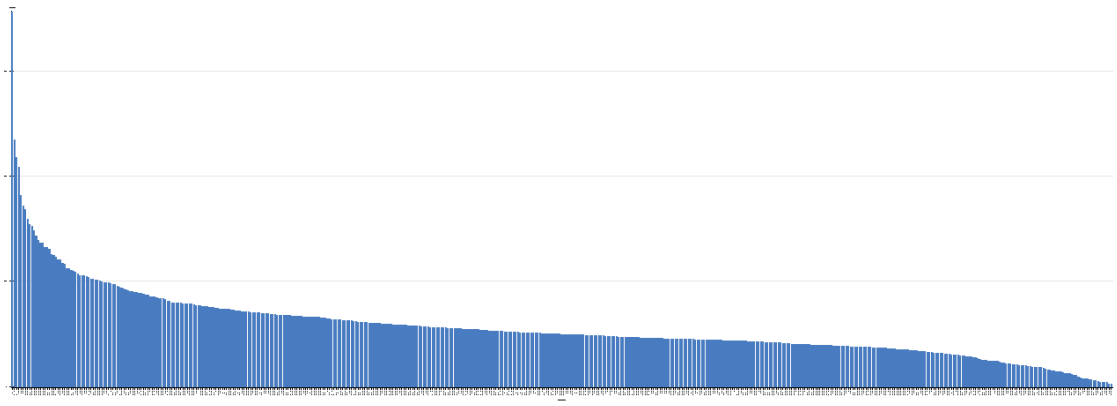


Figure 5.82: Distribution of terms in the 503 documents.

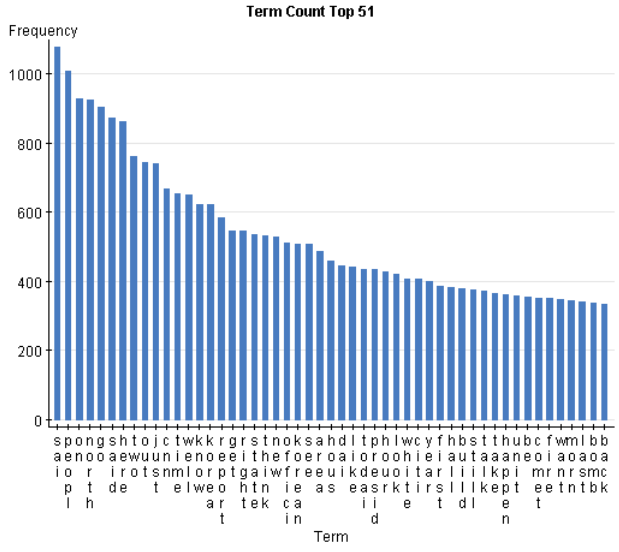


Figure 5.83: Top 51 terms.

There are in total 7143 unique stemmed and denoised terms associated with the 503 documents; therefore, the weighted term-document matrix is ${}^{b}TD_{7143\times503}$. Figure 5.84 shows how the terms relate to documents, a black square indicates that the given term belongs to the given document. As new documents are processed new terms are being add to the corpus at a high rate. Horizontal bars indicate that some terms show in multiple documents, those are the high frequency terms. Terms up in the list (the core of the network) are connected to more documents compared with terms down the list (the periphery of the network).

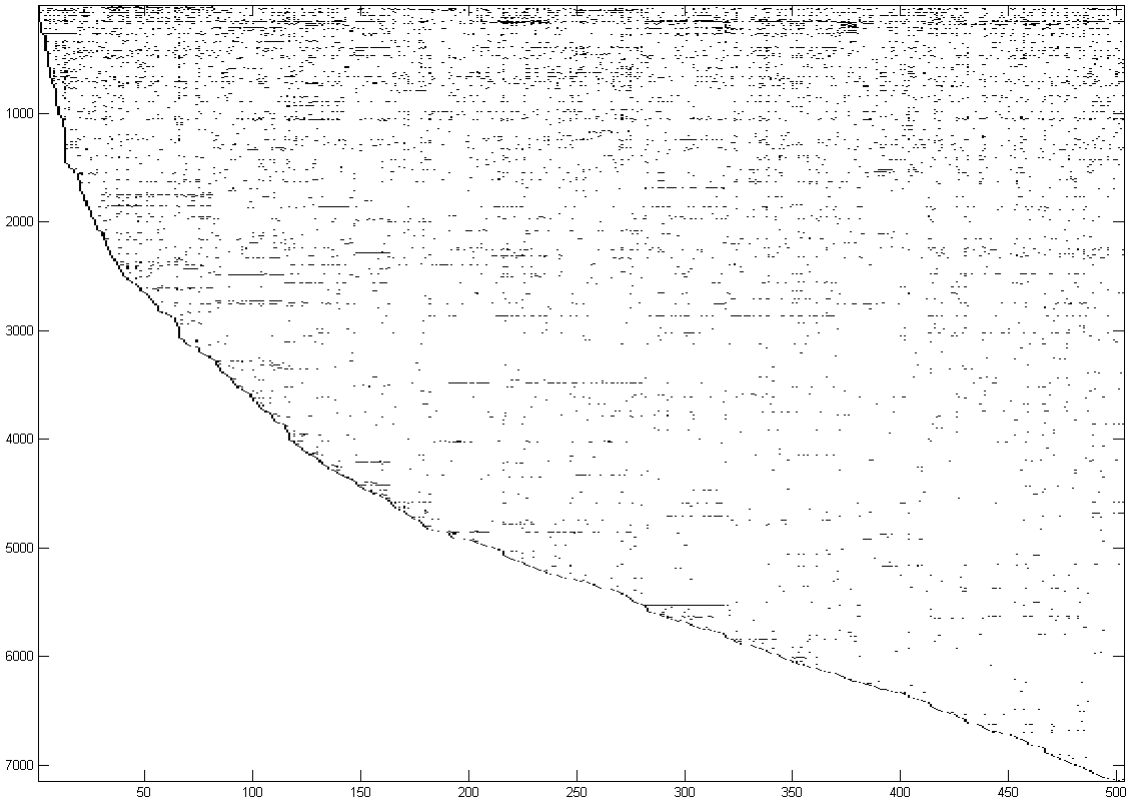


Figure 5.84: Term-Document binary matrix.

Figure 5.85 shows the binary co-occurrence matrix of the first 500 terms only. Figure 5.86

depicts the first 255 terms of the term-document bipartite network. The graph shows a cluster of highly ranked terms linked to many documents at the core of the network, whereas low ranking terms reside at the periphery of the network. The graph in Figure 5.87 focuses only edges having weight eight and up of the first 255 terms (the core). Low frequency terms at the periphery are now isolated.

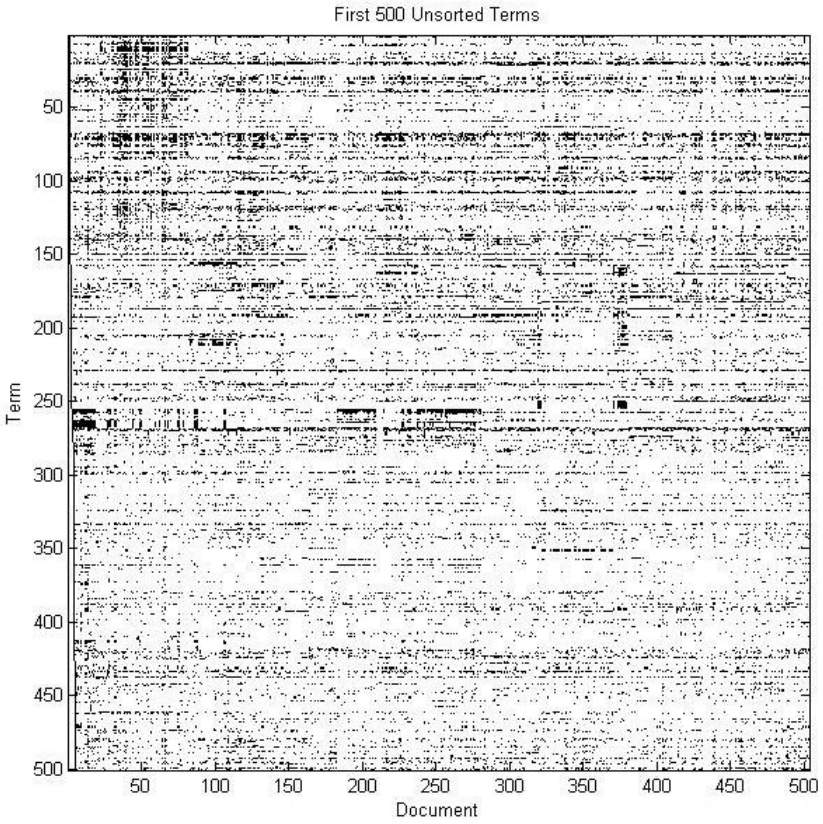


Figure 5.85: Term-Document binary matrix of top 500 unsorted terms.

The binary adjacency matrix of the text document application does not show the whole picture; it suggests whether a certain document contains a certain term. Nevertheless, because terms may be present more than once in a document, it is worthy to have record of term frequencies as well, which makes the matrix a weighted two-mode matrix. Figure 5.88

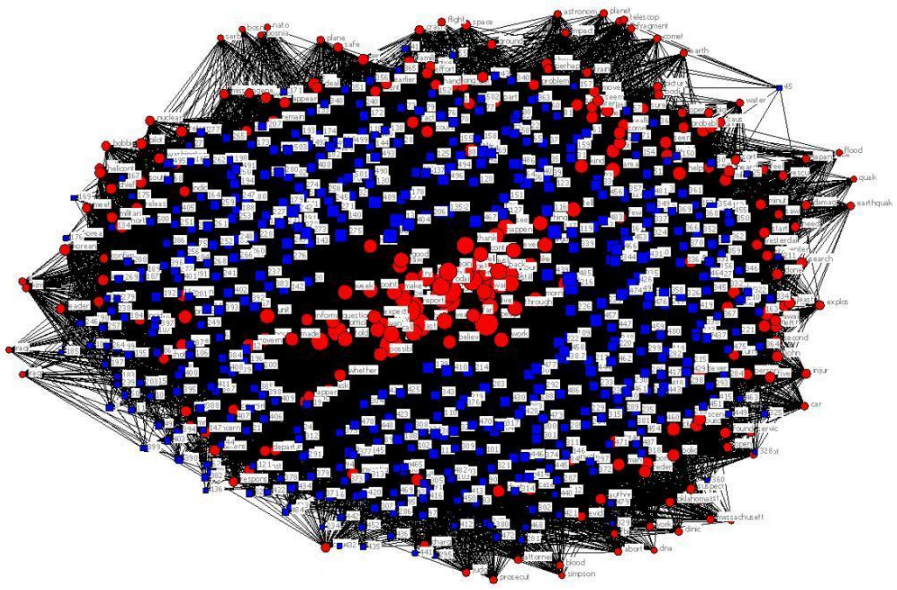


Figure 5.86: Term-Document network of top 255 terms.

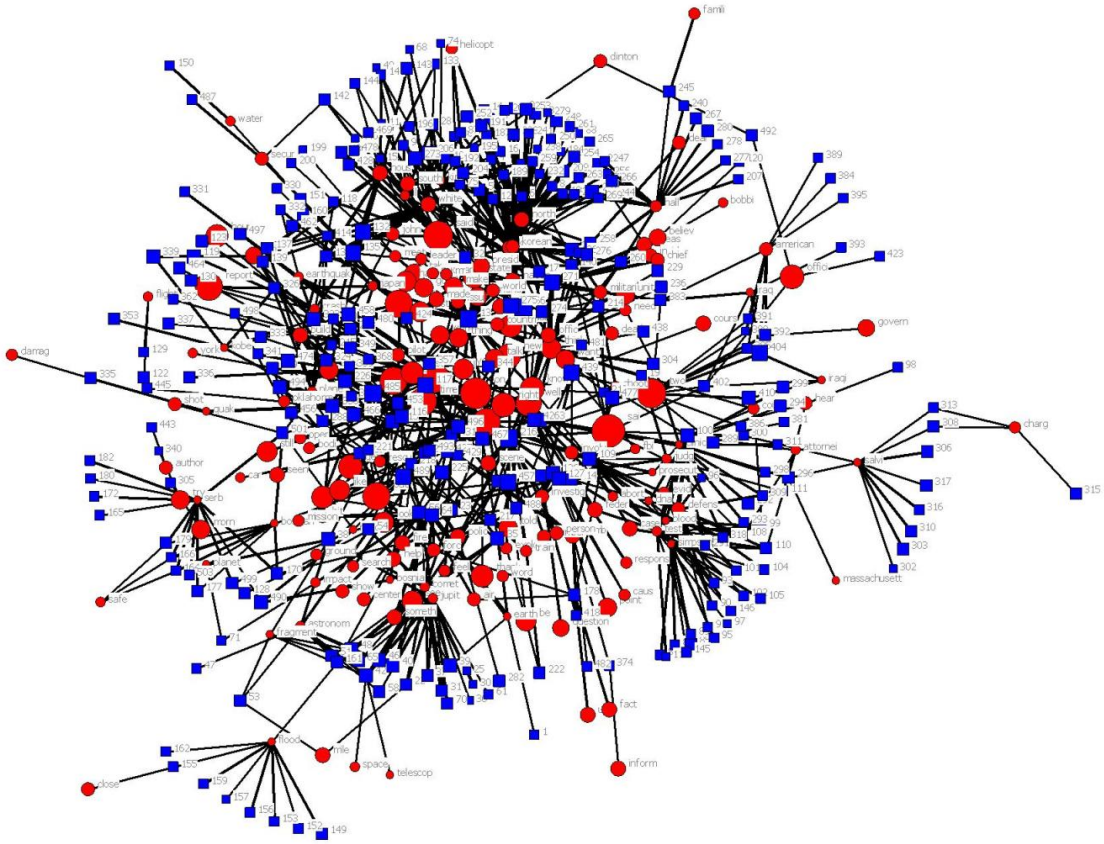


Figure 5.87: Showing strong ties of term-document network of the first 255 terms.

shows the top 300 highly ranked sorted terms of the bipartite weighted matrix ${}^{w}TD_{300\times503}$. We observe both vertical and horizontal patterns, but more of horizontal. Vertical patterns indicate that certain documents contain several highly ranked terms. Conversely, horizontal patterns reveal that certain popular terms show in multiple sequential documents. The patterns are made more visible by taking the natural logarithm of tie-strength.

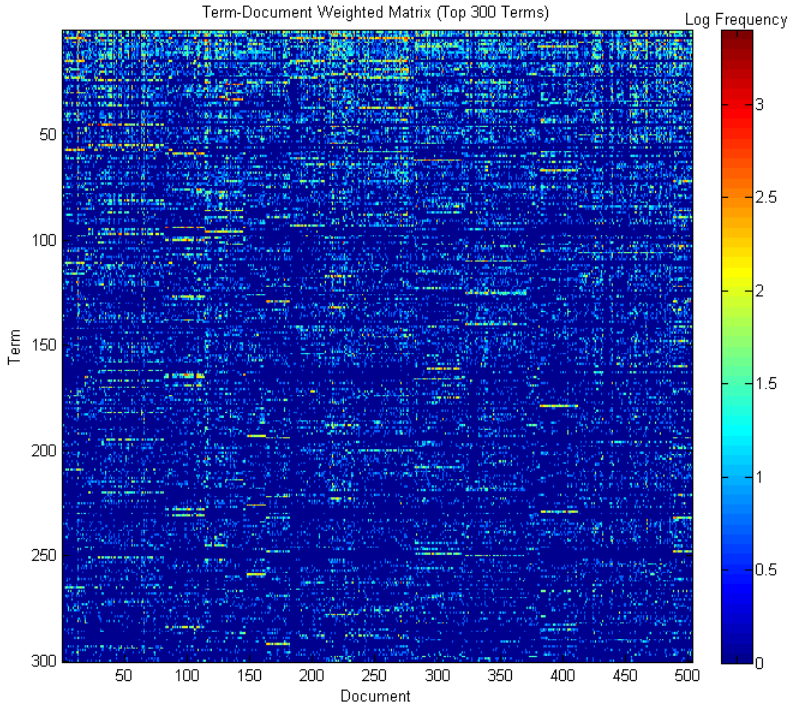
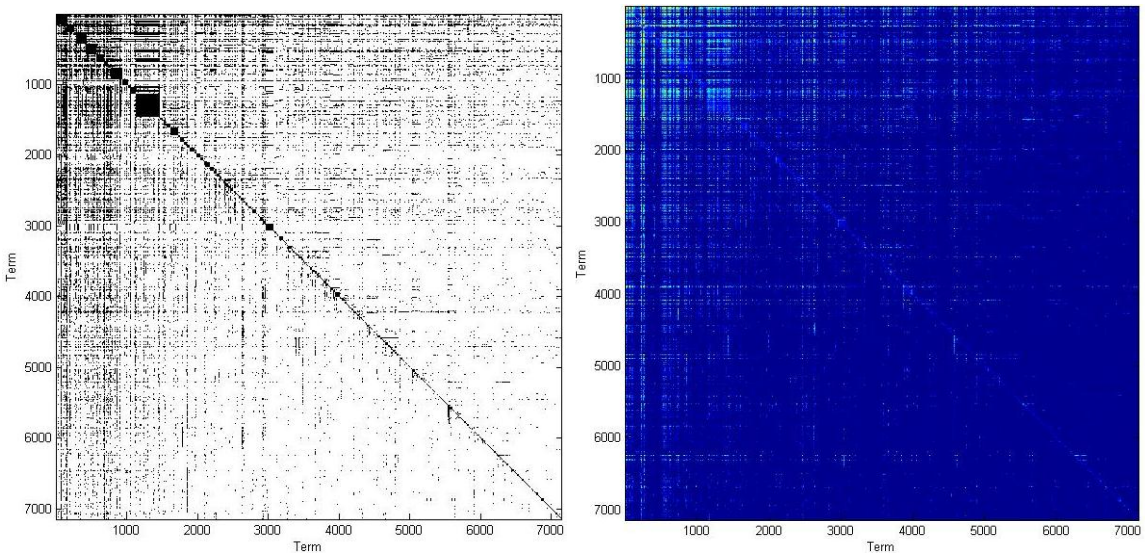


Figure 5.88: Term-Document weighted matrix of log frequency of top 300 sorted terms.

Continuing with the same spirit, I now look into how terms relate to other terms through documents and how documents related to other documents through terms. The graph in Figure 5.89 displays the one-mode term-by-term co-occurrence matrix ${}^{w}TT_{7143\times7143}$. The binary matrix ${}^{b}TT_{7143\times7143}$ in Figure 5.89(a) shows a cluster (block) of highly correlated terms, while the weighted matrix in Figure 5.89(b) exhibits a scatter of highly correlated terms located at the top left of the graph. Horizontal and vertical patterns suggest that the terms are connected to several other terms in the corpus.

Handling matrices with huge size is computationally problematic, any operation requires both storage and memory. Therefore, I reduced the number of terms to 2086 instead of 7143 after having them sorted in descending order. The first 2086 sorted terms have total row count greater than 10 appearances in a single document or multiple documents. The matrix in Figure 5.90 shows the log of the term-by-term weighted proximity matrix of terms having frequency ≥ 10, so that weak ties are discarded; and it is derived from the weighted two-mode term-document adjacency matrix ${}^{w}TD_{2086\times503}$.



(a) Binary ties. (b) Log weighted ties.

Figure 5.89: Term-Term structure matrix.

Figure 5.91 presents the one-mode document-by-document co-relationship matrix ${}^{w}DD_{503\times503}$ of documents related through the top 2086 sorted terms of the two-mode term-by-document weighted matrix. The top 2086 sorted terms are terms with total tie strength of ten or more. The plots show clusters of highly correlated documents through terms; the relation is based on the sum of the number of common terms a document and co-document share. Once

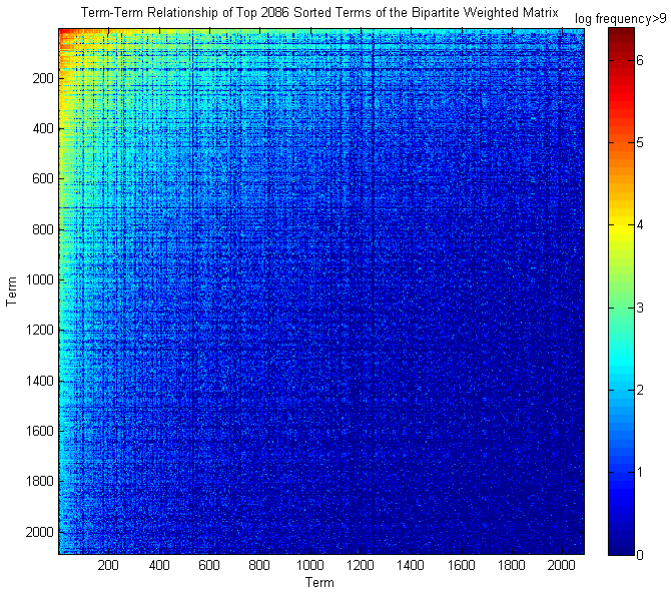
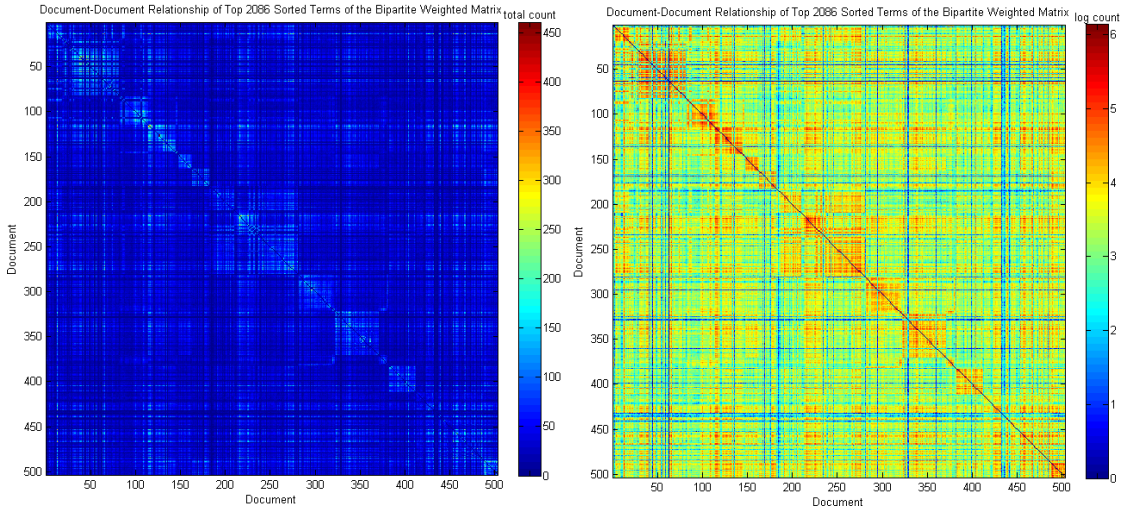


Figure 5.90: Terms related through documents weighted matrix.

again, horizontal and vertical patterns are present. In Figure 5.91(b), bluish patterns indicate that the document(s) are not highly co-related with other documents and yellowish patterns show the averagely related documents. Dark red squares point to the highly related documents.

If two or more documents have many terms in common then these documents are highly similar. The plot in Figure 5.92 portrays the document-document similarity matrix ${}^{w}DD_{503\times503}$ of the top 2086 sorted terms with tie strength of ten or more. Several blocks of averagely similar documents are visible along the main diagonal. The documents are listed in chronological order, this means that whenever there is a block there is a cluster of similar documents based on the vectors of terms associated with the documents. These clusters of documents more or less address the same issue. Another interesting observation concerns the block on the top left corner of the graph, this cluster of documents demonstrates patterns of similarity with other sets of documents in the corpus.



(a) Weighted ties.

(b) Log weighted ties.

Figure 5.91: Documents related through the top 2086 sorted terms.

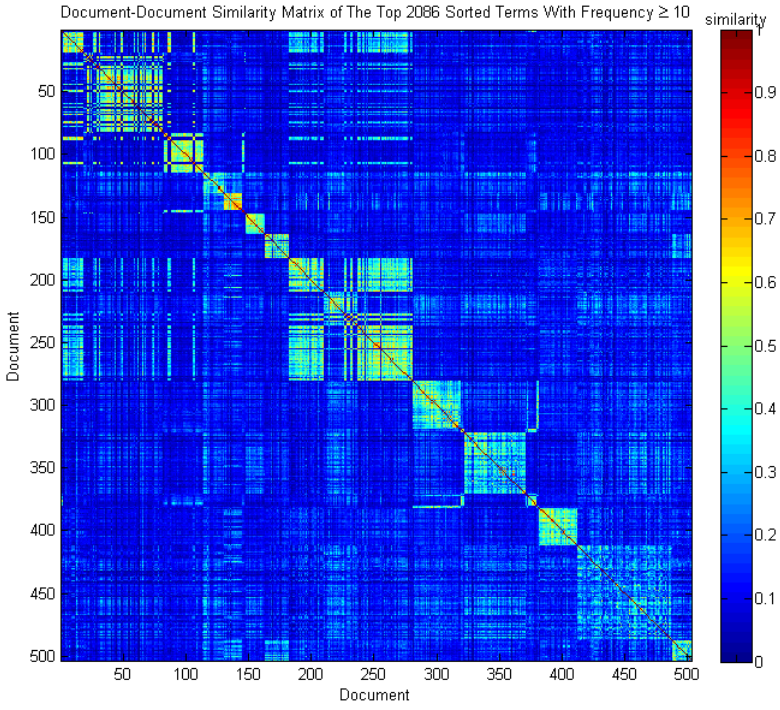


Figure 5.92: Document-Document similarity matrix with respect to the top 2086 sorted terms.

In an attempt to compare the document-document matrix resulted from the inner product similarity computations with the matrix resulted from the weighted matrix multiplication method I constructed the residual matrix corresponding to both matrices, the result is shown in Figure 5.93. The residual matrix excluded self-ties; i.e. entries of documents along the main diagonal, and gave a sum squares difference of 2849.1. The document-document co-relation matrix was normalized, so that the values be in the range $[0, 1]$ to match the values of the similarity matrix. The pattern that appears here is the same pattern obtained by both the inner product similarity and the matrix multiplication methods; however, the errors are higher along the main diagonal and among the clusters than the off-diagonal document-document association.

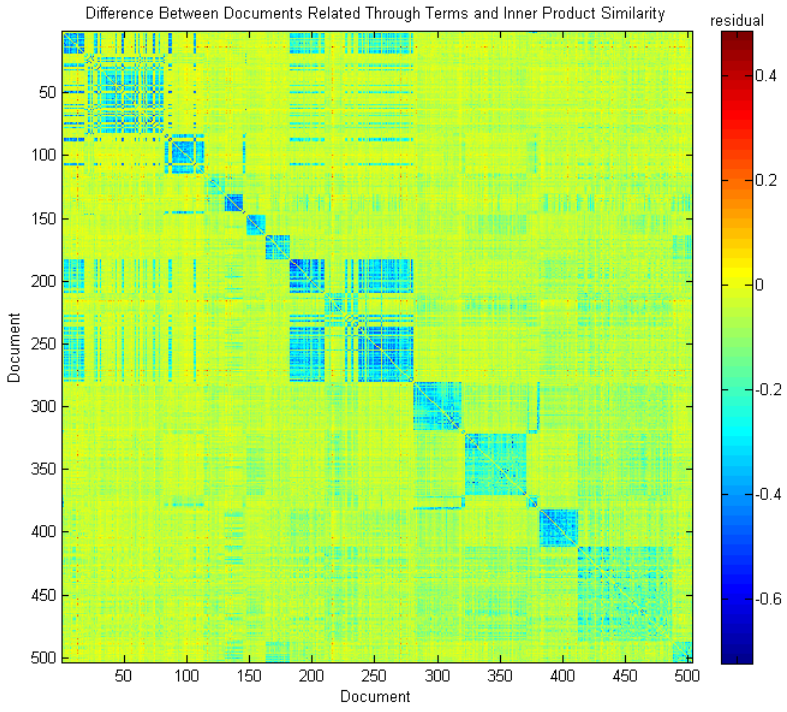


Figure 5.93: Document-Document residual matrix with respect to terms.

In the next stage, I incorporate an improved methodology for analyzing text within documents. The method integrates bigrams in the process as opposed to terms. A bigram contains two successive terms as one entity, which is a stricter criterion for forming multiple edges within documents or among documents. It is more likely for a term to appear in multiple documents than a bigram; therefore, documents related through bigrams and documents similarity with respect to bigrams are robust, but computationally expensive.

The bipartite relationship is called the bigram-document network, it is shown in Figure 5.94. There are 91709 unique bigrams contained in the 503 documents; thus, the two-mode bigram-by-document matrix is ${}^{w}BD_{91709\times503}$. Figure 5.95 shows the bar-plot for the top 60 bigrams, these are the bigrams with total bigram-document frequency of 15 or above. Among the top appearing bigrams in documents are “white-hous”, “north-korea”, “kim-il”, “il-sung”, “north-korean”, “kim-jong”, “jong-il”, “shoemak-levi”, “lo-angel”, “mile-brien”, “unit-state”, “space-telescop”, “hubbl-space”, “comet-fragment”.

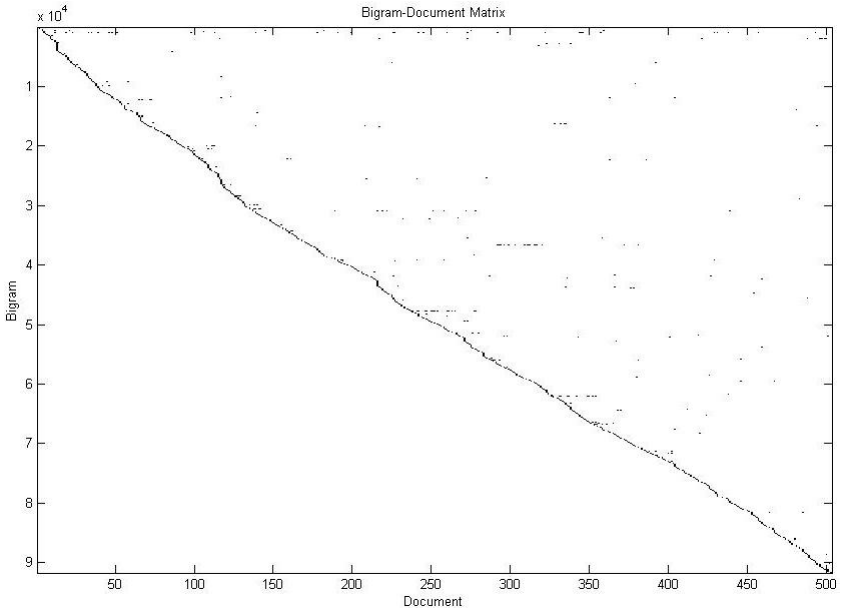
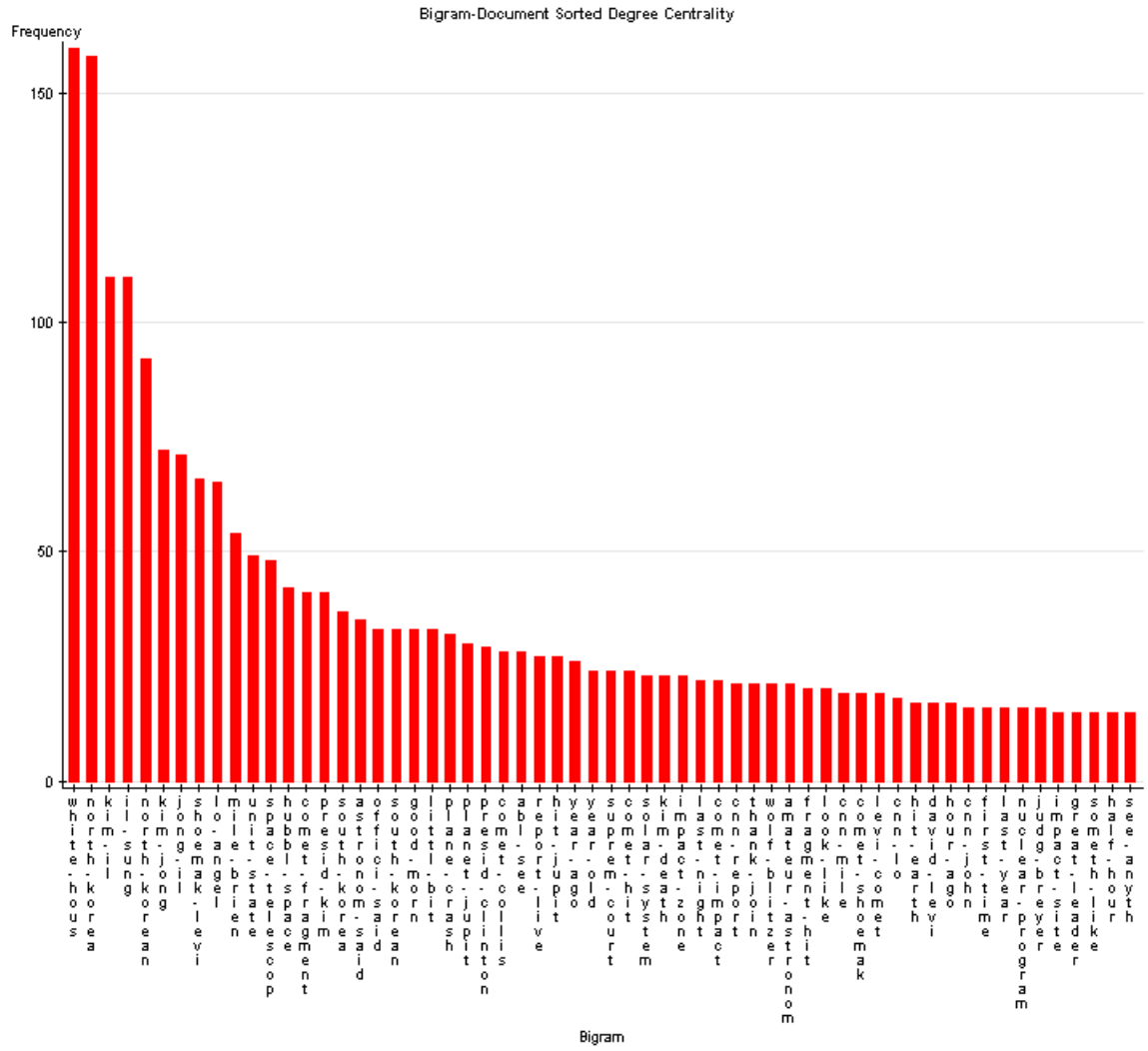


Figure 5.94: Bigram-Document binary structure matrix.



Bigram

Figure 5.95: Top 60 bigrams.

Because the bigrams related through documents matrix ${}^{w}BB_{91709\times91709}$ requires both large storage and memory to process, it was computational unfeasible to obtain and visualize in MatLab. Consequently, I decided to consider only 50 documents instead, which produced about 9000 bigrams. The plot in Figure 5.96 is the bigram-bigram network of the first 50 documents. The graph shows a giant block of highly co-related bigrams and several smaller blocks of co-related bigrams along the main diagonal. Interestingly enough, the bigram-bigram relation is not symmetric meaning bigram A is connected to bigram B does not necessarily imply bigram B is connected to bigram A.

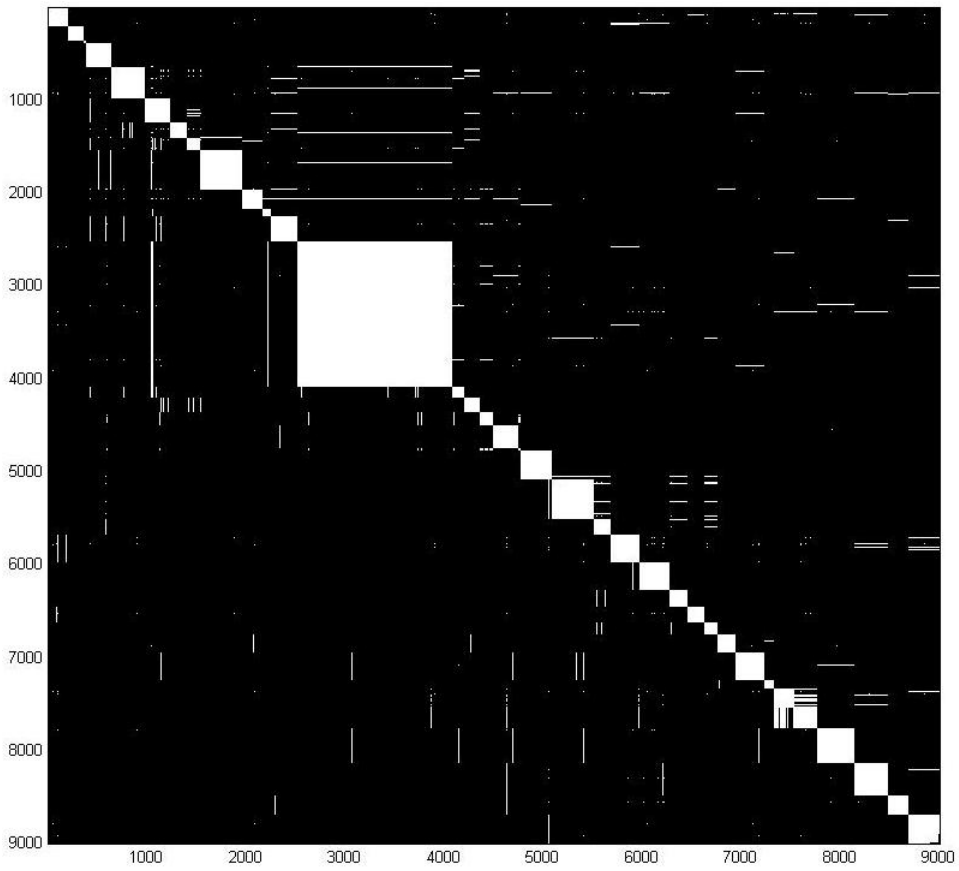


Figure 5.96: Bigram-Bigram matrix for 50 documents.

Alternatively, to tackle the issue of having a huge matrix size, I undertook a process of sorting the bigram-document matrix in a descending order with respect to row total and then discarding weak ties or bigrams having row total less than one. To break up the matrix even further, I took the first 252 bigrams of the bigram-document matrix for the 503 documents, the result is displayed in Figure 5.97; isolated bigrams with total tie strength < 5 and documents associated with those bigrams are deleted. The graph shows two main clusters of documents and bigrams. Figure 5.98 is the one-mode representation of bigrams related through documents. Once again, we observe two main subgroups of bigrams, which are shown separately after sorting all bigrams with respect to total tie strength in Figures 5.99 and 5.100.

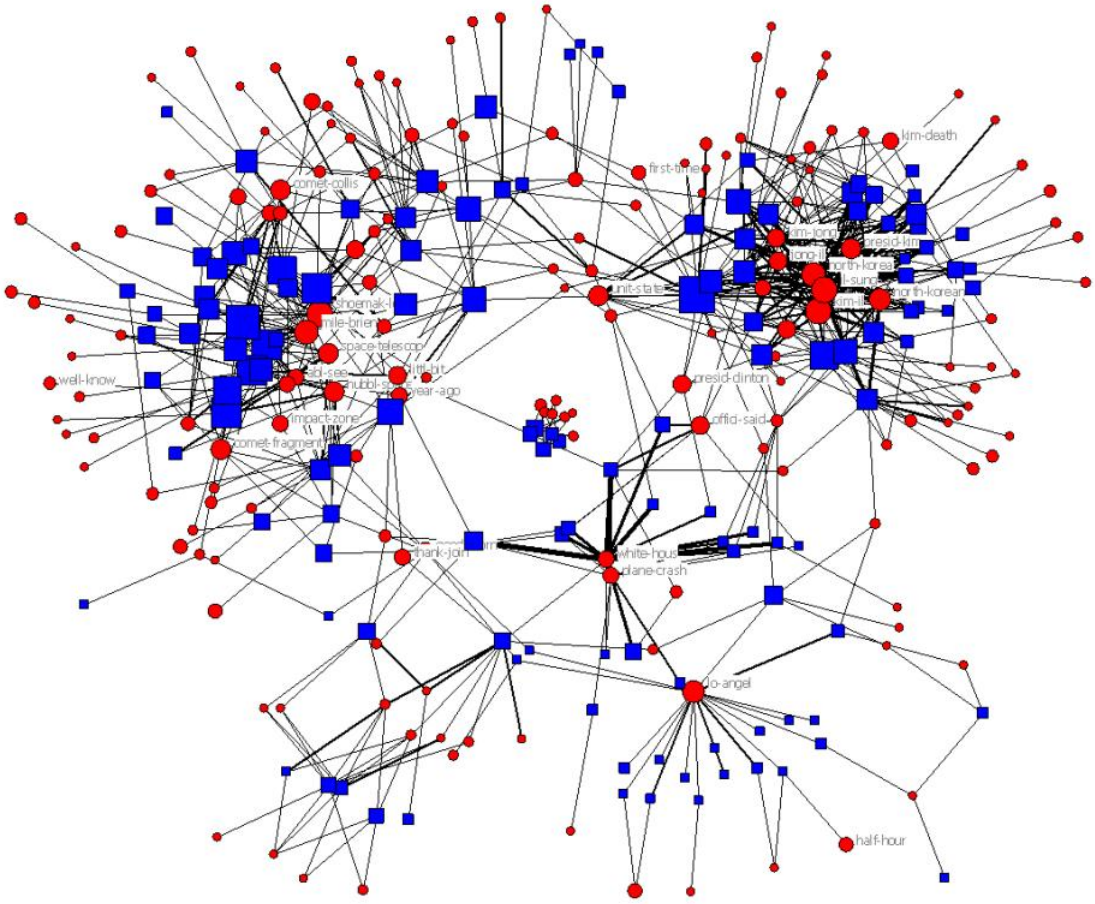


Figure 5.97: Bigram-Document two-mode network of top 252 bigrams.

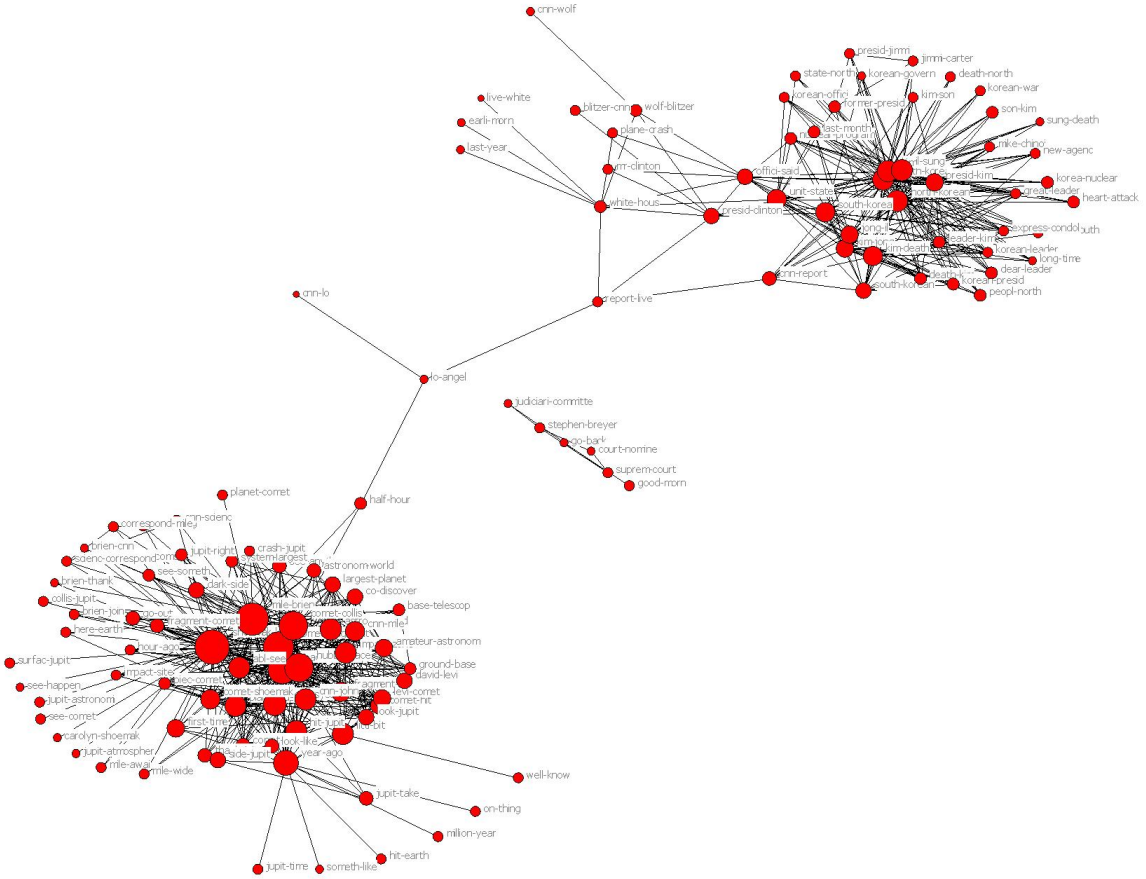


Figure 5.98: Bigram-Bigram network of top 252 bigrams having frequency ≥ 6.

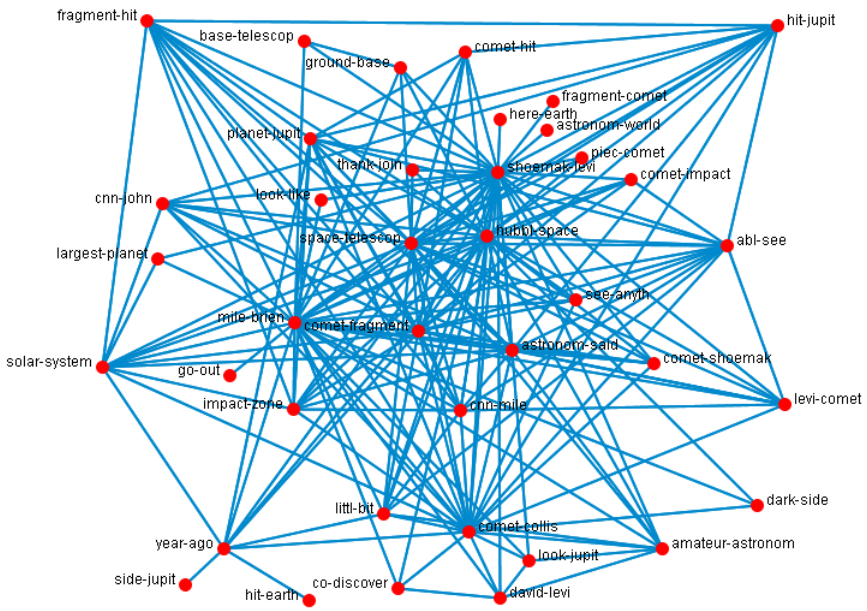


Figure 5.99: Bigram-Bigram subnetwork of astronomical bigrams.

Figures 5.101 and 5.102 display the bigram-bigram subnetwork of bigrams 253 through 503 and having total tie strength four and five only. Besides several blocks along the main diagonal, a pattern of related bigrams off the main diagonal is clear.

The plots in Figure 5.103 depict the structure matrices ${}^{w}BB_{1950\times1950}$ of the one-mode bigram-bigram related through documents network; Figure 5.103(a) shows binary ties, while Figure 5.103(b) presents the natural logarithm of tie strength plus one. Bigrams located at the top left corner are highly correlated, these bigrams show up in multiple documents and thus are connected. There is also a clear pattern of related bigrams off the main diagonal. Figure 5.104 display the first 300 bigram-bigram sub-matrix of the 1950 bigram-bigram structure matrix.

The top 1950 bigram-bigram similarity matrix is presented in Figure 5.105, to some extent it provides the same structural pattern as the top 1950 bigram-bigram co-related matrix;

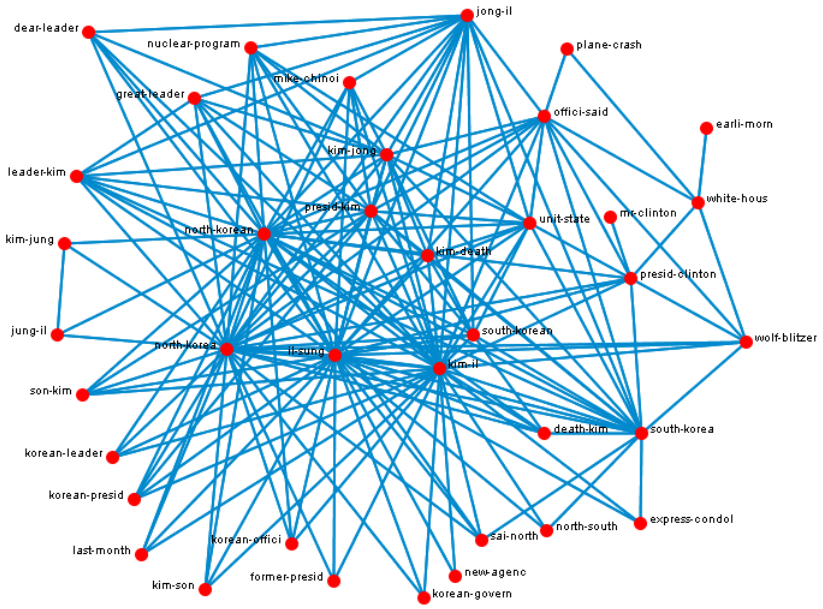


Figure 5.100: Bigram-Bigram subnetwork of political bigrams.

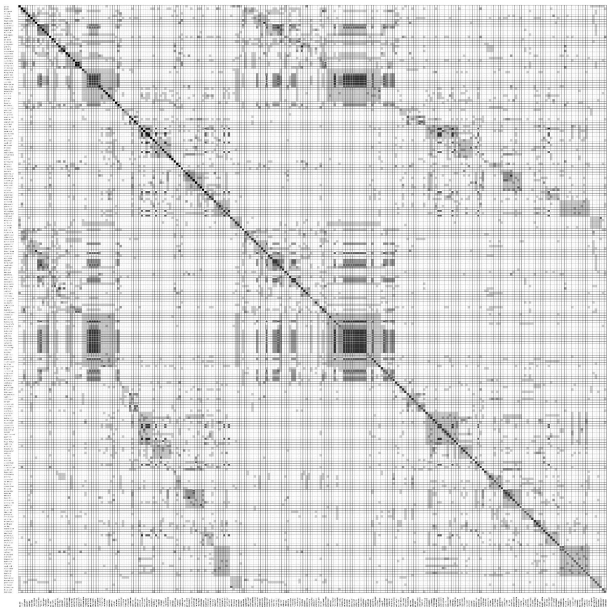


Figure 5.101: Bigram-Bigram matrix of bigrams 253 through 503.

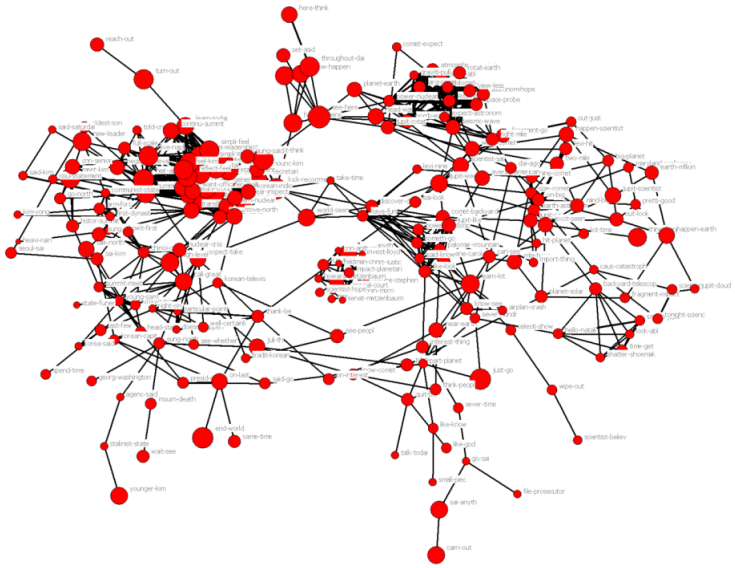
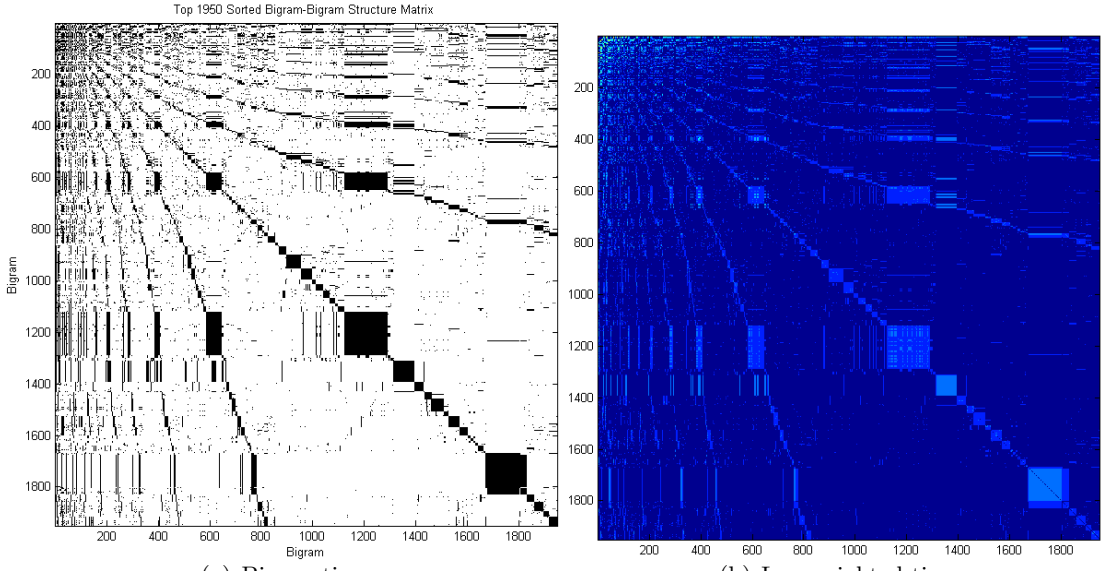


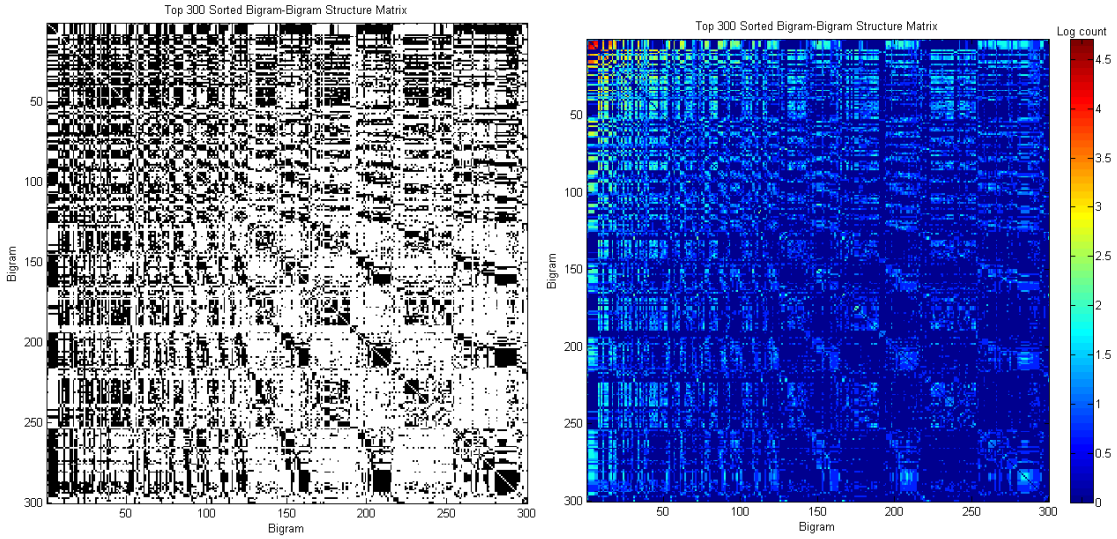
Figure 5.102: Bigram-Bigram network of bigrams 253 through 503.



(a) Binary ties.

(b) Log weighted ties.

Figure 5.103: Sorted Bigrams related through documents.



(a) Binary ties.

(b) Log weighted ties.

Figure 5.104: Sorted Bigrams related through documents for top 300 bigrams.

however, weights are slightly different specially clusters along the main diagonal. The two plots in Figure 5.103 and Figure 5.105 suggest two main clusters and several minor clusters of highly similar highly co-related bigrams. The two big clusters are the same clusters presented in Figures 5.99 and 5.100.

Continuing with the same ardor, I now turn to the document-document related through bigrams and document-document bigram similarity matrices and networks. I start by exploring the documents related through bigrams network, the graph is shown in Figure 5.106, which corresponds to the one-mode matrix ${}^{w}DD_{503\times503}$. There appears to be clusters of highly co-related documents.

After extensive computation in MatLab the bigram-document matrix ${}^{w}BD_{91709\times503}$ produced the documents related through bigrams proximity matrix ${}^{w}DD_{503\times503}$, the result is portrayed in Figure 5.107. To make the matrix structure more visible I applied a natural logarithm transformation. The documents related through bigrams graph density is 0.487.

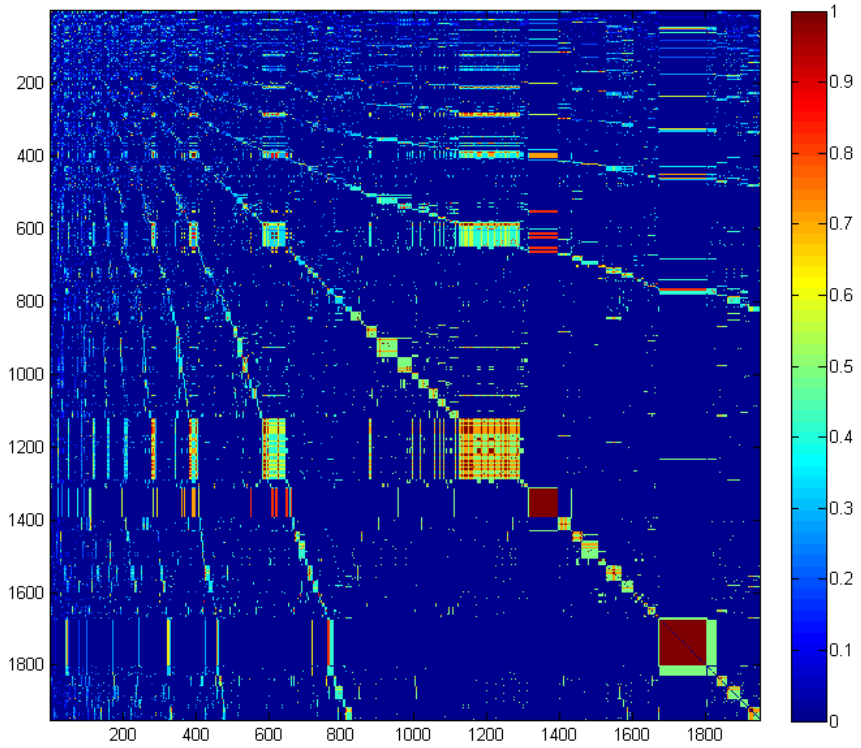


Figure 5.105: Sorted Bigrams related through documents similarity matrix.

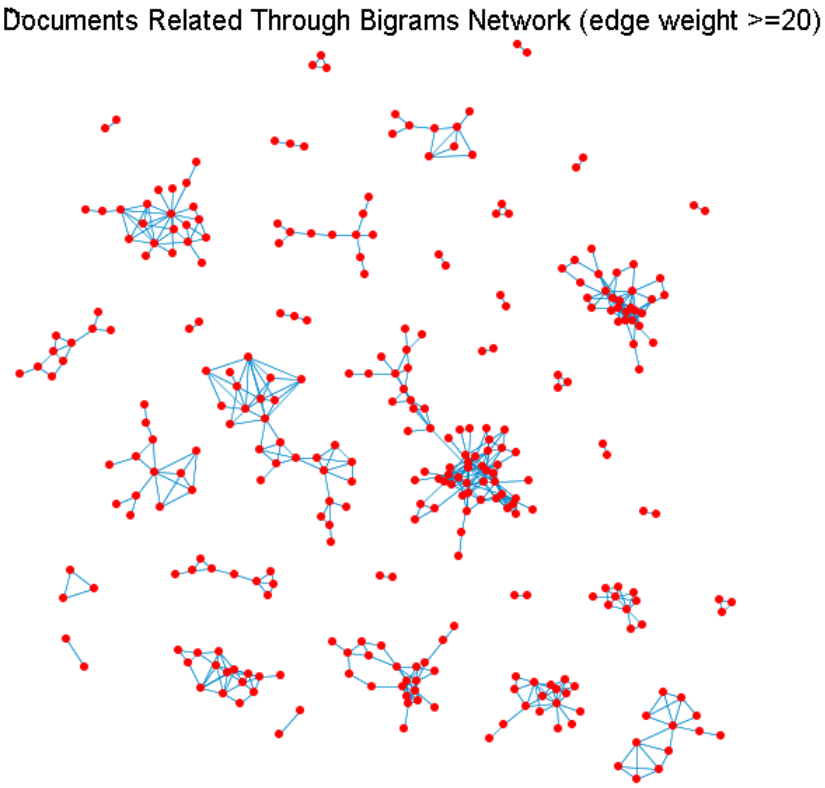
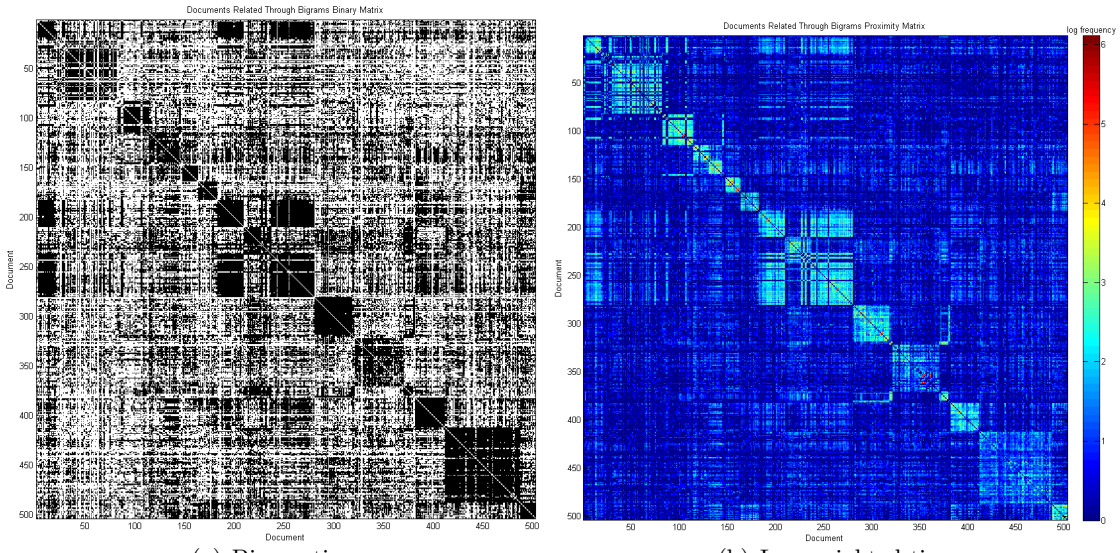


Figure 5.106: Documents related through bigrams network.

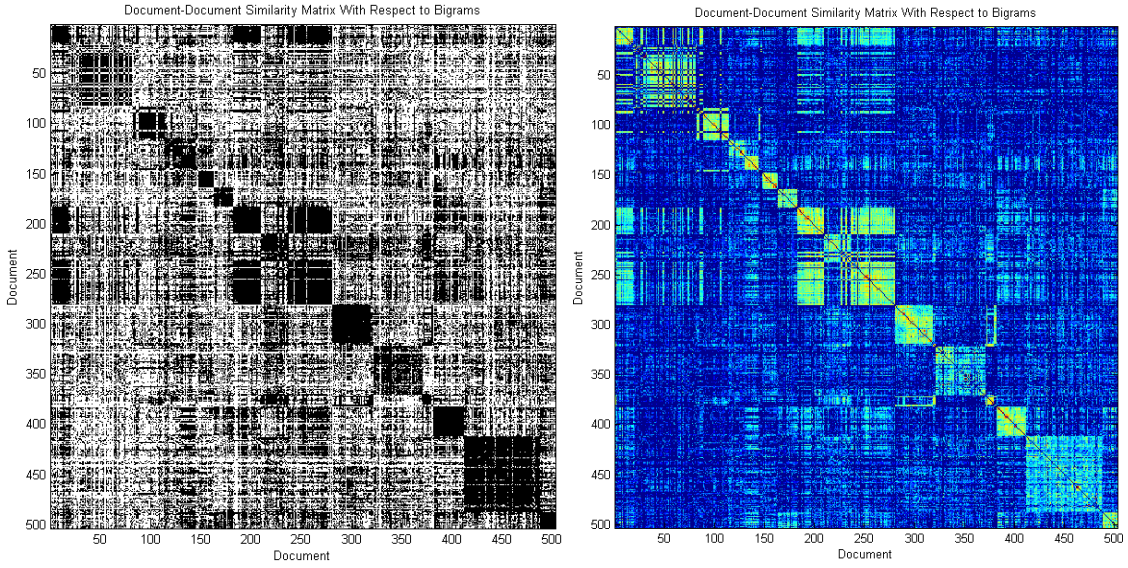


(a) Binary ties.

(b) Log weighted ties.

Figure 5.107: Documents related through bigrams matrix.

The document-document similarity matrix ${}^{w}DD_{503\times503}$ with respect to bigrams is shown in Figure 5.108. To some extent the similarity matrix depicts the documents related through bigrams structure matrix, which is confirmed by the document-document residual matrix with respect to bigrams presented in Figure 5.109.



(a) Binary ties. (b) Log weighted ties.

Figure 5.108: Documents similarity matrix with respect to bigrams.

In an attempt to compare documents related through terms matrix with documents related through bigrams matrix and documents similarity matrix with respect to terms with documents similarity matrix with respect to bigrams I constructed the document-document residual matrices for terms and bigrams respectively, the results are shown in Figure 5.110. The documents related through terms and bigrams difference matrix has relatively small residuals except for a very few set of documents with high residuals. The documents similarity difference matrix has also relatively small residuals except for some blocks along the main diagonal with higher residuals.

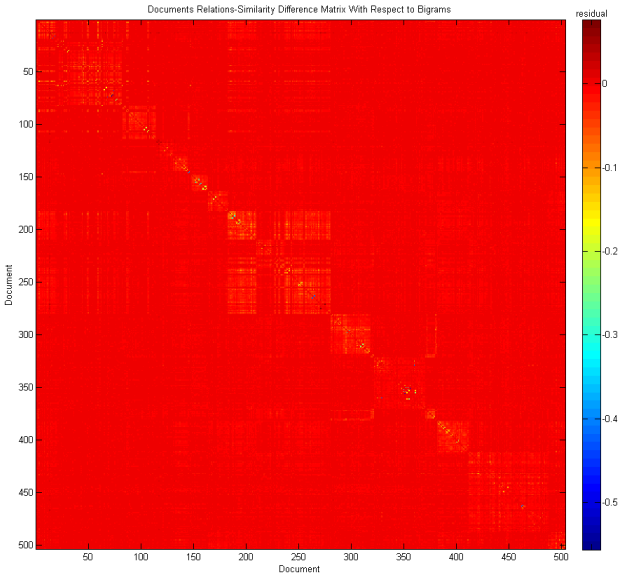
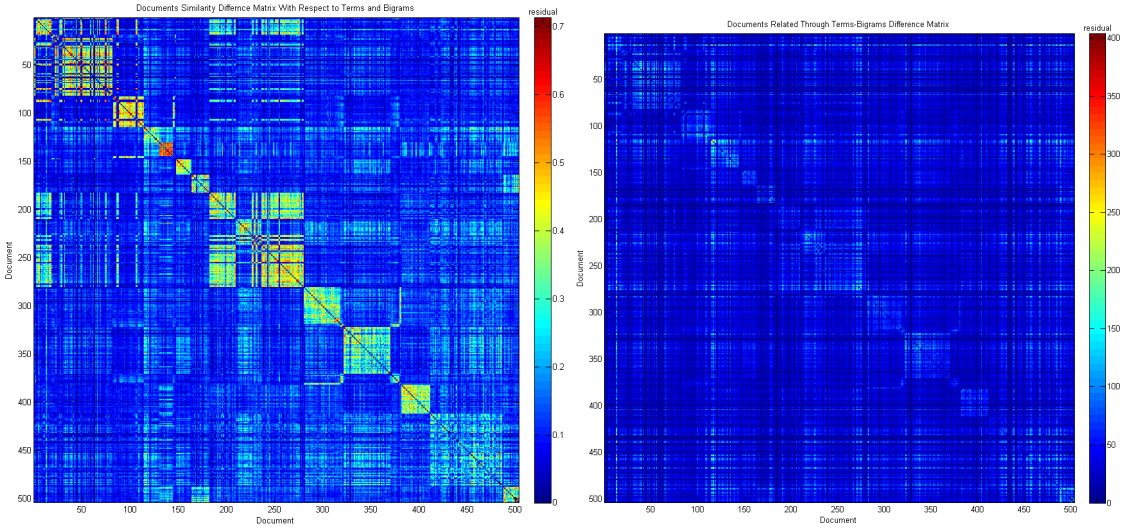


Figure 5.109: Document-Document residual matrix with respect to bigrams.



(a) Documents similarity difference matrix with respect to terms and bigrams.

(b) Documents related through terms-bigrams difference matrix

Figure 5.110: Term-Bigram comparison for document-document matrix.

5.6 Online Music Friendship Network

The following is a preferential attachment simulated network representing online friendship based on music tastes. The sample space consists of 8 music genres. An integer vector of music tastes of size 8 is randomly generated at each time step and assigned to each agent. The following rules are set to generate the network. Actors are allowed to choose at most 3 different music tastes out of the 8 available tastes (in some trials I set this to 2 music tastes). I used the MatLab built in function "And" to obtain a similarity measure between actors. If two actors share more than one music taste they have a chance of being friends (in some experiments I restricted this to two music tastes, which made it harder on agents to attach, see Figure 5.111). In addition, new actors are allowed to be friends – attach – with only one similar agent, (in some trials the restriction was relaxed to at most 2, 3, 4 or 5 friends). This relaxation of the number of friends an agent can be of generated different interesting network structures.

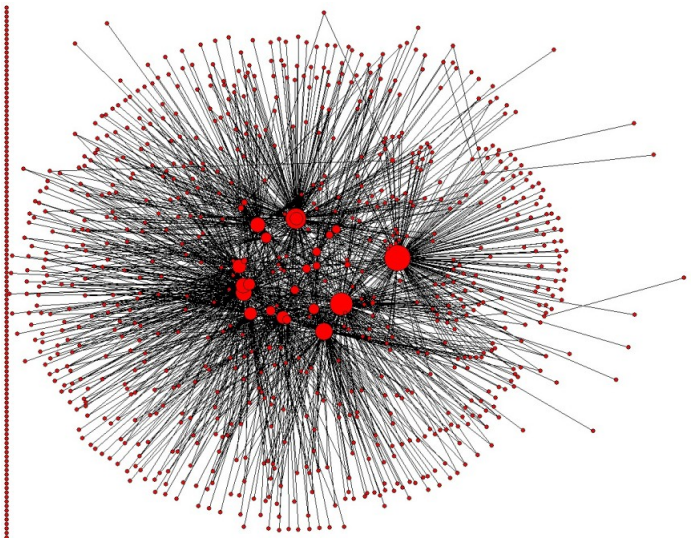


Figure 5.111: Generated music friendship network based on 2 tastes, 3 attachments, 2 matches.

When the criterion to make friendship is set to "attach only to one similar friend" the

generated network is a "star-like" network with few highly central agents in the middle representing the "core" of the network, see Figure 5.112. This structure converged after 100 iterations, it resembles a "scale-free" (self-similar) network; the network maintains shape and properties regardless to the number of actors and independent of the network size. The topology of the net after 1000 iterations is similar to 100 iterations. These types of networks cause stress on the "Ego" vertex resulted from the high cognitive load. With so many friends attached to "Ego" - the core, "Ego" struggles to keep and maintain friendships. On the other hand, there is less stress and maintenance required from actors on the periphery.

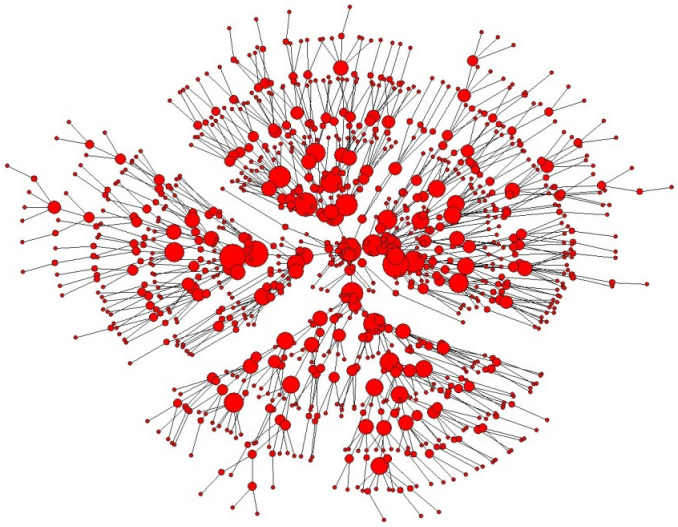
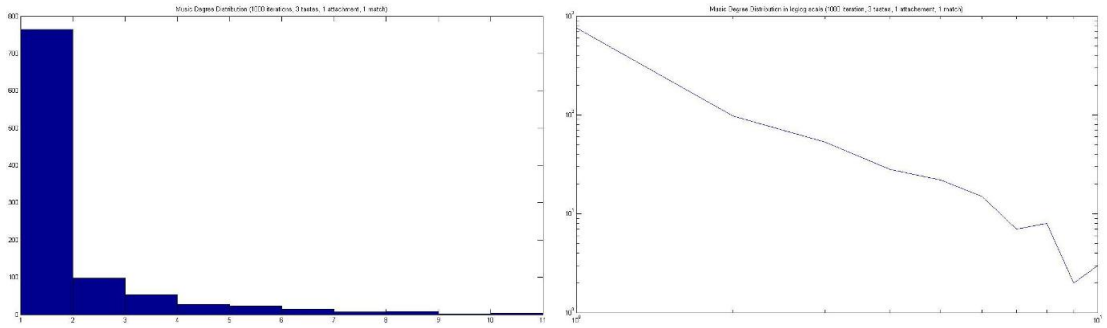


Figure 5.112: Simulated music friendship network based on 3 tastes, 1 attachment, 1 match.

Figure 5.113 shows the distribution of degree centrality, it is somewhat a perfect power-law with almost no noise reflecting the "Many-Some-Few" pattern. Actors generated at the early stage (first few iterations) attain the highest degree centrality at the core, while actors introduced at later steps form the periphery.

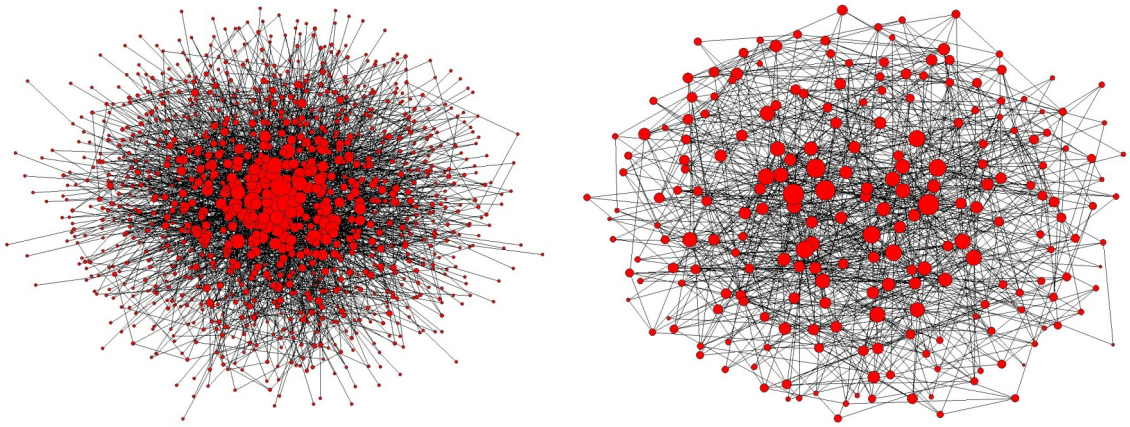
When the criterion of making friendship is relaxed to two or more friends at a time, the network unfolds unto itself after 1000 iterations generating a "hair-ball" very dense network with a possibility of having multiple cores (central agents) at the center with high degree



(a) Histogram. (b) Loglog scale.

Figure 5.113: Distribution of degree based on 3 tastes, 1 attachments, 1 match.

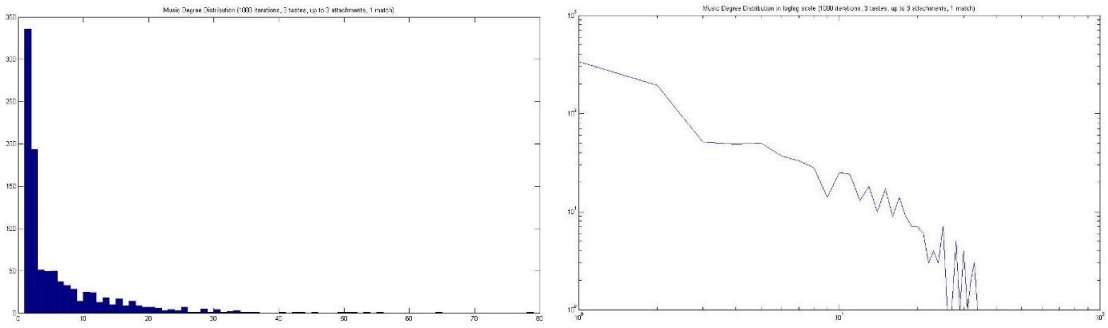
centrality. The degree distribution, however, is still power-law but has noise at the tail because the number of friends new actors can have is drawn from a uniform distribution with at most 3 friends in one experiment see Figure 5.115, and 4 friends in another.



(a) Dense network. (b) Zooming in the network.

Figure 5.114: Simulated network based on 3 tastes, up to 3 attachments, 1 match.

Multiple components are also present due to the fact that the criterion to attach and be friend to someone is strict; thus, some actors formed separate groups (multi-group network) that carry the same properties of the single-component highly centralized network, see Figures 5.116 and 5.117.



(a) Histogram. (b) Loglog scale.

Figure 5.115: Distribution of degree based on 3 tastes, up to 3 attachments, 1 match.

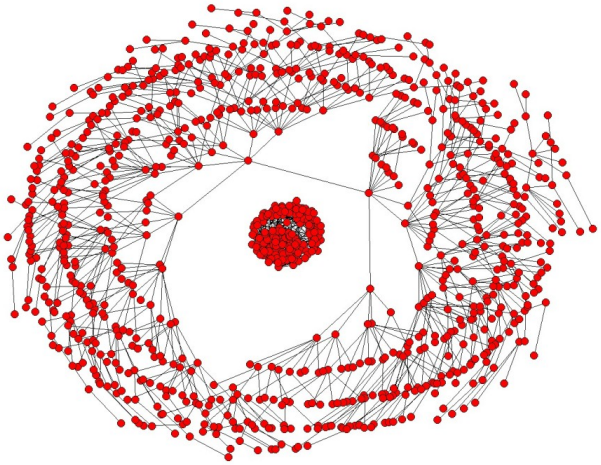
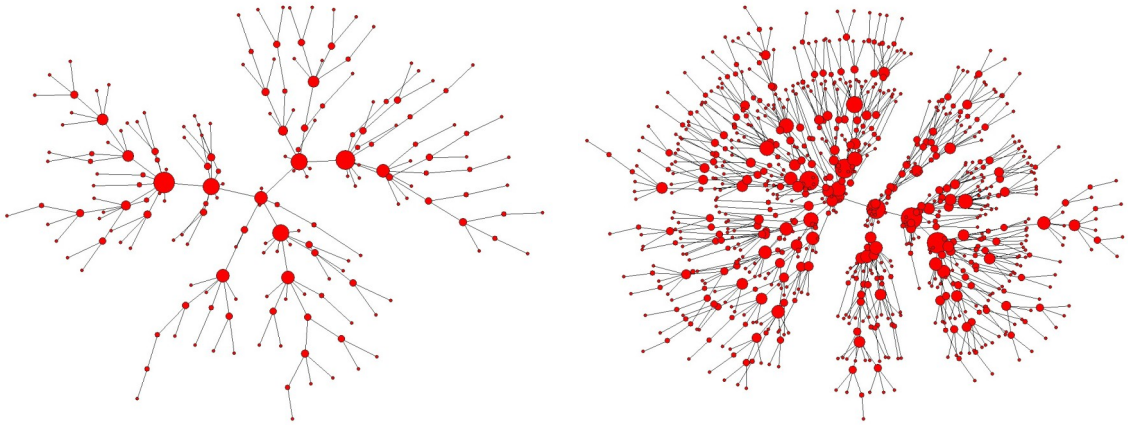


Figure 5.116: Music friendship network with two components based on 3 tastes, 1 attachments, 1 match.



(a) Component 1. (b) Component 2.

Figure 5.117: Two components based on 3 tastes, 1 attachments, 1 match.

To conclude, relaxing the criterion of attachment and the number of music tastes and the number of similar tastes an actor results in more attachments (ties), which generates dense network. In contrast restricting the rules of attachment makes it harder on agents to find mutual friends. It is possible though that if the rules are very strict, some actors may not find friends to attach to at and hence being isolated.

Chapter 6: Conclusions, Contributions and Future Work

6.1 Conclusions

In this dissertation, I presented a methodology to analyze networks – social and other types. To begin with, I have developed the mathematics underpinning networks, which involved the integration of several branches of mathematics such as matrix theory, graph theory, estimation, geometry and fuzzy logic. The approach addressed two types of networks, namely, stationary networks and evolutionary networks. The study focused on how these fields of mathematics can be utilized to address network issues. Then I have implemented the theory on real networks of different levels of interactivity. Finally, I have simulated two social networks based on the preferential attachment model.

One of the major research questions I addressed in this study is the fact that in evolving dynamic networks vertices and edges may be introduced at any time and thus the network size and order constantly change. To tackle this problem, I have invented a methodology that expands networks into infinite networks and matrices into infinitely dimensional matrices in which vertices and edges may be introduced at any given time without having to worry about the network and matrix dimensionality. I then addressed the issue of manipulating multi-mode networks to gain information and knowledge about networks on the different levels and modes. Followed by a considerable work that relates to the interchangeability and duality between vertices and edges in a graph in which vertices convert to edges and edges convert to vertices to estimate the probability of missing dyadic edges or to estimate the probability missing vertices using covariate information associated with vertices and edges.

I have examined and studied several network applications such as coauthorship social networks, road fatalities networks and news documents networks. In coauthorship social networks, I identified special groups of coauthors that are high in degree and tie strength in which I called elite group. The road fatalities network demonstrated how states are related through crash factors and how they are similar with respect to crash factors. And finally, in the news documents example I performed an assessment of the documents network derived from the term-document and bigram-document networks.

I believe this dissertation offers a valuable tool for analysts and credible literature for researchers. The theory and implementation serve as a solid foundation for further exploration and expansion of network theory.

6.2 Contributions

To briefly summarize my contribution in this dissertation, I worked out a mechanism to store and explore finite and infinite networks using primitive network blocks represented with sub-matrices stored in a global matrix. To this end, the tool expands on vertices by introducing a matrix of infinite dimension corresponding to an infinite network in which vertices are categorized as active or inactive. The infinite matrix offers a mechanism of observing the development of a network over time. The advantage of this matrix representation is the fact that performing matrix operations on such matrices is computationally cheap because the matrix is block-diagonal and the elements are sub-matrices whose entries are ones.

A substantial portion of the dissertation covered mathematical techniques that efficiently compute graph and network measures such as edge count, network diameter, graph density, degree centrality matrix, distance matrix. I provided detailed study to some special graphs and their properties and importance in network theory.

I undertook estimation of an edge and vertex in a graph using exogenous structure pertaining to the network, it is based on quantitative and qualitative covariate information and the inner product method and chi-square significance test. This part relates to the interchangeability and duality between vertices and edges in a graph. I have suggested a method that uses covariate information associated with vertices to estimate the probability of missing edges and covariate information associated with edges to estimate the probability of missing vertices. In order to predict missing vertices, I have utilized the line graph transformation to convert edges to vertices and vertices to edges and the problem now is to compute the probability of an edge in the line graph. I applied the inner product method on the vectors of covariates to estimate the probability of a dyadic edge. I have extended the methodology of predicting edges (dyadic ties) to predict edges in a triad (triadic edges). The method incorporates covariate information as well; however, the basis for this method is largely geometrical and through the use fuzzy logic rather than the inner product of two vectors.

Perhaps the most remarkable contribution of work in the field of network theory is the generalization of the $N-$mode networks and their implications and the manipulation of higher dimensional relational networks to extract information and gain knowledge about networks on the different lower dimensions and modes. To this end, I have developed an advanced approach to derive one-mode networks from weighted (valued) two-mode networks, then I extended the two-mode method to work for multi-mode networks. The algorithm expresses the weighted network as a combination of dichotomous (binary) networks that are used to obtain the one-mode weighted network.

I have ended the dissertation with two simulations of two evolutionary social networks demonstrating preferential attachment; the first model simulates the evolution of a coauthorship social network, while the second simulates the evolution of a online music friendship social network.

6.3 Future Work

Network theory is fairly a new topic although the early publication in the field dates back to the 1930s. Having articulated several components in the field of network theory, I believe there is more need to be done and consider this study a work in progress. I intend to extend this work in different directions. I plan on applying the theory of the generalized $N-$mode networks on real data. I plan to implement the theory presented on estimating missing vertices using both quantitative and categorical information pertaining to vertices on real social networks. Furthermore, I plan to deeply investigate the relationship between cliques and hypergraphs and develop mathematical models related to these special components of network theory. I also plan to work on better ways to visualize large scale networks. I need to improve the methods of analyzing evolutionary networks and try to incorporate advanced mathematical ideas for approaching such networks. I plan on researching networks both on macro and micro levels as dynamical systems. I plan to study the possibility of predicting the entire network's structure and linkages based on information associated with vertices with respect to a vector of global preference.

Appendix A: Degree and Closeness Centrality Measures Illustration

To illustrate the degree measure, consider the graph (network) in Figure 1.1. Table A.1 shows the results.

Table A.1: Degree Centrality Example.

Adjacency	Ego	A	B	C	D	E	F	G	Degree
Ego	0	1	1	1	1	1	1	1	7
A	1	0	0	1	0	0	0	0	2
B	1	0	0	0	0	0	0	0	1
C	1	1	0	0	0	0	0	0	2
D	1	0	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	0	1
F	1	0	0	0	0	0	0	0	1
G	1	0	0	0	0	0	0	0	1

The network in Figure 1.1 has the following closeness measures:

As an illustration, in Figure 1.1, the nodes {Ego, A, C} form a clique.

Table A.2: Closeness Centrality Example.

Distance	Ego	A	B	C	D	E	F	G	Sum	Closeness	Normalized
Ego	0	1	1	1	1	1	1	1	7	0.143	1.000
A	1	0	2	1	2	2	2	2	12	0.083	0.583
B	1	2	0	2	2	2	2	2	13	0.077	0.538
C	1	1	2	0	2	2	2	2	12	0.083	0.583
D	1	2	2	2	0	2	2	2	13	0.077	0.538
E	1	2	2	2	2	0	2	2	13	0.077	0.538
F	1	2	2	2	2	2	0	2	13	0.077	0.538
G	1	2	2	2	2	2	2	0	13	0.077	0.538

Bibliography

Bibliography

[1] R. Admiraal and M. Handcock, *Sequential importance sampling for bipartite graphs with applications to likelihood-based inference*, (2006).

[2] A. Barabasi and R. Albert, *Emergence of scaling in random networks*, *Science*, no. 5439 **286** (1999), 509–512, DOI: 10.1126/science.286.5439.509.

[3] S. Borgatti, *Centrality and network flow*, 2002.

[4] S. Borgatti, M. Everett, and L. Freeman, *Ucinet for windows: Software for social network analysis*, Analytic Technologies (2002).

[5] K. Borner, *Making sense of mankind's scholarly knowledge and expertise: collecting, interlinking, and organizing what we know and different approaches to mapping (network) science*, Environment and Planning B: Planning and Design (2007).

[6] K. Borner, L. Dallasta, W. Ke, and A. Vespignani, *Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams*, Indiana University (2005).

[7] K. Borner, J. Maru, and R. Goldstone, *The simultaneous evolution of author and paper networks*, PNAS **101** (2004), 52665273.

[8] K. Carley, *Smart agents and organizations of the future*, Carnegie Mellon University.

[9] B. Carlin, N. Polson, and D. Stoffer, *A monte carlo approach to nonnormal and nonlinear state-space modeling*, Journal of the American Statistical Association **87** (1992), 493–500.

[10] S. Chan, R. Pon, and A. Cardenas, *Visualization and clustering of author social networks*, (2006).

[11] C. Cioffi-Revilla, *Power laws in the social sciences: Discovering complexity and non-equilibrium dynamics in the social universe*, George Mason University, Fairfax VA, 2005.

[12] CIS, *Current index to statistics, http://www.statindex.org/.*

[13] DBLP, *The dblp computer science bibliography, http://www.informatik.unitrier.de/∼ley/db/.*

[14] W. de Nooy, V. Batageli, and A. Mrvar, *Exploratory social network analysis with pajek*, Cambridge University Press, 2004.

[15] I. Dhillon, *Co-clustering documents and words using bipartite spectral graph partitioning*, ACM (2001), 269–274.

[16] Matthew J. Dombroski and Kathleen M. Carley, *Netest: Estimating a terrorist network's structuregraduate student best paper award, casos 2002 conference*, Comput. Math. Organ. Theory **8** (2002), no. 3, 235–241.

[17] R. D'Souza, C. Borgs, J. Chayes, N. Berger, and R. Kleinberg, *Emergence of tempered preferential attachment from optimization*, (2007).

[18] J. Eaton, J. Ward, A. Kumar, and P. Reingen, *Structural analysis of coauthor relationships and author productivity in selected outlets for consumer behavior research.*

[19] S. Eubank, H. Guclu, V.S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, *Modelling disease outbreaks in realistic urban social networks*, Nature Publishing Group **429** (2004).

[20] M. Fafchamps, S. Goyal, and M. van der Leij, *Matching and network effects*, (2006).

[21] FARS, *Fatality analysis reporting system, http://www-fars.nhtsa.dot.gov/.*

[22] L. Fleming and M. Marx, *Managing creativity in small words*, (2006).

[23] Z. Ghahramani, T.L. Griffiths, and P. Sollich, *Bayesian nonparametric latent feature models*, *Bayesian Statistics 8.* Oxford University Press (2007).

[24] K. Gile and M. Handcock, *Model-based assessment of the impact of missing data on inference for networks*, (2006).

[25] S. Goyal, M. van der Leij, and J. Moraga-Gonzalez, *Economics: An emerging small world?*, (2004).

[26] M. Granovetter, *The strength of weak ties*, American Journal of Sociology **78** (1973), 1360–1380.

[27] M. Handcock and K. Gile, *Modeling social networks with sampled or missing data*, (2007).

[28] R. Hanneman and M. Riddle, *Introduction to social network methods*, Online textbook: http://www.faculty.ucr.edu/∼hanneman/nettext/, Riverside, CA, 2005.

[29] P. Hoff, *Multiplicative latent factor models for description and prediction of social networks*, (2006).

[30] Seock-Ho Kim, *An investigation of the likelihood ratio test, the mantel test, and the generalized mantel-haenszel test of dif*, Annual Meeting of the American Educational Research Association (2000).

[31] S. Kirkland, J. Molitierno, M. Neumann, and B. Shader, *On graphs with equal algebraic and vertex connectivity*, Elsevier: Linear Algebra and its Applications **34** (2000), 45–56.

[32] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, *Analyzing social networks within bibliographic data*, Lecture Notes in Computer Science **4080** (2006), 234–243.

[33] D. Knowles and Z. Ghahramani, *Infinite sparse factor analysis and infinite independent components analysis*, 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007). Lecture Notes in Computer Science Series (LNCS), Springer (2007).

[34] D. Krackhardt and K. Carley, *Pcans model of structure in organizations*, Carnegie Mellon University.

[35] S. Kuriyama, M. Ohira, H. Igaki, and Ken ichi Matsumoto, *A wearable interface for visualizing coauthor networks toward building a sustainable research community*, in Proceedings of the Working Conference on Advanced Visual Interfaces, 492–495.

[36] R. Lambiotte and M. Ausloos, *Collaborative tagging as a tripartite network*, arXiv (2005), PACS numbers: 89.75.Fb, 89.65.Ef, 64.60.Ak.

[37] B. De Malafosse, *An application of the infinite matrix theory to mathieu equation*, Elsevier: Computers and Mathematics with Applications **52** (2006), 1439–1452.

[38] Bradley Malin, *Unsupervised name disambiguation via social network similarity*, SIAM International Conference on Data Mining (2005), 93–102.

[39] D. Marchette and E. Leed-Hohman, *A dynamic graph model for analyzing streaming news documents*, Computational Intelligence and Data Mining (2007).

[40] D. Marchette and C. Priebe, *Modeling interstate alliances with constrained random dot product graphs*.

[41] _____, *Predicting unobserved links in incompletely observed networks*, Computational Statistics and Data Analysis (2007).

[42] P. Marsden, *Egocentric and sociocentric measures of network centrality*, Social Networks (2002), 407–422.

[43] Y. Matsuo1 and Y. Yasuda, *An analysis of researcher network evolution on the web*, (2005).

[44] V. Mukha, *The best polynomial multidimensional-matrix regression*, Cybernetics and Systems Analysis **43** (2007).

[45] C. Murray, W. Ke, and K. Borner, *Mapping scientific disciplines and author expertise based on personal bibliography files*, (2006).

[46] Mark Newman, Albert-László Barabási, and Duncan J. Watts, *The structure and dynamics of networks*, Princeton University Press., Princeton, NJ, 2006.

[47] F. Perez-Cruz, Z. Ghahramani, and M. Pontil, *Conditional graphical models*, MIT Press (September, 2007), Predicting Structured Data, Edited by G. H. Bakir, T. Hofmann, B. Scholkopf, A. J. Smola, B. Taskar and S. V. N. Vishwanathan.

[48] PubMed, *The nih pubmed database, http://www.ncbi.nlm.nih.gov/pubmed/.*

[49] C. Robertson, *A matrix regression model for the transition probabilities in a finite state stochastic process*, Applied Statistics **39** (1990), 1–19.

[50] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, *An introduction to exponential random graph (p^*) models for social networks*, (2006).

[51] C. Roth, *Generalized preferential attachment: Towards realistic social network models*, (2005).

[52] Y. Said, *Agent-base simulation of ecological alcohol systems*, dissertation submitted to George Mason University in partial fulfillment of the Ph.D. in Computational Statistics and Informatics, 2005.

[53] ———, *Bayesian social network models with acute outcomes*, (2005).

[54] Y. Said and E. Wegman, *A bipartite graph model of the interaction between alcohol users and institutions*, 2007.

[55] Y. Said, E. Wegman, W. Sharabati, and J. Rigsby, *Implications of co-author networks on peer review*, (2007).

[56] ———, *Social networks of author-coauthor relationships*, Computational Statistics and Data Analysis **52** (2007), 2177–2184, DOI: 10.1016/j.csda.2007.07.021.

[57] W. Simpson, *The quadratic assignment procedure (qap)*, (2007).

[58] M. Uchiyama, *Mixed matrix (operator) inequalities*, Elsevier: Linear Algebra and its Applications **341** (2002), 249–257.

[59] M. van der Leij and S. Goyal, *Strong ties in a small world*, (2006).

[60] M. van Duijn, K. Gile, and M. Handcock, *Comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models*, (2007).

[61] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press, New York, 1994.

[62] E. Wegman and Y. Said, *A directed graph model of ecological alcohol systems incorporating spatiotemporal effects*, Compstat 2008 (2008), 179–190.

[63] E. Wegman, D. Scott, and Y. Said, 2006, Ad-hoc Committee Report on the 'Hockey Stick' Global Climate Rconstruction, A Report to Chairman Barton, House Committee on Energy and Commerce and to Chairman Whitfield, House Subcommittee on Oversight and Investigations: Paleoclimate Reconstruction to United States House of Representatives.

[64] B. Wellman, *Computer networks as social networks*, Science, New Series **293** (2001), 2031–2034.

[65] D. Whitney and D. Alderson, *Are technological and social networks really different?*

[66] Han-Ming Wu, Yin-Jing Tien, and Chun houth Chen, *Gap: A graphical environment for matrix visualization and cluster analysis*, Computational Statistics and Data Analysis (2008).

[67] T. Zhou and R. Shumway, *One-step approximations for detecting regime changes in the state space model with application to the influenza data*, Computational Statistics and Data Analysis (2008), 2277–2291.

Curriculum Vitae

Walid Sharabati was born in Bethlehem, Palestine (West Bank). He went to Bethlehem University for his Bachelor's of Science majoring in mathematics with a minor in physics. While at Bethlehem University, he was awarded three honors certificates and he performed a research on the dimensions and self-similarity of fractals. Additionally, he was awarded a certificate of recognition for his services in the university's computer laboratories. After graduating from Bethlehem University in 1998, he was offered a position in the computer center at Bethlehem University in which he was responsible for the software and hardware installation and maintenance of the university's computers. In November 1999, he volunteered to serve on the preparatory committee of Bethlehem Youth Conference held at Bethlehem University.

In 2000, he moved to the United States to pursue his higher education. He received his masters of science in mathematics and computer science in 2003 from Minnesota State University. During the four year period at Minnesota State University, he performed a research on the reaction diffusion equation for the growth of Mycelium under the supervision of Dr. Ernest Boyd. At the same time, he worked as a teaching assistant for the Department of Mathematics and Statistics in which he taught college algebra and trigonometry.

Subsequent to his graduation from Minnesota State University, he moved to Washington, DC metropolitan area. In 2005, he enrolled the Ph.D. program at George Mason University. While working on his Ph.D., he conducted several research projects in network theory and data visualization that became the corner stone and foundation of his dissertation. He co-authored three papers entitled "Implications of co-author networks on peer review", "Social networks of author-coauthor relationships", and "Author-Coauthor Social Networks and Emerging Scientific Subfields". He gave an invited talk at Interface 2008 and two contributed talks at JSM 2007 and JSM 2008. He is a member of the American Statistical Association, American Mathematical Society, Institute of Mathematical Statistics, and the International Network for Social Network Analysis.

Between August 2004 and August 2008, he worked as an instructor for the department of mathematics and statistics at the American University. At American University, he taught a number of undergraduate and graduate courses as well, which include applied calculus, basic statistics, precalculus, finite mathematics, intermediate statistics and statistical methods. In August 2008, he accepted an offer to teach at Purdue University in West Lafayette, Indiana as a Visiting Assistant Professor.

Walid's future plan is to expand and develop his dissertation research on network theory

as well as teaching. His immediate goals are to implement the theory of multi-mode large scales networks and continue the study of hypergraphs and their relationship to cliques.