ANALYSIS AND APPLICATIONS OF SOCIAL NETWORK FORMATION

by

Daning Hu

A Dissertation Submitted to the Faculty of the

COMMITTEE ON BUSINESS ADMINISTRATION

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2009

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Daning Hu

entitled Analysis and Applications of Social Network Formation

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 07/15/2009
J. Leon Zhao

_____ Date: 07/15/2009
Jay F. Nunamker, Jr.

_____ Date: 07/15/2009
Zhu Zhang

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 07/15/2009
Dissertation Director: J. Leon Zhao

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

SIGNED: Daning Hu

ACKNOWLEDGEMENTS

My thanks and great appreciation first go to my advisor Professor J. Leon Zhao for his guidance, encouragement, patience and kindness over the years. He taught me how to be a researcher, helped me reach my full potential. I always think I am blessed to have him as my advisor. I am thankful for his trust in me and the opportunities he gave me to grow not just as a researcher but also as a person. Thank you, Dr. Zhao. Thank you for everything.

I am also grateful to my other dissertation committee members, Professor Jay F. Nunamaker, Professor Zhu Zhang, and my minor committee member Dr. Tong Yao, for their valuable suggestions and feedback on my research and continuous support during my doctoral program. I also thank other MIS faculty members for their support during my studies. In addition, my dissertation was partly supported by grants from the National Science Foundation.

I would like to thank my colleagues and friends, Xin Li, Jennifer Jie Xu, Siddharth Kaza, Wei Chang, Zhiying Sun, Shaokun Fan, Xinlei Zhao, Manlu Liu, Runpu Sun, Ping Yan, Yilu Zhou, Jiexun Li, Jialun Qin, Harry Jiannan Wang, Xiaoyun Sun, Yan Dang, Yulei Zhang, Yan An, Nichalin Suakkaphong, Tianjun Fu, Ying Liu, Chunju Tseng, Ming Lin, Saiwu Lin, Ted Elhourani, Hsin-min Lu, Ahmed Abbasi, G. Alan Wang, Zan Huang, Lin Zhu, Yiwen Zhang, Byron Marshall, Kris Chang, Lijun Yan, Yida Chen and Rob Schumaker. I have worked closely with Xin, Siddharth, Jennifer, Shaokun, and Xinlei. I greatly appreciate their help and support to my research.

My close friends Yujie, Lei, Jing, Kai, Xiaoling, thank you for helping me focus on other priorities in life.

Lastly, but most importantly, I greatly appreciate the support from my family, especially my mother. Her care and love is what keeps me going during hard times.

DEDICATION

This dissertation is dedicated to my mother, Qi Tang.

TABLE OF CONTENTS

TABLE OF CONTENTS - CONTINUED

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

Nowadays people and organizations are more and more interconnected in the forms of social networks: the nodes are social entities and the links are various relationships among them. Social network theory and methods of social network analysis (SNA) are being increasingly used to study such real-world networks in order to support knowledge management and decision making in organizations. However, most existing social network studies focus on the static topologies of networks. The dynamic network link formation process is largely ignored. This dissertation is devoted to studying such dynamic network formation processes to support knowledge management and decision making in networked environments.

Three challenges remain to be addressed in modeling and analyzing the dynamic network link formation processes. The first challenge is about modeling the network topological changes using longitudinal network data. The second challenge is concerned with examining factors that influence formation of links among individuals in networks. The third challenge regards link prediction in evolving social networks. This dissertation presents four essays that address these challenges in various knowledge management domains.

The first essay studies the topological changes of a major international terrorist network over a 14-year period. In addition, this paper used a simulation approach to examine this

network's vulnerability to random failures, targeted attacks, and real-world authorities' counterattacks.

The second essay and third essay focus on examining determinants that significantly influence the link formation processes in social networks. The second essay found that mutual acquaintances and vehicle affiliations facilitate future co-offending link formation in a real-world criminal network. The third essay found that homophily in programming language preference and mutual acquaintances are determinants for forming participation links in an online Open Source social network.

The fourth essay focuses on link prediction in evolving social networks. It proposes a novel infrastructure for describing and utilizing the discovered determinants of link formation processes (i.e., semantics of social networks) in link prediction to support expert recommendation application in an Open Source developer community. It is found that the integrated mechanism outperforms both user-based and Top-N most recognized mechanisms.

CHAPTER 1: INTRODUCTION

Social networks are becoming increasingly pervasive in organizations, with individuals leveraging the communication abilities of computing systems to interact and collaborate. A social network often refers to a set of nodes connected by links. The nodes are social entities that range from Open Source developers (in the case of Open Source developer networks) to criminals (in the case of narcotic networks). The links between nodes are various kinds of social relationships ranging from collaboration between OSS developers to co-offending relationships between criminals.

With the emergence of social networks and the availability of unprecedented amounts of online network data, an analytical method – social network analysis (SNA) – has been increasingly used by information systems (IS) researchers to study social networks in various organizational settings. SNA aims to uncover the patterns of relationships among social actors and help researchers to study and test theories in network environments. In addition, it is believed studying networks has led to a "new science of networks" (Newman et al. 2006). This new science has great potential in supporting knowledge management and decision making in diverse domains such as e-commerce, computer mediated communications, and security informatics. Empirical studies that are intended to apply this new science to such application domains need to be conducted to demonstrate the value and significance of this new field.

Social networks grow over time with the addition of nodes and the formation of links to survive and thrive in their environments. It is critical to study various dynamic network processes that govern the formation of social networks, especially the link formation processes. However, existing SNA research mainly viewed the networks as static structures rather than evolving dynamical systems. Modeling and analyzing the dynamic processes of social networks has faced several challenges.

The first challenge is about modeling the network topological changes using longitudinal network data. Various network measures have been developed in previous SNA studies to study the static topologies of networks. For example, the degree measure (i.e., the number of links a node has) is usually used to identify important individuals in a social network. However, few techniques have been developed to model the changes in network topologies over time.

The second challenge is concerned with the study of the factors that influence link formation process in social networks. Many real-world networks are the products of dynamic processes that form various links/relationships among individuals. For instance, a social network of collaboration changes as individuals make or break collaboration links with others. An individual with many collaboration links may be more likely to attract new collaborators than someone else less connected. Or individuals may be more likely to collaborate with people who share a mutual acquaintance. In these link formation processes, many social and technological factors significantly influence

people's behavior choices to form links (i.e., relationships) with one another, therefore affecting the formation of the social network. There is a great need to study such factors in order to better understand the drivers behind the evolution of the social network.

The third challenge regards/is related to link prediction in social networks. Network link analysis and link prediction algorithms have been widely used in various recommendation systems. Topology-based link prediction algorithms recommend nodes based on an aggregate opinion from the members of the network. However, for newly joined nodes without much network topology (global) information, such algorithms may not work well. In that case, (local) information such as individual attributes or shared affiliation can be used to complement topology information in link prediction. Therefore, link prediction algorithms which utilize both local information (i.e., individual node level) and global information (i.e., network topology) are greatly needed.

Facing these challenges, this dissertation aims to achieve the following research objectives:

- Develop and employ effective techniques to describe the topological changes of social networks using longitudinal data.
- Develop and employ effective modeling and analytical techniques to identify the determinants of social network processes in several application domains.

- Evaluate the performance of these modeling techniques in terms of their abilities to predict future link formation in order to support knowledge management and decision making.

Each of the four dissertation essays aims to achieve one or more of the objectives stated above with empirical studies in various knowledge management domains. Chapter 2 focuses on developing and employing network analysis techniques to describe and simulate the changes of topologies for a large global terrorist network. Nowadays terrorists usually work in network forms to conduct attacks. Terrorist networks remain active and can still function even after being severely damaged by authorities. Analyzing terrorist networks from a dynamic point of view can provide insights about the mechanisms responsible for the survival of terrorist organizations. This study analyzed the dynamics of a major international terrorist organization over a 14-year period – the Global Salafi Jihad (GSJ) terrorist network. It was found that a scale-free topology gradually emerged as new members joined the GSJ network based on operational needs. In addition, since the network has been experiencing member losses while it grows, the robustness of the GSJ network was also examined. The study proposed a simulation approach to examine GSJ's vulnerability to random failures, targeted attacks, and real-world authorities' counterattacks. It was found that authorities' counterattacks have been rather ineffective in disrupting the terrorist network.

Chapter 3 focuses on modeling the dynamic link formation processes in social networks and studying their facilitators. Social networks evolve over time with the addition and removal of nodes and links to survive and thrive in their environments. Previous studies have shown that the link formation process in such networks is influenced by a set of facilitators. However, there have been few empirical evaluations to determine the important facilitators. In a research partnership with law enforcement agencies, this study used dynamic social network analysis methods to examine several plausible facilitators of co-offending relationships in a large-scale narcotics network consisting of individuals and vehicles. Multivariate Cox regression and a two-proportion z-test on cyclic and focal closures of the network showed that mutual acquaintances and vehicle affiliations were significant facilitators for the network under study. It was also found that homophily with respect to age, race, and gender were not good predictors of future link formation in these networks. In addition, the social causes and policy implications for the significance and insignificance of various facilitators, including shared jails on future co-offending, were also examined. These findings provide important insights into the link formation processes and the resilience of social networks. In addition, they can be used to aid in the prediction of future links. The methods described can also help in understanding the driving forces behind the formation and evolution of social networks facilitated by mobile and web technologies.

Chapter 4 focuses on examining the determinants of link (relationship) formation in social networks, particularly in an online open source community. Successful open

source software (OSS) projects often require a steady supply of self-motivated software developers. However, little work has been done from a relational/network perspective to study the factors that drive the developers to participate (i.e., form a participate link) in OSS projects. This study investigates the participation dynamics in a social network, particularly in an online open source community called Ohloh. Through a REST-based API, information about 11,530 OSS projects involving 94,330 developers was collected from the Ohloh community. I examined a set of social and technical factors in the Ohloh dataset, which was defined as the determinants that significantly influence the developers' participation choices, using social network analysis and statistical analysis methods. It was found that the determinants include (1) homophily in programming language, (2) project mutual acquaintances, and (3) project age. In addition, our research findings provide the possibility of predicting developers' participation choices, and therefore can have important implications for OSS project management and in designing social network enabled recommendation systems.

Chapter 5 focuses on describing, modeling and utilizing the semantics of social networks - determinants of social networks - to support expert recommendation in OSS communities. The use of social network analysis (SNA) in the design of expert recommendation systems is becoming increasingly popular. However, the experts recommended from such systems often do not meet users' needs since the network semantic information is largely ignored. In this study, we used conditional logistic analysis to quantitatively examine the semantics of two social networks in a large open

source community called Ohloh. It was found that homophily in nationality, location, programming language preference, and reputation are determinants for forming evaluation and collaboration relationships among Ohloh members. Moreover, past collaborations and mutual acquaintances are also found to significantly affect the formation of evaluation links but not collaborations. In addition, we proposed an infrastructure to describe, model and utilize the discovered network semantics for integrating two expert recommendation mechanisms - user-based link prediction and Top-N most recognized mechanism.

Chapter 6 concludes the dissertation, stating its contributions to the social network and knowledge management domains, and presents future directions for further research/study.

CHAPTER 2: DESCRIBING THE TOPOLOGICAL CHANGES OF LARGE SCALE SOCIAL NETWORKS

2.1 Introduction

Terrorism and terrorist attacks seriously threaten national security and public safety in countries around the world. Authorities have been fighting terrorism for a long time, and numerous arrests have damaged major terrorist organizations such as Al Qaeda. However, the tragic events of September 11 in the U.S., the Madrid train bombing in Spain, and the London subway/bus bombings in the United Kingdom indicate that terrorist organizations remain active and can still function even after severe damage. How these terrorist organizations have survived disruption and attacks is a question that has long puzzled authorities and terrorism researchers.

It is conjectured that the structure of terrorist organizations greatly enhances their resistibility and robustness to attacks and damage (Klerks 2001; Krebs 2001). Traditionally, terrorist organizations were believed to have a centralized, hierarchical structure in which the leaders at the top of the hierarchy control the operation of the entire organization. Such a hierarchical structure is more vulnerable to attacks targeting the leaders. Contemporary terrorist organizations have adopted a network structure which is decentralized and more flattened (Klerks 2001; Milward et al. 2002). In these networks social ties between terrorists hold the organization together and the control of operations

is dispersed all over the network. As a result, the network can still function even if some parts of it are destroyed.

Although the conjecture is interesting, there have been few empirical studies that systematically verify it. The structural mechanisms responsible for the survival of terrorist networks remain unknown for two major reasons. First, nearly all theoretical and practical studies on terrorist networks suffer from the lack of empirical data. As terrorist networks are clandestine organizations that operate covertly, data about the individual members and their social ties are extremely difficult to gather. Anecdotal evidence from news stories and media sources is highly unreliable. Second, the dynamic nature of terrorist networks is largely ignored. Terrorist organizations are dynamic systems and undergo constant changes over time (Carley et al. 2003; Dombroski et al. 2002). On one hand, a network can grow by recruiting new members. New members may join the network through all sorts of social ties such as friendship, kinship, and religion. On the other hand, it may lose its members due to arrests and suicide bombings.

The purpose of this article is to analyze terrorist networks from a dynamic point of view in order to uncover the mechanisms responsible for their survival. Based on a relatively reliable dataset about a major international terrorist organization, Global Salafi Jihad, we aim to answer a series of questions: What is the topology of these networks? How have these networks evolved? How robust are these networks? How have these networks managed to survive? Have authorities' counterattacks been effective?

The remainder of this article is organized as follows. In the next section we review related literature about network structure and dynamics. Section 2.3 introduces the research testbed. Section 2.4 presents the research design used of this study. Section 2.5 discusses the results. Section 2.6 then summarizes our findings and concludes the essay.

2.2 Literature review

Recent development in the topological analysis of large networks (Albert et al. 2002) has provided a great opportunity for studying the dynamics of terrorist networks. Examples of networks are the World Wide Web (Broder et al. 2000; Huberman et al. 1999; Kumar et al. 1999), the Internet (Faloutsos et al. 1999), movie actor networks in which nodes are actors and links are their collaboration relationships in movies (Watts et al. 1998), coauthorship networks of academic authors who wrote papers together (Newman 2004), and metabolic pathways that consist of biochemical reactions occurring in a cell (Jeong et al. 2000). One important contribution of such development is its focus on the dynamics of scale-free networks (Barabási et al. 1999). In a scale-free network, a large percentage of the nodes have only a few links (low degree) while a small percentage of nodes have a very large number of links (high degree), where degree is defined as the number of links a node has.

A scale-free network's structure is significantly different from a random graph network, in which every node has roughly the same number of links. That is, nodes are randomly connected and the network is rather homogeneous in terms of node degree. However, in a

scale-free network, there are some "hubs" that connect to a large number of other nodes and hold the network together. The degree distribution, p(k), which plots the probability that an arbitrary node in a network has exactly k links, can clearly show the difference between the two types of network. The degree distribution of a random graph network is a bell-shaped Poison distribution peaking at the average degree, while that of a scale-free network is a highly skewed one that has no peak but a very long, flat tail, which is often called power-law distribution (Albert et al. 2002).

It is interesting to study what causes the emergence of the scale-free structure, which is common in many real networks. Various models (Faloutsos et al. 1999; Garlaschelli et al. 2003; Jeong et al. 2000; Newman 2004) have been proposed to uncover the mechanisms responsible for such a highly skewed power-law degree distribution. Among these mechanisms, growth and preferential attachment have been believed to be the two fundamental mechanisms in the evolution of scale-free networks (Barabási et al. 1999). Growth means that the size of a network is not fixed. Instead, a network can grow by including new nodes. Preferential attachment means that when a new node is added to the network the probability that an old node receives a link from the new node depends on the number of existing links of the old node, a phenomenon known as "the rich get richer." With both mechanisms the scale-free structure thus is the product of evolution, a dynamic process in nature.

Moreover, some research has studied the robustness of different topologies (random and scale-free) against failures and attacks (Albert et al. 2000; Cohen et al. 2000; Crucitti et al. 2003; Solé et al. 2001). Scale-free networks have been found to be very robust against random errors but highly vulnerable to attacks targeting the hubs. Because random errors remove nodes randomly from a network, the majority of the network can remain connected even if it loses a number of low-degree nodes. However, the removal of just a small number of hubs will easily break down the entire network, because the attacks remove not only the hubs but also their links to a large number of low-degree nodes, causing those nodes to be disconnected (Albert et al. 2000).

Although these findings about scale-free networks in general are illuminating, they cannot be applied to terrorist networks in a straightforward manner. Most of these studies assume that a network is either in a growing mode, where the network adds new nodes without losing existing nodes, or in a decaying mode, where some nodes are removed due to failure or attack. However, in reality terrorist networks are seldom in merely one mode. The survival of a terrorist network is actually the product of the mixture of both growing and decaying modes. Some studies have considered aging as a decaying factor during the growth of a network, when nodes naturally drop out of the network after a certain period of time (Amaral et al. 2000). However, for a terrorist network, attack from authorities rather than aging may be the most important decaying factor. No existing findings thus far can be applied to directly account for the survival of terrorist organizations. In addition, because of the lack of large reliable datasets, statistical analysis of the topology

of terrorist networks is almost impossible, let alone dynamic analysis which requires information about the time when changes occur to a network.

In this research we focus on the survival process of terrorist networks by studying their evolution and robustness, realizing that they can both grow and be under attack at the same time. We use the Global Salafi Jihad data to study the process and demonstrate our findings. Through this study we hope to contribute to the research on terrorism and counterterrorism policies and to provide insights into the survival mechanisms of large networks in hostile environments.

2.3 Research testbed

We study the Global Salafi Jihad (GSJ) terrorist network (Sageman 2004) which consists of 366 members, including those from Osama Bin Laden's Al Qaeda. These terrorists were connected by kinship, friendship, religious ties, and relations formed after they joined the GSJ network.

The Global Salafi Jihad is part of a violent worldwide Muslim revivalist movement. It is a new form of terrorism which is driven by a fanatical determination to inflict maximum civilian and economic damages. Although mainly targeting the West, the reckless operations of the GSJ have resulted in indiscriminate slaughter. The GSJ includes many terrorist groups from different countries forms a large global terrorist network. Through this network, the GSJ has successfully planned and launched many large-scale terrorist

attacks across the world, including the 9/11 tragedy in 2001, the bombing in Bali in 2002, and the bombing in Morocco in 2003.

The data about the GSJ network were provided by the author of a recently published book, *Understanding the Terror Networks* (Sageman 2004). The author is a former Foreign Service officer who worked closely with Afghanistan's mujahedin from 1987 to 1989. The network was constructed based entirely on open-source information. In decreasing degrees of reliability, the information sources include transcripts of court proceedings involving GSJ terrorists and their organizations, reports of court proceedings, corroborated information from people with direct access to the information provided, uncorroborated statements from people with the access, and finally, statements from people who had heard the information second-hand (Sageman 2004). Information about all the nodes (terrorists) and links (relations) was scrutinized and carefully cross-validated.

The final dataset consists of the profile information of 366 GSJ terrorists which includes a set of sociological features (e.g., geographical origins, original socio-economic status, education, occupation, etc.) and individual psychological features (e.g., mental illness, personality, pathological narcissism, etc.) that could explain why these people became terrorists. More importantly, the data also captures all known relationships and interactions among these 366 GSJ terrorists. These relationships and interactions include personal relationships (e.g., acquaintance, friend, relative, and family member), religious

relationships (following the same religious leader), operational interactions (participating in the same attacks), and other relationships. The dataset is presented in the form of a spreadsheet with each row containing the basic features of a certain GSJ member as well as all the other members that are related to this member through the various relationships or interactions mentioned above. Our network visualization provides an intuitive and clear view of the overall GSJ network (Fig. 1a).

However, as the author points out in the book, the data are subject to several limitations. First, the members included in the network may not be a representative sample of the Global Salafi Jihad as a whole. It is biased toward leaders and the members who have been captured or uncovered in executed attacks. Second, because most of the sources were based on retrospective accounts, the data may be subject to self-reported biases. Despite the limitations, the data have revealed stunning insights into the clandestine organizations of terrorists (Sageman 2004). More importantly, this dataset contains information about the time when each individual terrorist joined or left the network between 1989 and 2003, making it a good sample for studying the dynamic survival process of this terrorist organization.

2.4 Research design

To study the dynamics of the GSJ network during the 15-year period, we use both descriptive and simulation approaches. Like many descriptive studies on network

dynamics (Barabási et al. 2002; Csányi et al. 2004; Hajra et al. 2005), we aim to capture and observe the changes in the network over time based on two topological statistics: average degree and degree distribution (Barabási et al. 2002). The degree of a node is the number of links it has. The degree distribution of a network, $P(k)$, is the probability that a node has exactly k links. The changes observed in the statistics are then plotted with respect to time in order to examine the dynamic patterns. In particular, the growth of a network can be described by its average degree. It is used to compare the growing speeds of links and nodes: if the average degree increases over time, the number of links grows faster than nodes, indicating accelerated growth (Albert et al. 2002).

In addition, we studied the robustness of the GSJ network by examining its structural changes under random and targeted attacks. Unlike most existing robustness studies (Albert et al. 2000; Cohen et al. 2000; Crucitti et al. 2003; Solé et al. 2001) which test network robustness based on a snapshot of the network assuming a single decaying mode, we examine the "dynamic robustness" of the network by allowing it to include new nodes when some nodes are removed. That is, the network can be in both growth and decay modes during its entire life span. Such a dynamic test is more realistic because the GSJ has never been in a single mode.

To examine the dynamic robustness of the network, we use the diameter l (Watts et al. 1998) (Fig. 3b) to measure the interconnectedness of the network over time. The diameter of a network is defined as the average length of the shortest paths between any

pair of nodes in the largest connected component of the network. In general, the shorter the diameter is, the more interconnected a network is. Removing a node usually will increase the diameter, since it may eliminate a link which is in the shortest path of another pair of nodes.

In our dynamic robustness test, we adopt three different node removal strategies: (a) random node removal (random errors) in which randomly selected nodes are removed; (b) preferential node removal (targeted attacks) in which the most connected nodes are removed; and (c) real node removal (authorities' counterattacks) in which nodes are removed in the same order as authorities arrested the terrorists in reality. The third removal strategy was possible because our dataset contains information about the exact time each terrorist died or was arrested.

2.5 Experimental results

2.5.1 The growth pattern

To study the growth patterns of the GSJ network, we measured the changes in the average degree from 1989 to 2003 (Table 2.1). We found that the GSJ network experienced three stages during its evolution: (I) the emerging stage from 1989 to 1991, during which the network was under accelerated growth in that the average degree drastically increased from 9.86 to 14.48; (II) the maturing stage from 1992 to 2000, during which the average degree first decreased a little to 13.86 and stayed relatively

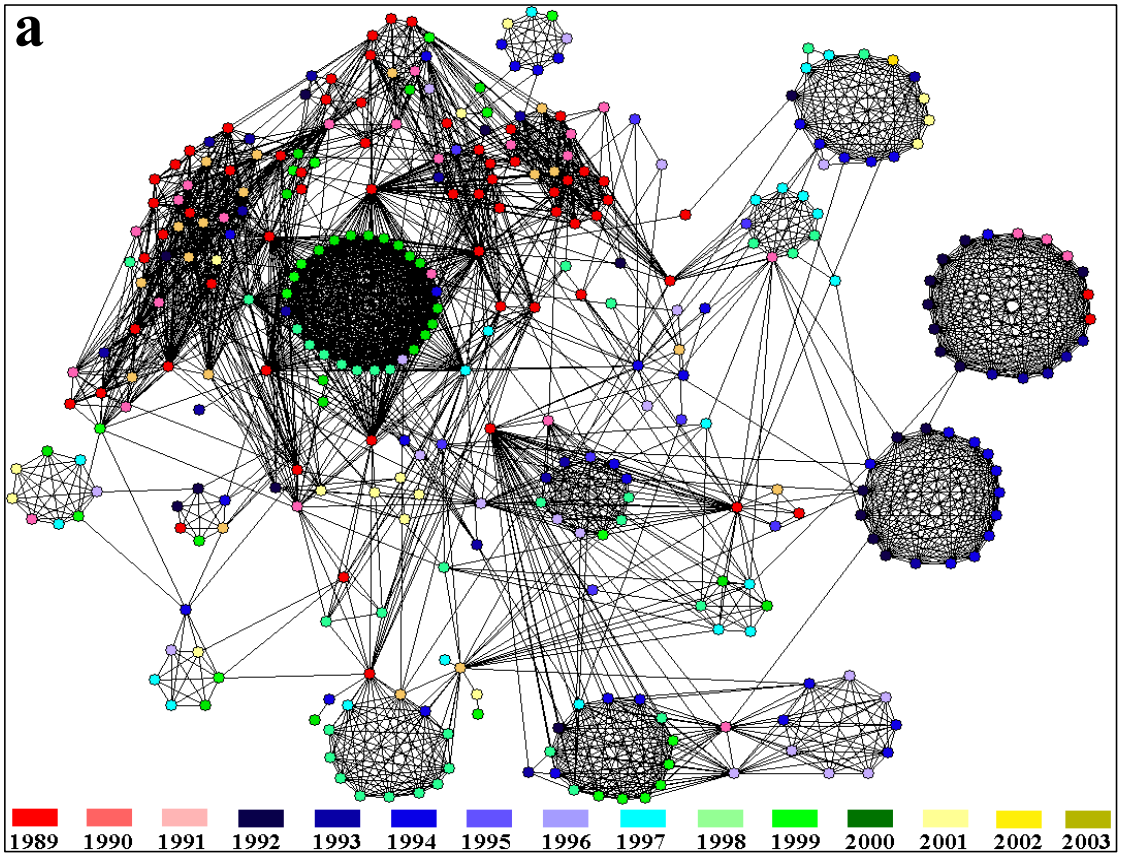
stable until 1997; it peaked at 14.54 in 2000; (III) the disintegrating stage from 2001 to 2003, during which the average degree dramatically decreased, indicating that the GSJ network started to fall apart as a large portion of nodes left the network. Figure 2.1 presents the visualizations of the network during the three stages.

Table 2.1. The statistics describing the structural changes of the GSJ network from 1989 to 2003.

Stage	Year	No. of nodes	No. of links	Average degree	R2 of regression analysis on the degree distribution
I	1989	61	301	9.86	0.05
	1990	79	476	12.06	0.07
	1991	102	739	14.48	0.06
II	1992	124	859	13.86	0.17
	1993	142	1026	14.46	0.22
	1994	170	1163	13.68	0.21
	1995	166	1079	13	0.52
	1996	164	1040	12.68	0.48
	1997	183	1135	12.40	0.47
	1998	197	1240	12.58	0.63
	1999	194	1264	13.04	0.62
	2000	206	1498	14.54	0.48
III	2001	151	714	9.46	0.54
	2002	103	386	7.50	0.45
	2003	48	92	3.80	0.67

During the emerging stage, three clusters emerged (Fig. 2.1b). The three clusters are defined mainly based on their geographical origins (Sageman 2004): the Central Staff cluster consisting of the leaders of Al Qaeda and the GSJ network including Bin Laden, the Southeast Asian cluster consisting of followers of Jemaah Islamiyan centered in

Indonesia and Malaysia, and the Core Arabs cluster consisting of terrorists from Arab states (e.g., Saudi Arabia, Egypt, Yemen, and Kuwait). These three clusters are the backbone of the GSJ network.



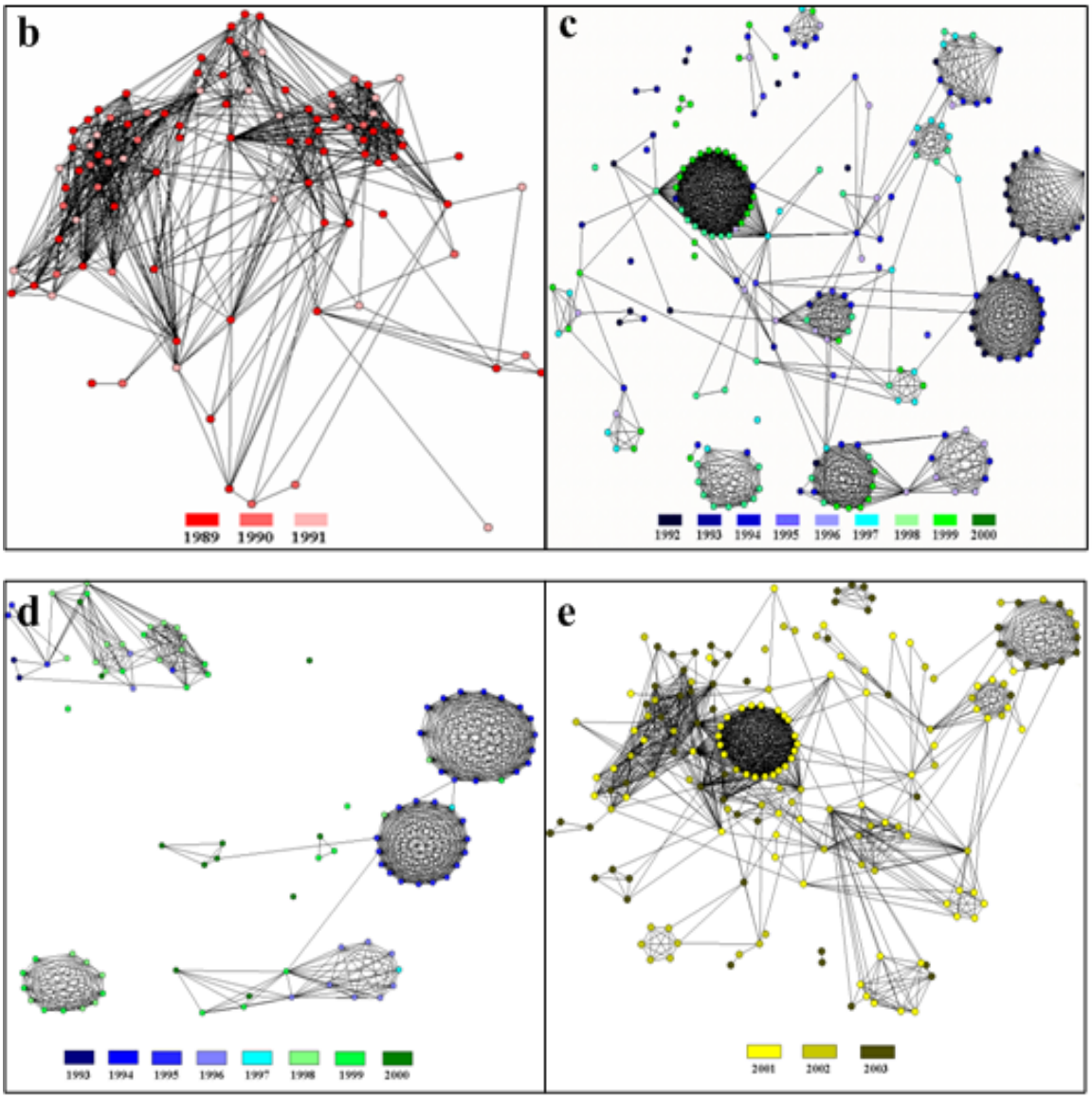


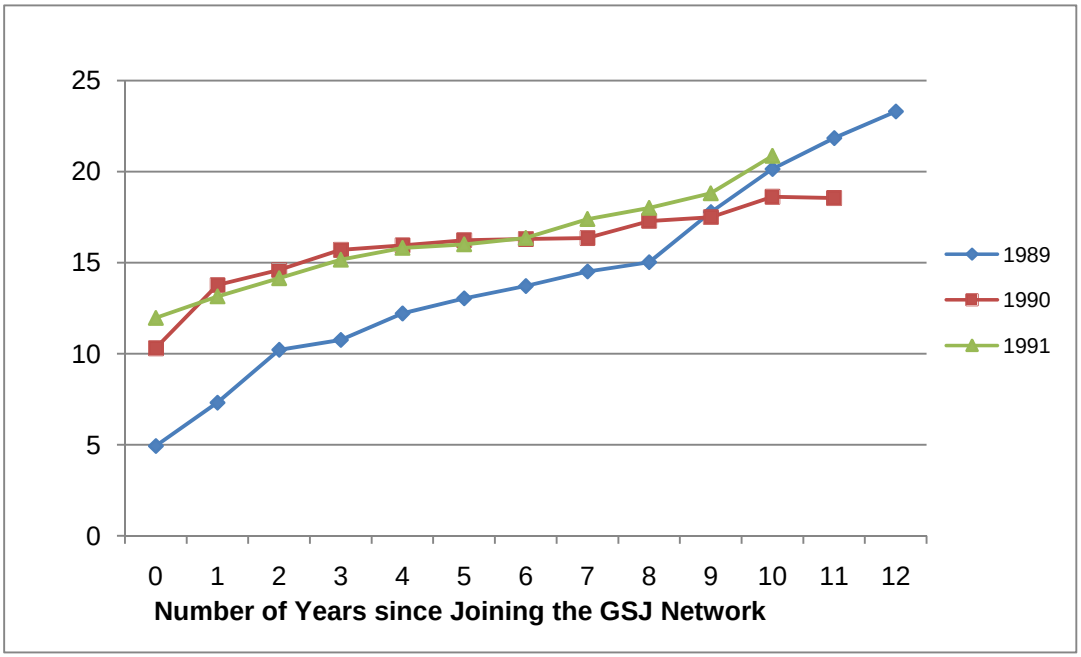
Figure 2.1. The dynamic view of the Global Salafi Jihad terrorist network:

In Figure 2.1, Nodes are color coded in terms of years. Figure 2.1 (a) shows the overall network with terrorists who joined the GSJ network from 1989 to 2003. Figure 2.1 (b) shows the network with terrorists who joined from 1989 to 1991. Figure 2.1 (c) presents the network with terrorists who joined from 1992 to 2000. Figure 2.1 (d) shows the one

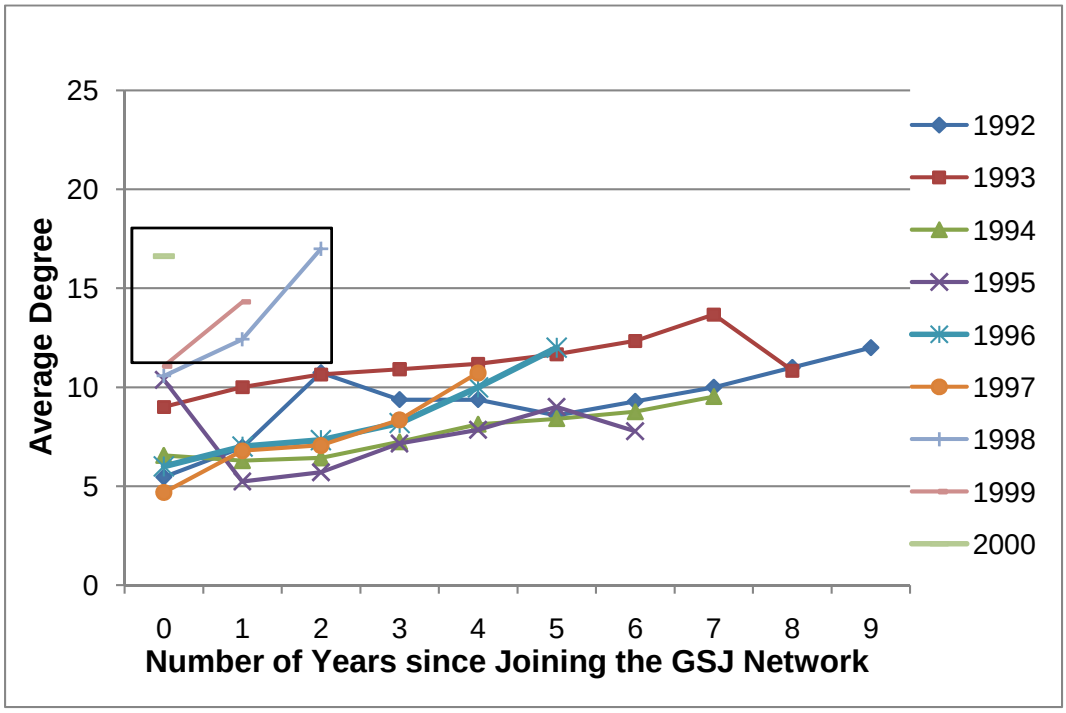
with terrorists who were arrested by authorities from 1993 to 2000. Figure 2.1 (e) shows the one with terrorists arrested from 2001 to 2003.

To study the growth patterns in depth, we divided all the nodes into 12 yearly groups from 1989 to 2000, according to the year they joined the GSJ network (few nodes joined after 2000). We then measured the changes in the average degree of each yearly group from the year they joined the GSJ network to 2000 (Fig. 2.2). We found that the average degrees of the three groups that joined during the emerging stage (Fig. 2.2a) were generally larger than those of the groups that joined during the maturing stage (Fig. 2.2b). This is consistent with the original scale-free model in which older nodes have more advantage over younger nodes in acquiring links (Barabási et al. 1999). This finding also shows that the aging effect (Albert et al. 2002) which accounts for the situation where older nodes become less capable of acquiring new links due to age, did not compromise the effect of growth. Terrorists who joined the network early still establish relationships with new members and stay active in operations.

Interestingly, the average degrees of the three most recent groups that joined during the maturing stage increased much faster than the average degrees of other groups who joined earlier in this stage (Fig. 2.2b). This is mainly because the GSJ network peaked in size around the end of the maturing stage (1998 - 2000). Hence those most recently joined nodes had more existing nodes available to connect to.



(a)



(b)

Figure 2.2. The changes in the average degrees of the 12 yearly groups (1989-2000) of terrorists from the year they joined the GSJ network to 2000.

Figure 2.2(a) shows the three yearly groups of terrorists who joined the GSJ network in 1989, 1990, and 1991. Figure 2.2(b) shows the nine yearly groups who joined the GSJ network from 1992 to 2000.

We also found that the GSJ network presented scale-free features. We conducted regression analysis on the degree distribution of the GSJ network for each year and measured the goodness of fit (R^2) of the power-law distribution. The changes in R^2 (Table 1) for the regression analysis indicate that the GSJ network was rather random at the beginning and displayed more and more scale-free features over time.

2.5.2 The evolution of GSJ network

We found that terrorists joined and left the GSJ network mainly on an operational base. An operation is a terrorist attack carried out by a group of terrorists, who are related to each other through operational links and formed an operational cluster. Examples of operations are the 9/11 attack in the U.S. and the Bali bombing in 2002. The maturing stage saw the network's most ambitious operations. During this stage, the network started recruiting new members and formed operational clusters to carry out terrorist attacks. In each year, most new members were involved in one or two terrorist attacks and most of these operational clusters were formed in one year or in two consecutive years. In addition, the node removal during this period followed a similar pattern. For example, 9 out of 13 members of the Aden (Yemen) terrorist attack in 1998 joined the GSJ network in the same year, and 12 of them were arrested immediately after the attack.

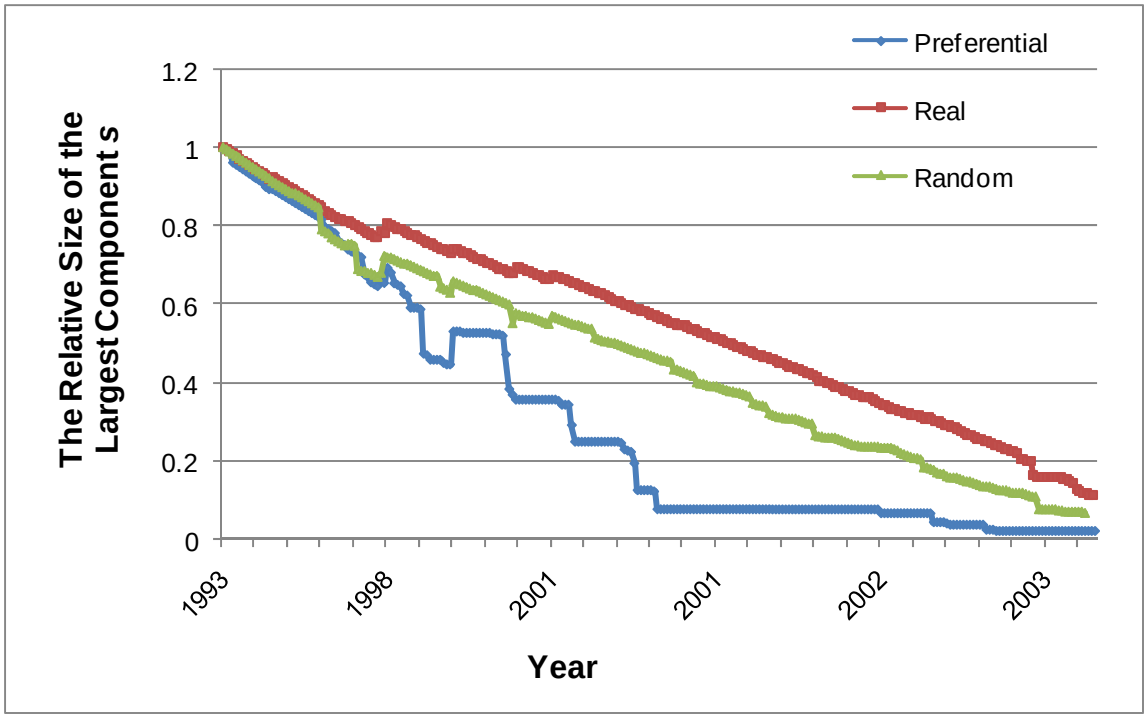
At the end of the maturing stage from 1998 to 2000, operations became more decentralized. According to Sageman (2004), field lieutenants and their local initiators, rather than the Central Staff took more responsibility for day-to-day operations during this period of time. Field lieutenants are important channels connecting operational clusters to the Central Staff. Most field lieutenants are highly connected hubs in the network. Moreover, the node removals (Fig. 1d) during this period became more severe: after a terrorist attack, most terrorist nodes in that operational cluster would be removed by authorities in a short time. Nevertheless, those field lieutenants usually tended to survive the first wave of counterattacks. This guaranteed the integrity of the main body of the GSJ network and diminished the effect of immediate counterattacks from authorities.

During the disintegrating stage (2001 to 2003) nearly 57% of the nodes in the GSJ network were removed. Unlike the maturing stage, the nodes removed in this stage included many highly connected hubs (Fig. 1e). The removal of these hubs caused the network to disintegrate into isolated cliques. This significantly weakened the network's communication ability and logistic support for large scale operations like the 9/11 attack (Sageman 2004).

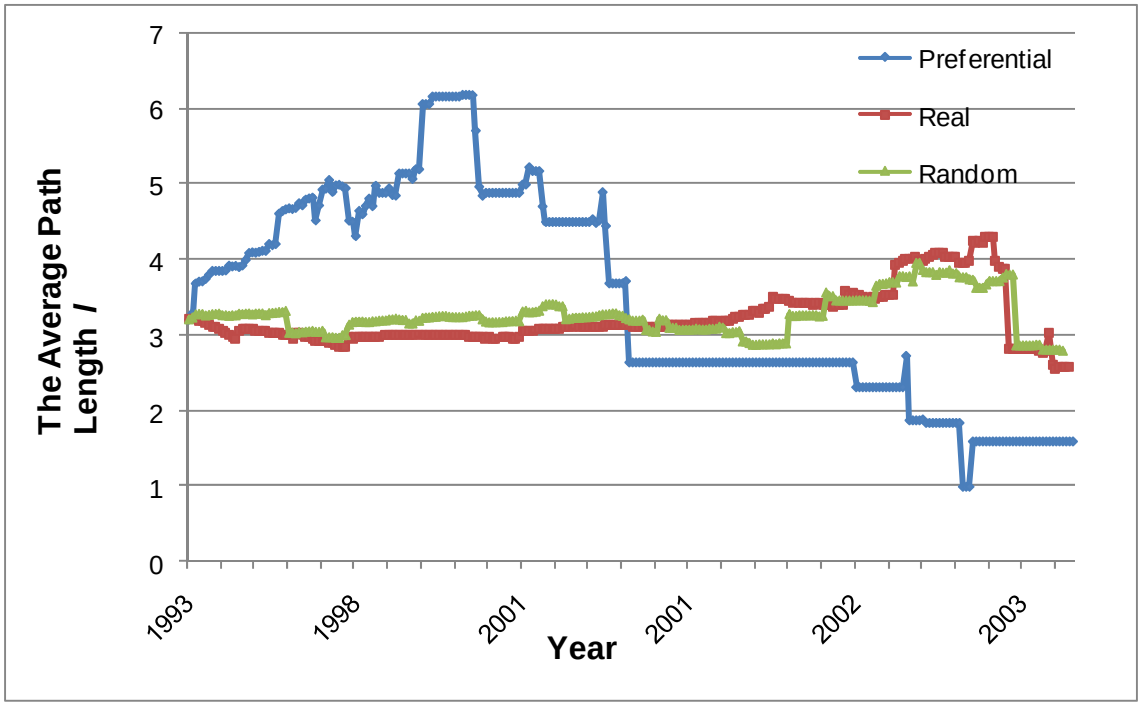
Note that few nodes joined the network during the disintegrating stage. There are two possible reasons, namely, the elimination of the training camps in Afghanistan by U.S. forces and the lack of current information about the network. The first reason may significantly prevent the new recruitment of terrorists. The latter one could be a limitation of our dataset.

2.5.3 The Robustness

As mentioned above, we found that the GSJ network displayed more and more scale-free features over time. Thus we expected similar error tolerance and attack vulnerability (Albert et al. 2000) of scale-free networks in our dynamic robustness test.



(a)



(b)

Figure 2.3. The changes in (a) the relative size of the largest component *s*, and (b) the average path length *l* from 1993 to 2003.

We simulated the changes of the GSJ network from 1993 to 2003 with three different node removal strategies. For each year we used the three node removal strategies to remove the same number of nodes as were removed in reality. At the same time, we added nodes to the network based on the data. We did not analyze the robustness before 1993 because few nodes were removed from the network before then. We then measured the average path length l of the three simulated networks from 1993 to 2003 to study their robustness (Fig. 2.3b).

The responses of the GSJ network to these three different strategies were quite different. The network displays a strong robustness against random errors and real attacks from authorities during the maturing stage. The average path lengths (Fig. 2.3b) of both random node removal and real node removal generally remained unchanged during the maturing stage. As more than 50% of the nodes were removed, the diameters started to increase to their peaks (3.95 and 4.31 for the random and real removals, respectively). They then decreased, indicating the breakdown of the network.

The network was more vulnerable to targeted attacks than both random errors and real attacks. For the preferential node removal, l increased to a more prominent peak of 6.18 much faster than the other two strategies. This implies that the network efficiency is rapidly reduced by the loss of its key members because on average, each node needs to go through more intermediate nodes to interact with other nodes.

The changes in the relative size (s) of the largest component of the network also confirm this finding (Fig. 2.3a): s decreased (the network fell apart) faster when using preferential node removal than using the random and real node removal. Moreover, we found that the real node removal was even less effective in disrupting the GSJ network than random node removal (Fig. 2.3a).

These distinct behaviors (error tolerance and attack vulnerability) of the GSJ network against different node removal strategies are rooted in the dynamical processes that

occurred in the network. In each year the most newly joined nodes form one or two operational clusters, in which these nodes were fully interconnected with each other by operational relationships (links). A small number of these nodes had links to the nodes outside their operational clusters and became hubs in the network. The degrees of these hubs usually were much larger than that of other nodes in their operational clusters. As more and more new nodes entered in this way, nodes in the GSJ network became more and more heterogeneous in terms of degree over time. As a result, low degree nodes at the peripheral of operational clusters are far more abundant than highly connected hubs. Random node removal is more likely to destroy these peripheral nodes without affecting the main structure of the network (error tolerance). In contrast, preferential node removal of these hubs can drastically degrade the network by isolating operational clusters from the largest component (attack vulnerability).

The ineffectiveness of the real attacks, on the other hand, may be because highly connected hubs such as commanders and coordinators usually are more experienced and better protected than average terrorists, and thus are more difficult to apprehend. The probability of their being captured is lower than random chance. The possibility of their survival from authorities' counterattacks is much higher than the average.

2.6 Conclusions

In this research we studied the evolution of the GSJ network to uncover the survival mechanisms of a terrorist organization. We found that three factors may have contributed to the survival of the network: growth, scale-free topology, and the ineffectiveness of the counterattack measures. The network experienced three distinct stages of growth from 1989 to 2003: emerging stage, maturing stage, and disintegrating stage. The network displayed different growth patterns in different stages.

We also found that the scale-free topology could partly account for the network's robustness, helping the network survive under constant counterattacks from authorities. The scale-free topology gradually emerged as new members joined in on an operational basis and the hubs acquired connections over time. On the other hand, the network could have remained active after numerous arrests of its members because the damages were localized within operations to a large extent. In addition, the leaders in the network are difficult to capture or remove and continue to function as hubs connecting members. Although numerous arrests and counterattacks have weakened the network, it still remains functional and has the potential to grow.

Note that findings in this study were obtained based only on the dataset about the Global Salafi Jihad and may not be generalized to other terrorist organizations. In addition, because of the possible data problems mentioned earlier, the results need further validation and verification. At this point we cannot draw definitive conclusions about the

exact means by which terrorist organizations have survived over time. More reliable data and further research are needed to gain deeper insights into the underlying mechanisms for the survival of terrorist networks.

CHAPTER 3: IDENTIFYING SIGNIFICANT FACILITATORS OF SOCIAL NETWORK EVOLUTION

3.1 Introduction

The notion of social networks and the methods of social network analysis (SNA) have received great interest from the information science community in recent years. The use of computer systems in the form of social networking and bookmarking websites (Mike 2008), cell phones and other mobile smart devices, and intra-organization virtual collaboration methods (Caroline 2006) have provided rich data sources for studying various large-scale social networks. There is a great need to understand the impact of these social networks and the extent to which they are facilitated by various social and technological factors. In this study, we focus on the facilitators that lead to link formation in social networks. These facilitators may be attributes of the individuals involved (e.g., age, gender), their network contexts (e.g., mutual acquaintances), and other kinds of nodes (e.g., use of vehicles, use of communication technology). We use an example of a dark network of criminals in this study and utilize novel methods to study its facilitators.

Illegal activities such as drug trafficking, money laundering, and terrorist attacks are usually conducted by individuals operating in networks (Chen 2006). These illegal and covert social networks are referred to as "dark networks" (Raab et al. 2003). Like "bright" organizational networks, dark networks also serve as communication, collaboration, and coordination mediums to better achieve goals. The networks consist of

a set of nodes/actors (e.g., criminals, terrorists) and links/relationships among them (e.g., crimes, terrorist operations, kinship). For instance, a network may consist of criminals who committed a bank robbery as nodes and their co-offending relationships in that robbery as links.

Networks evolve over time with the addition and removal of nodes and links. Often, the evolution is influenced by external factors. For instance, a terrorist network may grow in a conducive environment like a failed nation state with ample resources. A social networking-site friend network might grow with the addition of an instant messaging application. Previous studies in sociology (Kossinets et al. 2006; McPherson et al. 2001) have shown that the link formation process in such social networks is influenced by a set of facilitators. These facilitators may be individual attributes like age, race, and gender (Feld 1982; McPherson et al. 2001; Mike 2008; Reiss 1986) of the nodes/actors in the network or shared affiliations between actors like kinship and mutual acquaintances (Backstrom et al. 2006; Kossinets et al. 2006).

Previous studies on the facilitators of dark network evolution have primarily viewed them from a qualitative standpoint with few empirical evaluations (Coles 2001; McPherson et al. 2001; Milward et al. 2006; Sarnecki 2001). In this study, dynamic social network analysis (SNA) methods are used to examine several plausible link formation facilitators in a large real-world narcotics network. We use multivariate survival analysis using Cox

regression and a two-proportion z-test on the network cyclic and focal closure to identify significant facilitators. This study aims to answer the following questions:

- What are the facilitators of the link formation process in an evolving social network?
- How can we quantitatively identify the significant facilitators that influence the link formation process in a social network?

The knowledge of significant facilitators can be used to intervene in the formation of networks and focus efforts to encourage or discourage the formation of future links. The methodology used can also be generalized to study the effects of various facilitators on networks of individuals and machines. The remainder of the paper is organized as follows: Section 3.2 reviews previous literature on the study of networks and dynamic SNA methods. Section 3.3 introduces the testbed for this study. The research design is described in Section 3.4. Experimental results are presented and discussed in Section 3.5. In Section 3.6, we conclude and suggest directions for future work.

3.2 Literature review

SNA has been used to study various real-world networks including dark networks (Krebs 2001; Raab et al. 2003; Sarnecki 2001; Yang et al. 2007). There are mainly two kinds of SNA studies. One focuses on the static topology of social networks where the structural properties of the nodes and links are examined to describe and explain how network

topologies affect the functions and behaviors of complex systems (Albert et al. 2002). Such SNA studies have mainly focused on two types of social networks: 1) communication networks such as Internet (Albert et al. 1999), email (Newman et al. 2002) and phone (Abello et al. 1999) networks; and 2) collaboration networks such as co-authorship networks (Barabási et al. 2002; Newman 2001a) and movie-actor networks (Watts et al. 1998). However, these studies have largely ignored the dynamic network processes such as link formation.

Another line of SNA research has thrived by study various dynamic network processes and the mechanisms and determinants behind those processes. Such dynamic social network analyses mainly use statistical methods to model different network processes. These models are then tested to account for the structural changes of network topologies. The network processes studied include formations of friendship links (Leenders 1996; Snijders et al. 2007), collaboration links (Kaza et al. 2009; Lomi et al. 2006; Nerkar et al. 2005; Powell et al. 2005), and communication links such as emails (Kossinets et al. 2006) and phone calls (Palla et al. 2007).

The main issues with dynamic SNA include network recovery techniques, appropriate network measures, and statistical analysis methods for analyzing evolving networks. These issues are discussed in the following sub-sections.

3.2.1 Dynamic social network analysis

This section reviews the three major issues of dynamic SNA (Kossinets et al. 2006): network recovery, network measurement, and statistical analysis. There are various techniques to address these issues. The drawbacks and advantages of these techniques are also discussed.

3.2.1.1 Network recovery

Recovery is the process by which multiple instantaneous network representations are recovered from longitudinal data to model an evolving network. Recovery techniques can be classified into discrete and continuous according to their analytically different perceptions of time (Moody et al. 2005). Discrete network recovery techniques take multiple cross-sectional snapshots of the network at fixed points in time. However, these snapshots usually do not capture the processes that lead to network evolution. Most previous dynamic SNA studies (Leenders 1996; Leskovec et al. 2005; Xu et al. 2004) used the discrete technique due to the lack of methods for continuous network recovery and the high computational complexity for the few methods that existed.

Some recent dynamic SNA studies (Kossinets et al. 2006; Moody et al. 2005) have used continuous recovery techniques to extract multiple instantaneous networks from longitudinal data. These instantaneous networks account for the processes that lead to changes in link structure. In order to model the evolution in an efficient manner, Moody

et al. (2005) proposed a method to aggregate sequential events into larger time units. Based on this idea, Kossinets and Watts (2006) proposed a sliding window filter to smooth the link formation process. The time-span of the sliding window (called the relevancy horizon) defines which past events are relevant to the current state of the network. In other words, it is the amount of time in which the formation of the network stabilizes and thus past events can be considered as evolutionary events leading to the current state.

In order to determine the time interval between two consecutive instantaneous networks (called the sampling period), Kossinets and Watts (2006) suggested applying the Nyquist sampling theorem (Oppenheim et al. 1989) to the maximum rate of link formation. The theorem states that the continuous link formation process can be captured by a sampling period $\delta \leq c/(2.f_{max})$, where c is the number of independent clusters in the network and f_{max} is the maximum frequency of link formation.

The sliding window filter combined with the sampling period provides a novel and efficient method to recover the network. These techniques have a clear advantage over the random sampling periods used by previous studies that do not capture actual link formation processes.

3.2.1.2 Network measurement

Most empirical studies on longitudinal data plot descriptive measures over time to describe network changes. There are three main categories of descriptive measures used in dynamic SNA: deterministic measures, probabilistic measures, and temporal measures.

3.2.1.2.1 Deterministic measures

Deterministic measures include statistics like network size, measures based on the number of links (like degree, average degree) , closeness (Bavelas 1950), and betweenness (Freeman 1977). There have been various studies that have used these measures for static network analysis. Albert and Barabasi (2002) provide a comprehensive review.

A few dynamic SNA studies (Barabási et al. 2002; Leskovec et al. 2005) have also used these methods to measure network evolution. Barabási et al. (2002) studied the changes in the number of links, nodes, and average degree in a scientific co-authorship network. They found the network was growing since all three measures were constantly increasing during that time period. Leskovec et al. (2005) discovered that several evolving networks became denser over time by identifying their increasing average degrees. In the dark network problem domain, Xu et al. (2004) studied an evolving criminal network by analyzing and visualizing changes in degree, betweeness, and closeness.

However, these studies determined the measures over discrete (and sometime arbitrary) instances of time. Thus, the studies may not model the actual evolutionary changes in the network.

3.2.1.2.2 Probabilistic measures

The two most commonly used probabilistic measures are degree distribution and clustering coefficient. Nodes in a network have different number of links connecting them. Degree distribution is the probability (*P(k)*) that a randomly selected node has exactly *k* links (Albert et al. 2002). The degree distribution is used to classify networks to different topologies like random (Erdos et al. 1960) and scale-free (Barabási et al. 1999). The degree distribution can also be used to explain dynamic processes like growth and preferential attachment in scale-free networks (Barabási et al. 1999).

Clustering coefficient is the probability that two nodes with a common neighbor also link to each other (Newman et al. 2006). The measure was introduced to determine the small-world nature of a network. In a dynamic analysis study of scientific collaborations (Barabási et al. 2002), the clustering coefficient was found to decay over time, signifying that the network became less interconnected.

3.2.1.2.3 Temporal measures

Temporal measures have been developed to describe continuous network processes by taking the time variable into consideration. The triadic closure (Rapoport 1953) is defined as the empirical probability that two unconnected nodes (at time t) with a common neighbor will form a new link at time $t + \delta$. Intuitively, the triadic closure is a clustering coefficient within a time period (t, $t+ \delta$). A general form of the triadic closure is called cyclic closure (Kossinets et al. 2006), where the two nodes (i and j) may not share a neighbor but can be a certain geodesic distance ($d_{ij}>2$) apart.

Another temporal measure known as focal closure (Kossinets et al. 2006) defines the probability that two previously unconnected nodes that are some distance apart and share one or more affiliations will form a new link. Thus, the focal closure is essentially similar to the cyclic closure, however, in this case the probability is also influenced by the presence of an affiliation. If the focal closure of a network is consistently larger than the cyclic closure then it can be assumed that the link formation processes in a network are influenced by selected shared affiliations (Kossinets et al. 2006).

A drawback of the focal closure is that it cannot be used to study the effect of individual attributes (like gender or race). This is because by definition it measures the temporal changes due to shared affiliations like mutual acquaintances or inmates affiliations (shared affiliations are also known as interaction focuses, hence the name *focal* closure). However, individual attributes like race and gender are static over time, thus their effect cannot be measured by focal closure. Even so, these temporal measures provide distinct

advantages over deterministic and probabilistic measures. Even though the degree distribution can be used to model growth (using preferential attachment), it cannot be used to quantify the influences of network facilitators. The cyclic/triadic closure subsumes the clustering coefficient and enhances it to add temporal information. In this study, we use focal and cyclic closure for measurement. To the best of our knowledge, these measures have also never been used in the dark networks problem domain.

3.2.1.3 Statistical analysis

In network analysis, statistical methods are usually used to explain the emergence of network topologies (like random, scale-free, and small-world networks) (Albert et al. 2002); However, in this study we focus on statistical methods that have been used to identify significant link formation facilitators. A study by Leenders (1996) used a continuous-time Markov model on longitudinal network data where nodes were individuals and links were friendship between them. The study found that the gender affiliation significantly affected the link formation between children. However, the continuous time model used in this study assumed that only the state of the network at time *t-1* affects the current state (at time *t*). This may not be a valid assumption for most real-world networks.

Snijders (1996; 2001) developed a class of actor-oriented models to explain network evolution. The models are based on the assumption that nodes adjust their positions in the network based on certain parameters. However, their model assumes that the nodes are

aware of their positions with respect to the whole network. This assumption may not be true in dark networks since they are covert in nature. In addition, Snijders (2004) also proposed to use the independent arc model and the reciprocity model to represent different network effects in evolving networks. Together with the actor-oriented model, these three models are all based on the assumption that the observed networks are outcomes of a Markov process evolving in continuous time.

In order to handle networks with missing information, Carley et al. (2003) developed the meta-matrix approach to combine a set of networks of people, groups, knowledge, resources, events, or tasks to predict behavior. In a recent study, Kossinets and Watts (2006) used Cox regression analysis to identify significant facilitators in a university campus email communication network. They found that the mutual acquaintance and shared class affiliations (among others) had a statistically significant effect on future link formation. A similar survival analysis approach was also used by Nerkar and Paruchuri (2005) to determine that network centrality of inventors had a statistically significant effect on the intra-firm citation of their patents. The survival analysis approach lends itself well to the dark network problem domain since it does not make any assumptions about the underlying network.

Kossinets and Watts (2006) also compared the focal and cyclic closure to determine if shared affiliations play a role in link formation. However, they did not use a statistical

significance test in the comparison. In this study, we propose to use the two proportion z-test to compare the two measures.

3.2.2 Dark network studies

Two main streams of research can be identified in dark network literature: 1) the study of link formation facilitators and 2) the use of statistical methods to measure existing networks.

3.2.2.1 Link formation facilitators

Most studies on link formation facilitators in dark networks have been done in the fields of sociology and criminology. Raab and Milward (2003) studied organizational changes in two real-world dark networks: the Al-Qaeda terrorist network and the Columbian cocaine trafficking network. They found that a set of facilitators motivated individuals to form or deactivate links. A later study by the same group (Milward et al. 2006) suggested that prison might be the most effective facilitator for drug trafficking. They contended that individuals who are jailed in close proximity to each other are likely to form future links in the network (i.e., co-offend in future crimes).

Various criminology studies (Reiss 1986; Reiss et al. 1991) have suggested that individual attributes like age and race play important roles in co-offending. Such attribute-based homophily has also been suggested by social network studies in other

domains (Louch 2000; McPherson et al. 2001). A survey on male criminals in London showed that they were more likely to co-offend with individuals of the same age (Reiss et al. 1991). In addition, the same study found that co-offending by criminals of different races was rare.

A study on the dynamics of delinquent groups (Warr 1996) found that similarity in gender had a mixed effect on relationship formation. Male youth offenders generally followed males. Older females were more likely to co-offend with other males, whereas, younger females were likely to co-offend with members of the same sex. Propinquity of living/working was also considered as a significant factor leading to co-offending (Milward et al. 2006; Reiss et al. 1991).

However, most of the above studies use small-scale datasets that usually involve a few hundred criminals over short time periods. Moreover, the findings are based on basic descriptive statistics instead of systematic statistical and SNA methods.

3.2.2.2 Statistical Analysis of Dark Networks

There have been a few recent studies that have explored the use of SNA methods for dark networks. Most have focused on assigning roles to actors in the network. Sparrow (1991) explored the use of centrality measures to identify key actors in criminal networks. Following this approach, Krebs (2001) used centrality measures to identify the group leader of the September 11th hijackers. Another terrorist network study calculated the

average degree of the Jemaah Islamiyah terrorist network (Koschade 2006) and uncovered that the 2002 Bali bombing cell had a high density that allowed it to sustain member losses. Xu and Chen (2004) used SNA methods to determine the leader and gatekeeper role for individual nodes and used hierarchical clustering methods to identify sub-groups in criminal networks.

By primarily concentrating on the node properties, none of these SNA studies have focused on the facilitators of network link formation in dark networks.

3.3 Research testbed

The testbed for this study was consolidated from two related real-world crime related datasets: police incident reports from Tucson Police Department (TPD) and inmate information from the Arizona Department of Corrections (ADOC).

Tucson Police Department incident reports contain information on 2.03 million individuals and 1.34 million vehicles involved in illegal activity in the Tucson, Arizona metropolitan region (a population of about 1 million) from 1990-2005. This comprehensive dataset is representative of typical crime incident databases of mid-size cities in the United States. The dataset was used to extract a narcotics network with individuals as nodes and crime incidents as links. Individuals were included as nodes if they were wanted, suspected, or arrested for narcotics or related crimes. We also extracted information on individual attributes like age, race, and gender. Two individuals

in the network were connected by a link if they were in the same incident report involving a narcotics crime (such as sale or possession of drugs) or a narcotics related crime (such as homicide, aggravated assault or armed robbery). The time of the incident was also extracted which was used to trace the evolution of the network.

Six subsets of this narcotics network were used in this study. They were limited to two or three year time period since the domain experts suggested that a criminal cannot be considered as a part of the network if he/she does not commit a crime for more than 3 years. The data from 1990 to 1992 was not used because it contains much less nodes and links than other datasets due to missing data. The size of these subsets was comparable or bigger than previous network analysis studies (Leskovec et al. 2005; McPherson et al. 2001; Xu et al. 2005; Xu et al. 2004) and also allowed for efficient computational times. Table 3.1 lists the key statistics for the subset.

Table 3.1. Key statistics of network subsets

Time period	Number of nodes	Number of links
1993-1994	3796	4560
1995-1996	4588	5356
1997-1998	5232	6934
1999-2001	5140	6733
2002-2003	5076	7135
2004-2005	4101	5693

The Arizona Department of Corrections dataset contains information (such as names, date of births and inmate housing facility) for 165,540 jailed individuals in Arizona from 1986 to 2006. 43,220 individuals in this dataset were also found in the Tucson Police Department dataset. This overlap was used to extract the inmate affiliation between individuals in the network subsets. Two individuals were considered to have an inmate affiliation if they were housed in the same facility during the same time period.

3.4 Research design

Figure 3.1 shows the research design for identifying significant link formation facilitators of a dark network. The process consists of four components. The first component, facilitator identification, involves selecting the link formation facilitators that need to be tested for significance. These facilitators may be selected based on previous studies or theoretical conjectures on dark networks. The second component, network recovery, contains methods to extract instantaneous networks from longitudinal network data. A set of instantaneous networks represents the evolution of the dark network. The third component, network measurement, involves calculating SNA measures like focal and cyclic closure. The last component, statistical analysis, involves identifying the significant facilitators from individual attributes and shared affiliations using multivariate survival analysis and two-proportion z-test. The details of the design are introduced in the following sections.

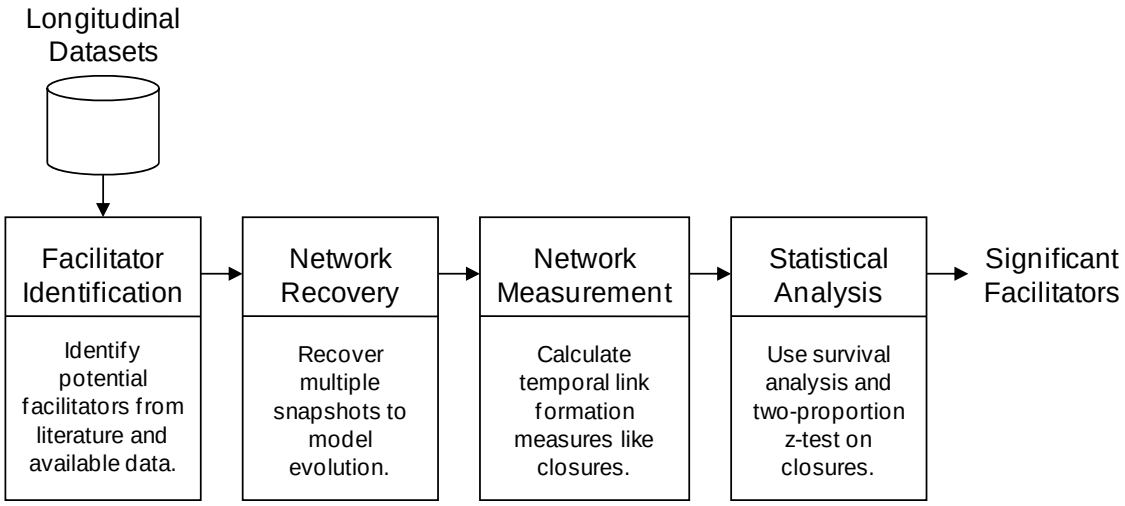


Figure 3.1. Research design - identifying significant link formation facilitators

3.4.1 Facilitator identification

In this study, the facilitators included three individual attributes and five shared affiliations. These were selected based on previous criminology and sociology studies in this domain.

3.4.1.1 Homophily in individual attributes

The individual attributes tested were homophily in age (Reiss 1986; Reiss et al. 1991), race (Reagans 2005; Reiss et al. 1991), and gender (Reiss 1986; Warr 1996). For the statistical analysis the homophily variables were operationalized as follows:

- Age: The age was set to '1' if the difference in two individuals' age was less than or equal to 1 year, and '0' otherwise.

- Race: The testbed contained five possible categories of ethnic origin: Caucasian, African American, Hispanic, Asian and Native Indian. The race was set to '1' if the individuals had the same ethnic origin and '0' otherwise.
- Gender: The gender was set to '1' if the individuals had the same gender and '0' otherwise.

3.4.1.2 Shared affiliations

The shared affiliations tested were mutual acquaintance, inmate affiliation, vehicle affiliation, phone affiliation, and residential address affiliation. The first three affiliations were included since they are suggested by previous studies (Coles 2001; Milward et al. 2006). The phone affiliation and residential affiliation were included since they indicated that the individuals worked or lived close to each other. For the statistical analysis these variables were operationalized as follows:

- Mutual acquaintance affiliation: The variable was set to one less than the number of common neighbors two individuals had (Kossinets et al. 2006).
- Inmate affiliation: set to '1' if the two individuals were housed in the same prison facility in the same period, and '0' otherwise.
- Vehicle affiliation: set to '1' if the two individuals were found to be associated to the same vehicle in different police reports and '0' otherwise. If two individuals were related to a vehicle in the same police report then this was not considered as a vehicle affiliation since such individuals would be directly linked.

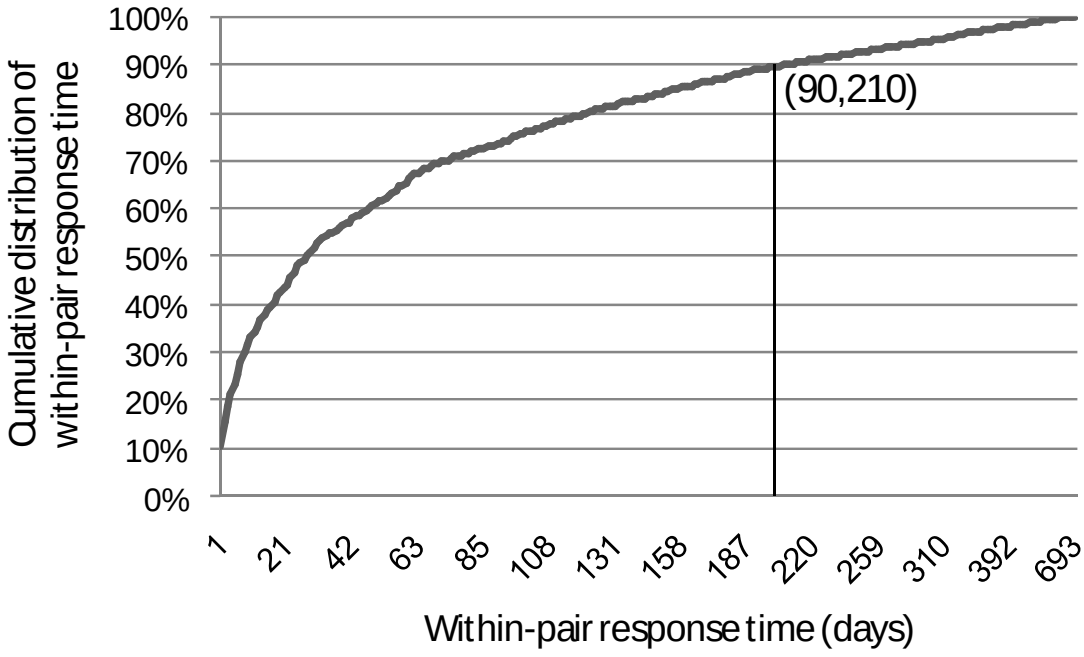
- Phone affiliation: set to '1' if the individuals had the same work or home phone number associated with them and '0' otherwise.
- Residential address affiliation: set to '1' if two individuals live in the same residential grid in Tucson. The Tucson city region is divided into 0.5 sq. mile administrative grids by Tucson Police Department. The residential grid can be derived from the address of an individual.

Not all the individuals in the dataset have all the information (e.g., phone or address) associated with them. In case an individual did not have data then the corresponding variable was set to zero.

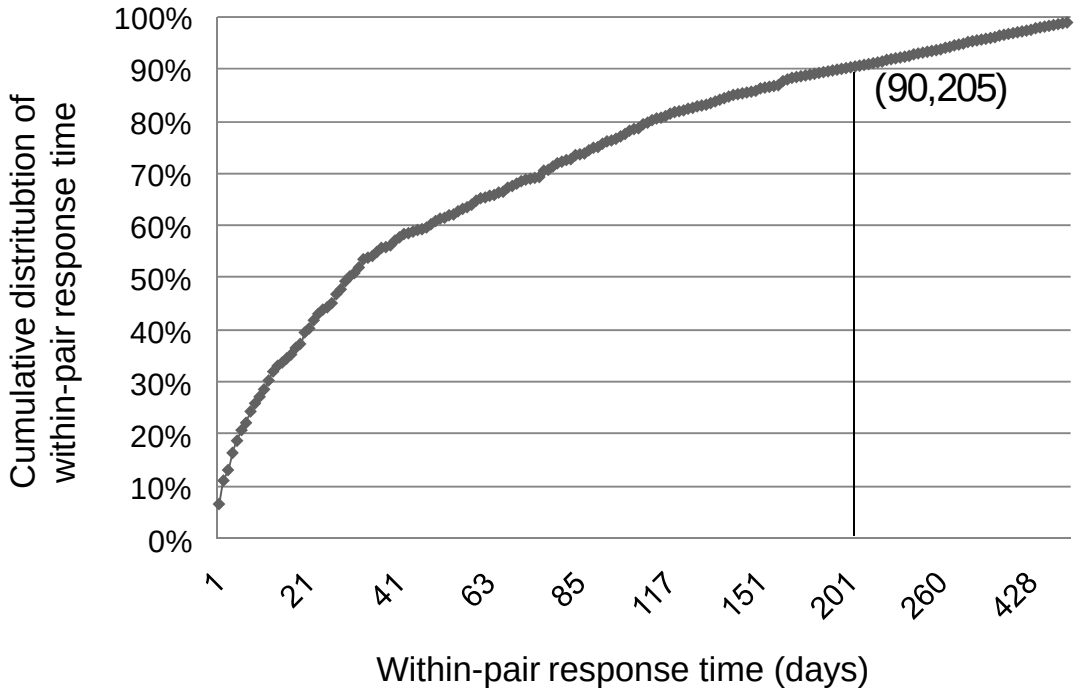
3.4.2 Network recovery

Recovery is the process by which multiple instantaneous network snapshots are recovered from longitudinal data to model the evolving network. We used the relevancy horizon (τ) and sampling period (δ) – described in the research background section – to recover the network. We used $\tau = 210$ days because the rate of the formation of new links between each pair of nodes stabilizes after approximately 210 days. This was determined using the within-pair response time t_{ij} between two individuals in the network (Kossinets et al. 2006). In this dataset, the within-pair response time is defined as the time interval between two subsequent police reports involving the same pair of criminals.

Figure 3.2(a) and 3.2(b) show the cumulative distribution of within-pair response time for the 2002-03 and 2004-05 subset respectively. We use these two datasets as examples to demonstrate how to estimate the relevancy horizon. On the X-axis, the within-pair response time (i.e., the number of days between subsequent co-offending) is plotted. The Y-axis shows the cumulative distribution of the response time. The graph can be used to determine the time-period within which most co-offending takes place. As shown in the figures, 90% of subsequent co-offending happened within $t_{90\%}$ = 210 days in the 2002-03 subset and $t_{90\%}$=205 days in the 2004-05 subset.



(a)



(b)

Figure 3.2. Cumulative distribution of within-pair response time for the (a) 2002-03 subset and (b) 2004-05 subset.

The sampling period was determined using the Nyquist theorem ($\delta = c/2.f_{max}$) as described in Section 3.2.1. For the network under study, c was approximated to be 1,794 and f_{max}=19 links/day, so δ was found to be equal to 47 days. Thus, the evolution of the network was modeled by $(730 - \tau)/\delta$ =11 instantaneous networks recovered from the dataset of 730 days (2 years).

3.4.3 Network measurement

Once instantaneous networks were recovered from the subset, we calculated the focal and cyclic closures. They were calculated using (Kossinets et al. 2006):

$$P_{new}(d_{ij}, s_{ij}) = \sum_{n=1}^{(T-\tau)/\delta} M_{new}(d_{ij}, s_{ij}, n) / \sum_{n=1}^{(T-\tau)/\delta} M(d_{ij}, s_{ij}, n)$$

where d_{ij} is the shortest path length between individuals i and j, and s_{ij} is the number of shared affiliations between them. $M(d_{ij}, s_{ij}, n)$ is the number of individual pairs who have s_{ij} shared affiliations and whose shortest path length is d_{ij} for the n^{th} recovered instantaneous network. $M_{new}(d_{ij}, s_{ij}, n)$ is the number of newly formed links that have the same d_{ij} and s_{ij} since the $(n-1)^{th}$ instantaneous network. $P_{new}(d_{ij},s_{ij})$ is the probability that a new link will form between any two previously unconnected individuals i and j who are d_{ij} distance apart and have s_{ij} shared affiliations.

The focal closure is $P_{new}(d_{ij},s_{ij})$ when $s_{ij} > 0$, while the cyclic closure is $P_{new}(d_{ij},s_{ij})$ when $s_{ij} = 0$. The triadic closure is then represented by $P_{new}(d_{ij}=2)$. The triadic closure is a special case of the cyclic closure that measures the probability of a new link formation between two previously unconnected individuals whose shortest path length $d_{ij} = 2$.

3.4.4 Statistical analysis

In this study, two methods were used to identify significant facilitators: (1) a comparative two-proportional z-test on the focal and cyclic closures and (2) a multivariate Cox

regression. As mentioned before, a disadvantage of the focal closure is that it can be calculated on affiliations and not individual attributes. Thus, it cannot be used to identify significant attributes. On the other hand, the Cox regression can be used to examine the significance all facilitators (including individual attributes and affiliations).

3.4.4.1 Two-proportional z-test on focal and cyclic closures

The focal and cyclic closures for each shared affiliation were compared to each other. Kossinets and Watts (2006) suggest that a higher focal closure (as compared to cyclic) between a pair of nodes indicates that an affiliation increases the probability of a link forming between them. We build on their study by incorporating statistical significance to test the difference in the values of focal and cyclic closure. Since the focal and cyclic closures deal with two different populations (one with affiliations and one without), the two-proportion z-test can be used to compare them. The test provides a greater validity to comparison and helps identify significant affiliations.

3.4.4.2 Multivariate Cox regression

There are two major types of survival regression models: log-linear and Cox regression models. Using goodness of fit tests and checking for proportionality assumptions, we found that the Cox regression model fit our data well. We used Cox regression model of the form (Afifi et al. 2003):

$$h(t, x_1, x_2, x_3 \ldots) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots)$$

where $h(t,x_1,x_2,x_3...)$ is instantaneous hazard - the probability that the event will happen at time *t,* given that the event has not happened up until time *t* with the observations of various independent variables $(x_1, x_2, x_3...)$. The event in this network is that two previously unconnected nodes with $d_{ij} = 2$ subsequently form a new link. In order to minimize possible correlations among the pairs of nodes with $d_{ij} = 2$, each node was included in only one pair for the analysis. The facilitators described in the previous section were the eight independent variables in the regression. On running the regression, the significance of each of the variables can be ascertained.

3.5 Experimental results

Table 3.2. Results of two-proportional z-tests on focal and cyclic closures

Time Period	Geodesic Distance	*p*-value			
		Inmate	Vehicle	Phone	Address
1993-94	2	>0.999	0.0041*	>0.999	>0.999
	3	0.9977	0.9977	>0.999	0.9977
1995-96	2	0.0823**	0.3336	>0.999	>0.999
	3	>0.999	0.0179*	>0.999	>0.999
1997-98	2	>0.999	0.0594**	>0.999	>0.999
	3	0.6628	0.0068*	>0.999	>0.999
1999-2001	2	0.8289	0.0014*	0.0059**	>0.999
	3	>0.999	>0.999	>0.999	>0.999
2002-03	2	>0.999	0.012*	>0.999	>0.999
	3	0.641	0.002*	0.058**	>0.999
2004-05	2	0.063**	0.004*	0.999	>0.999
	3	0.242	0.022*	>0.999	>0.999

Note: * $p < 0.05$, ** $p < 0.10$

Table 3.2 shows the results of the comparative z-test on the difference of focal closure and cyclic closure for the shared affiliations. We calculated the measures for pairs of nodes with geodesic distance (d_{ij}) of two and three. This was based on domain expert feedback which indicated that only individuals two or three crime-based links away from

a pair are likely to have an effect on their future link formation. In addition, our previous study (Kaza et al. under review) indicated that almost the entire narcotics network in our dataset could be reached within an geodesic distance ranging between four and five. Thus, using d_{ij} = 4 would have included almost all the possible node pairs in the network leading to impractically large computation times. The focal closure cannot be calculated for $d_{ij} = 2$, since by definition all pairs of individuals with $d_{ij} = 2$ shared at least one mutual acquaintance. Thus, focal and cyclic closure will be exactly the same for such individuals. In addition, at $d_{ij} = 3$, no pairs of individuals will have mutual acquaintances.

As can be seen from Table 3.2, the focal closure with the vehicle affiliation was found to be significantly larger than the cyclic closure for d_{ij} = 2 and d_{ij} = 3 for all subsets. This finding suggests that two previously unconnected criminals who were linked to a vehicle through different crimes are much more likely to co-offend in future crimes than criminals who do not have the vehicle affiliation. The inmate and phone affiliation were found to be mildly significant for two subsets, showing some support for the theory suggested by Milward and Raab (2006). Sections 3.6.1 and 3.6.2 contain more discussion on the implications of significant and insignificant facilitators.

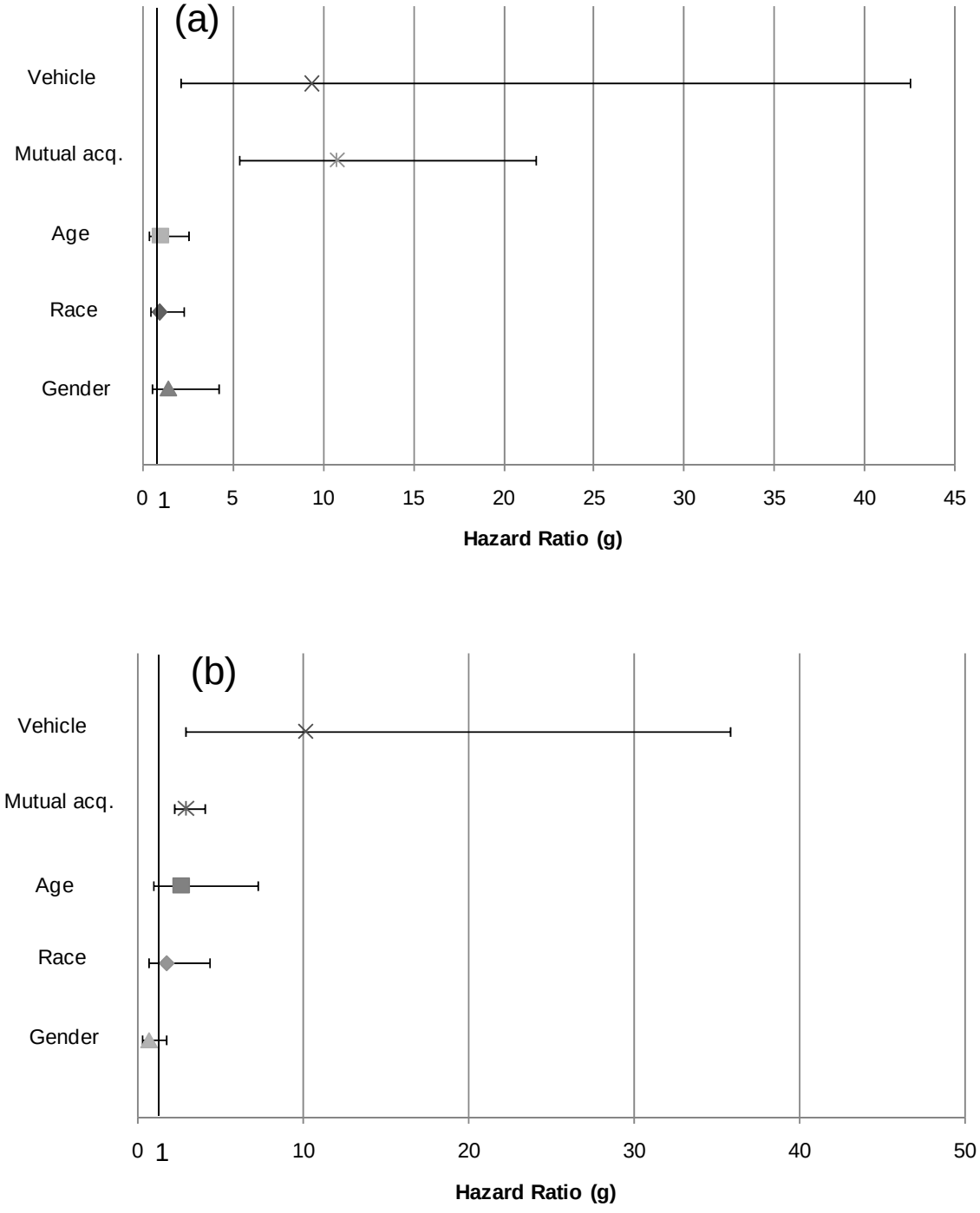


Figure 3.3. Results of multivariate survival analysis (Cox regression) of triadic closure for pairs of individuals from (a) 2002-03 and (b) 2004-05

Figure 3.3 shows the hazard ratios and their 95% confidence intervals for the independent variables examined for the two most recent network subsets. Each facilitator was represented by an independent variable, and the probability of the triadic closure would increase by a factor of hazard ratio (g) when the corresponding independent variable increases by one unit. The independent variables were considered to be significant only when the confidence intervals did not include the value 1, since a hazard ratio of 1 indicates that the independent variable has no effect on the dependent variable. Three variables representing inmate affiliation, phone affiliation, and residential affiliation were dropped from the Cox model due to collinearity. The confidence intervals in Figure 4 show that the vehicle and mutual acquaintance affiliation were found to be significant with hazard ratio ranges not including the value 1.

Table 3.3. Results of the Cox regression on facilitators

Time Period	Number of Nodes	*p*-value				
		Vehicle	Mutual acq.	Age	Race	Gender
2002-03	5,076	0.004*	<0.001*	0.995	0.943	0.492
2004-05	4,101	<0.001*	<0.001*	0.046*	0.212	0.514

Note: * $p < 0.05$

In addition, the p-values for each of the variables for the two most recent subsets are also shown in Figure 3. These results were consistent with the results of the two-proportional

z-test. The age, race, and gender were found to be insignificant facilitators of link formation. These results are discussed in the following sub-sections.

3.6 Discussions

3.6.1 Significant facilitators

3.6.1.1 Mutual acquaintance

The two-proportion z-test and the Cox regression found the mutual acquaintance and the vehicle affiliation to be significant link-formation facilitators. A significant mutual acquaintance affiliation implies that two previously unconnected individuals are likely to commit a crime together if they have committed crimes with one or more shared acquaintances before.

This facilitator has been well studied in sociology (Kossinets et al. 2006; McPherson et al. 2001) and criminology (Coles 2001) and studies from both domains have found that individuals tend to select new acquaintances who are 'friends of friends.' In the network under study, this social mechanism also suggests that criminals operate in groups of close acquaintances and are likely to form operational cliques.

According to domain experts and our previous study (Kaza et al. under review) this phenomenon is not unusual in narcotics networks, where individuals tend to have circles of trust that include friends and family members. These operational cliques enhance communication within the network and increase the capacity to act.

This phenomenon is also in line with the social closure theory (Coleman 1990) which suggests that the greatest value is obtained from networks that are densely connected with a high level of trust among actors. This property of criminal networks is advantageous to law enforcement because it helps them form strong conspiracy cases against members of the group. Conspiracy cases generally remove more criminals from the street for a longer time as compared to individual convictions.

3.6.1.2 Vehicle affiliation

A significant vehicle affiliation variable implies that two criminals who have used the same vehicle for different crimes before are likely to co-offend in the future. Though this affiliation has not been studied in previous dark network studies, domain experts suggested that individuals involved in narcotics crimes often use certain vehicles. These vehicles are common to a particular gang or may have been stolen for specific purposes like load vehicles (vehicle carrying narcotics) or scout/lookout vehicles. We included the vehicle affiliation in this study with the intuition that the affiliation may point to hidden social and operational links between two previously unrelated individuals. The statistical significance of the affiliation suggests the importance of including two-mode information in social networks. We believe that it is especially important to include such affiliation information in networks where relationships in one-mode data may be missing or incomplete. In the future, we plan to explore the use of such affiliations in the construction of social networks.

In addition to identifying significant facilitators, the Cox regression can also be used to determine the scale of influence of each of the facilitators. For example, sharing the same vehicle in different crimes increases the probability of triadic closure by a factor of 9.38 and each additional mutual acquaintance increases it by a factor of 10.79. Therefore, if two unconnected criminals have used the same vehicle in different crimes and have five mutual acquaintances then they are $9.38^{1} \times 10.79^{(5-1)} \approx 127141.88$ times more likely to co-offend in the future as compared to those who do not share these affiliations. Such calculations can be used as a prediction mechanism to identify individuals who are very likely to be associated in the future based on their past activity.

3.6.2 Insignificant facilitators

3.6.2.1 Homophily in age

Many of the facilitators suggested by previous studies in this problem domain were found to be insignificant. The Cox regression analysis found homophily in age to be an insignificant factor in the link formation process. This finding implies that individuals who are in the same age group were not necessarily likely to co-offend in future narcotics crimes. A recent study found that the criminals tend to co-offend with other individuals of a younger age (Sarnecki 2001). Since, in this study the age variable was defined based on similarity, this may have been the cause for the insignificance.

3.6.2.2 Gender affiliation

Similarly, the gender affiliation may be dependent on the age of criminals. Female criminals may tend to co-offend with the same sex at a younger age and then co-offend with males as they get older (Warr 1996). A change in the method of operationalizing the age and gender variables may lead to different results.

3.6.2.3 Race affiliation

The race affiliation was also not found to be significant, in contrast to previous studies (Reiss et al. 1991). We believe that this might be because Tucson is an immigrant city (about 1 hour from the southern border) where individuals of low social-economic status live in the same vicinity as immigrants of a different race. This may lead to criminal links being formed across race boundaries and thus similarity in race may not be a good predictor of future activity. This has also been suggested by a previous study dealing with crime in an ethnically mixed environment (Sarnecki 2001).

Even though the insignificance of individual attributes is contradictory to previous research suggesting attribute-based homophily (Louch 2000; McPherson et al. 2001), it is possible that their effect was seen through other variables. For instance, similarity in race and age may be reflected in the mutual acquaintance affiliation. Thus, individuals having the same age group and race might have the same mutual acquaintance which, in turn, affects their link formation likelihood.

3.6.2.4 Inmate housing affiliation

Previous research (Milward et al. 2006) suggested that prison is an ideal place to develop future co-offending relationships for criminals. However, we found that inmate affiliation was insignificant for co-offending in the network studied here. This may be attributed to the inmate custody classification system implemented by Arizona Department of Corrections that attempts to separate inmates with previous affiliations (e.g., same gang membership, same criminal records) to different housing facilities. Therefore, inmates in the same housing facility that are screened by this system may have lower chance to co-offend in the future. This finding has important implications which can be used in the policy decisions in other correctional facilities.

3.7 Conclusions

In this paper, we used dynamic SNA methods to examine the facilitators of link formation in a real-world social network. We studied several possible facilitators including homophily, mutual acquaintances, and various affiliations. The results showed that mutual acquaintance and shared vehicle affiliations were significant facilitators in the network under study. Homophily in age, race, and gender were not found to have a significant effect on the link formation process in a narcotics network. We also quantified the influences of the facilitators on the triadic closure by using the hazard ratios of Cox regression analysis and used the information to calculate the likelihood of future co-offending. In addition, we examined the social causes and policy implications for the

significance and insignificance of various facilitators. The set of generic dynamic SNA methods along with the corresponding statistical analyses used in our study may be applied to other types of networks to test the effect various facilitators. This research may help the academic and practitioner community better understand the dynamics of social networks and devise effective strategies to influence their growth.

In the future, we plan to explore several directions including (1) studying the evolving social networks with multiple relationships, (2) extending this dynamic analysis to other real-world social networks, and (3) adding more potential facilitators into our analysis such as psychological and economical factors.

CHAPTER 4: DISCOVERING DETERMINANTS OF LINK FORMATION IN SOCIAL NETWORKS

4.1 Introduction

In recent years a new model of production has emerged in which self-organizing individuals voluntarily collaborated online with one another to produce goods and service ranging from software to encyclopedias (Yochai 2006). A phenomenal success of this production model is open source software (OSS) development whose flagships include Linux, Firefox, MySQL, and many others. The OSS development is characterized by its voluntary developers distributed globally who mainly collaborated through the Internet. This software development approach, facilitated by today's digital network environment, has provided new ways for business to collaborate and interact with customers in software development and maintenance. A well-known business success that partially adopts the OSS approach is one of the most popular first-person shooter games, *Half Life: Counter-Strike*, which was developed by a group of public users/developers by modifying the original game *Half Life*. To fully realize OSS developers' potential in software development, it is important to understand why they choose to participate in an OSS project.

In this paper, we aim to study how voluntary software developers choose OSS projects to participate in a social network, within which their participation choices are significantly

influenced by various social and technological factors. We defined these factors as determinants of OSS project participation choices.

While the OSS phenomenon is not new, research interest in its social network perspective has only recently emerged, as more and more developers choose to collaborate and interact with each other in a network form through online communities (Hahn et al. 2008). In such online communities, an OSS project can be viewed as a complex social network in which developers are nodes and the various social relationships among them are links. These OSS developer social networks evolve over time with new developers (nodes) participating in projects and forming new social relationships (links) with existing members. Thus the research question of our study is how to discover the determinants of participation links in OSS developer social networks. To answer this question, we examined a set of plausible determinants extracted from a large online OSS developer community using social network analysis (SNA) methods and conditional (fixed-effects) logistic regression (McFadden et al. 1974). In addition, we developed a computational mechanism based on the conditional logistic model (CLM) to predict future OSS developers' project participation choices. Such a mechanism can be used to recommend possible collaboration opportunities for developers.

By studying the determinants of OSS developers' participation choices (links) in social networks, this research aims to better understand the developers' participation choice behaviors which may have significant practical impact on OSS project success.

Theoretically, this study enables us to examine how various determinants embedded in the social networks will affect individual nodes' choices to form new participation links. As the similarity of individual attributes (i.e., homophily) may significantly influence individual's social relationships such as marriage and friendship (Leenders 1996; McPherson et al. 2001), or as the shared activities and affiliations of college students may affect their email communication relationships (Kossinets et al. 2006), the underlying social and technological factors embedded in the OSS developer networks may influence developers' choices about which project to participate in (i.e., forming a participation link).

Based on findings from both social network studies and OSS studies, we theorize two types of factors – homophily in individual attributes and shared affiliations - will significantly affect OSS developers' choices about which project to participate. In addition, by participating in a project, OSS developers choose to collaborate with the group of existing developers and form various new relationships with them, further changing the structures and contents of the OSS developer social networks. Therefore, studying the OSS developers' participation choices is a natural context for studying how social networks evolve with new links formed under the influences of various determinants. However, despite the relevance of OSS developer networks, few studies have empirically examined the determinants of OSS participation links.

From a practical point of view, studying the determinants of OSS developers' project participation choices from a social network perspective will have many implications for the increasing number of organizations, both commercial and non-profit, that adopt the OSS approach in their software development projects. Firstly, while the traditional propriety software development approach usually manages the developers with a hierarchy headed by firm/project managers to work together effectively, the OSS approach mainly utilizes the collective efforts from volunteer developers to achieve project successes through peer collaboration and coordination. One recent study (Grewal et al. 2006) found that individuals' project participation choices will affect the network embeddedness of OSS projects and developers, which has been shown to have significant effects on the technical and commercial successes of OSS projects.

Secondly, despite a considerable amount of studies focusing on the developers' motivations for contributing to OSS projects in general, there is a paucity of research that examines the developers' choices as to which project to participate in when similar alternatives are available (Hahn et al. 2008). Studying the determinants of developers' participation choices may provide insights for OSS projects on how to increase the chances of attracting new members. Such knowledge can be used by researchers and practitioners to devise useful strategies or design collaborative information systems for OSS project management.

In terms of research methodology, the computational approach we used in this study combines both social network analysis and conditional logistic analysis to discover the link formation process determinants in evolving social networks. This approach can be generalized to studies of social networks in other domains.

The remainder of this Chapter is organized as follows. In the next section, we develop our theoretical framework by integrating empirical findings from previous OSS network studies with theories from social network analysis and sociology that examined the mechanisms behind the network link formation processes. The third section introduces the dataset we used and outlines our research methods. Then we present the results and discuss their implications. We also propose a mechanism to predict developers' future participation choices using the discovered determinants. Finally, we conclude the Chapter and suggest directions for future research.

4.2 Theoretical basis and research hypotheses

To study OSS developers' choices about which project to participate in, we first need to understand why they voluntarily contribute in OSS development in general. As summarized in (Lakhani et al. 2005), the motivations for developers to participate in OSS development include 1) intrinsic motivations such as a sense of enjoyment and creativity in task accomplishment, and a desire for community identification and reputation; and 2)

extrinsic motivations such as payment, career advancement, and the need to improve programming skills.

Hahn et al., (2008) argue that the realization of the above goals for developers mainly depends on the success of OSS projects. Therefore, it is natural that developers want to choose to participate in the projects which are most likely to succeed, both in terms of the project's outcome and development process. However, it is difficult for the prospective developers to correctly predict the probability of future success of OSS projects, especially newly joined, inexperienced ones, due to the information asymmetry problem (Ba 2001) (e.g., incomplete information) in online communities. Without actually participating and contributing to OSS projects, prospective developers may not be able to get first-hand information and tacit knowledge about the true quality of the projects.

4.2.1 A social network perspective of OSS developers and project participation process

Social network perspective arises as an alternative approach to help evaluate the quality and likelihood of future success of OSS projects, by providing information and analytical insights about the existing project members and their relationships. Since an OSS project can be viewed as a group of people collaborating together to create a software product, developer participation in OSS projects is essentially a process in which an individual chooses to work with a group of developers and in doing so forms relationships through

virtual interactions over the Internet (von Hippel et al. 2003). As a result of continuous developer participation, these relationships emerge and exist in a network form and in turn influence OSS developers' behaviors such as project participation choices and peer evaluations.

Therefore, a set of studies has adopted a social network perspective to investigate various OSS phenomena (Crowston et al. 2003; Grewal et al. 2006; Hu et al. 2008a; Hu et al. 2008b; Jin et al. 2005; Madey 2002; Wagstrom et al. 2005). In those studies, a social network is modeled as a graph with nodes representing developers or projects and links representing various relationships between the nodes such as project participations, collaborations or peer evaluations. Two types of methods were often used in SNA research: one focuses on modeling the static topologies of social networks; the other focuses on modeling the network evolution with various dynamic network processes such as link formation process.

The static topological properties of OSS social networks were examined to explain how topologies affect functions and behaviors of networks (Crowston et al. 2003; Hu et al. 2008a; Jin et al. 2005; Madey 2002). Three types of models have been developed to characterize the topologies of complex networks: random graph model (Erdos et al. 1960), small-world model (Watts et al. 1998), and scale-free model (Barabási et al. 1999). Among them, the scale-free model is noteworthy in our study because many empirically observed networks are found to be scale-free networks, including OSS-related networks.

It features power-law degree distributions, which means a small fraction of the nodes have a large number of links while a big fraction of nodes have just a few. Madey (2002) conducted one of the first empirical investigations from a social network perspective on OSS developers from SourceForge.net. He modeled those developers working in projects as a collaboration network and found this network displays features of the scale-free topology. A more recent analysis (Jin et al. 2005) of SourceForge collaboration networks has also discovered similar scale-free features. Moreover, small-world network features – large clustering coefficient and small average path length – were also found in those SourceForge developer collaboration networks.

Another stream of SNA research focused on studying the dynamic network processes, especially the link formation processes. A social network grows over time with the participation of new nodes and the formation of new links. In OSS network studies, while topological analysis mostly just captures snapshots of social structures in OSS developer communities, studying the mechanisms behind the participation link formation process can provide insights about how developers choose to participate in OSS projects. Hahn et al. (2008) examined how a developer's choice to participate in an OSS project is influenced by his past collaborations with existing project members. They found that a developer is more likely to participate in a project if he has strong collaboration links with its initiator. In this case, past collaboration in OSS projects is a factor that significantly influences the formation of participation links. Such factors are defined as

determinants of the participation link formation process in (OSS) social networks (Hu et al. 2008b).

4.2.2 Determinants of participation link formation in OSS developer networks

As discussed earlier, the information asymmetry as well as the anonymous nature of online communities made it very difficult for prospective developers to evaluate the true quality of OSS projects and the programming and management skills of existing project members, thus largely hindering the predictions of future project successes. To address this problem, developers need to rely on other types of information as proxies to evaluate the underlying values of existing project members and the likelihood of successful collaboration with them. From the social network perspective, the determinants of link formation can be used to serve this purpose. Prior studies suggest two types of determinants – 1) homophily and 2) shared affiliations – may significantly influence individuals' choices in forming relationships/links (e.g., collaborations) with unfamiliar people when the outcome of such link formation is uncertain (Albert et al. 2002; Hu et al. 2009; Kossinets et al. 2006; McPherson et al. 2001). Therefore, we hypothesize that these two types of factors are the determinants of participation link formation in OSS social networks. We then examine how these factors influence OSS developers' project participation choices using empirical data. In the following sections, we introduce these factors and develop our hypotheses in further detail.

4.2.2.1 Homophily factors

Research on homophily theory suggests that people are more likely to interact and form relationships with individuals with whom they share similar attributes (Byrne 1971; McPherson et al. 2001; Yuan et al. 2006). When choosing new OSS projects to participate in, a developer may heavily rely on homophily in individual attributes between him and the existing project members. These attributes are defined as homophily factors in this study. According to Turner's (Turner 1987) self-categorization process in homophily theory, firstly the developers may self-categorize themselves and the existing project members in terms of age, gender, education and various other attributes. Then, based on these categories, they further differentiate existing members as similar or dissimilar. For instance, developer A will perceive existing member B to be more similar with him or her if B and A both use the same primary programming language – Java – in OSS development. For developer A, such interpersonal similarity will increase the predictability of B's behaviors in the OSS project and thus may enhance A's judgment on the probability of project success and willingness to participate.

Findings of various empirical research have also supported the homophily theory. In the sociology literature, Leenders (1996) used a continuous-time Markov model to study the determinants of link formation in a children's friendship network. The results showed that gender homophily (i.e., being the same gender) significantly affects the friendship link formation among children. Marsden (Marsden 1987) studied core discussion networks and found that Americans tend to discuss important issues with people who

have similar education levels. In addition, McPherson et al. (1987) discovered homophily in socioeconomic status in the friendship networks of voluntary organizations in Nebraska. Finally, in a position paper, McPherson et al. (2001) summarized that various social relationships, including marriage, work, advice, and friendship, are influenced by the homophily principle - similarity breeds connection. Consequently, many social networks become homogeneous in terms of various individual attributes.

This research also employed the theory of homophily to explain OSS developer's project participation choices. However, it contributes to current OSS studies and homophily research in several important ways. Firstly, past research mainly focused on personal factors such as age (Feld 1982) or gender (Leenders 1996), while our study explores homophily not only in personal factors but also in contextual factors in OSS developer communities, such as programming language and community status. This is mainly because in OSS communities developers have limited access to information about other members' social and personal characteristics due to online anonymity. When choosing OSS projects and developers to work with, they mainly rely on contextual information such as community ranking scores.

Secondly, past homophily research mainly focused on face-to-face interaction or co-located groups (Yuan et al. 2006). Our research is one of the few that studied how the homophily factors influence the network link formation process in globally distributed groups. The globally distributed developers are very diverse in terms of nationality,

language, culture, and various other social and personal characteristics. Previous research suggested that the more diverse the group is, the greater the homophily effect in social networks (Mcpherson et al. 1987), which supports the following hypothesis we have proposed:

Hypothesis 1 (H1). *The likelihood that a developer will participate in an OSS development project is positively related to the similarity between this developer and the project's existing members in various homophily factors.*

4.2.2.2 Shared affiliation factors

Research suggests that individuals tend to work with people who they have prior affiliations with (Schachter 1959). When evaluating an OSS project consisting of prior affiliated individuals, a developer may use his or her personal experiences of shared affiliations to evaluate the benefits and risks in working with them, and eventually to determine whether to participate in this project. The shared affiliation factors include shared activities, group memberships, and mutual acquaintances that may significantly influence link formation processes in social networks (Kossinets et al. 2006). For example, Kossinets et al. (2006) analyzed an email communication network in a large university over the course of one academic year and found that the number of shared classes and mutual acquaintances between two students significantly influence the formation of their email communication links. In other words, the more classes and

mutual acquaintances the two students shared, the more likely they will email each other. Newman (Newman 2001b) studied a scientific collaboration network and also found that two scientists are more likely to co-author a paper if they previously had a third common collaborator.

For OSS developer networks, Hahn et al. (2008) studied the impacts of prior collaboration affiliations on developers' participation choices. He found that a developer is more likely to participate in an OSS project if he has a strong collaboration link (i.e., collaborated in another OSS project before) with its initiator. He argues that projects consisting of developers with prior collaborative affiliations may be more effective in the coordination of OSS developers' distributed expertise since they have knowledge of "who knows what" from past collaborations. Therefore, developers may be more likely to work with past collaborators in new projects. Such preferences will increase the strengths of the past collaborative links and lead to developing closer and more cohesive cliques in social networks (Granovet.Ms 1973). This in turn further increases the likelihood that these developers with past affiliations will collaborate in future OSS projects.

While shared affiliation through past collaborations gives certain indirect indications for the prediction of future participation links, we argue that a past evaluation relationship between two developers is a stronger indicator for the following two reasons. Firstly, the collaboration relationship between two developers is extracted from their joint participation in OSS projects (Hahn et al. 2008; Jin et al. 2005; Madey 2002). However,

the strengths of such relationships vary greatly among the pairs of developers in each project, especially for a large OSS project such as Linux which often involves thousands of developers. In such large projects, most of the times two members may not even know each other. Therefore, shared affiliation through past collaborations in OSS projects may only give limited indications of personal relationships between two developers. Meanwhile, a recent study (Hu et al. 2008a) has compared evaluation relationships in an OSS community against collaboration relationships using SNA methods. It found that evaluation of an OSS developer made by another is a direct relationship between these two and indicates acquaintance and stronger personal relationships than past collaborations.

Secondly, shared affiliation through collaborations without sentiment indicators may not be useful in predicting future participation links. Many OSS projects fail due to the conflicts among project members. Thus not all past collaboration experiences are positive and can reinforce future collaborations between two developers. Hahn et al. (2008) suggested that if the OSS developers intend to gain benefits from OSS project participation, they will be more likely to choose projects consisting of developers with positive past collaboration experiences. Negative experiences in past collaborations may even discourage developers from working with the same collaborators. However, past OSS research on collaboration relationships often lacks the sentiment information and may ignore the negative effects of failed collaborations on developers' project participation choices. On the other hand, evaluation relationships in OSS communities

contains information about one developer's sentiment towards another, often based on the experiences from their past interactions such as collaborating in the same OSS project or admiration of programming skills. Such sentiment information makes shared affiliation through evaluation a better choice than collaboration in predicting the OSS developers' project participation choices. Therefore, we hypothesize that:

Hypothesis 2 (H2). *A developer will be more likely to participate in an OSS development project if that project consists of one or more member(s) that are positively evaluated by this developer before.*

4.3 Dataset

4.3.1 Data characteristics

The dataset for this study was collected from a large online OSS community – Ohloh -- through its API (Allen et al. 2009), which provides information about 11,530 OSS projects involving 94,330 developers. This data source is unique compared with other major OSS communities such as SourceForge.net in two main/primary/particular ways. Firstly, it not only provides OSS project participation information but also evaluation information about/from Ohloh community members. Each Ohloh member can send any other member a link called "Kudo" which is a simple gesture of thanks, praise, or endorsement. Sometimes a "Kudo" link can be given to a co-developer in the same OSS project as positive evaluation for his or her contribution. Sometimes people receive

"Kudo" links from others as recognition of their programming skills or appreciation for their help. Therefore, the "Kudo" evaluation links may indicate many underlying social relationships among OSS community members. Moreover, Ohloh provides information about the attributes of registered developers and projects while Souceforge.net does not. This information includes developers' attributes, such as nationalities and locations, and projects statistics like total lines of codes and comment ratio. Such information is crucial for the determinant analysis of social networks.

Second, the Ohloh dataset covers a more comprehensive list of major OSS projects than Sourceforge.net because of its data sources. It retrieves OSS-related data from three major software revision control repositories – Subversion, CVS and Git - while SourceForge.net only has data from Subversion.

In addition, the Ohloh website provides several other types of information about OSS projects. For example, the project activity information keeps track of every change made about an OSS project, including what was changed, when it was changed, and who made the change. Other global statistics like programming language usage are also included. Such information coupled with the results from social network analysis may provide insights about the determinants of link formations in Ohloh networks.

4.3.2 Data collection and preprocessing

Ohloh web site provides a REST-based application programming interface (API) for users to access and query its data. We developed a set of Java programs to automatically query and retrieve OSS related data using this API. Figure 1 shows sample data for project Firefox retrieved through this API. Since all retrieved data items are in XML format, a customized parser program was developed to parse all the data into the Ohloh database. Figure 2 shows a sample of one Ohloh database table which stores the parsed OSS project information. The entire Ohloh database stores information about 11,530 projects and 94,330 developers.


```
<?xml version="1.0" encoding="UTF-8" ?>
- <response>
    <status>success</status>
  - <result>
    - <project>
        <id>9</id>
        <name>Mozilla Firefox</name>
        <created_at>2006-10-10T15:51:31Z</created_at>
        <updated_at>2009-05-05T15:31:19Z</updated_at>
        <description>The award-winning Web browser is now faster, more secure, and fully customizable to your online life. With more
          than 15,000 improvements, Firefox 3 is faster, safer and smarter than ever before. If you stack this project, you should also
          stack the Mozilla Core.</description>
        <homepage_url>http://www.firefox.com/</homepage_url>
        <download_url>http://getfirefox.com/</download_url>
        <url_name>firefox</url_name>
        <medium_logo_url>http://s3.amazonaws.com/attachments/6346/firefox_med.png</medium_logo_url>
        <small_logo_url>http://s3.amazonaws.com/attachments/6346/firefox_small.png</small_logo_url>
        <user_count>5243</user_count>
        <average_rating>4.49937</average_rating>
        <rating_count>1586</rating_count>
        <analysis_id>535485</analysis_id>
      - <analysis>
          <id>535485</id>
          <project_id>9</project_id>
          <updated_at>2009-05-02T06:33:47Z</updated_at>
          <logged_at>2009-05-02T06:20:56Z</logged_at>
          <min_month>2002-04-01T00:00:00Z</min_month>
          <max_month>2009-04-01T00:00:00Z</max_month>
          <twelve_month_contributor_count>33</twelve_month_contributor_count>
          <total_code_lines>63683</total_code_lines>
          <main_language_id>6</main_language_id>
          <main_language_name>JavaScript</main_language_name>
```


Figure 4.1. Sample data retrieved from Ohloh API

Project_id	Project_name	Project_URL	# of Users	Average Rating	Rating_count
9	Mozilla Firefox	http://www.firefox.com/	2307	4.535897435	390
1	Subversion	http://subversion.tigris.org/	2094	4.43884892	278
72	Apache HTTP Server	http://httpd.apache.org	1621	4.609589041	146
3141	Linux Kernel 2.6	http://www.kernel.org/	1230	4.806666666	150

Figure 4.2. Sample data parsed into the project table in the Ohloh database

4.4 Research design

To answer the research question, we proposed a set of analytical methods which combine both SNA topological analysis and conditional logistic analysis to discover determinants of participation links in the Ohloh social networks. The research design is presented in Figure 4.3. It consists of two steps. The first step involves two components: network construction and determinants extraction. We constructed a participation network from Ohloh dataset. Meanwhile, the potential determinants of participation links may be extracted based on literature and theoretical conjectures on social networks. The second step, network analysis, contains both SNA topological analysis and determinant analysis. The details of the design are introduced in the following sub-sections.

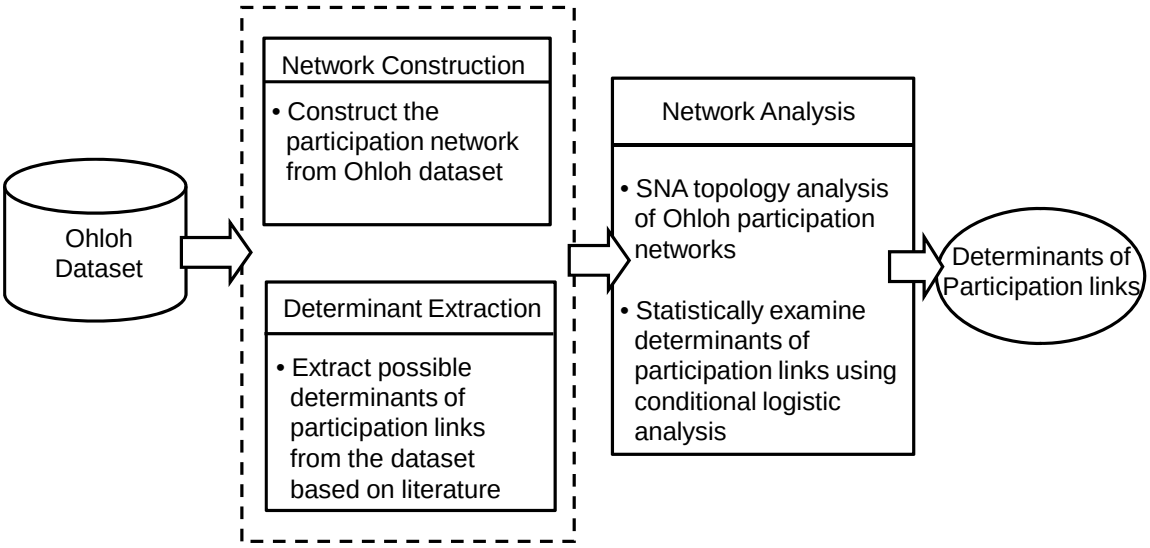


Figure 4.3. Discovering determinants of positive evaluation links in Ohloh networks

4.4.1 Network construction

In this section, we construct a social network by using the Ohloh participation dataset. The Ohloh participation dataset can be naturally represented as a network by treating the developers and projects as nodes, and developer's participation in projects as links. We defined such network as the participation network, which is a bipartite network consisting of two sets of nodes and no links within the same set. Such bipartite networks are also termed as affiliation networks (Davis et al. 2003) in social network literature. Figure 4 shows a sample participation network from the Ohloh dataset. The round nodes represent OSS projects in Ohloh community while the square nodes are the developers. In this network, a link from developer 3987 to the project Firefox represents that developer 3987 had participated in GNOME project.

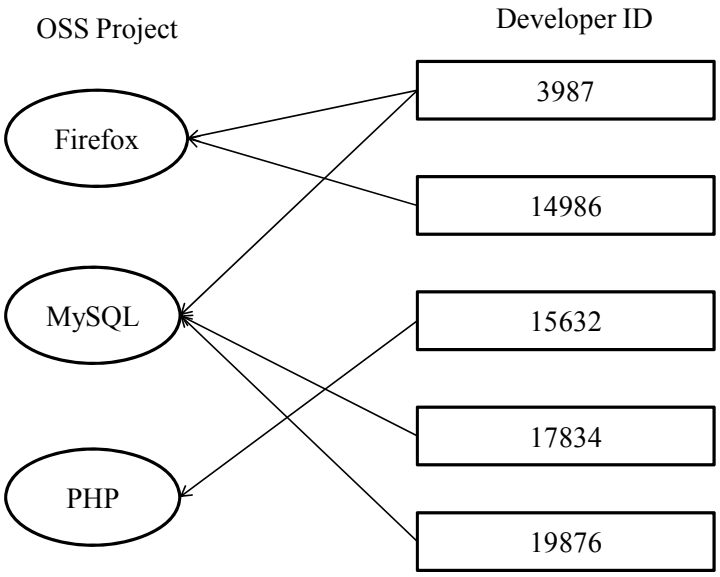


Figure 4.4. A sample Ohloh participation network

The participation dataset used in our analysis only include developers and projects which registered with complete attribute information such as developer's nationality and project's primary programming language. Table 4.1 shows the key statistics of the constructed Ohloh participation network. It contains 4,690 registered developers, 5,351 OSS projects, and 11,532 participation records among them.

Table 4.1. Key statistics of the Ohloh participation network

	Number
Developers	4690
Projects	5351
Participation Links	11532

4.4.2 Determinants extraction

In this study, the possible determinants of participation were extracted from Oholh dataset based on findings or conjectures from prior studies in both OSS domain (Crowston et al. 2006; Roberts et al. 2006) and social network domain (Kossinets et al. 2006; McPherson et al. 2001; Powell et al. 2005).

4.4.2.1 SNA Related Determinants

Prior SNA studies found that the determinants of social relationships may include homophily in individual attributes and shared affiliation such as mutual acquaintance (Kossinets et al. 2006; McPherson et al. 2001; Powell et al. 2005).

In Ohloh dataset, one possible determinant for project participation choices is homophily in programming language,

- **Homophily in Programming Language**: It indicates if the contributing developer's primary programming language is the main language for the OSS project. To our best knowledge, this proposed possible determinant for the first time examines if the developers' preference on programming language influence his or her project participation choices.

Shared affiliation factors among social entities usually refer to indirect social relationships such as mutual acquaintance. They are found to be the determinants to link formation processes in various social networks (Kossinets et al. 2006).

- **Project Mutual Acquaintance with Evaluation Relationships**: It indicates if the developer has positively evaluates another developer and they both participated in the same OSS project under study. This share affiliation factor is include is because a developer may tend to participate an OSS project with many of his or her acquaintance.

4.4.2.2 Control variables: project performance related determinants

Previous OSS studies (Roberts et al. 2006) show that developer and project performance may significantly enhance developers' motivation in project participation and to gain higher status. Crowston et al. (2006) applied the most commonly cited model for IS success (DeLone et al. 2003; DeLone et al. 2004) in the context of OSS development and developed a set of performance measures for OSS projects. These measures include system creation, system quality, system use and system consequences. We adopted and applied these measures in Ohloh dataset to extract the following possible participation determinants related with project performance.

System creation factors mainly measure the activity level and the overall efforts of the contributing developers on each project.

- **Total Lines of Codes**: Total lines of source code contributed by all the developers in an Ohloh project. Blank lines and comment lines are excluded.
- **Total Man Power Invested:** The cumulative total months of effort spent by all participating developers on this project. For instance, if 1 contributing developer has worked for 3 months and the other 2 have worked for 5 months, total man power invested will be 13.
- **Commits**: The total number of commits made by all contributing developers on an Ohloh project.
- **Project age**: In Ohloh dataset, the number of the years the OSS project existed may also affect participation choices. The assumption is that projects with older age are more likely to attract new developers to participate.

System quality factors measure the code and documentation quality of OSS projects, such as understandability, completeness, maintainability, usability, and efficiency (Crowston et al. 2006). In Ohloh dataset, we extracted the comment ratio for each OSS project.

- **Comment Ratio**: The fraction of total lines of code which are comments. Comments are usually very useful for the modifications and maintenance of source code. It measures the maintainability of source code.

System use factors measure usability and user satisfaction of the open source software. In Ohloh dataset, we extracted the number of users, average rating score and rating respond rate as system use factors.

- **Number of Users**: The number of users who used this OSS software project. Higher number of users indicates more popular projects.
- **Average Rating Score:** A floating point value from 1.0 to 5.0, representing the average value of all user ratings. 1.0 is the worst possible rating and 5.0 is the highest possible rating.
- **Rating Respond Rate:** The faction of the users of an OSS project who rated that project.

4.4.3 Network analysis

After we construct the Ohloh participation network and extract potential determinants, we conduct SNA topology analysis on these two networks and examine the extracted determinants using conditional logistic analysis.

4.4.3.1 Topology analysis

There are two goals for the topology analysis of Ohloh networks in our study. Firstly, it helps us uncover the structures of Ohloh open source software community, and better understand the nature of OSS participation relationships. Second, the determinants of link formation are found to significantly affect the network topologies (Kossinets et al. 2006; Powell et al. 2005). For instance, the preferential attachment factor – older project(s) – may attract more developers than younger project(s) in the OSS community, causing a scale-free topology in the participation network. Therefore, the results of topology

analysis can be used to find plausible network determinants for further statistical analysis on determinants of link formation.

We use SNA centrality measures to describe the topology of the Ohloh participation network and identify its key members. High degrees usually indicate high levels of activity and wide social influence. The average degree of a network is also calculated to measure how dense a network is. In addition, previous research (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks and have small-world network properties. Thus we examine the Ohloh participation network to see if it has these features. Several SNA measures are examined, including the average path length, the clustering coefficient, link density, and the degree distribution. These properties are then, checked against the small world and scale-free models.

4.4.3.2 Statistical analysis to examine determinants

Our choice of the statistical model for analyzing determinants of participation in Ohloh network is based on both theoretical and empirical considerations. Theoretically, our study intends to model human (participation) choice behaviors in social networks. The research question asks how to discover determinants account for differential (as opposed to random) patterns of the link formations in OSS participation networks. Empirically, we need to model the choice behavior of a dyad – developers to projects. For these reasons, we choose to use conditional logistic model that takes each choice as a unit of analysis,

which in our case is the formation of a participation link between a developer and an OSS project.

Conditional logistic model (CLM) and its variations (Jin et al. 2005; McFadden 1980; McFadden et al. 1974; Powell et al. 2005) have been widely used to model human choice behaviors and examine the determinants that influence those choices. In our study, for the Ohloh participation network, the probability of an OSS developer *I* choose to participate an OSS project *j* from the alternative set J_i, is specified as follows:

$$\Pr(y_i = j) = \frac{\exp(X_j \beta)}{\sum_j^J \exp(X_j \beta)}$$

where y_i is the observed choice for developer *I and* X_j is a vector of the characteristics of the project j. The unknown coefficients β are typically estimated by maximum likelihood methods.

We estimate the CLM for the Ohloh participation network data using *clogistic* command in Stata 10/MP. The dependent variable is a binary indicator of the outcome for participation link formation between an Ohloh developer and project. The independent variables are the selected potential determinants explained in the determinant extraction section. In addition, for the statistical analysis, these independent variables are operationalized as Table 4.2 shows.

Table 4.2: Variables selected for analysis

Label	Description
Dependent Variable	
Participation Link	Computed as a binary indicator of whether a developer participated in an OSS project
Independent Variables: **SNA Related Factors**	
Homophily in Programming Language	Set to '1' if the developer's primary programming language is the same with the main language of the project, and '0' otherwise
Shared affiliation through Evaluation Relationships	Set to N if the developer has positively evaluated (sent a Kudo link to) N (N>=1) developers in the OSS project under study
Control Variables: **Project Related Factors**	
Total Lines of Code	The total net lines of code, excluding comments and blanks, as of the end of this month.
Total Man Power Invested	The cumulative total months of effort expended by all contributors on this project
Commits	The cumulative total number commits to the project source control
Comment Ratio	The fraction of net lines which are comments

Number of Users	The total number of users of this open source software
Average Rating Score	A floating point value from 1.0 to 5.0, representing the average value of all user ratings. 1.0 is the worst possible rating and 5.0f is the highest.
Rating Respond Rate	The number of users who rated this open source software divided by the total number of users.
Project Age	No. of calendar months in which the OSS project existed.

4.5 Results

4.5.1 Topology analysis

We start the analysis from describing the topology of the overall participation networks. Table 3 shows that the results of SNA topology measures for the two types of nodes in the whole Ohloh participation network. The average degree of developer nodes is the average number of projects a developer participates. On the other hand, the average degree for project nodes is average developers a project has. In addition, since prior studies (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks, we then fit the degree distributions of the two networks to power-law distribution to test for scale-free topological features.

Table 4.3. Results of SNA measures of the Ohloh participation network

	Developer Nodes	Project Nodes
Number	4690	5351
Average Degree	2.60	2.28
Max Degree	55	168
Min Degree	1	1
Degree Distribution		
R^2	**0.96**	**0.89**
γ	2.27	1.81

As Table 4.3 shows, we found that the average degrees of these two types of nodes are similar while the max degree of project nodes is much larger than the developer nodes. This indicates there is more heterogeneity in the project nodes than the developer nodes. The last three rows of Table 3 show the results of linear regression on the degree distribution of the Ohloh participation network. It was found that this network follows the power-law degree distribution (Jin et al. 2005; Newman 2001a) for both developer nodes and project nodes, $p(k) \sim k^{-\gamma}$. The coefficients of determination R^2 is extremely large at 0.96 and 0.89 (ranging from 0 to 1) respectively, indicating high fitness of the power-law degree distribution model. These findings from topology analysis imply that the Ohloh participation network is a scale-free network.

4.5.2 Statistical analysis

Since the information about OSS projects is provided and updated monthly in Ohloh web site, our statistical analysis focused on developers who has participated in OSS projects in the most recent month data (May 2007). In addition, the conditional logistic analysis requires a set of alternative projects for the developer in each participation occurrence. The number of possible participation links is the possible combinations between developers and projects, calculated as the number of developers multiplies the number of alternative projects. Since the average age of OSS project in Ohloh dataset is around 2.5 years, we limited the alternative set of projects to the ones which started within the past three years (of May 2007). Table 4.4 summarized the key statistics of the data sample used in the statistical analysis.

Table 4.4. Key Statistics of the Ohloh data sample from conditional logistic analysis

Time Span of Analysis	Number of Developers	Number of OSS projects	Number of Possible Participation Links	Number of Actual Participation Links
05/2007	229	772	176,788	462

To facilitate the interpretation of the results from conditional logistic analysis, we present the odds ratios and the coefficients as Table 4.5 shows. The odds ratios are obtained from the coefficients by using exponential function. That is $X_i = \exp(\beta_i)$ where β_i is the coefficient. In the Ohloh context, the odds ratio measures the change of the probability

that a participation link is formed caused by each unit increase in an independent variable. This means that the probability of the link formation would increase by a factor of odds ratio when the corresponding independent variable increases by one unit. Odds ratio equals to one means there is no effect of the independent variable on the dependent variable.

Table 4.5. Results from conditional logistic regression analysis

	Odds Ratio X_i	Coefficient β_i
SNA Related Factors		
Homophily in Programming Language	**8.480***	**2.137***
Shared Affiliation through Evaluation relationships	**28.001***	**3.332***
Project Related Factors		
Total Lines of Code	1.000	-2.30e-08
Total Man Power Invested	1.001*	0.002*
Commits	1.000*	-.0001*
Comment Ratio	0.492	-0.709
Number of Users	1.005****	0.006*
Average Rating Score	0.968	-0.032
Rating Respond Rate	0.971	-0.028
Project Age	**0.771***	**-0.260***

* $p < 0.01$

Table 4.5 shows the odds ratios and coefficients for each independent variables from the CLM analysis on the Ohloh participation networks. Only independent variables with p-value less than 0.05 and odds ratios significantly different from 1 are determinants. The results show that the homophily in programming language and project mutual acquaintance are determinants of participation links with odds ratios larger than 1. In addition, the project age factor is also found to be the determinant with odds ratio smaller than 1. This means that the probability for project participation decreases as the age of this project grows.

4.6 Discussion

4.6.1 Determinants of OSS project participation choices

As the results of the conditional logistic analysis indicated, a developer in Ohloh community is more likely to participate in an OSS project if 1) his primary programming language is the same with the project main programming language, or 2) he has positively evaluated one or more developers in that OSS project before. Therefore the hypothesis 1 and 2 in our study are supported by the empirical analysis of Ohloh developer networks. Moreover, a developer is less likely to participate in an OSS project as its age grows.

These findings may partly be explained by the following conjectures. First, skills in Programming language is one of the core competencies needed by OSS developers and

projects (Crowston et al. 2002). OSS developers autonomously develop their skills of programming language by participating in projects. Skillful developers may tend to work in the projects that use the programming languages they are most familiar with, while junior developers tend to seize every opportunity to practice their primary programming language (Yunwen et al. 2003). In addition, some OSS project may require their participants equipped with knowledge and skills of certain programming language to maintain the quality of the development work.

Another determinant – (project) mutual acquaintance with evaluation relationship - has been studied in several other SNA research (Hu et al. 2008b; Kossinets et al. 2006; McPherson et al. 2001). It was found that individuals tend to select new acquaintances who are friends of a friend. In the Ohloh participation network, this finding suggests that OSS developers tend to form circles of close acquaintances and are attracted to OSS projects that their acquaintances are in. Another possibility is that the developers of an OSS project may tend to recruit new members from their close acquaintances. One implication of this determinant is that, recruiting developers with good reputation and high community status may also attract his close acquaintances to the same project. These close acquaintances participate in the projects usually for the opportunities to learn from or cooperate with the original developer.

Surprisingly, the project age factor was found to negatively influence the developers' project participation choices. This finding means in time the OSS projects gradually lose

its capability to attract new developers. One possible explanation is that, with the rapid development of information and software technology, developers newly joined in the Ohloh community are less likely to be interested in developing older software or have the skills of older programming languages required by such development.

4.6.2 Insignificant factors

Another surprising finding is that all project performance factors have little effect on the developers' choices to participate in OSS projects. It may be because that the project performance information, such as total lines of codes, is often not available for potential contributing developers. In addition, for new developers with little OSS experiences, it is often difficult for them to understand such project performance information. On the other hand, social factors such as mutual acquaintance may play a more important role in attracting such inexperienced developers. One very practical implication for the OSS project managers is that, simply advertising strong project performance may not be useful in attracting and recruiting developers. Personal connections or other customized information, such as an invitation from a respected developer friend is critical for such recruiting tasks in OSS project management.

4.6.3 Impacts of determinants of OSS project participation

In addition, the coefficients of the discovered determinants vary from 0.771 to 28.001. We found that the more personal the determinant's context, the larger the coefficient is.

This means, having a mutual acquaintance with evaluation relationship is a stronger personal connection between the developer and the project, than just using the same programming language, therefore weighs more in the developer's decision in participating in the project. This finding may indicate that social and personal factors have larger impacts on OSS developers' participation choices than the project performance factors.

4.6.4 CLM-based link prediction mechanism

In addition, the results from conditional logistic analysis can also be used to calculate the probability for a developer to participate in an OSS project. For instance, as Table 4.5 shows, in the Ohloh participation network, the coefficients for homophily in programming language and for having a project mutual acquaintance are 2.137 and 3.332, respectively. Assuming both developer a and project b uses Java as their main programming language. In addition, *a* has positively evaluated two other developers in Project b before. Then using the conditional logistic model, the probability for developer *a* to participate in project b in an alternative project set J_i, can be calculated as $\Pr(y_a = b) = \exp(1 \times 2.137 + 2 \times 3.332) / \sum_j^J \exp(X_j \beta)$. This calculation can be applied to any pair of developer and project in the Ohloh community, and therefore can be used to predict how likely an developer is going to participate in a specific OSS project.

This prediction mechanism can be used for OSS project management in two ways. Firstly, it can help OSS project managers to identify potential developers in the community and devise useful strategies to attract time. More specifically speaking, based on the homophily in programming language factor, the manager may need to put more advertising and recruiting efforts in technical communities and forums which are related to the programming language used by his project. Secondly, this mechanism can also be used to recommend inexperienced developers to the easy-to-do projects which are suitable for them.

4.7 Conclusions

Although there have been a fair amount of studies analyzing the motivations of OSS project participation, such work mainly focused on attributes of individual developer or project, but largely ignoring the relationships between these two entities and the underlying determinants that significantly affect the developers' participation choice. In this study, we used both SNA and conditional logistic analysis to study the determinants of developers' participation choices in a large online OSS community – Ohloh. It was found that the homophily in programming language, and shared affiliation through evaluation relationship are significant determinants of participation link formation in OSS developer networks. The set of methods used in this research can also be applied to study determinants and topologies of social networks in other domains.

We also explored the possible social causes and implications for these identified determinants. Our analysis may help researchers and practitioners in OSS community to better understand the developers' participation choice behaviors and devise useful strategies for OSS project management. For instance, as stated in the discussion, the project manager may improve the recruitment by advertising in online communities and forums that are related to the programming language his project uses.

In addition, based on the identified determinants and conditional logistic model, we proposed a prediction mechanism that can quantitatively estimate Ohloh developer's project participation choice at the dyadic level. This prediction mechanism may be used in the design of collaborative information systems to support OSS development.

Our future work consists of three research directions including (1) investigating other OSS participation networks to further validate the determinants found in this study, (2) applying the proposed methods to analyze determinants of links in other types of social networks, and (3) using the knowledge and the link prediction mechanisms derived from this study to design various information systems to support knowledge management and decision making.

CHAPTER 5: EXPERT RECOMMENDATION VIA SEMANTIC SOCIAL NETWORKS

5.1 Introduction

Social networks are becoming increasingly pervasive in today's society with individuals leveraging the communication abilities of computing systems to interact and collaborate. Advances in computing and networking technologies, especially the Web 2.0, have facilitated numerous social network based applications. Social networking websites such as MySpace, FaceBook and Orkut have revolutionized the way people socialize with each other. Another major trend is to develop social network based organizational applications such as locating and recommending knowledge experts (Henry et al. 1997; Raghavan et al. 2002). More recently, a novel Google initiative – Social Graph – has been launched aiming to describe the social network infrastructure of the public web by indexing websites with open standards such as XHTML Friends Network (XFN) and Friend of a Friend (FOAF).

The emergence of social networks has raised a great need to analyze and understand the knowledge they generated for various applications. To reach this goal, a research method - social network analysis (SNA) – is developed and attracts great interests in diverse domains such as organizational studies, computer science, and information system (IS) research, largely because of its powerful capabilities of modeling and analyzing real-

world complex systems ranging from scientific collaboration networks (Palla et al. 2007) to the World Wide Web (Albert et al. 1999). In SNA studies, a network is usually represented by a number of nodes (e.g., software developers) connected by links (e.g., evaluation relationships).

Advances in communication technologies, especially the Web, are making the world more and more interconnected. In this environment, computing applications are trying to realize various social functions which are based on social relationships. Berners-Lee (Berners-Lee et al. 1999) defined them as "social systems". For instance, consider the coupling of statistical analysis on couples' social profiles with Internet technologies that led to the matchmaking web site - eHarmony. Such social systems need to consider not only the underlying technologies but also the social relationships, rules and organizational structures among people (James et al. 2008).

Therefore, the use of SNA in expert recommendation systems is becoming increasingly popular, most notably in ReferralWeb (Henry et al. 1997), Yenta (Leonard 1997), and Expertise Recommender (David et al. 2000). These systems search for people with appropriate expertise in the users' social networks to answer questions, solve problems or provide collaborations (Jun et al. 2005). However, previous research (David 2003) found that the experts recommended by such SNA based systems often do not match users' needs and the perceptions of their personal social networks.

This is mainly because social network analysis embedded in those expert recommendation systems largely ignored semantic information of social networks and only focused on topological features. For instance, a software developer is looking for an expert to answer his questions about java programming. However, the topological analysis of his email network only identifies the individual who receives most emails but cannot tell why. Therefore, the network semantic information such as individuals' java programming experiences coupled is required in finding such experts.

These network semantics can lead to significant differences in the perceptions of individuals' expertise, therefore affecting the results of expert recommendations. In this study, we focused on the semantics of social network links – the determinants of the link formation. According to prior research (Kossinets et al. 2006; McPherson et al. 2001; Powell et al. 2005), these determinants include both individual attributes and shared affiliations:

- People in a social network vary in their attributes such as age, gender and experiences. Unlike nodes in theoretically constructed networks, a person in real-world networks form relationships (links) by considering individual attributes of the candidates.

- The links (i.e. relationships, interactions) among people in a real-world social network are highly meaningful and vary greatly (Wasserman et al. 1994). The link

formation process is largely driven by the shared affiliations of their members (e.g., attending the same class) (Kossinets et al. 2006).

However, SNA in existing expert recommendation systems largely ignored network semantics mainly due to the lack of 1) effective methods to quantitatively identify network semantics, 2) large data sets about semantics of social networks, and 3) methods to embed network semantics into the system design. These challenges naturally lead to two research questions:

- How to quantitatively discover semantics in social networks?

- How to embed the discovered semantics into the design of expert recommendation systems?

To answer these two questions, we used multiple methods to study the semantics of the social networks in an online open source software (OSS) community – Ohloh, which hosts more than 11,530 projects involving over 94,330 developers. Firstly we conducted topological analysis on two social networks in Ohloh community. Secondly, we statistically examined the determinants of link formations in these two networks using conditional (fixed-effects) logistic model (CLM). We then integrate the identified semantics (determinants) to corresponding network links to construct semantic social

networks. Such semantic social networks provide contextual meanings to its links such as sharing mutual acquaintance.

In addition, we propose a framework which aims to utilize semantics of social networks for various business applications. In this framework, we demonstrated how to use the discovered network semantics in the designs of expert recommendation systems through combining two network recommendation mechanisms. The first mechanism is used to quantitatively reflect individual users' perspectives on the recommended experts. The second mechanism is to construct semantic social networks which provide contextual information for more accurate expert recommendation. The approaches we proposed in this study can also be used to study the semantics of social networks in other domains.

The main contribution of this paper is to propose a computational approach to that derives useful semantics of social networks and embed these semantics into the design of expert recommendation systems through two mechanisms. The approach we used in this study may be generalized to the designs of other SNA based information systems.

The remainder of this paper is organized as follows. In the next section, we provide a review of literature on SNA, expert recommendation and OSS community analysis. The third section introduces the dataset for this study. Then we describe the research design and the experimental results. After that, we demonstrate how to embed the discovered

network semantics into the designs of expert recommendation systems. At last, we conclude and suggest directions for future work.

5.2 Related work

5.2.1 Social network analysis

Social network analysis is originally developed and used in sociology research to analyze patterns of relationships and interactions among social actors, aiming to discover the underlying social structure. It has been widely used to study various real-world networks (Albert et al. 1999; Boissevain 1974; Kossinets et al. 2006; McPherson et al. 2001), including networks in open source software communities (Crowston et al. 2003; Grewal et al. 2006; Jin et al. 2005; Madey 2002; Wagstrom et al. 2005).

There are mainly two types of SNA studies. One focuses on the topological characteristics of social networks. In such studies, the structural properties of the nodes and links are examined to describe and explain how network topologies affect the functions and behaviors of complex systems (Albert et al. 2002). Another line of SNA research studies the mechanisms and determinants behind the network dynamics. They mainly use statistical methods to model different network mechanisms. These models are then tested to account for the structural changes of network topologies.

However, while both types of SNA studies have made great efforts in analyzing the network topologies, little attention is given to the semantics of social networks mainly due to the lack of appropriate analytical methods and network semantic information. The existing analytical methods of SNA are introduced in the following sub-sections.

5.2.1.1 Topology analysis

Several quantitative SNA measures are developed to describe network topologies at both individual and network level. At the individual node level, network centrality measures are used to identify key members and interaction pattern between sub-groups. One of the most commonly used centrality measure –a node's degree– is defined by Freeman (1979) as the number of direct links this node has. It measures how active a particular node (individual) is. A network member with a high degree can be the leader or "hub" in a network.

On the other hand, several network level SNA measures such as average degree, clustering coefficient, average path length, and degree distribution are developed to describe and distinguish different network topology models. Three models have been employed to characterize complex networks: random graph model (Erdos et al. 1960), small-world model (Watts et al. 1998), and scale-free model. In random networks (Barabási et al. 1999), each node has roughly the same number of links which equals to its average degree.

Clustering coefficient is usually used to determine the small-world nature of social networks. It is the probability that two nodes with a common neighbor also link to each other. A small-world network usually has a significantly larger clustering coefficient (Watts et al. 1998) than its random model counterpart, indicating a high tendency for nodes to form clusters and communities. A small-world network also usually has a relatively small average path length (i.e., average number of steps along the shortest paths for all possible pairs of network nodes) (Watts et al. 1998).

Degree distribution *P(k)* is the probability distribution of a node has exactly k links. Power-law degree distribution is used to characterize scale-free networks (Wasserman et al. 1994). In such networks, a small fraction of the nodes have a large number of links while a big fraction of nodes have just a few. This scale-free topology may be caused by the newly joined nodes' preferential attachment to the existing nodes with high degrees (Albert et al. 2002).

In general, network topological analysis is good at describing the structure of complex network systems but lack the capabilities to explain the emergence of such topologies and analyze the determinants of various network processes such as link formation. As a result, statistical methods are used in SNA studies to complement the insufficiencies of topological analysis.

5.2.1.2 Statistical analysis on determinants of social network links

Statistical analysis has been widely used to model topological changes of various networks (Albert et al. 2002). In such analysis, it is assumed that network structural changes are caused by certain stochastic processes of network effects such as reciprocity, transitivity, and balance. Thus several network topology models have been developed based on these network effects. They were fitted to empirical data to identify which network effects account for the observed structural changes (Snijders 2001).

However, in this study, we focus on another type of statistical analysis that has been used to identify and examine the determinants of network link formation processes. Such analysis is widely adopted in many domains such as organizational studies, sociology, and network analysis. For instance, in organizational studies, Beckman et al. (2004) studied inter-organizational network change by statistically examining factors that affect the firms' choices of partners. They analyzed data on alliance networks for the 300 largest U.S. firms from 1988 to 1993. The results showed that the stability of a firm's alliance network structure depends on the type of uncertainty it experienced. The greater the uncertainty (risks) that a firm faces alone, the more likely this firm will expand its alliance network. Likewise, the greater the uncertainty (risks) that a firm's market or industry faces, the more likely that firm will strengthen the ties it presently has with others.

In the sociology literature, Leenders (1996) used a continuous-time Markov model to study the determinants of link formation in a children's friendship network. The results showed that the homophily in gender (i.e. being the same gender) significantly affects the link (friendship) formation among children. The Markov model assumes that only the state of the network at time t-1 affects the current state (at time t). However, this assumption may not be valid for most real-world networks.

Not only limited to friendship, McPherson et al. (2001) argue that various other social relationships, including marriage, work, advice are also influenced by the homophily principle - similarity breeds connection. In addition, Snijders (2001) developed a class of actor-oriented models to examine if the nodes adjust their linking choices in the network based on certain parameters such as their degrees. However, these models assume that the nodes are aware of their positions with respect to the whole network which is often not true in large complex networks.

Another study done by Powell et al. (2005) examined the determinants of the partner selection process for biotechnology firms in 1990s. They examined several types of determinants such as preferential attachment and homophily (i.e. people tend to interact with others having similar characteristics) using McFadden's (1980; 1974) discrete choice model, a variant of the conditional logistic model. This econometric model is usually used to statistically model and analyze the human behaviors of making choices in various settings.

In this model, a subject is presented with choice alternatives and asked to choose the best alternative. In addition, the explanatory variables are alternative-specific or subject-specific. One limitation of this model is that it requires detail personal information of the subjects and the alternatives. Such information is usually not provided in existing empirical data sources.

In addition, longitudinal network data were employed to study network determinants too. Kossinets and Watts (2006) used Cox survival analysis to identify determinants of the email link formation in a university campus over a year time period. They found that the mutual acquaintance (i.e. two individuals are acquainted with a common person) and shared class affiliations (i.e. attending the same class) significantly affect the future email link formation between two students.

In addition, a similar survival analysis approach was also used by Nerkar and Paruchuri (2005) to determine that if network centrality of inventors had a statistically significant effect on the intra-firm citation of their patents. Survival analysis lends itself well to the longitudinal analysis of network data since it involves the modeling of time to event data. In the context of semantic analysis, an event is the formation of a link. However, sometimes the time information is not available or inaccurate in the network datasets, which makes survival analysis not suitable for identifying semantics.

In general, most existing statistical models for analyzing the determinants of network link formation are limited by specific assumptions or data completeness issues. There is a lack of statistical techniques which are general enough and can be applied on empirical network datasets with missing information.

5.2.1.3 Expert recommendation and social network analysis

Systems that help find individuals with expertise required by users are defined as expert recommendation systems. Recently the use of social network analysis in such system is becoming increasingly popular. Henry et al. (1997) developed such a system called ReferralWeb which provides referrals via chains of named individuals. The users may choose to search for referrals to people who are closely linked with famous, trusted experts to help them. Another referral based system Yenta (Leonard 1997) was designed to find experts having similar interests with the users. In addition, Expertise Recommender (David et al. 2000) used the distance between the user and the expert in their social network to filter recommended experts. If the distance is less than a threshold, the recommended expert is added to the final recommendation.

However, David (2003) found that the experts recommended by SNA based systems often do not match users' needs and the perceptions of their personal social networks. That is mainly because the SNA methods used in existing expert recommendation systems mainly use topological information such as nodes' degrees, distances. The

semantics of social networks are largely ignored and needed for expert recommendation systems.

5.2.2 Open source software community and social network analysis

5.2.2.1 Open source software community

Nowadays, OSS communities have emerged as a major place for software developers to seek help and share knowledge. Thus many researchers have begun to study the OSS community, aiming to find out how it is related to the success of OSS software development. Such studies mainly focus on two topics. The first topic is the composition of the OSS community. Koch et al. (2002) analyzed the logs of source code changes for an OSS project and identified a core set of developers who produce most of the source code output. Such core OSS community members are also found to have most intensive communications in a project (Robertsa et al. 2006).

Another line of OSS community composition research focuses on the participation process of members. For instance, von Krogh et al. (2003) found that new OSS community members gain benefits from specializing their initial contributions. Roberts et al. (2006) have developed a theoretical model and evaluated it using empirical data from the Apache projects, trying to understand how participations, motivations and performance of OSS community members interrelate. The results showed that people with higher status motivations are more likely to contribute.

Another empirical study (Bagozzi et al. 2006) surveyed 402 active members from 191 Linux User Groups (LUG) in 23 countries and found that the participation to LUG is positively related with the person's experience level in Linux. In our study, the participation of an OSS project is modeled as joining the collaboration network of Ohloh community. Therefore, according to these prior studies, the link formation process of Ohloh collaboration network is likely to be influenced by members' attributes such as community status and experiences.

The second research topic is studying various relationships among OSS community members. Most such studies focused on the collaboration relationship. Ducheneaut (2005) observed that successful OSS developers progressively enroll into a collaboration network of human and material allies to support each other. Another descriptive study (Yutaka et al. 2000) found that the communication in OSS development collaborations heavily relies on electronic media (e.g., forum, mailing lists) rather than face-to-face contact. In addition, Bergquist et al. (2001) found that OSS community members gain trust from others by actively giving out high quality source code and answering questions. However, the above research mainly focused on the relationships at a micro level. The overall network effects on OSS communities caused by the aggregation of multiple relationships are largely ignored.

To address this problem, a stream of literature using social network analysis methods studied the topologies of social networks in OSS communities. We introduce these studies in the following section.

5.2.2.2 Social network analysis on OSS communities

Social network analysis has been widely used to modeling and analyzing various relationships in OSS communities, especially the collaboration relationship. Madey (2002) uses SNA methods to study as a collaboration network of OSS developers in SourceForge.net and found it displays the scale-free network features. The small fraction of the developers with a large number of collaboration links can be explained by people's tendency to collaborate with high-profile, skillful members.

A more recent empirical analysis (Jin et al. 2005) of SourceForge data has discovered similar scale-free features in the collaboration network. Moreover, small-world network features – large clustering coefficient and small average path length – were also found in those SourceForge networks. Crowston (2003) have studied the topology of OSS collaboration networks using data from bug reports of 122 projects. It was found that the network topologies of bigger projects are less centralized. This may be caused by the modularization process of large OSS projects.

Another SNA study (Wagstrom et al. 2005) used empirical data from blog links and mailing lists to simulated OSS network evolution, aiming to develop and validate a model

which can explain developers' choices in their project participations. In addition, Grewal et al. (2006) examined OSS collaboration network embeddedness and discovered it has more influence on the technical success than the commercial success of OSS projects.

5.2.3 Research gap

Existing SNA based expert recommendation systems recommend people that do not match users' perceptions of social networks since network semantic information is largely ignored. Although there is research arguing the need of derive semantics from social networks (John et al. 2007), few analytical methods was developed for such semantic analysis. Previous SNA research mainly focused on topological analysis which lacks the capabilities to analyze the semantics (determinants) of various network processes such as link formation. Several attempts have been made to study the determinants of network link formation. However, the statistical models used in those studies are limited by specific assumptions or data incompleteness.

In general, there is a lack of formal analytical methods and appropriate dataset for semantic analysis of social networks. To address these problems, we used a novel approach which includes SNA and conditional logistic analysis to study two social networks in a large OSS community – Ohloh, aiming to answer the following research questions:

- How to identify and examine the semantics of the social networks?

- What are the semantics of the social networks in Ohloh community?
- How to utilize the semantic social networks for expert recommendation?

5.3 Dataset

5.3.1 Data characteristics

The dataset for this study was collected from a large online OSS community – Ohloh through its API (Allen et al. 2009), which provides information about 11,530 OSS projects involving 94,330 developers. This data source is unique comparing with other major OSS communities such as SourceForge.net mainly in two ways. Firstly, it not only provides OSS project participation information but also evaluation information among Ohloh community members. Each Ohloh member can send any other member a link called "Kudo" which is a simple gesture of thanks, praise, or endorsement. Sometimes a "Kudo" link can be given to a co-developer in the same OSS project as positive evaluation for his or her contribution. Sometimes people receive "Kudo" links from others as recognition of their programming skills or appreciation for their help. Therefore, the "Kudo" evaluation links may indicate many underlying social relationships among OSS community members.

Moreover, Ohloh provides information about the attributes of registered developers and projects while Souceforge.net does not. These information include developers' attributes

such as nationalities and locations, and projects statistics like total lines of codes and comment ratio. Such information is crucial for the determinant analysis of social networks.

Second, Ohloh dataset covers a more comprehensive list of major OSS projects than Sourceforge.net because of its data sources. It retrieves OSS related data from three major software revision control repositories – Subversion, CVS and Git while SourceForge.net only has data from Subversion.

In addition, Ohloh website provides several other types of information about OSS projects. For example, the project activity information keeps track of every change made about an OSS project, including what was changed, when it was changed, and who made the change. Other global statistics like programming language usage are also included. Such information coupled with the results from social network analysis may provide insights about the determinants of link formations in Ohloh networks.

5.3.2 Data collection and preprocessing

Ohloh web site provides a REST-based application programming interface (API) for users to access and query its data. We developed a set of Java programs to automatically query and retrieve OSS related data using this API. Figure 5.1 shows sample data for project Firefox retrieved through this API. Since all retrieved data items are in XML

format, a customized parser program was developed to parse all the data into the Ohloh database.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <response>
    <status>success</status>
  - <result>
    - <project>
        <id>9</id>
        <name>Mozilla Firefox</name>
        <created_at>2006-10-10T15:51:31Z</created_at>
        <updated_at>2009-05-05T15:31:19Z</updated_at>
        <description>The award-winning Web browser is now faster, more secure, and fully customizable to your online life. With more
          than 15,000 improvements, Firefox 3 is faster, safer and smarter than ever before. If you stack this project, you should also
          stack the Mozilla Core.</description>
        <homepage_url>http://www.firefox.com/</homepage_url>
        <download_url>http://getfirefox.com/</download_url>
        <url_name>firefox</url_name>
        <medium_logo_url>http://s3.amazonaws.com/attachments/6346/firefox_med.png</medium_logo_url>
        <small_logo_url>http://s3.amazonaws.com/attachments/6346/firefox_small.png</small_logo_url>
        <user_count>5243</user_count>
        <average_rating>4.49937</average_rating>
        <rating_count>1586</rating_count>
        <analysis_id>535485</analysis_id>
      - <analysis>
          <id>535485</id>
          <project_id>9</project_id>
          <updated_at>2009-05-02T06:33:47Z</updated_at>
          <logged_at>2009-05-02T06:20:56Z</logged_at>
          <min_month>2002-04-01T00:00:00Z</min_month>
          <max_month>2009-04-01T00:00:00Z</max_month>
          <twelve_month_contributor_count>33</twelve_month_contributor_count>
          <total_code_lines>63683</total_code_lines>
          <main_language_id>6</main_language_id>
          <main_language_name>JavaScript</main_language_name>
```


Figure 5.1. Sample data retrieved from Ohloh API

Figure 5.2 shows a sample of one Ohloh database table which stores the parsed OSS project information. The entire Ohloh database stores information about 11,530 projects and 94,330 developers.

Project_id	Project_name	Project_URL	# of Users	Average Rating	Rating_count
9	Mozilla Firefox	http://www.firefox.com/	2307	4.535897435	390
1	Subversion	http://subversion.tigris.org/	2094	4.43884892	278
72	Apache HTTP Server	http://httpd.apache.org	1621	4.609589041	146
3141	Linux Kernel 2.6	http://www.kernel.org/	1230	4.806666666	150

Figure 5.2. Sample data parsed into the project table in the Ohloh database

Figure 5.3 shows a sample of data about Ohloh developers. This sample includes four developers and information about seven attributes. These attributes include account name, the time the account was created, location (city and state), Country, Kudo_rank which is an indicator of community status, preferred programming language, and total number of commits to OSS projects. These seven attributes are just a part of Ohloh dataset and in one database table. More information was parsed from Ohloh web pages into other database tables.

Name	Created_at	Location	Country	Kudo_rank	Programming Language	Total Commits
Jason Allen	2006-09-15T02:23:01Z	Sammamish WA	US	9	Java	789
Robin Luckey	2006-09-15T02:23:01Z	Seattle WA	US	9	Java	11358
Scott Collison	2006-09-15T02:23:01Z	Seattle WA	US	8	C	254
The Ohloh Slave	2006-09-15T02:23:02Z	Redmond WA	US	9	Php	13

Figure 5.3. Sample data parsed into the developer table in the Ohloh database

5.4 Research design

5.4.1 Discovering semantic of social networks in Ohloh community

To address the first sub research question, we proposed to use a set of analytical methods include SNA and conditional logistic analysis to discover semantics in Ohloh social networks. The design is presented in Figure 5.4. It consists of two steps. The first step involves two components, network construction and semantic extraction. The second step, network analysis, contains both SNA topological analysis and semantic analysis.

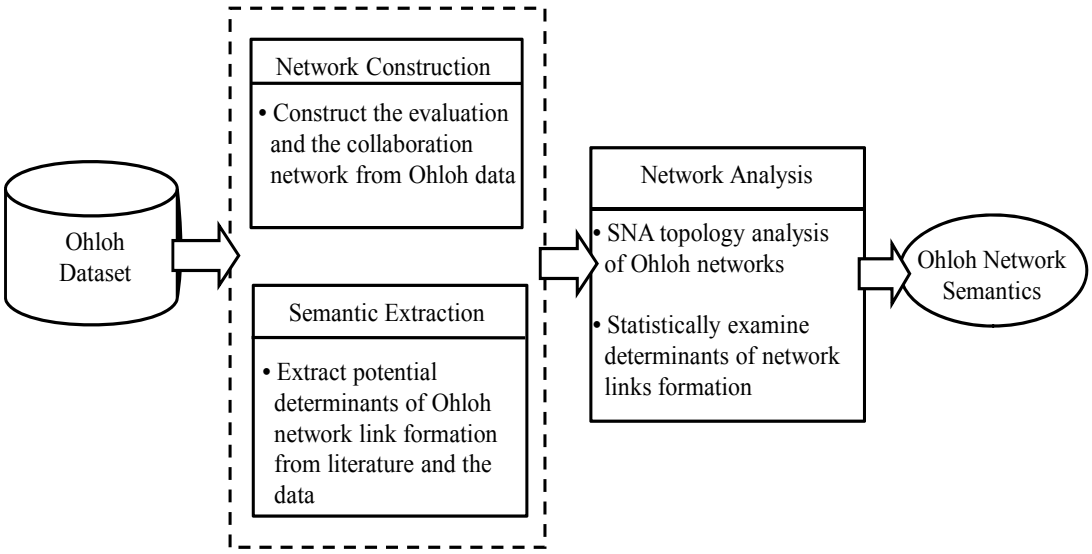


Figure 5.4. Discovering semantic of social networks in Ohloh community

In the first step, we constructed two social networks from Ohloh dataset – an evaluation network and a collaboration network – based on the "Kudo" evaluation links and past project participation information respectively. At the same time, the potential network semantics may be selected based on literature or theoretical conjectures on social networks in OSS communities.

Then the second step - network analysis involves examining the determinants of link formation processes in the two constructed social networks using conditional logistic analysis. The details of the design are introduced in the following sub-sections.

5.4.1.1 Network construction

As Figure 5.4 shows, to construct social networks from Ohloh dataset, we first need to identify the network nodes and links. Among the 94,330 developers listed in Oholh web site, 14,075 of them registered with detail information such as location and nationality. The rest only have OSS development activity information retrieved from revision control repositories. Since only registered users are allowed to send or receive "Kudo" links in Ohloh community, the constructed evaluation network only contain 3,451 developers as nodes and 9,827 evaluation links among them.

In addition, the network analysis requires detail information about individual attributes and shared affiliations which is only contained in registered accounts of Ohloh community. Therefore, we also just include the registered developers in the Ohloh collaboration network. Each collaboration link indicates that the pair of developers has worked in the same OSS project before. The resultant collaboration network includes 3,798 registered developers with 77,513 collaboration links.

5.4.1.2 Semantic (Determinants) extraction

In this study, the semantics are determinants of link formation in the Ohloh evaluation network and the collaboration network. These potential determinants were selected based on findings or conjectures from prior OSS studies (Kossinets et al. 2006; McPherson et al. 2001; Powell et al. 2005). They may include:

- **Individual attributes**: People in a social network vary in their attributes such as age, gender and experiences. Unlike nodes in theoretically constructed networks, a person in real-world networks forms relationships (links) by considering individual attributes of the candidates.

- **Shared affiliations**: The links (i.e. relationships) among people in a real-world social network are highly meaningful and vary greatly (Wasserman et al. 1994). The link formation process is largely driven by the shared affiliations of their members (e.g., attending the same class) (Kossinets et al. 2006).

In our study, we include six individual attributes and three shared affiliations. The individual attributes selected were OSS experience, coding experience (Bagozzi et al. 2006), developer degree, homophily (McPherson et al. 2001) in country, location, programming language, and community reputation (Okoli et al. 2007), while the share affiliation determinants are participation in the same OSS project, mutual acquaintance

for the evaluation relationship. These potential determinants are explained in detail as follows.

5.4.1.2.1 Individual attributes

- Coding experience: the total number of commits made by this developer for all OSS projects in Ohloh dataset (one commit refers to making changes for one time in the source code of an OSS project).
- OSS experience: the total number of months in which this developer made at least one commit to OSS projects.
- Developer degree: the number of positive evaluations the developer has (incoming or outgoing).

These individual attributes are selected assuming that a person's OSS experience level and existing evaluation links are positively correlated to the number of evaluation links he or she will receive. In sociology, such phenomenon is referred as preferential attachment or accumulative advantage (Powell et al. 2005). In our study, they are reflected in the developers' degrees and experiences of the developers.

- Homophily in country: the developer's claimed country in his or her registration file.

- Homophily in location: the developer's claimed living location (city level) in his or her registration file.
- Homophily in programming language: the programming language most often used by this developer, measured by the total number of his commits in that language.
- Homophily in community reputation (status): a score called KudoRank ranging from 1 to 10 calculated based on the number and quality of the "Kudo" links this developer received. In other words, a high KudoRank comes from not only receiving a lot of Kudo links, but also receiving Kudo links from highly ranked people.

Homophily is the phenomenon that people with similar attributes are more likely to form various social relations such as friendship and collaborations (McPherson et al. 2001). In this study, it is assessed with the above attributes. We measure the country and location differences between two developers because the homophily in geographic location is found to be a determinant for forming various relationships in prior research. We also measure homophily in primary programming language since knowing the same programming language is a usually a prerequisite for developers to collaborate in the same OSS project.

5.4.1.2.2 Shared affiliations

- Participation in the Same OSS project: a binary indicator if two developers have worked in the same OSS project before current evaluation/collaboration link forms.
- Mutual acquaintance from past collaboration(s) in OSS projects: a binary indicator if two developers both collaborated with the same developer in OSS projects before.
- Mutual acquaintance for evaluation: a binary indicator if two developers both link to a common node in the Ohloh evaluation network, indicating they both evaluate the same developer before.

The reason to include past OSS project collaborations is because past positive collaboration experiences between two developers may facilitate the positive evaluations on each other. Moreover, mutual acquaintance is included because it usually serves a bridge for two individuals to get to know each other and further form various relationships (Kossinets et al. 2006).

5.4.1.3 Network analysis

After we construct the evaluation network and extract potential determinants, we conduct SNA topology analysis on these two networks and examine the extracted determinants using conditional logistic analysis.

5.4.1.3.1 Topology analysis

There are two goals for the topological analysis of Ohloh networks in our study. Firstly, it can help us to uncover the structure of Ohloh open source software community, and better understand the nature of OSS collaboration and evaluation relationships. Secondly, the determinants of link formation are found to significantly affect the network structural changes (Kossinets et al. 2006; Powell et al. 2005). For instance, experienced developers may attract more collaborators causing a scale-free network topology. Therefore, the results of network topology analysis can be used to verify network determinants examined by statistical analysis.

Firstly, we use SNA centrality measures to describe the topology of the Ohloh collaboration and evaluation networks and identify its key members. High degrees usually indicate high levels of activities and wide social influences. Therefore, the OSS community members with large number of degrees are likely to be the leaders of their networks. The average degree of a network was also calculated to measure how dense a network is.

In addition, previous research (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks and have small-world network properties. Thus we examine both Ohloh evaluation and collaboration networks to see if they have these features. Several SNA measures were examined, including the average path length, the

clustering coefficient, link density, and the degree distribution. These properties then were checked against the small world and scale-free models.

5.4.1.3.2 Statistical analysis to examine semantics (determinants of link formation process)

Our choice of a statistical model for analyzing determinants of Ohloh network link formation is based on both theoretical and empirical considerations. Theoretically, our study intends to model human choice behavior in social networks. The research question asks what determinates account for differential (as opposed to random) patterns of the link formations in OSS evaluation and collaboration networks. Empirically, for the evaluation relationship, we need to model the choice behavior of Kudo senders to receivers. For the collaboration networks, the choices are modeled as bi-directional between two developers who collaborated in an OSS project. For these reasons, in our study, we choose to use conditional logistic model that takes each choice as a unit of analysis, which in our case are the formation of an evaluation or collaboration link between two developers.

Conditional logistic model and its variations (Jin et al. 2005; McFadden 1980; McFadden et al. 1974; Powell et al. 2005) have been widely used to model human choice behavior and examine the determinants which affect the choices. In this study, for both Ohloh evaluation and collaboration networks, the probability of an OSS developer *i* choose to

send a Kudo evaluation link to another developer j from the alternative set J_i, is specified as the following formula:

$$\Pr(y_i = j) = \frac{\exp(X_j \beta)}{\sum_j^J \exp(X_j \beta)}$$

where y_i is the observed choice for developer *i and* X_j is a vector of the characteristics of the developer j. The unknown coefficients β are typically estimated by maximum likelihood methods.

We estimate the CLM for both the evaluation and collaboration network data using *clogistic* command in Stata 10/MP. The two dependent variables are binary indicators of the outcome for link formations in the two Ohloh networks. The independent variables are the selected potential determinants explained in the semantic extraction section. In addition, for the statistical analysis, these independent variables are operationalized as Table 5.1 shows.

Table 5.1. Variables constructed for CLM analysis

Label	Description
Dependent Variables	
Evaluation (Kudo) link	Computed as a binary indicator of whether an Kudo evaluation link is sent by a OSS developer to another
Collaboration link	Computed as a binary indicator of whether a developer chooses to collaborate with another one in an OSS project
Independent Variables	
Coding experience	No. of commits made by the developer for all OSS projects in Ohloh dataset
OSS experience	No. of calendar months in which the developer made at least one commit
Developer degree	No. of the links the developer has just prior to the current link formation
Same Country	Set to '1' if the pair of individuals are from the same country and '0' otherwise
Same Location	Set to '1' if the pair of individuals are from the same city and '0' otherwise
Same Programming Language	Set to '1' if the pair of individuals' primary programming language is the same and '0' otherwise
Same KudoRank	Set to '1' if the pair of individuals have the same KudoRank and '0' otherwise

Past OSS project(s)	Set to '1' if the pair of individuals have worked in the same OSS project before and '0' otherwise
Mutual Acquaintance in the Evaluation Network	Set to '1' if the pair of individuals both have evaluated at least one same person before and '0' otherwise
Mutual Acquaintance in the Collaboration Network	Set to '1' if the pair of individuals both have collaborated with at least one same person before and '0' otherwise

To demonstrate how to construct the variables for statistical analysis from the data provided by Ohloh dataset, we use the example of developer Jason Allen and Robin Luckey in Figure 5.3. As Table 5.2 shows, first we checked if either one of them have sent kudo to the other. For example, if Allen has sent Luckey a Kudo, the constructed dependent variable – positive evaluation is 1.

Then we extracted information for the preferential attachment factors – the coding experience, OSS experience and Developer degree from Ohloh dataset. Since the conditional logistic model only considers characteristics of the developer who is chosen to be evaluated, the constructed variables for these factors will take values only from Robin Luckey.

After that, we construct variables for homophily in country, location, programming language, and Kudo rank. For instance, since Jason Allen's location is not the same as Robin Luckey's. The constructed variable for the homophily in location is set to 0.

At last, the variables for shared affiliations among developers – past project(s), mutual acquaintance by evaluation or collaboration relation are also constructed. If Jason have participated in two OSS projects - Linux and PhP while Robin Luckey only joined Firefox, the constructed variable needs to be set to 1 since they had no past collaborations. In addition, if Jason and Robin both know Scott Collision through collaborations in OSS projects. The variable for mutual acquaintance by collaboration is 1.

Table 5.2. Constructed variables from Ohloh dataset

	Jason Allen	Robin Luckey	Constructed Variable
Positive Evaluation (Kudo)	N/A	N/A	1
Coding experience	789	11,358	11,358
OSS experience	25	25	25
Developer degree	34	168	168
Same Country	United States	United States	1
Same Location	Sammamish, WA	Seattle, WA	0
Same Programming Language	Java	Java	1
Same KudoRank	9	9	1
Past OSS project(s)	Linux, PhP	Firefox	0
Mutual Acquaintance by the Evaluation Relation	Scott Collison	The Ohloh Slave	0
Mutual Acquaintance by the Collaboration Relation	Scott Collison	Scott Collison	1

After constructing those variables from the 3,451 (3,798) developers in Ohloh evaluation (collaboration) network, there are 11,905,950 (14,424,804) pairs of developers with the dependent variable and the independent variables. We then applied them to the conditional logistic analysis to determine which variable is significant in forming positive evaluations between two OSS developers.

5.4.2 Constructing semantic social networks for supporting knowledge management applications

Using the discovered semantics of social networks, we then focus on how to utilize such knowledge to support various knowledge management applications such as expert recommendation systems. This is because such knowledge is mostly needed in the applications that facilitate real-world decision making.

However, it is difficult to utilize such semantics of social networks for real-time knowledge management applications due to two reasons. Firstly, the discovered semantics of a social network are usually domain specific and each domain usually has a unique set of social factors, rules and constraints. The analysis of social network semantics heavily relies on manual work done by domain experts. Secondly, without contextual information, mathematical SNA methods often provide meaningless or inaccurate results to applications. For example, experts recommended purely based on

degree measure often do not match users' specific needs due to the lack of network contextual information (David 2003).

To address these problems, we present a semantic web based infrastructure for representing, and utilizing the semantics discovered from social networks for various real time applications. We define this infrastructure using the notion of "semantic social network". Firstly, it includes an ontology-based framework aiming to embed domain specific contextual information into mathematical representations of social networks. Secondly, a web service based architecture is presented for utilizing the discovered semantics for various social network based applications. At last, in the discussion section of this paper, we demonstrate how to utilize the discovered semantics of Ohloh evaluation and collaboration networks for an expert recommendation application in the proposed semantic web based infrastructure.

5.4.2.1 An ontology for representing social network semantics

The analytical framework of social networks, distinct from other perspectives, uses relational information to study or test theories in social and behavior sciences. The goal of social network measurement is to precisely and quantitatively translate social concepts into formal definitions (Wasserman et al. 1994). For example, the popularity of an individual in a community can be represented by the number of social relationships he has, which is termed as degree in network measurements.

However, when it comes to complex contextual information – the semantics of social networks, traditional SNA representations usually fails to precisely capture and convey such information. For instance, the reasons why an individual has large number of relationships are difficult to be represented by pure mathematical network measurements. In social network theory, such information is defined as the determinants of the social relations (2005) and vital for many SNA-based applications such as relationship prediction and expert recommendation. Therefore, a richer representation mechanism is required to take into account the semantics of social networks.

One suitable environment for supporting semantic representation of social networks is the use of Semantic Web technology, more specifically, the ontology. The use of ontology provides a formal and common understanding of critical concepts of social networks and the relationships among them. Such mechanism allows various automated SNA applications to understand, discover and utilize the semantics of social networks.

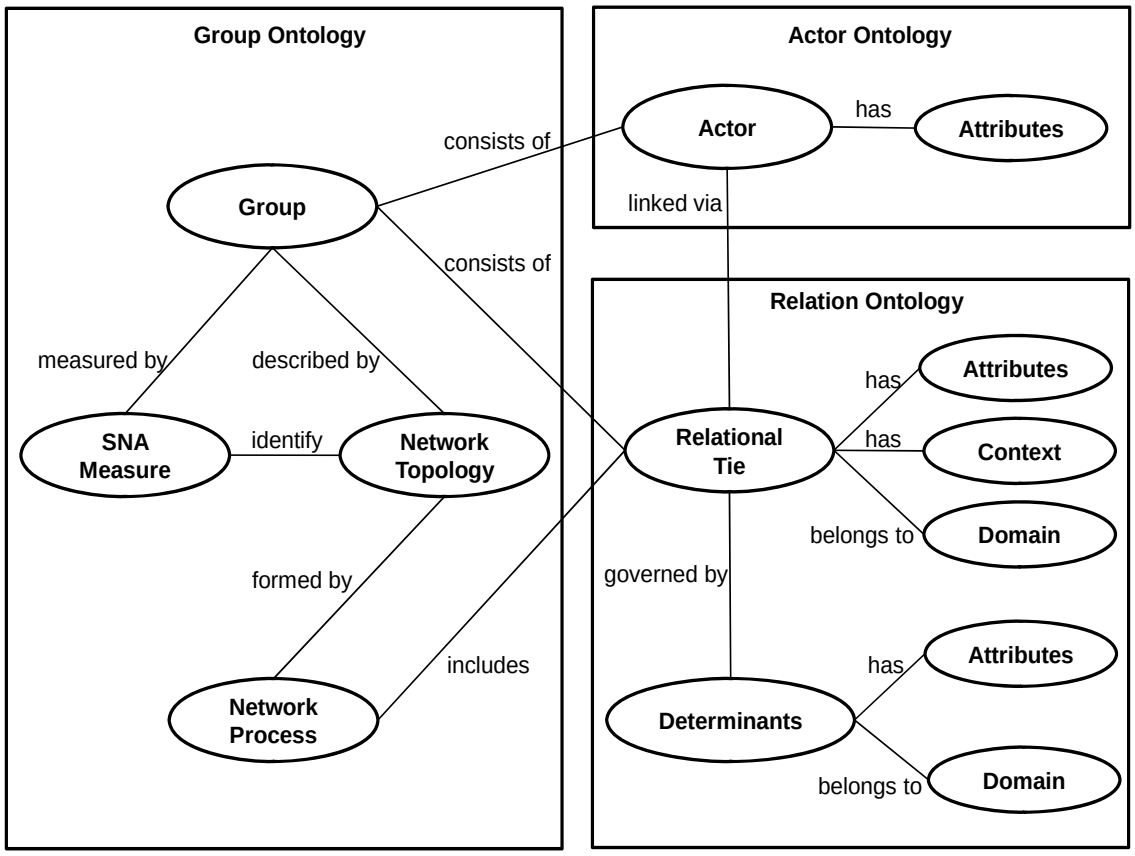


Figure 5.5. An ontology for representing social network semantics

In this research, we then develop a multifaceted reference ontology to model the semantics of social networks (See Figure 5.5). In this ontology, we use the terminologies defined by Wasserman and Faust (1994) since they are widely used in social network analysis research. The unit of analysis in social networks is a pair of actors which have many individual attributes and are linked by one or more relational tie(s). A group is the collection of a set of actors and all the relational ties among them (Wasserman et al. 1994). Therefore, this ontology mainly includes three interconnected sub-ontologies that

describe the key entities in social networks: an actor ontology, a relation ontology, and a group ontology.

In the relation ontology, we model the semantics of relational tie class with three facets: attribute, context and domain. Attribute describes the contents of the relational tie (e.g., collaboration, evaluation). Context facet represents the application context of the tie, providing information such as when this tie formed and under what circumstances (e.g., both actors prefer to use the same programming language). In addition, the domain facet provides domain-specific vocabularies for contextual information such as the name of the programming language both actors prefer to use.

The determinant class represents the factors that govern the formation of relational ties in a social network. They are the semantics discovered from social networks by using various statistical analysis methods (Kossinets et al. 2006; Powell et al. 2005). Previous SNA research identified homophily (Kossinets et al. 2006; McPherson et al. 2001) in individual attributes (i.e. tendency to associate with similar others) and shared affiliations (Kossinets et al. 2006; Powell et al. 2005) (e.g., shared memberships) are two major types of determinants of tie formation. The attribute facet of this class describes the nature (i.e. homophily or shared affiliations) and the content (e.g., attributes such as age and nationality) of the determinant. The domain facet of determinant provides information about the application domain (e.g., law, finance, medical) that allow for classification of determinants. For instance, in an open source community, the domain facet of

determinants can contain information such as developers' coding experiences or the preference of programming languages.

At last, in the group ontology, a group consists of a set of actors and the relational ties among them. The group can be analyzed by SNA structural measures such as shortest path length and degree distribution. These measures can then be used to identify the network topology of the group under study. Moreover, network processes are the different sequences of relational tie formation influencing by various social factors, therefore shaping the network topologies. For instance, preferential attachment (i.e. the rich get richer) process will generate a scale-free network topology with a power law degree distribution (Albert et al. 2002).

With this ontology, a real-world social network can be conceptualized as an approach to provide access to social network based knowledge for various applications. Semantic social networks use the context and domain ontology for two reasons: to support automatically discovering domain-specific semantics of social networks and to be discovered and utilized as a service by SNA applications, namely to determine its identity. In another word, by incorporating domain-specific concepts into a Semantic Web service of social networks, various applications can recognize and utilize the social network based knowledge it provides.

5.4.2.2 A web service based architecture for utilizing semantics of social network for various applications

In order to allow organizational applications to automatically discover and access the knowledge from various SNA-based knowledge sources (e.g., social networking sites, online communities), we developed a web service architecture called semantic social network service and enhanced it with the proposed ontology in the previous section. Similar with traditional web services, this architecture includes service providers, service requesters, and a service broker component (See Figure 5).

The SNA knowledge service providers include social network systems from various domains. For example, they may be social networking web sites, aiming at facilitating various relationships among people such as matchmaking. Another example is online software developer communities. They may offer services to recommend experts based on social network analysis of their communities to software development projects. These service providers will conduct social network analysis using standard measures and statistical analysis to discover the semantics of social networks – SNA-based knowledge. Through the interface of knowledge service broker, such domain-specific knowledge can then be embedded into Web Services Description Language (WDSL) documents by using the reference ontology developed in this paper.

On the other hand, the service requesters are usually individual actors or organizational applications such as expert recommendation systems. They may search the SNA-based knowledge services through the service broker by using ontology-based queries.

In this web service architecture, the service broker component is central for providing SNA-based knowledge services to service requesters. It has three sub-components – a semantic annotation component, an enhanced s-UDDI registry and a semantic service discovery and composition component.

The semantic universal discovery, description, and integration (s-UDDI) is a registry which hosts and provides access to the enhanced Web Services Description Language (WSDL) documents. The semantic enrichment component allows the SNA knowledge service providers to annotate their services inusing the reference ontologies and publish them in s-UDDI. These enhanced WSDL documents not only describes the protocol bindings required to interact with the SNA knowledge services, but also provides unique contextual information of each service listed in its directory, in order to facilitate semantically enhanced service discovery and invocation.

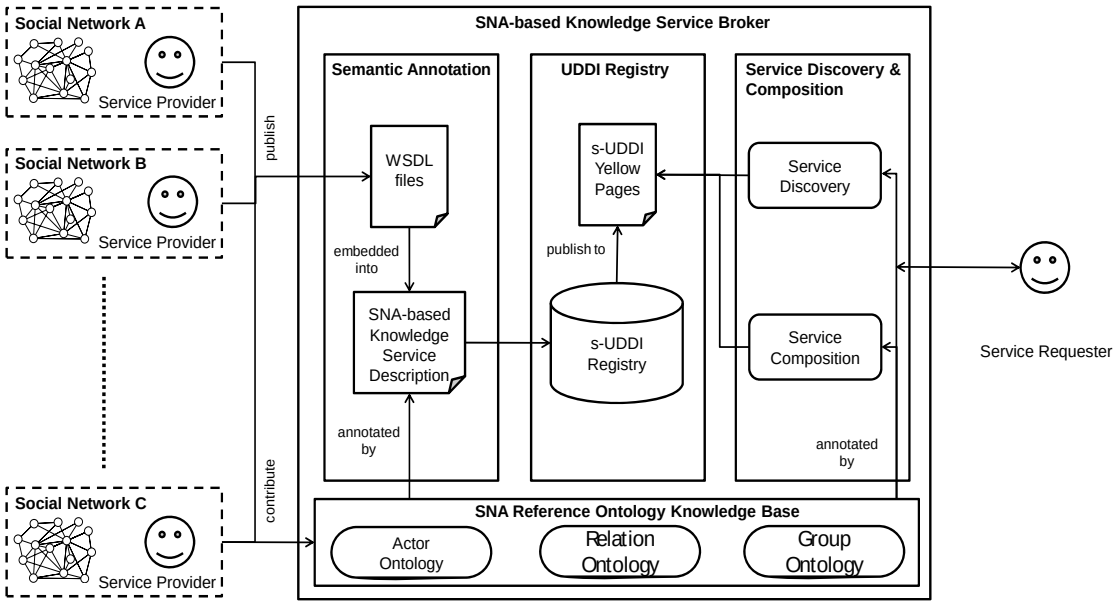


Figure 5.6. A web service based architecture for semantic social network based applications

5.5 Results for discovering semantics of Ohloh social networks

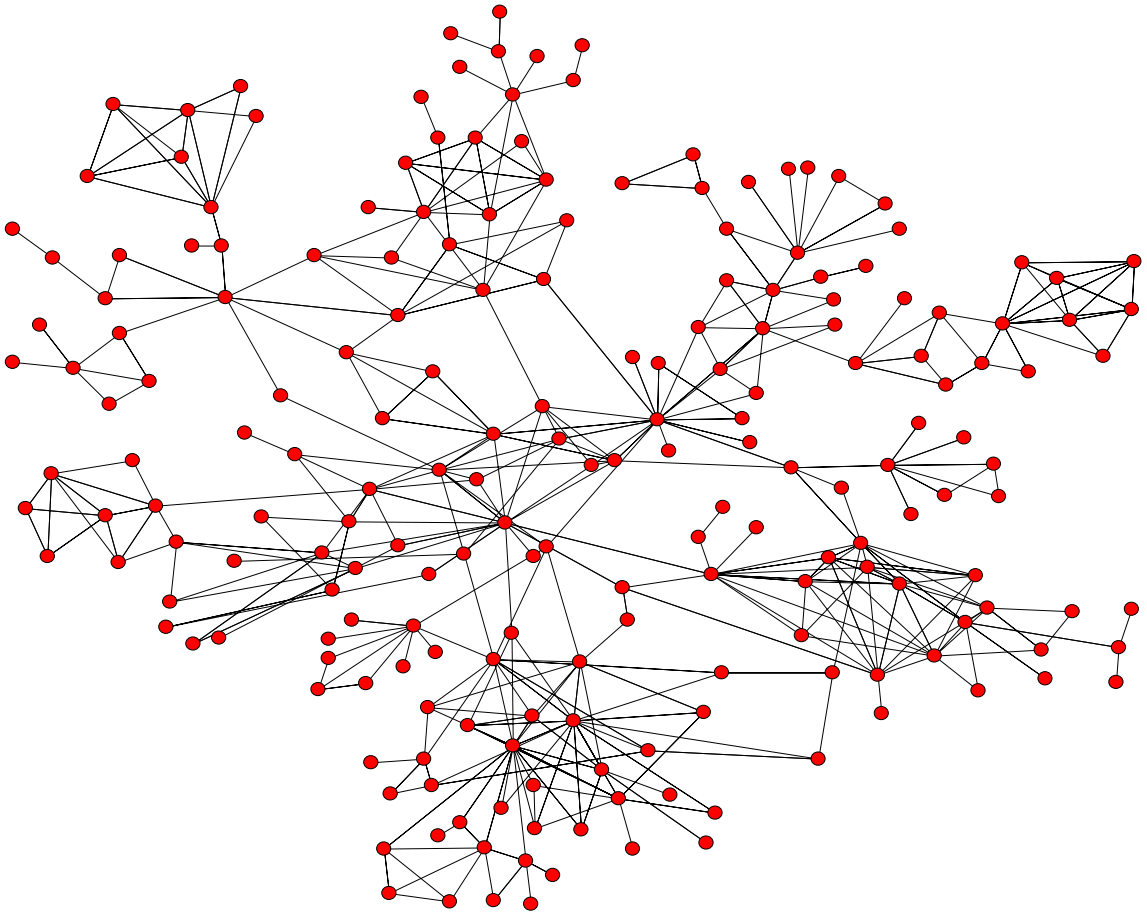
5.5.1 Topological analysis

We start from describing the basic statistics of the two networks in Table 5.3. Among the 9,827 evaluation relationships, about 93.1% of them have corresponding collaboration relationships before. This may imply that the past project collaborations may facilitate future evaluation relationships among OSS developers.

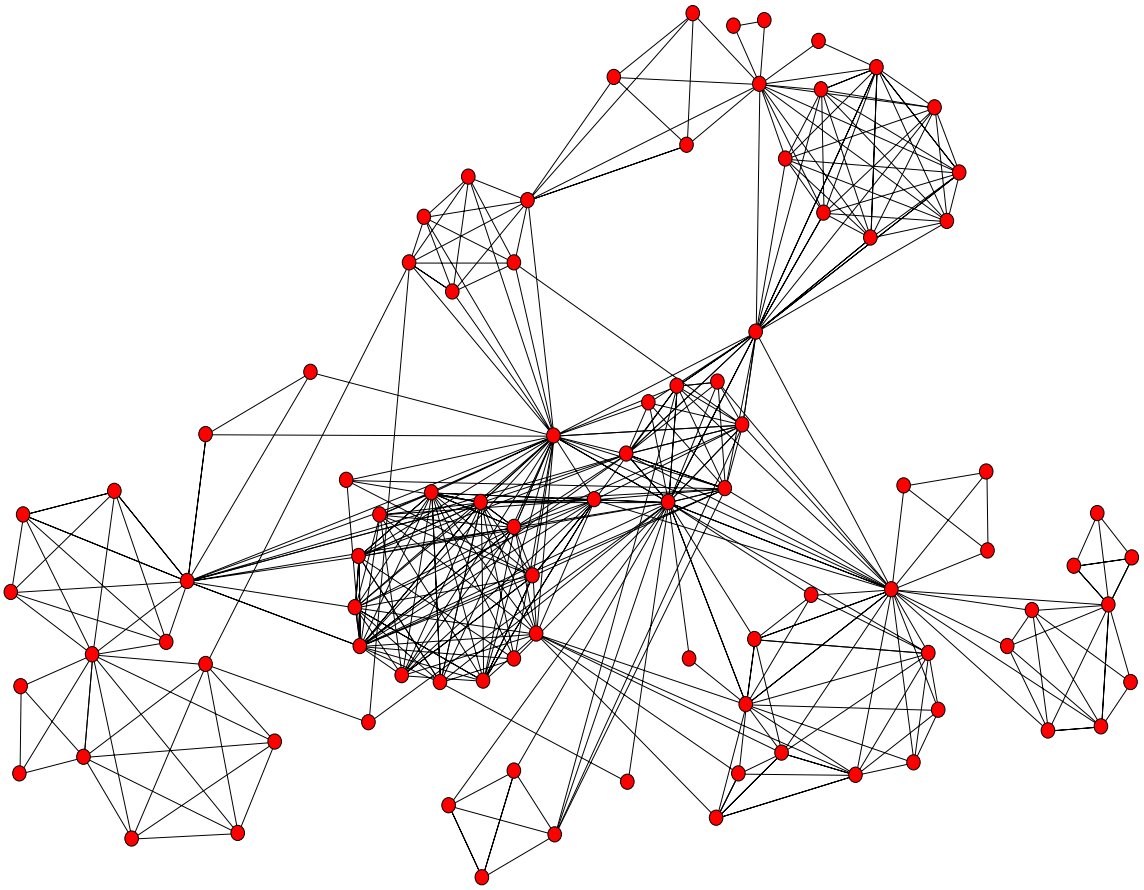
Table 5.3. Key statistics of Ohloh networks

	Evaluation Network	Collaboration Network	Overlap
No. of Nodes	3,451	3,798	2,456
No. of Links	9,827	77,513	9,124

Figure 5.7 shows a part of evaluation network and a part of the collaboration network in Ohloh community. We observe that the sample collaboration network is comprised of dense, fully connected local clusters which represent OSS projects. In contrast, for the evaluation network, it appears less interconnected inside each cluster but there are more nodes and links serving as hubs and bridges among different clusters, thus creating more critical links.



(a)



(b)

Figure 5.7. (a) Sample Ohloh evaluation network; (b) Sample Ohloh collaboration network

Then SNA centrality measures are used to describe the topologies of both Ohloh networks. These measures include the average degree, average path length, the clustering coefficient, and link density. Firstly, average degree was calculated for both networks. It was found that the collaboration network has a much larger average degree than the evaluation network, indicating denser network structure. This structural difference may be caused by the different natures of these two types of links. Collaboration links indicate

a developer's work relationships with all the project members while evaluation links are more selective and personal relationships.

In addition, prior studies (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks and have small-world network properties. Table 5.4 shows that the results of these measures on the Ohloh evaluation network and a random network with similar link density. These measures were examined for all three networks, including the average path length, the clustering coefficient, and link density. The results then were checked against the small world and scale-free features. In addition, the degree distributions of the two networks were fit to power-law distribution using linear regression technique to test for scale-free topological features.

Table 5.4. Results of SNA measures for Ohloh networks and a random network

	Evaluation Network	Collaboration Network	Random Network
Average Degree	4.44	27.60	N/A
Average Path Length	**5.643**	**4.187**	5.110
Clustering Coefficient	**0.455**	**0.876**	0.002
Link Density	0.0017	0.0077	0.0017
Degree Distribution			
R^2	**0.91**	**0.92**	N/A
γ	1.95	1.86	N/A

In our topological analysis, we focused on the largest connected clusters in the two Ohloh networks. Firstly, we found that both networks are small world networks. Their average path lengths are small with respect to their sizes. Thus, an Ohloh member can reach any other member in both networks through just 4 or 5 mediators. Another small-world property, high clustering coefficient (comparing with their random network counterpart), is also found for both networks. The clustering coefficients are significantly higher than their random graph counterpart in the fourth column.

In addition, the evaluation network is very sparse with a very low link density (Wasserman et al. 1994) of 0.0017. This property has important implications for the cost of sharing codes and other resources in OSS communities. Since such cost increases as more people join in and their relationships become denser in one project (cluster), the small average path length and link sparseness can help lower costs and enhance communication efficiency for the overall network.

The last three rows of Table 5.4 show the results of linear regressions of the degree distributions for both evaluation network and collaboration network. It was found that the evaluation network follows the power-law degree distribution (Newman 2001a), $p(k) \sim k^{-\gamma}$, with exponent $\gamma = 1.95$. The coefficient of determination R^2 of the regression for evaluation network is extremely large at 0.91 (ranging from 0 to 1), indicating high fitness of the power-law degree distribution. The collaboration network also has similar results and shows scale-free features with R^2 at 0.92 and $\gamma = 1.86$.

5.5.2 Statistical analysis

To facilitate interpretation of results from CLM analysis, we present the odds ratios and the coefficients as Table 5.5 shows. The odds ratios are obtained from the coefficients by exponentiation. That is $X_i = \exp(\beta_i)$ where β_i is the coefficient. The odds ratio measures the change of the odds that a link is formed caused by each unit increase in an independent variable (odds ratios equal to one means there is no effect and less than one

reflects negative effect). This means that the probability of the link formation would increase by a factor of odds ratio when the corresponding independent variable increases by one unit.

Table 5.5. Results from conditional (fixed-effects) logistic regression analysis

	Evaluation Network		**Collaboration Network**	
	Odds Ratio	Coefficient	Odds Ratio	Coefficient
Coding experience	1.000**	0.001*	1.001	0.001
OSS experience	1.000*	0.001*	1.000*	0.005**
Developer degree	1.049**	0.048**	1.005**	0.005**
Same Country	**5.343****	1.676**	1.009**	0.069**
Same Location	**9.017****	2.199**	**9.426***	2.243*
Same Programming Language	**3.345****	1.208**	**5.579****	1.719**
Same KudoRank	**1.488****	0.398**	**1.319****	0.277**
Past OSS project(s)	**3354.612****	8.118**	**4.347****	1.470**
Mutual Acquaintances in the Evaluation Network	**26.288****	3.269**	***	***
Mutual Acquaintances in the Collaboration Network	2.74e-10	-22.01	***	***

Note * $p < 0.05$ ** $p < 0.01$ *** the independent variable is dropped because of collinearity

Table 5.5 shows the odds ratios and coefficients for each independent variables from the conditional logistic regression analysis on both Ohloh networks. For the evaluation network, the homophily in country, location, programming language, KudoRank socre are found to be significant determinants with odds ratios larger than value 1. In addition, two shared affiliations – past OSS projects and mutual acquaintance in the evaluation network – are also found to be determinants and have larger odds ratios than other determinants.

On the other hand, in the analysis of the collaboration network, mutual acquaintance for the evaluation network and collaboration network variables are dropped from the conditional logistic analysis due to collinearity. Seven independent variables are found to be statistically significant. Four of them – same location, same country, same programming language, and past OSS projects – have odds ratios that are larger than 1.01. Therefore they are found to be significant determinants of link formation in the Ohloh collaboration network. These results with the ones for the evaluation network are discussed in the following section.

5.5.3 Determinants of link formation

For both Ohloh evaluation and collaboration network, the conditional logistic regression analysis found that the homophily in location, programming language, KudoRank, and past OSS projects are significant determinates of link formation. Those determinants implies that two previously unconnected developers are likely to evaluate or collaborate with each other if they have lived in the same city, mainly use the same programming language, have similar community reputation, or have worked in the same OSS project before.

These findings may partly be explained by the following conjectures. Geographic propinquity indicates that such developers may have more opportunity to meet each other in person and form stronger personal relationships. Consequently this may increase the likelihood of future collaborations and then evaluations.

Moreover, an OSS project usually requires a primary programming language. Therefore, developers may take that into their considerations for choosing collaborators. In addition, developers with similar OSS community reputation may be at the same stage of their OSS activities. They may have more common experiences which bring them together for collaboration or evaluation. At last, it is not surprising that past collaborations in OSS projects is also a determinant, considering 93.1% of the evaluation links have corresponding past OSS collaboration links.

Homophily in the same country is found to be a significant determinant of link formation processes only for the evaluation network but not for the collaboration network. This may be caused by the personal nature of the evaluation network. A Kudo evaluation link means a developer explicitly select another one for positive evaluation. This indicates the Kudo sender must know the receiver in some depth. However, a collaboration link only indicates two developers have worked in the same OSS project but not necessarily know each other in person. In addition, global collaborations in OSS development are becoming more popular and common due to the emergence of various communication technologies such as Internet. Therefore, homophily in nationality as a personal attribute may have more weights in influencing link formation processes in the evaluation network than the collaboration network.

Another significant determinant – mutual acquaintance in evaluation network – implies that two previously unconnected individuals are likely to evaluate each other with one or more shared acquaintances. This determinant has been well studied in social network analysis research (Kossinets et al. 2006; McPherson et al. 2001). It was found that individuals tend to select new acquaintances who are friends of friends. In the evaluation network under study, this suggests that OSS developers tend to have circles of trust that include close acquaintances. They are likely to form operational cliques which enhance communication within the network and increase the capacity to act. This is also in line with the social closure theory (Coleman 1990) which suggests that the greatest value is

obtained from networks that are densely connected with a high level of trust among actors.

5.5.4 Insignificant factors

One surprising finding is that all preferential attachment factors of OSS developers – OSS and Coding experiences and OSS developer (Kudo Receiver) degree – have little effect on the positive evaluations. This finding is congruent with a proposition in another OSS study (Stewart 2005) that a developer's ability to attract positive evaluations becomes increasingly institutionalized or stable and as such, other community members are less likely to evaluate this developer. Therefore, one very practical implication for an OSS project manager is that experienced developers who already are highly recognized in the community may not have guaranteed positive evaluations when including them to his or her project. The managers may not want to waste their efforts or resources to attract such developers. On the other hand, for developers who desire to gain many positive evaluations, they need to do so quickly. Because a developer's ability to attract positive evaluations tends to become stable over time, he or she may run the risk of falling to an inert position with few positive evaluations in the OSS community.

5.5.5 Impacts of determinants on positive evaluations

In addition, the values of various determinant coefficients vary from 8.118 to 1.208. We found that the more personal the context for the determinant, the larger its coefficient

becomes. For example, two developers participating in the same OSS project indicates a higher chance that they may know each other in person than just being in the same country. Accordingly, past OSS collaboration experience has a larger coefficient (impact) on forming positive evaluation relationship than homophily in the nationality. This phenomenon may be caused by the personal nature of the evaluation links in OSS community.

5.6 Expert recommendation mechanisms based on semantics of social networks

Expert recommendation is an important aspect of effective collaboration in networked organizations. Existing expert recommendation methods mainly use topological measures to locate experts in a social network. Such systems mainly focus on reflecting network members' aggregated views on experts but largely ignore individual users' perspectives. In this paper, we then develop a new mechanism for expert recommendation by utilizing both topological measures and semantics of social networks, thus improving the quality of expert recommendation.

In addition, the effectiveness of the new method is evaluated using data from Ohloh community and compared with both mechanisms that reflect individual user's perspective and the aggregated views of network members. All three approaches use the positive evaluation relations among Ohloh members to infer and recommend experts. The

evaluation compares the effectiveness of the recommendation mechanisms based on group perspective, individual user's perspective and a hybrid perceptive which combine the first two. The focus of this contribution is on examining the effectiveness of different social network perspectives in the context of OSS expert recommendation. The empirical insights derived from this evaluation may help design better social network based expert recommendation systems.

5.6.1 User-based link prediction mechanism

Recently, the use of social networks in expert recommendation systems is becoming increasingly popular, most notably in ReferralWeb (Henry et al. 1997), Yenta (Leonard 1997), and Expertise Recommender (David et al. 2000). These systems search for people with appropriate expertise in the users' social networks to answer questions, solve problems or provide collaborations (Jun et al. 2005). However, oftentimes users find that the experts recommended by such social network based systems do not match their needs and are difficult to interact and collaborate with. This may be because such systems focus on reflecting network members' aggregated views on experts but largely ignore individual users' perspectives. The implications of adopting different social network based designs in expert recommendation systems have rarely been studied. There are many factors needs to be evaluated from both an individual user and group perspective (David 2003).

Our analysis on presents a novel approach to discover semantics of social networks and quantify their impacts on network link formation processes. This approach actually studied the determinants of link formation from the individual users' perspective since the unit of analysis is a dyad. Then the second research question naturally arises – how to use the discovered semantics in the design of expert recommendation systems that reflect individual user's perspective. The infrastructure proposed in the research design section partly answers this question. To demonstrate how to implement that infrastructure in the application domain, we design a computational mechanism for expert recommendation application to predict users' positive evaluation choices in OSS social networks based on the discovered semantics.

As Figure 5.8. shows, the proposed mechanism consists of three steps: data processing, conditional logistic analysis, and expert ranking. Firstly, the semantic information – homophily and shared affiliations - is extracted by matching users' profiles with candidate experts' information. Secondly, the conditional logistic model is used to calculate the probability for a user to positively evaluate a candidate expert based on the discovered network semantics. For example, as Table 5.5. shows, in the Ohloh evaluation network, the coefficient for homophily in location (city) is 2.199 and coefficient for using the same programming language is 1.208. If two developers a and b both live in New York City and use Java as their primary programming language (without any other homophily and shared affiliations), the probability for a to positively evaluate b from an alternative set J_i can be calculated as $\Pr(y_a = b) = \exp(1 \times 2.199 + 1 \times 1.208) / \sum_j^J \exp(X_j \beta)$. This

calculation can be applied to any pair of members in Ohloh community. Then a list of all candidate experts can be ranked based on this link prediction probability for each user. The more likely a candidate expert is positively evaluated by the user, the higher this expert is ranked in the recommendation list.

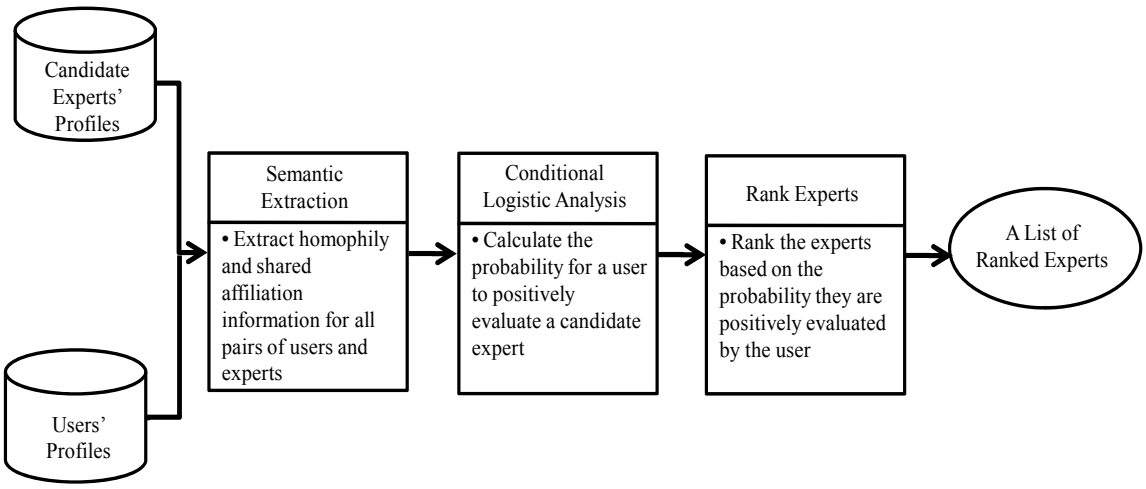


Figure 5.8. User-based link prediction mechanism for expert recommendation

This mechanism can quantitatively reflect individual users' evaluative opinions on each recommended expert. By embedding this user-based mechanism into the design of SNA based expert recommendation systems, the recommendation results will be more acceptable for users.

5.6.2 Support top-N most recognized mechanism with semantics of social networks

While the link prediction mechanism has reflected individual user's perspective, our analysis can also enhance expert recommendation by providing semantic (contextual) information from the network perspective. A very common mechanism in SNA based expert recommendation is to measure the number of links (i.e. degree) an expert receives in the network. The underlying assumption is that the more links an expert receives, the more recognized and popular this person is. This mechanism reflects an aggregated opinion on the candidate expert by other members in this community. In this research, we defined it as the *top-N most recognized* mechanism.

The *top-N most recognized* mechanism actually reflects an aggregated view on the identified node by most members in this community. Advanced link analysis algorithms with similar functions include PageRank and HITS. They not only consider the number of links but also the importance of the nodes which send the links. However,

However, the experts identified by top-N most recognized mechanism do not always meet users' needs in terms of required expertise. This is mainly because the empirical network datasets usually lack of links with expertise information. Social network analysis on such datasets mainly relay on collaboration and communication links (e.g., emails), and largely ignore network semantics. For instance, a software developer is looking for an expert to answer his questions about java programming. However, the topological

analysis of his email network only identifies the individual who receives most emails but cannot tell why.

Our analysis can address this problem by integrating the discovered network semantics with corresponding links to by constructing semantic social networks proposed in the research design section. Figure 5.9. shows an example of using semantic social networks for expert recommendation. It uses Ohloh community and the results from our semantic analysis as the setting. We assume that there are only five developers with ID number ranging from 1 to 5 in the dataset. They have two types of relationships – collaborations and evaluations.

In this example, a user Josh is starting an OSS project using XML languages. He cannot run his first XML-based Java application and needs some help from others. He wants to use expert recommendation systems to find an expert who is most recognized by others for his 1) XML-related knowledge and 2) mutual acquaintance connections. The first criterion proves the recommended individual's expertise in related technical area. The second one increases the probability of reaching such expert through mutual acquaintances and guarantees the expert himself is well connected to reach out for help.

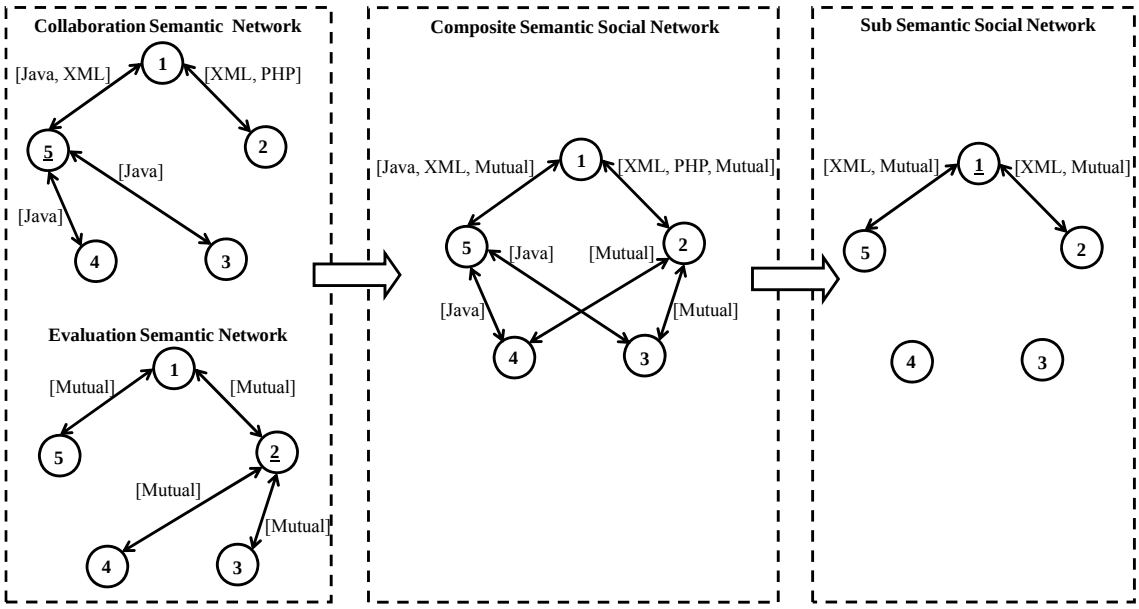


Figure 5.9. An example of using semantics social networks to support top-N most recognized expert recommendation mechanisms

There are three steps in this example. The first step shown on the left panel of Figure 4 presents the semantics identified from our analysis for both Ohloh collaboration network links and evaluation network links. For example, the identified semantic of the collaboration links is homophily in the programming language. Therefore, in the collaboration network, the semantic of the link between developer 1 and 5 is noted as [Java, XML], indicating they both use Java and XML as their programming languages. The second step shown in the middle involves aggregating the two types of links into one composite link and integrating all its identified semantics to this link. This composite semantic social network uncovers the structure of all types of relationships among the developers with their corresponding semantics.

At the third step, a sub semantic social network is extracted from the composite one and based on the semantics queried by the user Josh – homophily in using XML and mutual acquaintances. The recommended developer from the sub semantic network is developer 1 since it has the largest number of links with these two semantics.

Without the identified semantics of social networks, the structural analysis based expert recommendations will yield distinct results from the recommended developer. In the collaboration network, developer 5 is recommended since it has the largest number of collaboration links. For the evaluation network, developer 2 with the most evaluation links is selected. However, those recommendation results are far less accurate than the one from the sub semantic social network since they just simply count the number of links without considering the semantic information required by the users. Therefore, semantic social networks may help provide more accurate expert recommendations to match users' specific needs.

5.6.3 An integrated approach: combining both user-based link prediction mechanism and top-N most recognized mechanism

While the traditional SNA approach characterizes a community perspective on the expertise recommendation problem, the link prediction approach reflects an individual user's perspective. We believe that the accuracy and efficiency of expert recommendation can be improved by combining both perspectives. Thus we propose an integrated

approach that leverages the strength of both approaches discussed above. Given a query q, we generate two ranked lists of experts:

- L_1, computed using traditional SNA methods. Specifically, each expert e_i is given a score s_i^1 equal to the number of evaluation links he or she receives. Then the list is ranked by s_i^1 in descending order. Each expert gets a rank r_i^1.

- L_2, computed using the link prediction method described in section 3. Specifically, each expert e_i is given a score s_i^2 equal to the predicted probability of the CLM model. Then the list is ranked by s_i^2 in descending order. Each expert gets a rank r_i^2.

We then combine the two rankings by generating an overall score s_i for each e_i, where $s_i = r_i^1 + r_i^2$. All experts are then re-ranked by this score in ascending order. The top experts in this final list will be returned as recommendations.

5.6.4 Experimental evaluation

5.6.4.1 Experiment setting

We conducted an experiment using the Ohloh data set to evaluate performances of the three algorithms mentioned in Section 2, 3 and 4. One of the biggest Ohloh project – GNOME was used as testing dataset. Its 168 members are treated as users of expert

recommendation services and evaluation “Kudo” links as “unknown” future links. The rest of the Ohloh is used as training dataset by the three algorithms to generate ranked lists of recommended experts for the users. The algorithms are set to generate a ranked list of recommendations of γ experts. For each user, the recommendation quality was measured based on the number of hits (i.e. recommendations that match the people actually being positively evaluated by the user) and their positions in the ranked list. We adopted the following recommendation-quality metrics and ranking quality of the ranked list recommendation from the literature (Breese et al.):

$$\text{Precision: } P_d = \frac{Number \text{ of hits}}{\gamma}$$

$$\text{Recall: } R_d = \frac{Number \text{ of hits}}{Number \text{ of developers in this OSS project}}$$

$$\text{F measure: } F_d = \frac{2 \times P_d \times R_d}{P_d + R_d}$$

$$\text{Rank score: } RS_d = \frac{q_{dj}}{2^{(j-1)/(\alpha-1)}}$$

where d is the user and j is the index for the recommended experts in the ranked list; α is the viewing halflife (the rank of the expert on the list such that there is a 50% chance the user will positively evaluate that expert); and $q_{dj}\begin{cases}1 & \text{if j is in the testing dataset} \\ 0 & \text{otherwise.}\end{cases}$

The precision, recall, and F measure are standard performance measures for the relevance and coverage of the recommended experts relative to the users' potential choices of positively evaluating those experts. Moreover, an aggregated rank score for all developers in the testing dataset was also derived as $RS = 100\frac{\sum_d RS_d}{\sum_d RS_d^{\max}}$, where $RS_d^{\max}$ was the maximum achievable score for developer d if all positively evaluated developers are on the top of the ranked recommendation list. The rank score is used to evaluate the ranking quality of the recommendation list. For instance, assume that two lists based on algorithms A and B respectively have both included the same set of correct recommendations, they will have achieve the same precision, recall, and F measure. However, if all the correct recommendations in the list based on A ranked higher than the one in B, algorithm A will have a higher rank score.

5.6.4.2 Evaluation results and discussions

In the experiment, the number of recommended experts γ was set to 5, 10, and 20, respectively, for different experimental configurations. Table 5.6. shows the results of evaluation measures for a list of 10 recommended experts using the three approaches

(lists with 5, 15 and 20 experts have similar evaluation results). We observe that no one approach significantly outperformed the other two in precision, recall, F-measure. It indicates that these three expert recommendation algorithms do not perform significantly differently in terms of relevance and coverage. This may be due to the limited size of the testing dataset. In a relatively small community, an individual user's view on experts may not differ much from the overall community's view. We plan to extend the size of testing dataset in our future research.

Table 5.6. Performances of three recommendation approaches

	Top-N Most Recognized	User-based	Integrated
Precision	**0.0685**	0.0661	0.0631
Recall	0.1313	**0.1367**	0.1340
F-measure	0.0720	**0.0733**	0.0706
Rank score	7.0182	3.7461	**12.9659**

However, the Rank score for the integrated approach is much larger than the other two recommendation approaches. It means that the ranking quality of the integrated approach significantly outperforms both the Top-N most recognized and user-based link prediction approach.

The findings from evaluation experiments advocate the integration of the group/network perspective and individual perspective in order to improve the quality of expert

recommendation results. We demonstrated empirically how the user-based (individual) approach may greatly enhance the traditional Top-N most recognized (group) approach in terms of ranking quality.

5.7 Conclusions

Although there have been a fair amount of studies analyzing the various relationships in OSS social networks using SNA, most of such work has typically focused on network topologies or processes, but without specifying empirically the underlying determinants that significantly affect them. In this paper, we use both SNA topological analysis and conditional logistic analysis to discover the semantics (determinants) of link formation in two real-world OSS social networks. The results showed that both networks show features of scale-free and small world topologies. We also found that the homophily in country, location, programming language, and KudoRank score are significant determinants for both networks under study. Homophily in the same country and mutual acquaintance are found to be significant only for the evaluation network. The set of methods for the semantic analysis of SNA used in our study may be applied to other types of networks.

We also explored the possible social causes and implications for these identified determinants. Our analysis may help researchers and practitioners in the OSS community to better understand the evaluation-induced motivation for OSS developers and devise

useful strategies to enhance such motivation. For example, as stated in the discussion, the developers who intend to get many good evaluations should do it quickly before his ability to do so become institutionalized.

Then, from the design science perspective, we proposed an infrastructure - semantic social networks - for the representation, discovery and utilization of knowledge based on analysis of social relationships, is critical for the success of social systems in the future interconnected world. This infrastructure will allow various social networks (service providers) to provide SNA-based knowledge to social systems through a universal ontology and a web service framework. With the advancing information technology, semantic social networks will help facilitate a human-machine environment that combines the physical world, social world, and digital virtual world together (Hai 2005), to support various SNA-based applications such as expert recommendation, generating interest group, and human resource management.

In addition, within this infrastructure, we discussed how to embed network semantics into the designs of three expert recommendation mechanisms. Firstly, based on the identified determinants and conditional logistic model, we proposed a link prediction mechanism that can quantitatively reflect Ohloh developers' evaluative opinions on one another at the dyadic level. This prediction mechanism may be used in the design of information systems to enhance motivation and support collaborations for OSS developers. It can quantitatively calculate the likelihood for a user to positively evaluate a recommended

expert based on their homophily and shared affiliations. For the top-N most recognized mechanism, we showed an example of constructing and utilizing semantic social networks to provide contextual information to meet users' specific needs. Moreover, we developed a new expert recommendation mechanism by combining both Top-N most recognized and user-based link prediction mechanisms, thus improving the quality of expert recommendation. The effectiveness of all three mechanisms is evaluated. The initial results indicate that the new expert recommendation method is more effective.

Our analysis may help the research and practitioner in design science community better understand the semantics of social networks and devise various business applications. Our future work consists of several directions including (1) investigating other empirical OSS evaluation networks to further validate the determinants found in this study, (2) applying the proposed methods to analyze social networks in other domains, and (3) using the knowledge and the link prediction mechanisms derived from this study to design various collaborative information systems such as expert recommendation systems. Our efforts will open a new venue of research in the expert recommendation systems and semantic social network analysis.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

Nowadays organizations are more and more connected and form various networks ranging from strategic alliance networks to customer-supplier networks, aiming to better manage their members, resources, information and knowledge assets in business operations. These organization networks evolve over time with the addition/removal of nodes (i.e., organizations or individuals) and the formation of various links (i.e., relationships). The formation of such networks shapes the structures, functions and behaviors of these networked organizations. The knowledge gained from analyzing the structure and formation of these networks can be used to support knowledge management and decision making in various domains such as open source software project management, e-commerce, and security informatics. In addition, knowledge of the mechanisms behind the network evolution can also be used in designing information systems to support various business applications such as expert recommendation.

This dissertation aims to describe, model, and predict the link formation processes of various social networks for supporting knowledge management and decision making in organizations. The four essays focus on one or more of these goals with empirical studies in various domains. These essays collectively contribute to social network theory, open source software development, and knowledge management in security informatics by describing changes in structures of terrorist and criminal networks, modeling and examining determinants of link formation processes in OSS developer participation,

collaboration and evaluation networks, and proposing a novel semantic social network framework for supporting business applications such as expert recommendation. In this Chapter, I summarize the main conclusions and contributions of this dissertation, discuss the relevance of this research to Management Information Systems research, and suggest future directions.

6.1 Conclusions and contributions

The conclusions and contributions for each of the four dissertation essays can be summarized as follows.

- The first essay studied the topological changes of a global terrorist network to uncover the survival mechanisms of a terrorist organization. It found that three factors may have contributed to the survival of the network: growth, scale-free topology, and the ineffectiveness of the counterattack measures. The network experienced three distinct stages of growth from 1989 to 2003: emerging stage, maturing stage, and disintegrating stage. The network displayed different growth patterns in different stages. It also found that the scale-free topology could partly account for the network's robustness, helping the network survive under constant counterattacks from authorities. The scale-free topology gradually emerged as new members joined in on an operational basis and the hubs acquired connections over time. On the other hand, the network could have remained active after

numerous arrests of its members because the damages were localized within operations to a large extent. In addition, the leaders in the network are difficult to capture or remove and continue to function as hubs connecting members. Although numerous arrests and counterattacks have weakened the network, it still remains functional and has the potential to grow.

- The second essay developed a set of dynamic SNA methods to model and examine the facilitators of link formation in a real-world criminal network. It studied several possible facilitators including homophily, mutual acquaintances, and various affiliations. The results showed that mutual acquaintances and shared vehicle affiliations were significant facilitators in the network under study. We also quantified the influences of the facilitators on the triadic closure using the hazard ratios of Cox regression analysis and used the information to calculate the likelihood of future co-offending. In addition, this essay examined the possible social causes and policy implications for the significance and insignificance of various facilitators. The set of generic dynamic SNA methods along with the corresponding statistical analyses used in our study may be applied to other types of networks to test the effect of various facilitators. This research may help the academic and practitioner community better understand the dynamics of social networks and devise effective strategies to influence their growth.

- The third essay focused on analyzing the motivations of OSS project participation choices. Previous research mainly focused on attributes of individual developers or projects, largely ignoring the relationships between these two entities and the underlying determinants that significantly affect the developers' participation choices. In this study, we used both SNA and conditional logistic analysis to study the determinants of developers' participation choices in a large online OSS community – Ohloh. It was found that homophily in programming language and project mutual acquaintances are significant determinants. The set of methods used in this research can also be applied to study determinants and topologies of social networks in other domains. This essay also explored the possible social causes and implications for these identified determinants. Our analysis may help researchers and practitioners in OSS communities to better understand the developers' participation choice behaviors and devise useful strategies for OSS project management.

- The fourth essay uses both SNA topological analysis and conditional logistic analysis to discover the semantics (determinants) of link formation in two real-world OSS social networks. The results showed that both networks show features of scale-free and small-world topologies. We also found that homophily in country, location, programming language, and KudoRank score are significant determinants for both networks under study. Homophily in nationality and mutual acquaintances are found to be significant only for the evaluation network. The set

of methods for the semantic analysis of SNA used in our study may be applied to other types of networks. The study also explored the possible social causes and implications for these identified determinants. Our analysis may help researchers and practitioners in the OSS community to better understand the evaluation-induced motivation of OSS developers and devise useful strategies to enhance such motivation.

Then, from the design science perspective, this essay proposed an infrastructure - semantic social networks - for the representation, discovery and utilization of knowledge about the semantics of social networks. This infrastructure will allow various social networks (service providers) to provide SNA-based knowledge to social systems through a universal ontology and a web service framework. In addition, within this infrastructure, we discussed how to embed network semantics into the design of three expert recommendation mechanisms. The effectiveness of all three mechanisms is evaluated. The initial results indicate that the new expert recommendation method is more effective. Our analysis may help researchers and practitioners in the design science community better understand the semantics of social networks and devise various business applications.

6.2 Contributions and relevance to business and management information systems research

As individuals and business firms become increasingly connected in a network form through various communication and collaboration technologies, the science of (social) networks has brought a new perspective to study various business and MIS problems. These networks evolve over time and influence the business and managerial functions, behaviors, and performances of the networked firms. While previous network studies focused on the static topology of such networks, largely ignoring their evolution, this dissertation provides a framework to analyze, model and predict the evolution of these business networks and design business applications. In particular, researchers and practitioners may find a number of opportunities in the business and MIS domains to adopt the semantic social network framework proposed in this dissertation.

- Studying the structure and formation of customer-supplier networks in various industries. This may help management to devise and examine strategies to identify key customers (suppliers) at different times to improve supply chain management and customer relation management. The methods used to describe the evolution of social networks in Chapter 1 are applicable to the challenges in this domain.

- Studying the determinants of firms' choices in making allies in business alliance networks. This may help firms to choose the most suitable partner firms to

improve their efficiency in various business operations. The methods in Chapter 4 are useful for solving this issue/challenge.

- Studying the mechanism behind link formation processes in social networks can also facilitate the design of various knowledge management systems within business firms such as expert recommendation systems. Such systems may improve the efficiency of knowledge transfer within or across different departments in business firms. The infrastructure for discovering and utilizing semantics of social networks for knowledge applications proposed in Chapter 5 are applicable in this area.

In addition, studying the evolution of social networks may also provide opportunities in other research areas such as e-commerce, recommendation systems, and collaboration systems.

6.3 Future directions

Besides the above-mentioned research opportunities in business and MIS domains, there are several potential future directions that will build on the methods and framework presented in this dissertation. An important direction of my future work is the area of inter-organizational information systems (IOIS). Analyzing the structure and evolution of organizational alliance networks may provide insights in designing more efficient inter-organizational information systems. Another immediate area of interest is designing more

efficient recommendation algorithms in other application domains such as online marketing based on the semantics of social networks. A third planned direction is to extend the semantic social network infrastructure to a platform that includes the representation, modeling and utilization of social network-based knowledge to support various business applications. Such a platform can integrate network-based information from different data sources. It can also facilitate analysis of the network information and provide universal representation of the knowledge generated from such analysis. Then this platform can provide web service-based interface for various applications to use the network-based knowledge.

REFERENCES

Abello, J., Pardalos, P., and Resende, M.G.C. "On maximum clique problems in very large graphs," in: *External memory algorithms*, American Mathematical Society, 1999, pp. 119-130.

Afifi, A., Clark, V., and May, S. *Computer-Aided Multivariate Analysis*, (Fourth ed.) Chapman & Hall, 2003.

Albert, R., and Barabási, A.L. "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* (74:1) 2002, pp 47-97.

Albert, R., Jeong, H., and Barabasi, A.L. "Error and attack tolerance of complex networks," *Nature* (406:6794) 2000, pp 378-382.

Albert, R., Jeong, H., and Barabasi, A.L. "Internet - Diameter of the World-Wide Web," *Nature* (401:6749) 1999, pp 130-131.

Allen, J., Collison, S., and Luckey, R. "Ohloh Web Site API," http://www.ohloh.net, 2009.

Amaral, L.A.N., Scala, A., Barthelemy, M., and Stanley, H.E. "Classes of small-world networks," *Proceedings of the National Academy of Science of the United States of America* (97) 2000, pp 11149-11152.

Ba, S. "Establishing online trust through a community responsibility system," *Decision Support Systems* (31:3) 2001, pp 323-336.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. "Group Formation in Large Social Networks: Membership, Growth, and Evolution," 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, Philadelphia, PA, USA, 2006.

Bagozzi, R.P., and Dholakia, U.M. "Open source software user communities: A study of participation in Linux user groups," *Management Science* (52:7), Jul 2006, pp 1099-1115.

Barabási, A.-L., and Albert, A.-L.R. "Emergence of scaling in random networks," *Science* (286:5439) 1999, pp 509-512.

Barabási, A.-L., Jeong, H., Zéda, Z., Ravasz, E., Schubert, A., and Vicsek, T. "Evolution of the social network of scientific collaborations," *Physica A* (311) 2002, pp 590-614.

Bavelas, A. "Communication Patterns in Task-Oriented Groups," *The Journal of the Acoustical Society of America* (22:6) 1950, pp 725-730.

Beckman, C.M., Haunschild, P.R., and Phillips, D.J. "Friends or Strangers? Firm-Specific Uncertainty, Market Uncertainty, and Network Partner Selection," *ORGANIZATION SCIENCE* (15:3), May 1, 2004 2004, pp 259-275.

Bergquist, M., and Ljungberg, J. "The power of gifts: organizing social relationships in open source communities," *Information Systems Journal* (11:4) 2001, pp 305-320.

Berners-Lee, T., and Fischetti, M. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor* Harper San Francisco, 1999.

Boissevain, J. *Friends of Friends: Networks, Manipulators and Coalitions* Oxford: Blackwell, 1974.

Breese, J., Heckerman, D., and Kadie, C. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," 1998, pp. 43-52.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., State, R., Tomkins, A., and Wiener, J. "Graph structure in the web," *Computer Networks* (33:1-6) 2000, pp 309-320.

Byrne, D.E. *The attraction paradigm* Academic Press, New York, 1971.

Carley, K.M., Dombroski, M., Tsvetovat, M., Reminga, J., and Kamneva, N. "Destabilizing dynamic covert networks," the 8th International Command and Control Research and Technology Symposium, Washington DC., VA, 2003.

Caroline, H. "Learning and knowledge networks in interdisciplinary collaborations," *Journal of the American Society for Information Science and Technology* (57:8) 2006, pp 1079-1092.

Chen, H. "Intelligence and Security informatics: Information Systems Perspective," *Decision Support Systems* (41:3) 2006, pp 555-559.

Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. "Resilience of the internet to random breakdowns," *Physical Review Letters* (85:21) 2000, pp 4626-4628.

Coleman, J.S. *Foundations of Social Theory* Harvard University Press, Cambridge, MA, 1990.

Coles, N. "It's not What you know - It's Who you know that counts. Analyzing Serious Crime Groups as Social Networks," *British Journal of Criminology* (41:4), FAL 2001, pp 580-594.

Crowston, K., and Howison, J. "The social structure of open source software development teams," *OASIS 2003 Workshop (IFIP 8.2 WG)*) 2003.

Crowston, K., Howison, J., and Annabi, H. "Information systems success in free and open source software development: theory and measures," *Software Process: Improvement and Practice* (11:2) 2006, pp 123-148.

Crowston, K., and Scozzi, B. "Open source software projects as virtual organisations: competency rallying for software development," *Software, IEE Proceedings* - (149:1) 2002, pp 3-17.

Crucitti, P., Latora, V., Marchiori, M., and Rapisarda, A. "Efficiency of scale-free networks: Error and attack tolerance," *Physica A* (320) 2003, pp 622-642.

Csányi, G., and Szendroi, B. "Structure of a large social network," *Physical Review E* (69) 2004, p 036131.

David, W.M. "Recommending collaboration with social networks: a comparative evaluation," in: *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, Ft. Lauderdale, Florida, USA, 2003.

David, W.M., and Mark, S.A. "Expertise recommender: a flexible recommendation system and architecture," in: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, Philadelphia, Pennsylvania, United States, 2000.

Davis, G.F., Yoo, M., and Baker, W.E. "The Small World of the American Corporate Elite, 1982-2001," *Strategic Organization* (1:3), August 1, 2003 2003, pp 301-326.

DeLone, W.H., and McLean, E.R. "The DeLone and McLean model of information systems success: a ten-year update," *Journal of Management Information Systems* (19:4), Spr 2003, pp 9-30.

DeLone, W.H., and McLean, E.R. "Measuring e-commerce success: Applying the DeLone & McLean information systems success model," *International Journal of Electronic Commerce* (9:1), Fal 2004, pp 31-47.

Dombroski, M.J., and Carley, K.M. "NETEST: Estimating a terrorist network's structure," *Computational & Mathematical Organization Theory* (8) 2002, pp 235-241.

Ducheneaut, N. "Socialization in an Open Source Software Community: A Socio-Technical Analysis," *Computer Supported Cooperative Work (CSCW)* (14:4) 2005, pp 323-368.

Erdos, P., and Renyi, A. "On the Evolution of Random Graphs," *Bulletin of the International Statistical Institute* (38:4) 1960, pp 343-347.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. "On power-law relationships of the Internet topology," Annual Conference of the Special Interest Group on Data Communication (SIGCOMM '99), Cambridge, MA, 1999, pp. 251-262.

Feld, S.L. "Social Structural Determinants of Similarity among Associates," *American Sociological Review* (47:6) 1982, pp 797-801.

Freeman, C.L. "A Set of Measures of Centrality Based on Betweenness," *Sociometry* (40:1) 1977, pp 35-41.

Freeman, L.C. "Centrality in Social Networks: Conceptual Clarification," *Social Networks* (1) 1979, pp 215-240.

Garlaschelli, D., Caldarelli, G., and Pietronero, L. "Universal scaling relations in food webs," *Nature* (423:6936) 2003, pp 165-168.

Granovet.Ms "Strength of Weak Ties," *American Journal of Sociology* (78:6) 1973, pp 1360-1380.

Grewal, R., Lilien, G.L., and Mallapragada, G. "Location, location, location: How network embeddedness affects project success in open source systems," *Management Science* (52:7), Jul 2006, pp 1043-1056.

Hahn, J., Moon, J.Y., and Zhang, C. "Emergence of New Project Teams from Open Source Software Developer Networks: Impact of Prior Collaboration Ties," *INFORMATION SYSTEMS RESEARCH* (19:3), September 1, 2008 2008, pp 369-391.

Hai, Z.G. "The future interconnection environment," *Computer* (38:4), Apr 2005, pp 27-+.

Hajra, K.B., and Sen, P. "Aging in citation networks," *Physica A* (346) 2005, pp 44-48.

Henry, K., Bart, S., and Mehul, S. "Referral Web: combining social networks and collaborative filtering," *Commun. ACM* (40:3) 1997, pp 63-65.

Hu, D., Kaza, S., and Chen, H. "Identifying Significant Facilitators of Dark Network Evolution," *Journal of the American Society for Information Science and Technology* (60:4), March 2009, pp 655-665.

Hu, D., and Zhao, J.L. "A Comparison of Evaluation Networks and Collaboration Networks in Open Source Software Communities," Americas Conference on Information Systems, Toronto, ON, Canada, 2008a.

Hu, D., and Zhao, J.L. "Expert Recommendation Via Semantic Social Networks," International Conference on Information Systems, Paris, France, 2008b.

Huberman, B.A., and Adamic, L.A. "Growth dynamics of the World-Wide Web," *Nature* (401) 1999, p 131.

James, H., Nigel, S., Wendy, H., Tim, B.-L., and Daniel, W. "Web science: an interdisciplinary approach to understanding the web," *Commun. ACM* (51:7) 2008, pp 60-69.

Jeong, H., Tombor, B., Albert, R., Oltval, Z.N., and Barabási, A.-L. "The large-scale organization of metabolic networks," *Nature* (407:6804) 2000, pp 651-654.

Jin, X., Yongqin, G., Christley, S., and Madey, G. "A Topological Analysis of the Open Souce Software Development Community," Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005, pp. 198a-198a.

John, B., and Stefan, D. "The Future of Social Networks on the Internet: The Need for Semantics," *IEEE Internet Computing* (11:6) 2007, pp 86-90.

Jun, Z., and Mark, S.A. "Searching for expertise in social networks: a simulation of potential strategies," in: *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, ACM, Sanibel Island, Florida, USA, 2005.

Kaza, S., and Chen, H. "Effect of Inventor Status on Intra-organizational Innovation Evolution," Hawaii International Conference on System Sciences (HICSS - 42), Big Island, HI, 2009.

Kaza, S., Xu, J., Marshall, B., and Chen, H. "Topological Analysis of Criminal Activity Networks: Enhancing Transportation Security," *IEEE Transactions on Intelligent Transportation Systems*) under review.

Klerks, P. "The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands," *Connections* (24:3) 2001, pp 53-65.

Koch, S., and Schneider, G. "Effort, co-operation and co-ordination in an F/OSS software project: GNOME," *Information Systems Journal* (12:1) 2002, pp 27-42.

Koschade, S. "A Social Network Analysis of Jemaah Islamiyah: The Applications to Counterterrorism and Intelligence," *Studies in Conflict & Terrorism* (29:6), SEP 2006, pp 589-605.

Kossinets, G., and Watts, D.J. "Empirical Analysis of an Evolving Social Network," *Science* (311:5757), JAN 6 2006, pp 88-90.

Krebs, V.E. "Mapping Networks of Terrorist Cells," *Connections* (24:3) 2001, pp 43-52.

Kumar, S.R., Raghavan, P., Rajagopalan, S., and Tomkins, A. "Trawling the web for emerging cyber-communities," *Computer Networks* (31:11-16) 1999, pp 1481-1493.

Lakhani, K., Wolf, R., Feller, J., Fitzgerald, B., Hissam, S., and Lakhani, K.R. "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects," in: *Perspectives on Free and Open Source Software*, 2005.

Leenders, R.T.A.J. "Evolution of Friendship and Best Friendship Choices," *Journal of Mathematical Sociology* (21:1-2) 1996, pp 133-148.

Leonard, N.F. "Yenta: a multi-agent, referral-based matchmaking system," in: *Proceedings of the first international conference on Autonomous agents*, ACM, Marina del Rey, California, United States, 1997.

Leskovec, J., Kleinberg, J., and Faloutsos, C. "Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations," The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Chicago, Illinois, USA, 2005.

Lomi, A., and Pattison, P. "Manufacturing Relations: An Empirical Study of the Organization of Production Across Multiple Networks," *Organization Science* (17:3) 2006, pp 313-332.

Louch, H. "Personal Network Integration: Transitivity and Homophily in Strong-tie Relations," *Social Networks* (22:1) 2000, pp 45-64.

Madey, G., Freeh, V., and Tynan, R. "The Open Source Software Development Phenomenon: An Analysis Based on Social Network Theory," AMCIS, Dallas, TX, 2002, pp. 1806-1813.

Marsden, P.V. "Core discussion networks in America," *American Sociological Review* (52:1) 1987, pp 122-113.

McFadden, D. "Econometric Models for Probabilistic Choice Among Products," *The Journal of Business* (53:3) 1980, pp S13-S29.

McFadden, D., and Zarembka, P. "Conditional Logit Analysis of Qualitative Choice Behavior," in: *Frontiers in econometrics*, Academic Press, 1974, pp. 105-142.

Mcpherson, J.M., and Smithlovin, L. "Homophily in Voluntary Organizations - Status Distance and the Composition of Face-to-Face Groups," *American Sociological Review* (52:3), Jun 1987, pp 370-379.

McPherson, M., Smith-Lovin, L., and Cook, J.M. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology* (27) 2001, pp 415-444.

Mike, T. "Social networks, gender, and friending: An analysis of MySpace member profiles," *Journal of the American Society for Information Science and Technology* (59:8) 2008, pp 1321-1330.

Milward, H.B., and Raab, J. "Dark network: The structure, operation, and performance of international drug, terror, and arms trafficking networks," International Conference on the Empirical Study of Covernance, Management, and Performance, Barcelona, Spain, 2002.

Milward, H.B., and Raab, J. "Dark Networks as Organizational Problems: Elements of a Theory," *International Public Management Journal* (9:3) 2006.

Moody, J., McFarland, D., and Bender-deMoll, S. "Dynamic Network Visualization," *American Journal of Sociology* (110) 2005, pp 1206-1241.

Nerkar, A., and Paruchuri, S. "Evolution of R&D Capabilities: The Role of Knowledge Networks Within a Firm," *Management Science* (51:5) 2005, pp 771-785.

Newman, M., Barabási, A.-L., and Watts, D.J. *The Structure and Dynamics of Networks*, (illustrated ed.) Princeton University Press, 2006

Newman, M.E.J. "Scientific collaboration networks. I. Network Construction and fundamental results," *Physical Review E* (64) 2001a, p 06131.

Newman, M.E.J. "The structure of scientific collaboration networks," *Proceedings of the National Academy of Science of the United States of America* (98) 2001b, pp 404-409.

Newman, M.E.J. "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Science of the United States of America* (101) 2004, pp 5200-5205.

Newman, M.E.J., Barabasi, A.L., and Watts, D.J. *The Structure and Dynamics of Networks* Princeton University Press, Princeton, NJ, 2006.

Newman, M.E.J., Forrest, S., and Balthrop, J. "Email networks and the spread of computer viruses," *Physical Review E* (66:3) 2002, p 035101.

Okoli, C., and Oh, W. "Investigating recognition-based performance in an open content community: A social capital perspective," *Information & Management* (44:3), Apr 2007, pp 240-252.

Oppenheim, A.V., and Schafer, R.W. *Discrete-Time Signal Processing* Prentice-Hall, Englewood Cliffs, NJ, 1989.

Palla, G., Barabasi, A.-L., and Vicsek, T. "Quantifying social group evolution," *Nature* (446:7136) 2007, pp 664-667.

Powell, W.W., White, D.R., Koput, K.W., and Owen-Smith, J. "Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences," *American Journal of Sociology* (110:4), Jan 2005, pp 1132-1205.

Raab, J., and Milward, H.B. "Dark Networks as Problems," *Journal of Public Administration Research and Theory* (13:4), OCT 2003, pp 413-439.

Raghavan, P., and Raghavan, P. "Social networks: from the Web to the enterprise
Social networks: from the Web to the enterprise," *Internet Computing, IEEE* (6:1) 2002, pp 91-94.

Rapoport, A. "Spread of Information through a Population with Socio-structural Bias: II. Various Models with Partial Transitivity," *Bulletin of Mathematical Biology* (15:4) 1953, pp 535-546.

Reagans, R. "Preferences, Identity, and Competition: Predicting Tie Strength from Demographic Data," *Management Science* (51:9) 2005, pp 1374-1383.

Reiss, A.J. "Co-offender Influences on Criminal Careers," in: *Criminal Careers and Career Criminals*, National Academy Press, 1986.

Reiss, A.J., and Farrington, D.P. "Advancing Knowledge about Co-Offending - Results from a Prospective Longitudinal Survey of London Males," *Journal of Criminal Law & Criminology* (82:2) 1991, pp 360-395.

Roberts, J.A., Hann, I.H., and Slaughter, S.A. "Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects," *Management Science* (52:7), Jul 2006, pp 984-999.

Robertsa, J., Hann, I.-H., and Slaughter, S. "Communication Networks in an Open Source Software Project," in: *Open Source Systems*, 2006, pp. 297-306.

Sageman, M. *Understanding Terror Networks* University of Pennsylvania Press, Philadelphia, PA, 2004.

Sarnecki, J. *Delinquent Networks: Youth Co-offending in Stockholm* Cambridge university press, 2001.

Schachter, S. *The psychology of affiliation* Stanford University Press, 1959.

Snijders, T.A.B. "Stochastic Actor-oriented Models for Network Change," *Journal of Mathematical Sociology*:21) 1996, pp 149-172.

Snijders, T.A.B. "The Statistical Evaluation of Social Network Dynamics," in: *Sociological Methodology,* M.E. Sobel and M.P. Becker (eds.), Basil Blackwell, London, 2001.

Snijders, T.A.B. "Models for Longitudinal Network Data," in: *Models and Methods in Social Network Analysis*, Cambridge University Press, New York, 2004, pp. 215-246.

Snijders, T.A.B., Steglich, C., and Schweinberger, M. "Modeling the co-evolution of networks and behavior," in: *Longitudinal models in the behavioral and related sciences,* H.O.a.A.S. Kees van Montfort (ed.), Lawrence Erlbaum, Mahwah, NJ, 2007, pp. 41-71.

Solé, R.V., and Montoya, J.M. "Complexity and fragility in ecological networks," *Proceedings of the Royal Society B* (268) 2001, pp 2039-2045.

Sparrow, M.K. "The Application of Network Analysis to Criminal Intelligence - an Assessment of the Prospects," *Social Networks* (13:3), SEP 1991, pp 251-274.

Stewart, D. "Social Status in an Open-Source Community," *American Sociological Review* (70) 2005, pp 823-842.

Turner, J.C. *Rediscovering the social group : self-categorization theory* B. Blackwell, Oxford, UK; New York, NY, USA, 1987.

von Hippel, E., and von Krogh, G. "Open source software and the "private-collective" innovation model: Issues for organization science," *Organization Science* (14:2), Mar-Apr 2003, pp 209-223.

von Krogh, G., Spaeth, S., and Lakhani, K.R. "Community, joining, and specialization in open source software innovation: a case study," *Research Policy* (32:7), Jul 2003, pp 1217-1241.

Wagstrom, P., Herbsleb, J., and Carley, K. "A social network approach to free/open source software simulation," First International Conference on Open Source Systems, 2005, pp. 16-23.

Warr, M. "Organization and Instigation in Delinquent Groups," *Criminology* (34:1), FEB 1996, pp 11-37.

Wasserman, S., and Faust, K. *Social network analysis: methods and applications* Cambridge University Press, 1994.

Watts, D.J., and Strogatz, S.H. "Collective Dynamics of 'Small-World' Networks," *Nature* (393) 1998, pp 440-442.

Xu, J., and Chen, H. "CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery," *ACM Transactions on Information Systems* (23:2) 2005, pp 201-226.

Xu, J., Marshall, B., Kaza, S., and Chen, H. "Analyzing and Visualizing Criminal Network Dynamics: A Case Study," in: *IEEE Intelligence and Security Informatics (ISI)*, Springer LNCS, 2004, pp. 359-377.

Yang, C.C., and Li, K.W. "An Associate Constraint Network Approach to Extract Multi-lingual Information for Crime Analysis," *Decision Support Systems* (43:4) 2007, pp 1348-1361.

Yochai, B. *The Wealth of Networks: How Social Production Transforms Markets and Freedom* Yale University Press, 2006.

Yuan, Y.C., and Geri, G. "Homophily of Network Ties and Bonding and Bridging Social Capital in Computer-Mediated Distributed Teams," *Journal of Computer-Mediated Communication* (11:4) 2006, pp 1062-1084.

Yunwen, Y., and Kishida, K. "Toward an understanding of the motivation of open source software developers," Software Engineering, 2003. Proceedings. 25th International Conference on, 2003, pp. 419-429.

Yutaka, Y., Makoto, Y., Takeshi, S., and Toru, I. "Collaboration with Lean Media: how open-source software succeeds," in: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, Philadelphia, Pennsylvania, United States, 2000.