The Pennsylvania State University
The Graduate School

A MULTI-LEVEL ANALYSIS OF INFORMATION AND SUPPLY FLOWS IN SOCIAL AND BUSINESS NETWORKS

A Dissertation in
Information Sciences and Technology
by
Kang Zhao

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The thesis of Kang Zhao was reviewed and approved* by the following:

John Yen
University Professor of Information Sciences and Technology
Dissertation Advisor
Chair of Committee

Réka Albert
Professor of Physics and Biology,
Affiliated Faculty of Information Sciences and Technology

C. Lee Giles
David Reese Professor of Information Sciences and Technology

Akhil Kumar
Professor of Information Systems

Mary Beth Rosson
Professor of Information Sciences and Technology,
Graduate Program Director for Information Sciences and Technology

*Signatures are on file in the Graduate School.

Abstract

Networks are ubiquitous in both the physical world and the cyberspace. They enable the flow of information, materials, services, etc. The common thread running through my dissertation is the macro-level and micro-level analysis of information and supply flows in social and business networks. The overarching research question is "How do network flows relate to network structures and individual entities' behaviors in networks?"

My dissertation approaches this question in the context of supply-chain networks, online social networks, and inter-organizational networks. Specifically, I explore how changes in supply-chain network topologies affect supply flows, and propose strategies to improve supply-chain networks' robustness against disruptions. I analyze how the flow of information through individuals' interactions influences their sentiment in online health communities, and utilize the sentimental influence to identify influential users. I model how the flow of information via organizations' interactions impacts the structure of their collaboration network, and provide suggestions on how to promote collaboration. To support the research, the dissertation uses various computational and quantitative methodologies, such as network analysis and modeling, data and text mining, agent-based simulation, and statistical analysis.

The goal of my dissertation research is to achieve a deeper understanding of network flows, better prediction of network performance and individual behaviors, as well as new insights into ways to improve the design, management, and utilization of networks. Specifically, the outcome of this dissertation has implications for network disruption management, health care, online community building, humanitarian relief, and so on.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would not have finished this dissertation without the guidance from my committee members, the help from my collaborators, the support from my friends, and the love from my family.

I would like to express my deepest gratitude to my advisor, Dr. John Yen, for the guidance, mentorship, and providing me with a flexible atmosphere and excellent opportunities for doing research. I would also like to thank Dr. Albert and Dr. Giles for their insightful comments to my research, and their help during my job search. Special thanks goes to Dr. Kumar, who is also a close collaborator, for introducing to me the world of information system research in business schools and opening a new door for my academic career.

I am also grateful to my collaborators and co-authors, including Dr. Michael Batty, Dr. Cornelia Caragea, Dr. Matthew P. Evett, Dr. Annemijn van Gorp, Greta E. Greer, Dr. Terry P. Harrison, Dr. Carleen Maitland, Dr. Edgar Maldonado, Dr. Prasenjit Mitra, Dr. Louis-Marie Ngamassi, Razvan Orendovici, Dr. Kenneth Portier, Dr. Baojun Qiu, Harry Robinson, Dr. David J. Saab, Massimiliano Spaziani, Dr. Andrea H. Tapia, Tom Wagner, Dr. Dinghao Wu, and Dr. Yichun Xie. It has been an honor for me to work with them. My research would not have been possible without their help.

I would also like to thank my friends in both China and the U.S. for supporting me and being part of my life.

Finally, I express my love and appreciation to my beloved family–my father, wife, grandmother, parents-in-law, aunts, uncles, and cousins–for their understanding, patience, support, and endless love through the duration of my doctoral study. I also thank my lovely daughter. My life has been full of joy and excitement since her birth.

Dedication

This dissertation is dedicated in loving memory to my mother, who passed away before I began my doctoral study. Words cannot describe my love for you. I miss you and hope I have made you proud.

谨以此论文献给我亲爱的母亲。

Chapter 1

Introduction

Networks are pervasive in our life. Social networks, transportation networks, power grids, and supply-chain networks are just some of the important networks around us. Therefore, the emerging multi-disciplinary network science, which "develops theoretical and practical approaches and techniques to increase our understanding of natural and man-made networks" [1], has drawn the attention of many researchers from various disciplines, including mathematics, sociology, physics, biology, computer science, information science, operations research, and so on.

Research on networks started in the 18th century and revived in the late 1990s. Its inception can be traced back to 1736, when Euler proposed the famous Seven Bridges of Konigsberg problem [2]. Since then, networks have been studied in areas related to graph theory, discrete mathematics, and optimization. Network metrics (e.g., node centrality [3]) and algorithms (e.g., finding shortest path [4]) have also been developed to aid the analysis of networks. However, lacking data of large-scale networks, research was often limited to relatively small networks. It was also often believed that random network [5] is the dominant topology for many real-world networks.

The rising of Internet, advances in computing technologies, and the recent prevalence of online social networks and social media all contributed to the revitalization of research on networks. Internet has made it possible and easier to collect "big data" of large-scale networks, such as the World Wide Web, citation networks, email networks, and so on. Moreover, such "big data" also includes details on how individuals interact with each other through these networks, and

even the content of the interactions. As a result, researchers are now able to study networks at a scale and granularity that were not possible before.

While today's network research covers many different topics, studies in this area can be grouped into micro-level and macro-level analysis, although it is difficult to find a clear-cut boundary between research at the two levels. At the micro level, the focus is on individuals in networks. Research topics include centrality measures for nodes [6], reciprocity [7], link prediction [8], and so on. Up to the macro level, research emphasizes more on network structures and phenomena at the system level. For example, scientists have built models for the topologies and evolution of networks, such as the beta re-wiring model for small-world networks [9], and the preferential-attachment model for scale-free networks [10]. Also, algorithms have been developed to discover community structures in networks [11], optimize or reveal flows in networks [12][13], etc.

One of the important topics related to research at both levels is the flow through a network. Examples include the flow of goods through supply-chain networks, the flow of information through Internet, the flow of influence through social networks, and so on. At the macro level,to understand how the structure of a network is related to the network's performance, it is often necessary to examine how flows through the network are enabled, facilitated, or constrained by the network structure. At the micro level, to study the dynamics or evolution of individuals' behaviors or opinions, one may need to focus on the flow of information or influence among individuals in a network context.

This dissertation conducts a multi-level analysis of information and supplies flows in social and business networks. It not only studies flows at both macroscopic and microscopic levels, but also tries to connect network structures at the macro level and the individual behaviors at the micro level using network flows.

First, at the macro level, the dissertation studies how topological changes caused by disruptions affect the flows of supplies in supply-chain networks, which are responsible for distributing or disseminating supplies [14]. With a special focus on the heterogeneous roles that different types of nodes play (as providers, distributors, and consumers) and the characteristic of supply flows in supply-chain networks, this study first proposes new performance metrics at the topological level, such as supply availability, network connectivity, delivery efficiency, etc.

Then two customizable heuristic strategies are proposed to improve or balance the robustness of distribution networks against disruptions. The first strategy, *Degree-Locality-based Attachment*, is a hybrid network growth model to build or design supply-chain networks. The second strategy, *Randomized Local Re-wiring*, adjusts topologies of existing supply-chain networks that may be difficult or expensive to re-build. In simulations based on synthesized and real-world supply-chain networks, both strategies lead to supply-chain network topologies that can preserve more supply flows than traditional topologies after random (e.g., natural disasters and power outages) and targeted (e.g., from terrorists and cyber-attacks) disruptions, especially when it is not possible to predict which type of disruption will occur.

Second, at the micro level, my research identifies influential individuals in an online health community (OHC) by leveraging the flows of information among OHC members. While flows in a supply-chain network may be planned or managed, the flows of information among autonomous individuals in social networks are often driven by distributed inter-personal interactions. Such flows may influence people's behaviors, attitudes, or emotions. Thus to identify influential users in OHCs, this research leverages the large-scale data of individual users' interactions in a popular OHC among more than 27,000 cancer survivors.

Two complementary approaches are used, compared, and eventually integrated. The first approach builds classifiers to incorporate multiple metrics, which could reflect one's influence in different ways. The structure of the online social network among OHC users is also utilized to generate neighborhood-based and cluster-based features that help to improve the performance of classifiers. The second approach tries to measure one's influence directly by focusing on the flow of information and support among OHC users. A new and intuitive metric, the number of influential responding replies, is proposed to identify influential users. The influential user ranking by the new metric outperforms rankings by many traditional metrics based on users' contributions (e.g., numbers of posts and active days) and social network centralities (e.g., degree centrality and PageRank). The performance of the new metric also dominates the classification-based approach, which utilizes 60 metrics of users' contributions and centralities. Integrating the two approaches further improves the accuracy of the identification of influential users.

Last, the dissertation connects micro-level interactions among individuals and macro-level network topologies using the flow of information. Using a case study of inter-organizational networks [15] among humanitarian agencies, the research starts with a network influence model for how organizations influence others' decisions about collaborations. Combining such exogenous network influence with organizations' endogenous evaluation of incoming information, a novel model is proposed for how the information about candidate collaborative projects disseminates through organizations' interactions. The dissemination of project information allows organizations to learn about, evaluate, and keep in their to-do lists several projects supported by others. When several organizations identify a mutually beneficial project and agree to collaborate on it, an event-based approach adds multiple edges simultaneously among them in the collaboration network. Agent-based simulations are used to predict how changes to the communication network affect the flow of information, which in turn alters the topology of the resulting collaboration network. Simulations also suggest a strategy to promote collaboration by encouraging interactions among organizations at the peripheral of the communication network, which will better facilitate the flow of information and collaboration.

Major contributions of the dissertation are two-fold.

First, it contributes to the utilization and management of social and business networks in various domains. For example, the study on supply-chain networks offers strategies to improve or balance network robustness against disruptions. The work on online health communities helps to build an active, supportive, and sustainable community for patients and their caregivers. The research on inter-organizational networks provides suggestions on how to promote collaboration among humanitarian agencies, which will eventually benefit disaster victims.

Second and more importantly, research presented in this dissertation has general implications for the study of online and offline networks in social and business contexts. The research on supply-chain network robustness illustrates that incorporating the heterogeneous roles of nodes can provide a new perspective to the analysis of network flows. Considering such heterogeneity can also have a significant impact on the evaluation of a network's performance. The exploration on online health communities shows analyzing the content of individuals' interactions

and conducting sentiment analysis on large-scale data can help to understand individuals' behavioral dynamics and the impact of social influence in online settings. The investigation on inter-organizational networks demonstrates the potential to use network flows to connect micro-level and macro-level network phenomena. It also advocates the multi-relational perspective of network analysis by showing how dynamics in one network could affect another.

The remainder of the dissertation is organized as follows. In Chapter 2, I will introduce the topological study on supply flows and the robustness of supply-chain networks against disruptions. Chapter 3 will cover the individual-level analysis of information flows in OHCs and the identification of influential users. In Chapter 4, the dissertation will describe the research on how information flows connects individual interactions and network topologies in inter-organizational networks. Finally, Chapter 5 provides a summary of the dissertation, as well as implications and contributions of this dissertation.

Chapter 2

Supply Flows in Supply-Chain Networks–A Topological Analysis

From a topological perspective, this chapter studies the flow of supplies in supply-chain networks and the robustness of supply-chain networks against disruptions. The chapter will start with an introduction to the research problem, followed by a brief survey of related work. Then I will introduce new supply-chain network performance metrics that can better capture the characteristics of supply flows. Two heuristic strategies to build, design, or adjust supply-chain networks are proposed and evaluated against random and targeted disruptions. Implications of this research and future research directions are discussed in the end. The outcome of this research has been published in 1 conference paper [16] and 2 journal papers [17][18].

2.1 Introduction

Our daily lives rely heavily on the distribution of goods and services, such as groceries, water and electricity, through supply chains. With globalization and the development of information technology, supply chain systems are becoming more complex and dynamic. Today's supply chain systems often feature a network of interacting entities of different types, such as suppliers, manufacturers, retailers, customers, and so forth. Supply chains, which represented linear flows of goods from suppliers to customers, are evolving into supply-chain networks [19]. Because

entities may take different forms in various application domains, I refer to them simply as *nodes* in a network.

In this research, I consider a supply-chain network as a graph of nodes, where the nodes are generally of two types: *supply* (or *supplier*) and *demand* (or *requester*) nodes. In this sense, these networks have *heterogeneous* nodes as opposed to conventional networks where all nodes are of one type and hence homogeneous. Such heterogeneity has important implications for the flow of supplies in these networks. Most research on the robustness of complex networks has focused on homogeneous networks. However, it is important to note that notions of network connectedness are different in these two types of networks. Therefore, metrics of connectedness, such as *path length and the size of the largest component, may mean something different in a heterogeneous supply-chain network* than they do in a homogeneous network. Whereas connectedness in a homogeneous network is measured by how a given node is connected to other nodes, in a supply-chain network connectedness should focus on the flow of supplies, which is determined by how a demand node is connected to any supply node. For instance, if a demand node is connected to other demand nodes only but isolated from a supply node, then it should not be considered as connected because the flow of supplies cannot reach the demand node.

Moreover, large global supply chains or networks are often embedded in dynamic environments and may face disruptions, such as natural disasters, economic recessions, unexpected accidents or terrorist attacks. A disruption may initially affect or disable only one or a few entities in the system, but its impact may propagate, sometimes even with amplifications [20], among inter-connected entities. Such cascading disruptions will thus affect the normal operations of many other entities. Occasionally, failures in a small portion of the system may cause the catastrophic failure of the whole system [21]. Those events may seriously disrupt or delay the flow of people, goods, information, and funds, leading to higher costs or reduced sales [22] and affecting a company's long-term stock performance [23]. Therefore, designing supply chains that are robust against disruptions is a high priority concern, and it has drawn a lot of attention from managers, shareholders and researchers [24][25].

Traditional research on supply chain disruptions often adopts the risk man-

agement perspective and focuses on strategies and technologies to identify, assess, and mitigate risks and problems caused by disruptions [22][24][25]. However, even though research has revealed that the topology of a supply-chain network will affect its robustness [26], subsequent research in this direction is lacking.

In this research, I adopt the complex-network view of supply chains and study the flow in supply-chain networks from a topological perspective. In particular I am interested in understanding how the network topology affects the flow of supplies and various metrics of connectedness in a supply-chain network in the presence of both *random* and *targeted* disruptions. In addition to developing new metrics for supply-chain networks, another goal in this research is to find network designs that can preserve supply flows under both random and targeted disruptions.

A "full-spectrum" supply-chain network is a large "eco-system" with many actors, such as raw material providers, manufacturers, warehouses, distribution centers, and retailers. In this research, I will focus mainly on the distribution network, in which manufacturers, distributors and retailers have close interaction with each another. Thus the managers of these organizations often have good knowledge of, as well as control over, the network's structure. Therefore, compared with the network from other parts of a supply-chain network such as the procurement network, it is easier to implement another design for distribution networks, or change the topology of an existing distribution network.

2.2 Related work

The literature contains ways to evaluate and optimize the flow of networks when disruptions occur [27][28][29]. However, these network flow optimization problems are generally NP-hard [30]. As a result, when the network is complex and evolving, finding the optimal network configuration is computationally expensive. While techniques, such as simulated annealing [31], genetic algorithms [32], and heuristic procedure [33], have been used in the design of supply chains, this stream of work focuses mainly on location of facilities, capacity planning and minimizing transportation costs, instead of the robustness of supply-chain networks when some nodes fail to operate. Most of the related work on robustness also does not consider the twin goals of maintaining supply flows under both random failures

and targeted attacks.

Moreover, the optimization-based approaches are generally centralized in nature, because a supply chain manager is required to have complete knowledge and control of the whole network to optimize it. However, it is often difficult to meet this requirement in the real world. Researchers have found that the structure of a supply-chain network often emerges from the distributed decisions of individual nodes, regardless of the centralized nature of the design [34], because some heuristic strategies are often used when a node decides which nodes to connect to. Further, a manager only knows and controls the network of its corporation or organization, but the supply-chain network at a macroscopic level consists of networks from different corporations or organizations. Corporations' or organizations' optimization of their own networks may not necessarily lead to a globally robust network. Therefore, managers' awareness of how network design strategies at the microscopic level affect the robustness of the whole supply-chain network at the macroscopic level is also important.

Complex networks are defined as networks whose "*physical or logical structure is irregular, complex and dynamically evolving in time*" [35]. Research on complex networks has drawn growing interest during the past decades [36]. Researchers have found that real-world networks, such as online social networks, the World Wide Web, and biological cellular networks, often have non-trivial and complex topologies that are different from lattice or random graph structures (with a Poisson degree distribution). Scale-free [10] and small-world [9] are two well-known network structures that were proposed to represent the topologies of many real-world complex networks. Moreover, some supply-chain networks also show characteristics similar to those of a scale-free network [37][38].

In addition, the research on network growth models reveals distributed strategies and heuristics that underlie many real-world networks. Growth models study the evolution of complex networks by specifying how new nodes connect with existing ones through a process of "attachment". As new nodes enter a network, the network topology emerges from the distributed attachment decisions of individual nodes. Different growth models follow different attachment rules and lead to different network topologies. For instance, the *random-attachment model* randomly connects two pair of nodes with a pre-defined probability and generates an

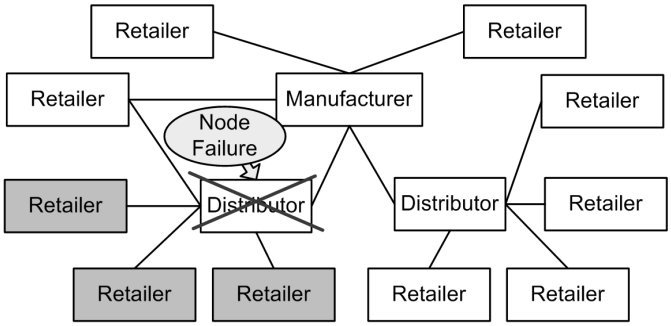


Figure 2.1: A hierarchical supply-chain network.

ER random network [5]. The growth models are often simple, yet they can produce complex network topologies [39]. For example, in the *preferential-attachment model* [10], the probability that a new node attaches to an existing node is proportional to the existing node's degree. This leads to the well-known scale-free topology that can explain many real-world networks, such as the Internet. It is also one of the key underlying principles that guides the evolution of supply-chain networks [40].

Specifically, the robustness of complex networks against failures and attacks has been analyzed. It has been found that scale-free networks, which are a kind of complex networks where the degree distribution follows a power law, have very high tolerance against random failures but are fragile to disruptions that target the most connected nodes [41][42]. Compared with scale-free networks, random networks [5] are less robust against random disruptions but often have better performance when important nodes fail. Reference [43] provided a review of approaches to assess network vulnerability and robustness.

The research of Thadakamalla et al. [26] introduced the *topological perspective* into the study of the robustness of supply-chain networks. The research argued that supply flows intraditional supply chains with hierarchical topologies are subject to disruptions. For example, in the hierarchical supply chain in Figure 2.1, *the failure of a single distributor disconnects about 25% of the retailers from supplies*. Using a military logistic network as an example, the study compared the performance of supply-chain networks with various complex network topologies in two types of node-removal disruption scenarios. The simulation results showed that the survivability of supply-chain networks is improved by concentrating on

the network topology. The authors proposed a new network growth model (henceforth referred to as the **Hierarchy+** model) that extends the hierarchical model by allowing edges between nodes of the same type. However, the new network design proposed in this research is an ad-hoc model for military logistic networks only. This research will further extend this work and try to find a general design to improve the robustness of supply-chain networks.

2.3 A strategy to design supply-chain networks

In this section, I first present a taxonomy of robustness metrics for supply-chain networks. With a focus on the flow of supplies from supply nodes to demand nodes, the new metrics reflect the heterogeneous roles (such as supply, relay and demand) of different types of nodes in supply-chain networks. I then introduce a new hybrid and tunable network growth model to generate a new supply-chain network, and evaluate multiple supply-chain network topologies' robustness against disruptions using the new metrics.

2.3.1 New robustness metrics

Robustness of a network is its ability to maintain operations and connectedness when some structures or functions are lost [44]. A robust supply-chain network, whose main function is to deliver supplies in response to demands, should be able to maintain the flow of supplies despite disruptions. Take the two illustrative networks in Figure 2.2 as examples. There are two supply nodes, four demand nodes, and six edges in both networks. A demand node can get supplies from either supply node. However, the two networks have different robustness against disruptions. For instance, if a supply node, say S2, fails because of disruptions, all demand nodes in network 2.2a can still access supplies from S1. Meanwhile, in network 2.2b, the failure of S2 will interrupt the flow of supplies to demand nodes D3 and D4.

To measure the robustness of a complex network, one must first develop appropriate metrics. In most previous research on complex network robustness, the evaluation of robustness focuses on the largest connected component (LCC), in

(a) (b)

Figure 2.2: Two sample supply-chain networks with the same number of supply (S) and demand (D) nodes and edges. Network in (a) can maintain flow of supplies better than the network in (b).

Table 2.1: Some standard metrics for network robustness.

Name of the metric	**Brief explanations of the metric (based on graph theory)**
Characteristic path length	The average of the shortest path length between any two nodes.
Size of the largest connected component of a network	The number of nodes in the largest connected component of a network.
Average path length in the largest connected component	The average of the shortest path length between any two nodes in the largest connected component of a network.
Maximum path length in the largest connected component	The maximum path length between any two nodes in the largest connected component of a network.

which there is a path between any pair of nodes [26][41]. Most of the existing network metrics are standard topological metrics from graph theory, including *characteristic path length*, *size of the largest connected component (LCC)*, *average path length in the LCC*, and *the maximum path length in the LCC*. Costa et al. provided a comprehensive survey of existing measurements for complex network [45]. Table 2.1 explains some of the commonly used metrics.

These metrics assume that entities in a supply-chain network perform homogeneous roles and supply flows can move from any one node to another. However, in real-world supply-chain networks, different types of entities play different roles in the system. Often times, the normal functioning of *downstream* entities may be highly dependent on the operations of *upstream* entities. As mentioned above, one of the fundamental purposes of a supply-chain network is to deliver supplies from

the supplier to the consumer. This type of "Supply-Demand" connection is the prerequisite for the flow of goods and services. As a result, preserving this type of connection in disruptions is critical for preserving supply flows and maintaining the operations of the whole supply-chain network.

Take the military logistic network in [26] as an example. The network consists of battalions and support units, including *forward support battalions* (FSB) and *main support battalions* (MSB). Support units play a different role from battalions in this network. Naturally, battalions cannot perform their military duties for long without receiving supplies from support units. Thus a large connected component, in which there is no support unit, or even where battalions are far from support units, should not be considered robust as there is none or limited supply flow in such a sub-network. Similarly, the distance between battalion and support units is generally more important for a robust supply-chain network than the distance among battalion units. Therefore, the heterogeneous roles (as supply and demand nodes) of different types of entities in a supply-chain network must be recognized when evaluating the flows in and the robustness of a supply-chain network. Although Albert et al. considered the roles of power plants and substations in a power-grid network when defining their connectivity metric [46], more systematic formalization and analysis are still needed to find metrics to reflect the robustness of supply-chain networks from more than one perspective. Also, in their research, only the robustness of a power grid was analyzed without offering any remedies.

The proposed taxonomy consists of system-level and topology-level metrics. Neely et al. proposed to use quality, flexibility, time, and cost as four major performance metrics for supply chains [47]. Although relying on topological measures, the new metrics can also reflect some aspects of those four performance metrics.

First, I introduce *supply availability* as a critical robustness metric for supply-chain networks because it shows whether entities in the network can get its requisite supplies to maintain normal operations. The inability to deliver goods to those who need them is a failure, which will affect the quality of the supply-chain network [48]. At the topological level, availability can also be interpreted as *supply availability rate*, which is the percentage of demand nodes that have access to supply nodes through the network.

Consider a supply-chain network as an un-directed, un-weighted graph $G(V, E)$

with node set V and edge set E, where $e_{i,j} \in E$ denotes an edge between nodes $v_i, v_j \in V$. As shown in Equation 2.1 below, V is the union of two non-overlapping subsets of demand and supply nodes (sets V_D, V_S), assuming a node cannot play both roles in the supply-chain network. Next, Equation 2.2 below defines the set of demand nodes that have access to supply nodes in the network, where $path_{i,j}$ denotes a path between nodes v_i and v_j. Thus $V_D^{'}$ is the set of demand nodes that have access to supply nodes through the supply-chain network. Consequently, the supply availability AV for a supply-chain network is the ratio between the cardinalities of sets $V_D^{'}$ and V_D (Equation 2.3).

$$V = V_D \cup V_S, \ where \ V_D \cap V_S = \phi \tag{2.1}$$

$$V_D^{'} = \{v_i \in V_D \mid \exists v_j \in V_S : \ \exists path_{i,j}\} \tag{2.2}$$

$$AV = \left|V_D^{'}\right| / |V_D| \tag{2.3}$$

The second metric is *network connectivity*. Clearly, the connectivity of the whole network is very important. On one hand, a well-connected network provides better access flexibility [49] because there are more options for routing, or re-routing, the delivery of goods to more customers if supply or demand patterns change or unexpected events occur. On the other hand, flows of supplies are often limited and less fluid in networks that are partitioned into small components. In complex network research, network connectivity is usually measured by the size of the LCC, in which there is a path between every pair of nodes. Here, the idea of availability is incorporated into this metric–the size of the largest *functional* sub-network is used instead of the size of the LCC.

For a supply-chain network $G(V, E)$, a functional sub-network $G'_m(V'_m, E'_m) \subseteq G_{sub}$ (G_{sub} is the set of all functional sub-networks) satisfies the two requirements in Equation 2.4. Thus the LFSN $G'_L(V'_L, E'_L)$ is one with the largest size in G_{sub} (Equation 2.5).

$$\forall v_i, v_j \in V_m^{'} : \exists path_{i,j} \ \text{ and } \ \exists v_k \in V_m^{'} : v_k \in V_s \tag{2.4}$$

$$G'_L(V'_L, E'_L) = \{G'_m(V'_m, E'_m) \in G_{sub} | \forall G'_n(V'_n, E'_n) \in G_{sub}(n \neq m) :$$

$$|G'_m(V'_m, E'_m)| \geq |G'_n(V'_n, E'_n)|\} \tag{2.5}$$

The difference between the old and the new connectivity metrics is that there must be at least one supply node in the LFSN. A sub-network of a military logistic network cannot function and maintain the flow of supplies without a support unit in it. When nodes fail during disruptions, a supply-chain network that features a larger functional sub-network can maintain a higher level of connectivity and is considered more robust.

I also describe two metrics for *accessibility* of supplies. Higher accessibility means that supplies are closer to consumers, and supplies can be delivered with lower cost or in lesser time. As mentioned earlier, from the perspective of supply-chain network robustness, the distance between a pair of demand nodes is not as important as that between a supply and a demand node. Consequently, I propose *supply path length* as the length of the path between a demand and a supply node (denoted by $dist(x, y)$). The length also reflects how far away the supply has to go in order to reach the destination. Thus, the *average supply-path length* in the LFSN is the average of the minimum supply-path lengths between all pairs of supply and demand nodes in the LFSN (see Equation 2.6). Moreover, the *maximum supply-path length* in the LFSN is the longest supply-path length between any supply and demand pair in the sub-network (see Equation 2.7). Naturally, shorter average and maximum supply-path lengths mean better average- and worst-case accessibility.

$$C_{avg} = \frac{\sum_{v_i \in V_{LS}} \sum_{v_j \in V_{LB}} dist(v_i, v_j)}{|V_{LS}| * |V_{LB}|}, \text{where } V_{LS} = V'_L \cap V_S, V_{LB} = V'_L \cap V_B \tag{2.6}$$

$$C_{max} = max(dist(v_i, v_j)), \text{where } v_i \in V_{LS}, v_j \in V_{LB} \tag{2.7}$$

However, there is a caveat when using the two accessibility metrics. In general, the comparison between the average and maximum supply-path lengths of different supply-chain networks (or sub-networks) are fair and meaningful only when the networks are of similar sizes. I have to take the sizes of the LFSNs into consideration because a larger supply-chain network with more nodes will often have longer average paths than a network with similar topology but fewer nodes. The existence of a few supply nodes that are far away from some demand nodes may

Table 2.2: Taxonomy of the new robustness metrics for supply-chain networks

Name	**Topology-level metric**	**Description**
Availability	Supply availability rate	The percentage of demand nodes that have access to supplies (Equation 2.3).
Connectivity	Size of the largest functional sub-network (LFSN)	The number of nodes in the LFSN, in which there is a path between any pair of nodes and there exists at least one supply node (Equation 2.5).
Accessibility	Average supply-path length in the LFSN	The average of the shortest supply-path length between all pairs of supply and demand nodes in the LFSN (Equation 2.6).
	Maximum supply-path length in the LFSN	The maximum path length between any pair of supply and demand nodes in the LFSN (Equation 2.7).

increase the average and maximum supply-path length in a large sub-network. I will illustrate this with experiments in the following sub-sections.

Overall, these metrics reflect the heterogeneous roles of different types of entities in supply-chain networks, capture the characteristics of supply flows, and improve our ability to measure supply-chain network robustness. Thus, I believe the new taxonomy is more systematic and realistic as compared to the metrics used in previous work [26]. Table 2.2 summarizes the new metrics.

One might also combine these metrics into a single objective function in order to optimize the overall performance of a network if the context in which a specific supply-chain network operates is known. For instance, a weighted linear combination of the four metrics may serve as an overall robustness metric. However, I decided to use multiple separate metrics instead of a single composite one, so as to gain a better understanding of a supply-chain network's performance from different perspectives.

2.3.2 The new hybrid design strategy

As I noted earlier, the topology of a complex network depends on its growth model. In the context of supply-chain networks, growth models represent distributed connection strategies that a node uses to decide which other nodes to connect to. For example, in the preferential-attachment model [10], new nodes prefer to connect to existing high-degree nodes. Thus, they can access other nodes efficiently and at lower cost. The random-attachment model is a strategy that randomly selects nodes to attach a new node to. The Hierarchy+ [26] model allows connections between nodes at the same level in the hierarchy.

While building a new supply-chain network from scratch is often a response to major disruptions [50], such as the Haiti earthquake and Hurricane Katrina, an organization may also take initiatives in rebuilding its supply-chain network, because today's corporations have to respond to very dynamic global markets. In order to be competitive, they may need to "burn themselves down every few years and rebuild their strategies, roles, and practices" [51]. This includes rebuilding the corporation's supply-chain network [52]. Such active re-building of supply-chain networks has been found in online retailers, chain retail stores, apparel manufactures, etc [53][54][55]. The overhaul often aims at saving cost, improving responsiveness, raising customer satisfaction level, and so on.

While Hierarchy+ uses ad-hoc attachment rules for different types of nodes, I propose a more general model called *Degree and Locality-based Attachment (**DLA**)* growth model, which is inspired by locality-based network growth models [56]. In the new model, a node considers not only the connectedness, but also the distance of a candidate node when establishing connections. This model does not require special attachment rules for different types of nodes and may be applied to other types of complex networks in general. The DLA growth model starts with a small number of disconnected nodes, say N_0. I assume that when a new node enters the system it initiates edges to connect to k existing nodes ($k < N_0$). The *attachment rules* for a new node are as follows:

The first edge connects to a node i of degree k_i with probability P_i where:

$$P_i = {k_i}^u / \Sigma_i({k_i}^u), \text{ where } u >= 0 \tag{2.8}$$

The remaining edge(s) will connect to a node j, which has a shortest distance of d_j to the new node, with probability P_j where:

$$P_j = {d_j}^{-r} / \Sigma_j({d_j}^{-r}), \text{ where } r >= 0 \tag{2.9}$$

Equation (2.8) describes the degree-based attachment preference of the first edge of a new node, and u is the customizable degree preference parameter. Given the same u, the new node will prefer connecting to existing higher degree nodes. When $u = 0$, the rule is similar to that of random networks, as every other node in the network has the same probability of being connected with the new node. When $u = 1$, the connection occurs by the preferential-attachment model of a scale-free network. A larger u gives even higher P_i to high-degree nodes. Therefore, as u becomes larger, it is more likely that the new node will connect to an existing node with higher connectedness. It is worth noting that, at the very beginning of the growth process, all the existing nodes are disconnected from each other, i.e., $\forall k_i : k_i = 0$. In this case, when the first new node enters the system, it will randomly choose an existing node to connect to.

On the other hand, Equation (2.9) gives preference to locality for the attachment of a new node's remaining edges if it is allowed to initiate more than one connection. As the node is already connected to the network through the first edge, I can then calculate the shortest distances from this node to all the other nodes. In (2.9), the non-negative integer distance d_j and the customizable locality preference parameter r constitute the attachment rule for the remaining edges of a new node. Given the same r, candidate nodes with smaller d_j will have a higher probability of being connected with new nodes. In other words, *the new node prefers nodes in its neighborhood over distant nodes*. When $r = 0$, every other node in the network has the same probability of being connected by the new node as in a random network. A larger r will reinforce the relative advantage of nearby nodes, while a smaller r will increase the new node's chance of connecting to more distant nodes. Additionally, in the special case of a sparse network, where no existing node is connected to the new node via any path, a node is chosen randomly. Lastly, and for obvious reasons, multiple links to the same node are disallowed. Also, because a new node does not need the exclusive access to other nodes when establishing

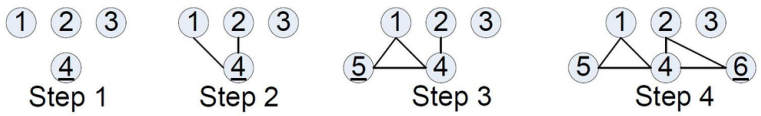


Figure 2.3: An example of the DLA growth model. ID for the new node is underscored.

connections, deadlocks will not happen in the DLA model.

Figure 2.3 illustrates a simple example for the DLA growth model with $u = 1$ and $r = 1$. In this example, each new node will establish two edges. Initially, the network starts with three disconnected nodes 1, 2, and 3. Node 4 is the first new node that enters the system and it randomly chooses two nodes to connect to, say, nodes 1 and 2. When node 5 comes in, its first edge will prefer existing high degree nodes, and thus node 4 has the highest probability to be chosen. The second edge of node 5 will prefer nodes that are close to it. Nodes 1 and 2 then have equal probability of being chosen. In this example, node 5 chooses node 4 for the first edge and node 1 for the second. Similarly, node 6 connects to nodes 4 and 2 in Step 4. As more nodes are added, a DLA network will emerge from this attachment process. As noted above, the "hybrid" DLA is more general than Hierarchy+. In fact, the Hierarchy+ model is a special case of the DLA growth model and can be realized with a suitable choice of parameters. In the next sub-section, I will evaluate the robustness of DLA networks, and compare it with other topologies using the new robustness metrics.

2.3.3 Simulation setup

As it is very difficult to construct a real-world supply-chain network and generate disruptions within it to evaluate its robustness, I have to rely on computer simulations. In this section, I describe the method for evaluating and comparing the robustness of different supply-chain network topologies using simulations. The results from the simulation of a military logistic network are illustrated and discussed.

The simulation study based on the method described in [57]. The main steps are: design a valid conceptual model, develop a program so simulate the model, design and run experiments, and perform output data analysis. I simulated the

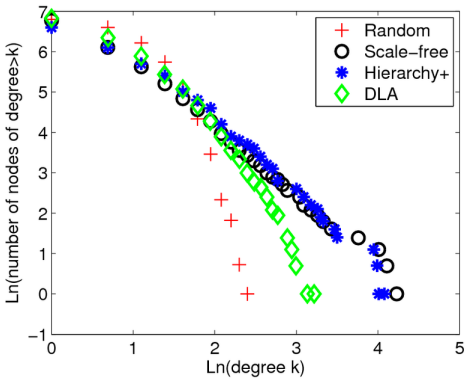
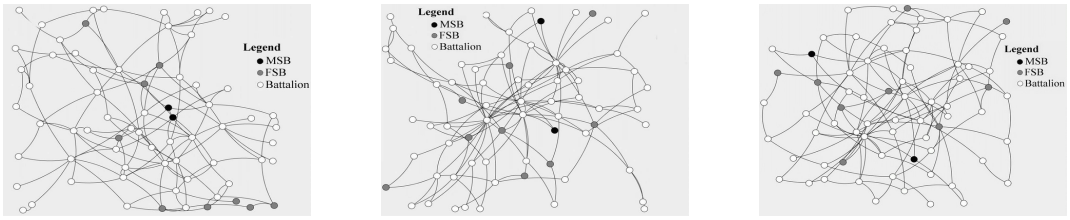


Figure 2.4: Cumulative degree distributions of supply-chain networks (each has 1000 nodes).

military logistic network example in [26], so as to directly compare Hierarchy+, which is designed for military logistic networks, with other topologies. The network is based on a real-world military logistic system. It consists of 1000 nodes and 1815 edges (the average degree is about 3.6). Battalions, FSB and MSB units enter the system following the ratio of 25:4:1, which was estimated from a real-world military logistic system. In other words, for every 25 newly-deployed battalions, the simulation adds 4 FSBs and then 1 MSB into the supply-chain network. In reality, other deployment schemes, such as deploying all support units before any battalions are deployed, are also possible and may lead to different supply-chain network topologies. However, I choose this scheme to ensure a consistent comparison with the previous work. I will then compare the robustness of supply-chain networks with *Random, Scale-free, Hierarchy+ and DLA topologies*. Each topology is generated with the corresponding growth model and the military logistic network configuration. For the DLA growth model, $u = 1/2$ and $r = 2$ is used.

Figure 2.4 illustrates the log-log degree distributions of four simulated 1000-nodes networks. A scale-free network is characterized by the famous power-law distribution. Fig. 2.5a, 2.5b, and 2.5c show the snap shots of simulated small-scale supply-chain networks with random, scale-free and DLA topologies (the legend is shown alongside). The DLA supply-chain network features some hub nodes and also some highly connected clusters, but connections to the hubs are not as concentrated as in the scale-free network.

The next step is to simulate disruptions. In the literature on network ro-



(a) A Random supply-chain network.

(b) A Scale-free supply-chain network.

(c) A DLA supply-chain network with $u = 1/2$, $r = 2$.

Figure 2.5: Simulated 70-node supply-chain networks with various topologies.

bustness, two types of disruption scenarios based on node removal are commonly studied: *random* and *targeted*. In random disruptions, each node has the same probability of failure. This scenario often corresponds to accidents (e.g., fires and power outrage), and unexpected economic events (e.g., recessions and bankruptcy). To simulate random disruptions, the simulation randomly removes nodes from the network. Edges that are connected to them are also removed.

On the other hand, in targeted disruptions important nodes are more likely to be disrupted than unimportant ones. Examples of targeted disruptions include terrorist and military attacks, which often target critical entities in the system such as network hubs. Among many metrics to measure a node's importance in a network, this study choses the widely-used *degree centrality* in line with previous research [26][41]. In other words, it is assumed that the higher the node degree is, the more important it is. The reason for picking degree as the indicator of importance is that node degree is easier for attackers to find. High-degree nodes are often more visible because they are in contact with many other nodes [58]. Other centrality measures, such as *closeness*, *betweenness*, and *eigenvector centrality* [59], require knowledge of the network topology, which is usually difficult for attackers to obtain. To simulate targeted disruptions, the simulation removes nodes in the order of decreasing node degree. In addition, because the removal of one node will affect the degree of some other, the simulation re-calculates the degrees of remaining nodes after each node removal. This dynamic update ensures the removal of the node with the highest degree in the remaining network.

While simulating disruptions with node removals has limitations (e.g., it assumes that disruptions will stop a node's operation, while in the real world an

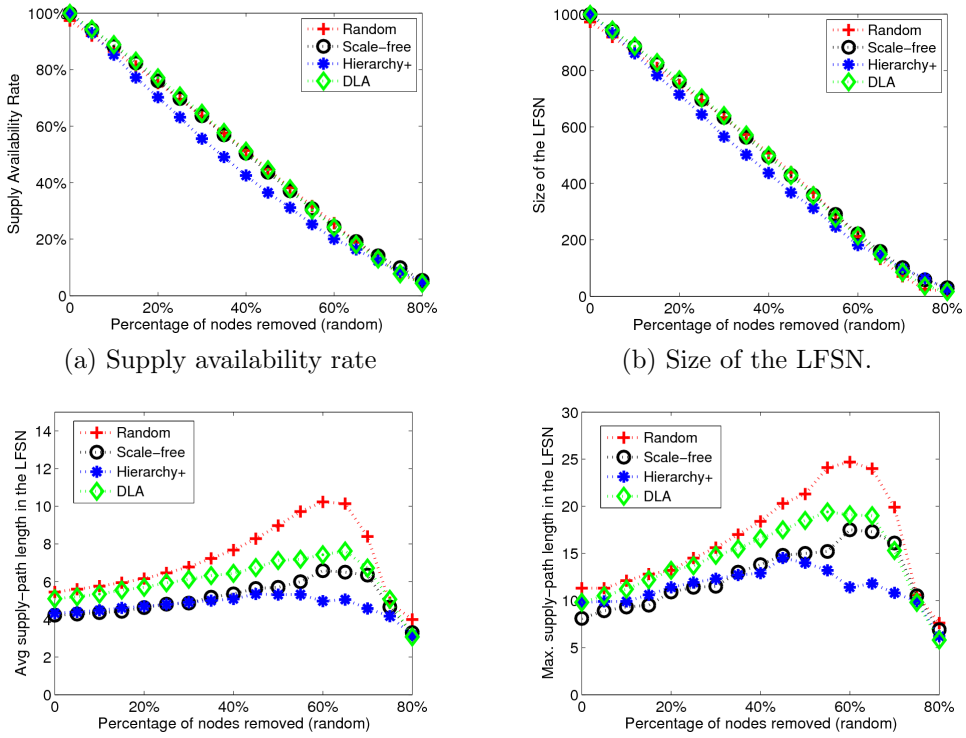
entity may only lose some of its capacities after disruptions), a great number of previous studies on complex network have shown that this simple and intuitive approach can help to reveal important properties on a network's robustness against disruptions [26][41].

In the simulation, 50 nodes (5% of the total nodes) are removed between successive observations to correspond to previous research [26] and to balance the simulation running time and the granularity of the results. During node removal, the robustness metrics are tracked for each network topology. In the end, the networks' robustness is compared using the metrics. To ensure a fair comparison, each network has the same number of nodes and edges. On average, each new node initiates 1.8 new edges to correspond with the military logistic network in [26]. Thus the total number of edges will be around 1800 and the average degree is 3.6 edges per node.

2.3.4 Simulation results for random disruptions

Fig. 2.6 shows the responses of the four network topologies to random disruptions. The horizontal axes denote the percentage of nodes that were removed, while the vertical axes are values of the topology-level supply-chain network robustness metrics from Table 2.2. The graphs are not extended beyond the 80% mark on the horizontal axis because they converge after this point. Fig. 2.6a and 2.6b show that for all the four network topologies, supply availability rate and the size of the LFSN decrease almost linearly as nodes are removed from the network. In terms of availability and connectivity, the performance of random, scale-free and DLA networks is very close, while the robustness of Hierarchy+ is slightly worse than of the other three, as indicated by its steeper slope. The 95% confidence intervals for the four networks' availability in Table 2.3 confirm the observation that Hierarchy+ falls a little behind in terms of maintaining supply availability. In addition, since the four networks are similar in the size of their LFSN, it allows us to make a fair comparison of accessibility next.

Nevertheless, the accessibility metrics in Fig. 2.6c and 2.6d point toward different conclusions. In general, when nodes are removed, the accessibility of supplies worsen because the average and maximum supply-path lengths in the LFSN in-



(a) Supply availability rate

(b) Size of the LFSN.

(c) Average supply-path length in the LFSN.

(d) Maximum supply-path length in the LFSN.

Figure 2.6: The four networks' responses to random disruptions. Each data point is the average of 20 runs.

Table 2.3: 95% confidence intervals for supply availaibility rates(%) in random disruptions

Nodes removed	Random	Scale-free	Hierarchy+	DLA
5%	91.7-92.3	94.0-94.3	92.5-93.3	94.1-94.5
10%	86.1-86.8	88.0-88.6	84.7-85.9	88.5-89.2
15%	80.8-81.5	82.2-82.6	76.5-77.9	82.7-83.3
20%	74.7-75.6	75.7-76.3	69.4-71.0	76.7-77.5
25%	69.1-70.0	69.4-70.2	62.3-64.0	70.3-71.1
30%	63.1-64.0	63.2-64.2	54.0-57.1	64.0-65.1
35%	57.2-57.9	56.1-57.8	47.9-50.2	57.1-58.4
40%	51.0-51.8	49.7-51.2	41.5-43.6	50.4-51.8

crease as more nodes are removed. This is intuitive, since disruptions make it increasingly difficult for demand nodes to receive supplies. At the 60% node removal point, almost all supply-path lengths reach their peak values and then start

to fall. The decreases are most likely caused by the fragmentation of the network and the isolation of "hard-to-reach" demand nodes.

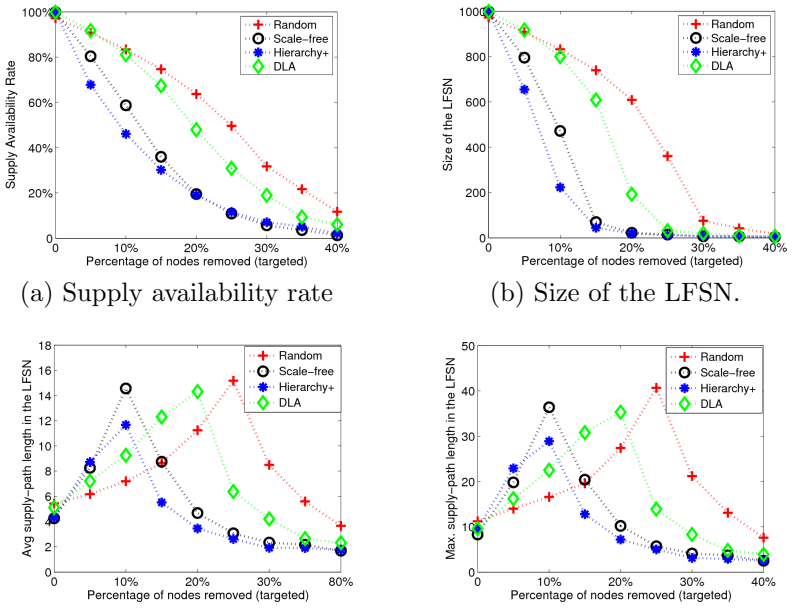
Before disruptions happen, the network is well-connected. A demand node may have multiple supply paths available, leading to different supply nodes. When some nodes fail, supply can still reach many demand nodes through alternative, albeit longer, supply paths. The existence of those "hard-to-reach" demand nodes that are still connected in the LFSN leads to the increase in supply path length. However, the longer a demand node's supply path is, the more this demand node depends on other nodes to receive supplies. As each node has the same probability to fail in random disruptions, the more dependent a demand node is, the more likely it is to get disconnected from supplies when additional nodes are removed. Thus a "hard-to-reach" demand node has higher probably to be isolated from the LFSN as more nodes are removed. After more than 60% of node removal, many "hard-to-reach" demand nodes are no longer in the LFSN. Instead, the remaining demand nodes in the LFSN are close to supply nodes. Meanwhile, the supply-chain network becomes very fragmented. The LFSN only has fewer than 200 nodes (20% of the original size). I believe the smaller LFSN and the disappearance of "hard-to-reach" demand nodes in the LFSN contribute to the decrease in supply path length when more than 60% nodes are removed.

In any case, the supply-path lengths of Hierarchy+ are the shortest, indicating that the Hierarchy+ network is able to preserve good supply accessibility. Even when 40-50% of nodes are removed, there is no dramatic increase in its supply-path length. On the other end of the spectrum, accessibility in the random network has the most serious degradation. The DLA is better than the Random network, but not as good as the Scale-free network, on accessibility.

Considering all robustness metrics, I believe the Hierarchy+ supply-chain network is generally the most robust against random disruptions. The robustness of other three network topologies against random disruptions can be ranked in a descending order as Scale-free, DLA and Random networks.

2.3.5 Simulation results for targeted disruptions

Arguably, robustness against targeted disruptions is more important than against random disruptions, because military logistic networks often face more targeted attacks than random attacks from opponents. Also, targeted attacks are generally more damaging than random attacks. Fig. 2.7 shows the responses of the four network topologies to targeted disruptions. Similar to Fig. 2.6, the horizontal axes denote the percentage of removed nodes, and the vertical axes display the robustness metrics. The graphs are not shown beyond the 45% mark on the horizontal axis because they converge after this point. As expected, robustness of all the supply-chain networks suffers different levels of deteriorations when compared with the case of random disruptions.



(a) Supply availability rate

(b) Size of the LFSN.

(c) Average supply-path length in the LFSN.

(d) Maximum supply-path length in the LFSN.

Figure 2.7: The four networks' responses to targeted disruptions. Each data point is the average of 20 runs.

Availability is examined first. Unlike the uniformly near-linear decreases found for random disruptions, the four supply-chain networks show significant differences on this metric in targeted disruptions. In Fig. 2.7a, all four networks have very low supply availability when 45% of the nodes are removed, while in random disruptions, availability of all networks is around 40% at this point. Among the four

Table 2.4: 95% confidence intervals for supply availaibility rate(%) in targted disruptions

Nodes removed	Random	Scale-free	Hierarchy+	DLA
5%	90.4-91.1	79.7-81.1	67.0-68.6	91.4-92.1
10%	83.0-83.8	57.6-59.9	45.1-47.1	80.7-81.5
15%	74.1-75.2	34.3-37.6	29.2-31.2	66.7-68.0
20%	63.0-64.4	17.9-21.1	18.2-20.1	46.9-49.0
25%	47.8-51.5	9.8-12.0	10.9-12.3	28.1-29.8
30%	30.0-33.5	5.2-6.3	6.7-7.5	18.0-19.9
35%	20.3-23.2	3.2-4.0	4.7-5.6	8.7-10.1
40%	10.8-12.7	1.1-1.5	1.7-2.5	5.5-6.6

networks, only the random network and the DLA network can still maintain near linear decreases. DLA's availability is close to that of the random network at the early stage of the disruption (from 0 to 15%), but drops faster than for the random network after 20% of nodes are removed. Meanwhile, the Scale-free and Hierarchy+ networks see significant decay in availability from the very beginning of targeted disruptions. For example, *10% node removal in the Hierarchy+ network leads to a near 60% drop in availability.* When 20% of the nodes are removed, only 20% of the battalions can still get supplies in the scale-free and the Hierarchy+ supply-chain networks. By comparison, *the Random and the DLA networks can still maintain their supply availability at 64% and 48% respectively.* 95% confidence intervals in Table 2.4 further illustrate that the four networks' availabilities are very different in targeted disruptions than in random disruptions.

Now let us look at connectivity. As demonstrated in Fig. 2.7b, the deterioration in robustness is even worse in terms of connectivity. Even the random network, the best performer on this metric, cannot maintain a linear decrease. The size of its LFSN is only 8% of its original size when 30% of the nodes are removed. *Very poor performance is observed from the Hierarchy+ network, whose LFSN drops to 22% of its original size with only 10% nodes removed.*

Clearly, the Random and DLA supply-chain networks show a considerable advantage over the Scale-free and Hierarchy+ supply-chain networks in availability and connectivity when facing targeted disruptions. What about accessibility? Similar to Fig. 2.6c and 2.6d, the plots of average and maximum supply-path lengths in Fig. 2.7c and 2.7d are also bell-shaped. Actually, Fig. 2.7c and 2.7d have very

similar graphs, except that they use different vertical axes scales. Intuitively, one would like to compare values of each network's supply-path lengths in the LFSN, but such a comparison is not representative if the LFSNs have various sizes. While this condition was satisfied in the results from random disruptions, however, as shown in Fig. 2.7b, the sizes of the LFSNs in the four networks differ significantly for the same disruption rate (especially when the disruption rate lies between 0% and 25%). For example, when 15% of the nodes are attacked, the supply-path lengths in the LFSN of the Hierarchy+ network are about 26% shorter than those of the DLA network. However, at the same point, *the size of Hierarchy+'s LFSN turns out to be only 7% of the size of DLA network's LFSN.* Therefore, one cannot conclude that Hierarchy+ has better accessibility since this likely advantage may be due to the much smaller size of its LFSN.

To better highlight the issue of accessibility comparison in the 0% to 25% disruption range, I draw the distributions of the shortest supply-path lengths in the four networks' LFSN at the 15% disruption point in Fig. 2.8. The horizontal axis denotes the shortest supply-path lengths from a battalion to its nearest supply unit, and the vertical axis represents the numbers of battalions that can access its nearest supply unit with a given shortest supply-path length in the LFSN. In Hierarchy+'s LFSN with 45 nodes, a nearest supply unit is always within a distance of 1 to 3 to a battalion. Meanwhile, in DLA's much larger sub-networks, more battalions can access the nearest supply unit within a distance of 1 or 2 than in Hierarchy+. However, a number of battalions are far away from a supply unit, with distance up to 7, thus contributing to DLA's higher average and maximum supply-path length. Yet, compared with Hierarchy+, DLA has far more battalions that can easily obtain supplies within the same distance as Hierarchy+ can.

Fig. 2.7c and 2.7d still contribute to the analysis, because they illustrate the rate at which the supply-path lengths increase in the LFSN, i.e., how fast the accessibility deteriorates when disruptions occur. Although supply-path lengths of scale-free and Hierarchy+ start with relatively lower values, they increase faster than of DLA and Random networks. The supply-path lengths of Scale-free and Hierarchy+ networks also reach their peaks very early. By the 10% point, the supply-path lengths have almost tripled. On the other hand, the peaks of DLA and Random come at 20% and 25% points respectively. In other words, *the supply*

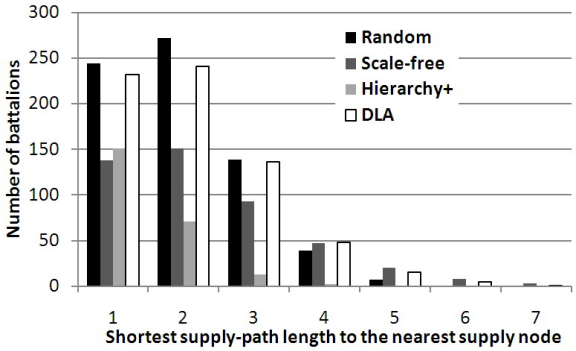


Figure 2.8: The distributions of shortest supply-path length in the four networks' largest functional sub-network (15% targeted disruption).

accessibility of the Scale-free and Hierarchy+ networks deteriorates faster than of the DLA and the Random networks. It is found that supply-path lengths reach peak values when another round of node removal makes the size of LFSNs drop below 100, which is about 10% of their original sizes. For example, in the Hierarchy+ network, the supply-path lengths reach peak values when 10% of nodes are removed and the LFSN has a size of 222. If an additional 50 nodes are removed, the LFSN retains only 45 nodes, while the supply-path lengths see significant drops. Here I argue that by the time the size of the LFSN falls below 10% of its original size, the network has already been decomposed into many isolated sub-networks, and the normal operations of the whole supply-chain network have also been seriously disrupted. Consequently, supply-path lengths after their peak values are not as meaningful to consider as those before the peaks. Therefore, I believe the random network has the best accessibility followed by the DLA network. The Scale-free and Hierarchy+ networks show similar accessibility.

Overall, the Random network is the most robust against targeted disruptions, with DLA a close second. Hierarchy+ is the least robust against targeted disruptions among the four, while the Scale-free network is only slightly better.

2.3.6 Discussion

Scale-free supply-chain networks are very vulnerable to targeted disruptions. This is largely because their operations and supply flows rely heavily on hub nodes, which have very high degrees and are removed at early stages of targeted dis-

ruptions. Recall that scale-free networks' preferential-attachment growth model corresponds to the strategy that nodes focus mainly on efficiency and cost. The results suggest if nodes make connection decisions solely based on efficiency and cost, the resulting network may be vulnerable to targeted disruptions.

Besides confirming previous research results, the study based on new robustness metrics also provides some new and surprising insights. For example, earlier work showed that Hierarchy+ supply-chain networks were reasonably robust against both random and targeted disruptions [26]. By the new robustness metrics, they still retain very good robustness against random disruptions. However, their robustness against targeted disruptions, which are more likely than random attacks in a military environment, is very disappointing. They have the worst availability and connectivity among the four supply-chain networks. Removal of only a small percentage of nodes will disastrously fragment the network and disrupt its flows and operations. Hence, they are unsuitable against targeted disruptions which may well arise in military logistic systems.

The reason for the Hierarchy+'s vulnerability against targeted disruptions lies in its growth model, which intentionally assigns more connections to support units. Therefore, support units will naturally become topological hubs in the supply-chain network. When target disruptions strike, support units will have higher probabilities of being attacked. The failures of support units, which act as both functional hubs and topological hubs, will inevitably hurt the availability and connectivity. The implication is that always assigning more connections to supply nodes may not be the best strategy to improve a supply-chain network's robustness against targeted disruptions.

The merit of DLA supply-chain networks is that they show good, although *not necessarily the best*, robustness against both types of disruptions. The robustness of the DLA network often lies in between that of the Random and the Scale-free networks. Specifically, in random disruptions, DLA is more robust than the Random supply-chain network, while in targeted disruptions it is more robust than the Scale-free network. Thus, it offers a balanced option when one cannot predict the probability of either type of disruption.

While previous research often improves robustness by introducing redundant capacities in manufacturing, transportation, and storage, etc, the approach taken

by DLA is slightly different because it does not require a node or edge to reserve redundant capacities. Instead, it changes the way that a supply-chain network is constructed and consequently the network topology. *Admittedly, DLA is not a clearly dominant method.* Similar to redundancy-based approaches, it also requires more investment to operate and maintain a supply-chain network. This is because, when no disruption occurs, DLA's accessibility is not as good as that of Scale-free and Hierarchy+, which implies that it may cost more to deliver supplies to demand nodes in a DLA network.

DLA's attachment rules also reflect the distributed nature of supply-chain network evolution. A new node entering a supply-chain network has only local information. Therefore, connecting with nearby nodes in its neighborhood is much easier and less expensive. In fact, the performance of DLA suggests that when nodes consider both efficiency and distance while making connection decisions, the resulting supply-chain network will have balanced robustness against both types of disruptions. Thus DLA represents a cost-effective strategy to build efficient yet robust supply-chain networks.

Moreover, by tuning the degree preference parameter u and the locality preference parameter r, it is possible to generate different supply-chain network topologies. Generally, a larger u leads to stronger preference for high degree nodes. Consequently, the resulting supply-chain network will deviate farther from randomness and rely more heavily on few "super hub" nodes that have very high-degrees. Larger r means more local connections. The resulting supply-chain network will feature more clusters and fewer connections that bridge nodes that previously have long distance between them. A simple numerical analysis is conducted to understand how the tuning of parameters u and r affects the resulting DLA supply-chain network's robustness.

As an example, this research analyzes the effect of changing u and r on supply availability rates when 10% of the nodes are removed in each disruption scenario. Node removal higher than this is not likely to occur in practice. For random disruptions, the supply availability rate falls between 86% and 89% and is only slightly affected by changes in u and r. On the other hand, for targeted disruptions availability is much more sensitive to u and r values. Supply availability rates of DLA supply-chain networks with 10% targeted node removal are summarized in

Table 2.5: Numerical analysis for supply availability rate (10% targeted node removal)

	r = 0	**r = 0.5**	**r = 1**	**r = 2**	**r = 3**	**r = 4**
u = 0	86.12%	85.87%	85.68%	85.13%	84.50%	82.88%
u = 0.5	83.55%	83.21%	82.96%	82.24%	81.24%	79.32%
u = 1	78.65%	78.05%	77.37%	76.29%	74.72%	72.36%
u = 1.5	64.43%	63.36%	61.97%	59.63%	57.25%	52.93%
u = 2	31.69%	31.65%	31.80%	31.50%	30.80%	30.58%
u = 3	28.11%	29.11%	28.74%	28.61%	28.62%	28.59%
u = 4	28.57%	28.50%	28.48%	28.84%	28.74%	28.93%

Table 2.5. Each value in the table is the average of results from 100 runs.

From Table 2.5 one can see that as u increases, the availability decays. Meanwhile, given the same degree preference u, higher preference for locality-based attachment, i.e. increasing r, will generally lead to slightly lower availability. Some may notice that when u is high, r has little impact on availability. This may be explained by the very strong preference for high-degree nodes. As a result, there emerge one or few "super hub" nodes that are connected to almost all the other nodes. When a new node enters the network, it will most likely establish the first connection with a "super hub" node. Then most of the other nodes have the same distance of 2 to this new node, so the preference for local nodes becomes less important. The results also agree with my previous finding that random supply-chain networks (i.e. DLA with $u = 0$ and $r = 0$) have better availability than scale-free and DLA networks with $u = 1/2$ and $r = 2$ in targeted disruptions.

It should also be noted from my analysis that the impact of locality on robustness is weaker than that of degree. This can be attributed to the fact that on average each new node initiates only 1.8 new edges in the experiments in order to ensure a fair comparison with the previous work [26]. In the DLA model, the first edge attachment is based on degree, while subsequent edge attachments are based on locality. This means that attachments based on locality are about 20% fewer than those based on degree. It is believed that given more attachments based on locality, the impact of locality on availability might even be more pronounced.

The numerical analysis shows that one can tune the DLA model to generate supply-chain network topologies with different levels of robustness against different

types of disruptions. This customizability has important implications for the design and management of supply chains. As there is no single supply-chain network topology that dominates all others in both random and targeted disruptions, one needs to seek a balance or trade-off between the robustness against random disruptions and targeted disruptions. The DLA network growth model provides the opportunity for nodes to make connection decisions based on what type of supply-chain network they are in and the balance of possible disruptions the supply-chain network will face. For instance, a DLA model with lower u is more appropriate for military logistic networks, which often handle targeted attacks from the enemy. A DLA model with higher u may be better for an automaker or a retailer as targeted disruptions are less likely for this type of supply chain.

2.4 A re-wiring strategy for supply-chain networks

The DLA approach proposed in the previous section is a strategy to build supply-chain networks. As I mentioned before, although building a supply-chain network does happen in the real world, it is not very frequent. What a supply-chain network may experience more often is to adjust the current topology to improve its robustness. Therefore, this section proposes a re-wiring strategy, which adjust connections among nodes in a supply-chain network to get nice robustness properties. I will evaluate the effectiveness of this re-wring strategy by applying it to military logistic networks with scale-free and small-world topologies, as well as a retailer's distribution network.

2.4.1 Updated robustness metrics

In this section, I keep using robustness metrics of *availability* and *connectivity* defined in the previous section (see Equations 2.3 and 2.5), as well as the symbols and notions. However, two new new metrics are developed to measure *delivery efficiency*.

While availability describes whether supplies can be delivered to a demand node from a supply node, it does not measure how "efficient" the delivery or the flow

is, in terms of lead time and transportation cost. One way to measure delivery efficiency is by the distance between supply and demand nodes. Intuitively, a shorter distance often means that goods can be delivered faster and cheaper, and thus the network is more efficient. For networks with homogeneous nodes, reference [60] proposed the network efficiency metric that is based on characteristic path length. However, in a heterogeneous supply-chain network, the distance between demand nodes is not as important as *the distance between supply nodes and demand nodes*.

Similar to the accessibility metrics used in the previous section, the two new delivery efficiency metrics are also based on *supply path lengths*, rather than **all** path lengths as in [60]. In a supply-chain network, *a supply path for a demand node refers to the path between this demand node and a supply node*. Then the average (across all demand nodes) of a demand node's shortest supply-path length to its nearest supply node, defined as AVG_DIST in Equation 2.10, is a measure of how fast supplies can be delivered to demand nodes across the supply-chain network. The reciprocal of the average is used as a metric for efficiency (defined as $BEST\_DEF$ in Equation 2.11), so that a higher value corresponds to supplies being nearer and the network being more efficient.

$$AVG\_DIST = \frac{1}{|V_D^{'}|}\sum_{i=1}^{|V_D^{'}|} min\{distance(v_i, v_j), \forall v_j \in V_S\}, \text{ where } v_i \in V_D^{'} \quad (2.10)$$

$$BEST\_DEF = \frac{1}{AVG\_DIST} = \frac{|V_D^{'}|}{\sum_{i=1}^{|V_D^{'}|} min\{dist(v_i, v_j), \forall v_j \in V_S\}}, \text{ where } v_i \in V_D^{'} \quad (2.11)$$

Efficiency $BEST\_DEF$ answers the question "how far away are the nearest supply nodes from demand nodes in the network?" This is essentially the best case scenario for a demand node to access supplies. What about the average case, since all demand will clearly not be filled by the nearest source of supply? Perhaps, one might be tempted to consider the average of shortest supply-path lengths between all pairs of connected demand nodes and supply nodes. However,

such an average fails to capture the number of supply nodes that are accessible from a demand node or the flexibility of supply flows. In fact, with this average, demand nodes that can access more supply nodes often look worse than those that can access fewer supply nodes, as illustrated in the next example.

(a) (b)

Figure 2.9: Two simple supply-chain networks with supply nodes (denoted with S) and demand nodes (denoted with D). Each edge represents a distance of 1.

Figure 2.9 shows two simple supply-chain networks. In network in Figure 2.9a, demand node D1 has access to two supply nodes S1 and S2, with supply path length of 1 and 2 respectively. D2 is in the same situation. In the network in Figure 2.9b, both demand nodes D3 and D4 can access only one supply node S3 with supply path length of 1. There are four demand-supply pairs in the network in Figure 2.9, namely D1-S1, D1-S2, D2-S1 and D2-S2. The average shortest-supply-path length is $(1+2+2+1)/4 = 1.5$. In the network in Figure 2.9b, there are two such pairs, D3-S3 and D4-S3, with an average shortest-supply-path length of 1. Although the network in Figure 2.9a has a longer average supply path length, its supplies are often considered more "accessible," because if a demand node is able to access more supply nodes it clearly suggests that average delivery efficiency is greater. Of course, there is a trade-off between having access to two supply nodes each at a distance of 10 (or 100) versus one supply node at a distance of 1.

Thus, this section introduces a new metric called *average delivery efficiency* that combines both the number of supply nodes that can be accessed by a demand node and also the distance at which each supply node is located. It is based on inverse supply path length calculation. For demand node D_i, its average delivery efficiency $AVG\_DEF_{D_i}$ is defined in Equation 2.12 below, where k is the total number of supply nodes that D_i can access. The supply nodes are sorted based on their shortest distance to the demand node, ties being broken arbitrarily. Thus $dist_{i,j}$ represents the shortest path length from D_i to its j-th nearest supply node.

For example, $dist_{i,2}$ is the shortest supply path length from demand node D_i to its second nearest supply node. The exponent $f(j)$ is a weighting factor to specify the relative importance of shortest supply paths j among the k shortest supply paths to k supply nodes.

$$AVG\_DEF_{D_i} = \sum_{j=1}^{k} [(1/dist_{i,j})^{1/f(j)}] \tag{2.12}$$

While one could define a weighting factor $f(j)$ in many ways, it is reasonable to argue that $f(j)$ should be *a non-increasing function of* j, because the distance to a nearer supply node is often considered more important than that to a farther supply node. As $1/dist_{i,j} \leq 1, \forall i, j$, the lower the exponent $f(j)$ is, the lower is $(1/dist_{i,j})^{1/f(j)}$. For example if $f(1) = f(2) = 1$, then equal importance is assigned to both the first and second shortest supply paths. However, if $f(2)$ is reduced to 0.5 and $f(1)$ is kept as 1, then it suggests that the weighting factor for the second path is lower, reflecting the lesser importance of the second shortest path as compared to the shortest one. In general, different $f(j)$ may be customized for different supply-chain networks in various domains. For example, the function $f(j)$ may vary from an exponential to a linear or sub-linear decreasing function. (Please see Appendix A for desirable properties of a metric like AVG_DEF.)

Consequently, the *average delivery efficiency* for the whole supply-chain network, defined as AVG_DEF in Equation 2.13, is the average of $AVG\_DEF_{D_i}$ over all demand nodes in the network. Naturally, a higher AVG_DEF value corresponds to greater efficiency.

$$AVG\_DEF = \frac{1}{|V_D|} \sum_{i=1}^{|V_D|} AVG\_DEF_{D_i} \tag{2.13}$$

Here I show how AVG_DEF is calculated for the network in Figure 2.9a. I use $f(j) = 1/j$, thus $f(1) = 1$ and $f(2) = 0.5$. Then $AVG\_DEF = (1 + 0.5^2 + 1 + 0.5^2)/2 = 1.25$.

Table 2.6 summarizes the updated new metrics. Those marked with * are the same as in Table 2.2

Table 2.6: Taxonomy of the updated robustness metrics for supply-chain networks.

System-level metric	**Topology-level metric**	**Brief explanations of topology-level metrics**
Supply Availability*	Supply availability rate	The percentage of demand nodes that have access to supplies from at least one supply node.
Network Connectivity*	Size of the largest functional sub-network	The number of nodes in the largest functional sub-network, in which there is a path between any pair of nodes and there exists at least one supply node.
Best Delivery Efficiency	Inverse of average minimum supply path length across all demand nodes	The reciprocal of the average of each demand node's shortest supply path length to its nearest supply node.
Average Delivery Efficiency	Adjusted average inverse supply path length across all (supply, demand) path pairs	The average inverse supply path length for all possible demand-supply node pairs, adjusted by a weighting factor for each path.

2.4.2 The rewiring model

The new model, based on the rewiring of a supply-chain network, is called *Random Local Rewiring (RLR)*. To rewire a supply-chain network using RLR, all edges are iterated, and consider the pairs of nodes at both ends of an edge. With a predetermined rewiring probability p_r, an edge will disconnect from one of its two endpoints-the one with higher degree. Then the other endpoint of the rewired edge will rewire the edge to connect with a randomly chosen node within a radius of d_{max}. Unlike the Thadakamalla model [26], this model does not require special attachment rules for different types of military units and may be applied to existing supply-chain networks with various topologies.

The rewiring probability p_r essentially determines how much rewiring will occur. On one hand, if $p_r = 0$, no rewiring will take place and the network remains unchanged. On the other hand, if $p_r = 1$, all edges will go through this random rewiring process. The other parameter, maximum rewiring radius d_{max}, imposes a practical constraint for rewiring. In the context of a supply-chain network, a

node may not be able to connect to any randomly selected node at will because the establishment of a connection between two nodes is often associated with cost. Connecting two nodes that closer to each other is generally more economical. Thus the upper limit on the rewiring radius reflects this preference on locality. When implementing the algorithm, the radius can be either physical distance (say, in miles), or topological distance (in number of hops/edges).

In general, the rewiring process aims at adding randomness in a controlled way to a supply-chain network. A higher p_r value will lead to more rewiring, and thus introduce more randomness, while a lower d_{max} will impose more control over rewiring. Figure 2.10 shows the pseudo-code for rewiring a network with RLR. It uses a function called "*Rewire*" that disconnects an edge from an existing node, and reconnects the remaining node (v_{toKeep}) of the edge to a new node (v_{new}) chosen at random, provided v_{new} is not a direct neighbor of v_{toKeep}.

In the remainder of this section, I will generate different supply-chain networks with various topologies and configurations and to apply the RLR approach to various network topologies to see whether rewiring can help to improve the robustness of supply-chain networks. Similar to the previous section, two types of disruptions are simulated: random and targeted. Targeted disruptions are also based on node degrees. However, note that this study only considers disruptions at the supply nodes, or supply disruptions. Disruptions at demand nodes are not considered in this section.

2.4.3 Rewiring scale-free networks with RLR

This sub-section analyzes how a RLR-rewiring affects the robustness of supply-chain networks with the well-known scale-free topology. Using an experiment for a military logistic network, this subsection describes the method for evaluating and comparing the robustness of different military logistic network topologies using simulations. The results will be illustrated and discussed.

Given a supply-chain network $G(V, E)$;
Given a rewiring probability p_r and a re-wiring radius d_{max}.
foreach (edge $e_i \in E$) {
    $rnd = Random(0, 1)$; //generate a random number rnd
    If ($rnd < p_r$) {
        v_i^1 = endpoint 1 of edge e_i ($v_i^1 \in V$);
        v_i^2 = endpoint 2 of edge e_i ($v_i^2 \in V$);
        Case Switch{ //rewire an edge
            degree(v_i^1) > degree(v_i^2):
                Rewire(e_i, v_i^1, v_i^2);
            degree(v_i^1) < degree(v_i^2):
                Rewire(e_i, v_i^2, v_i^1);
            degree(v_i^1) == degree(v_i^2):
                Rewire(e_i, v_i^1, v_i^2) or Rewire(e_i, v_i^2, v_i^1);
        }
    }
}
Function Rewire (e_{rewire}, $v_{toDisconnect}$, v_{toKeep}){
    $E = E - \{e_{rewire}\}$
    Identify $V_r \subset V$, such that $\forall v_j \in V_r$, 0<distance(v_{toKeep},v_j)$\leq d_{max}$;
    v_{new} = Random($v_j \in V_r$);
    If ($v_{new} \neq v_{toKeep}$ and $v_{new} \notin$ direct neighbors of v_{toKeep}){
        e_{rewire} = Connect(v_{toKeep},v_{new});
        $E = E + \{e_{rewire}\}$;
    }
}

Figure 2.10: Pseudo code for the algorithm to generate the RLR network topology.

2.4.3.1 Simulation setup

As in the previous section, the simulation scenario is also based on the military logistic network example in [26]. In this network, demand nodes are battalions. MSBs and FSBs are considered as supply nodes for simplicity. The supply-chain network consists of 1000 nodes, including 166 supply and 834 demand nodes. The supply-demand ratio was estimated from a real-world military logistic system. RLR-rewiring is performed with different probabilities and the robustness of the so called *RLR scale-free networks* is analyzed. To better illustrate the robustness of different supply-chain networks, an ER-random network is also included

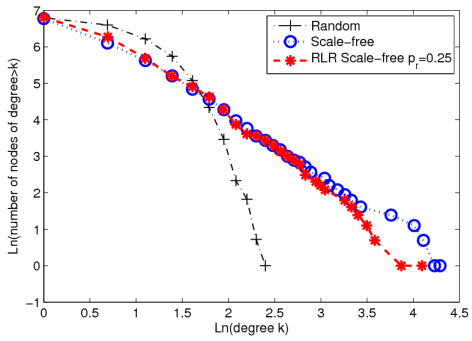


Figure 2.11: The distribution of the three simulated supply-chain networks with 1000 nodes each.

in the comparison. For each supply-chain network topology, the network is first constructed using the military logistic network configuration. Then disruptions to those networks are simulated. Their responses on robustness metrics proposed above are observed.

The Hierarchy+ model [26] is not included in the comparison because it often makes battalions the network hubs, i.e. nodes with very high degrees, in the resulting supply-chain network. This is not desirable because although a battalion can occasionally forward supplies to other battalions, it is not the primary task of a battalion and a battalion should not forward supplies more than MSBs and FSBs do. Thus such a configuration is expected to be rarely used in real-world military logistic systems.

When RLR-rewiring the military logistic network, it is assumed that all the military units are deployed in a battlefield and close to each other. Thus, when generating a RLR scale-free network, the distance between units is not a significant factor for selecting nodes as rewiring destinations. In other words, the maximum rewiring radius in the RLR approach is set to infinity ($d_{max} = \infty$), so that a rewired edge can randomly select a new destination node among all other nodes.

Figure 2.11 illustrates the ln-ln degree distributions of 3 simulated networks. For illustrative purposes, I also show the snapshots of a 70-node, 120-edge military logistic networks with RLR scale-free (with $p_r = 0.25$) topologies in Figure 2.12.

The simulation removes 8 nodes, about 5% of all the supply nodes, between successive observations. Because most attacks in a real world network will disrupt only a relatively small number of nodes, this research simulates disruption

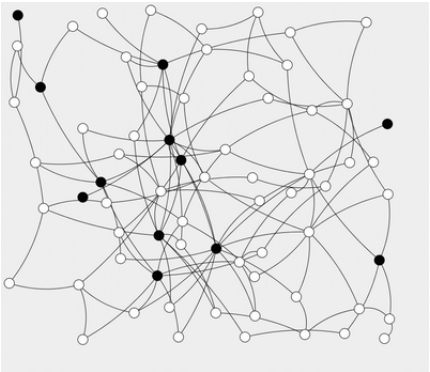
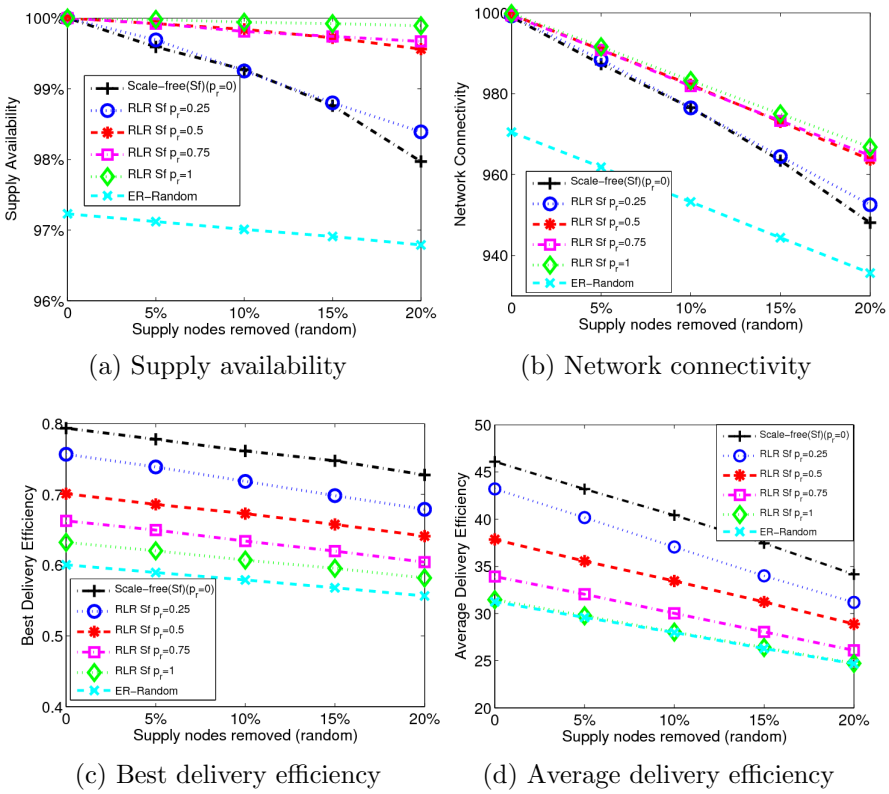


Figure 2.12: Snapshots of a simulated 70-node 120-edge supply-chain networks with the RLR scale-free topology (Supply nodes are denoted in black and demand nodes are in white.)

scenarios where the percentage of supply nodes removed lies in the 0% to 20% range. During the process of node removal, I track the robustness metrics for each network topology. When evaluating the average delivery efficiency of supply-chain networks, $f(j) = 1$ is used for Equation 2.12, i.e., assigning equal importance to all supply paths. This experiment is repeated for various topologies in order to facilitate a comparison among them. To ensure a fair comparison, each network topology will have the same number of nodes and edges. The average degree is kept at 3.6 edges per node in my simulations so as to correspond with the military supply-chain network in [26].

2.4.3.2 Simulation results for random supply disruptions

Figure 2.13 shows the responses to random disruptions of six network topologies, including scale-free, ER-random and four RLR scale-free networks with various p_r. The horizontal axes denote the percentage of supply nodes removed, while the vertical axes are values of the topology-level robustness metrics in Table 2.6. As one would expect intuitively, the performance of all networks decreases when nodes are removed from the network. The ER-random has the least value for all four metrics, with small slopes and lower initial values. The scale-free supply-chain network excels in delivery efficiency (both, best and average). Higher rewiring probabilities for RLR scale-free networks lead to better availability and connectivity, but at the cost of delivery efficiency.



(a) Supply availability (b) Network connectivity

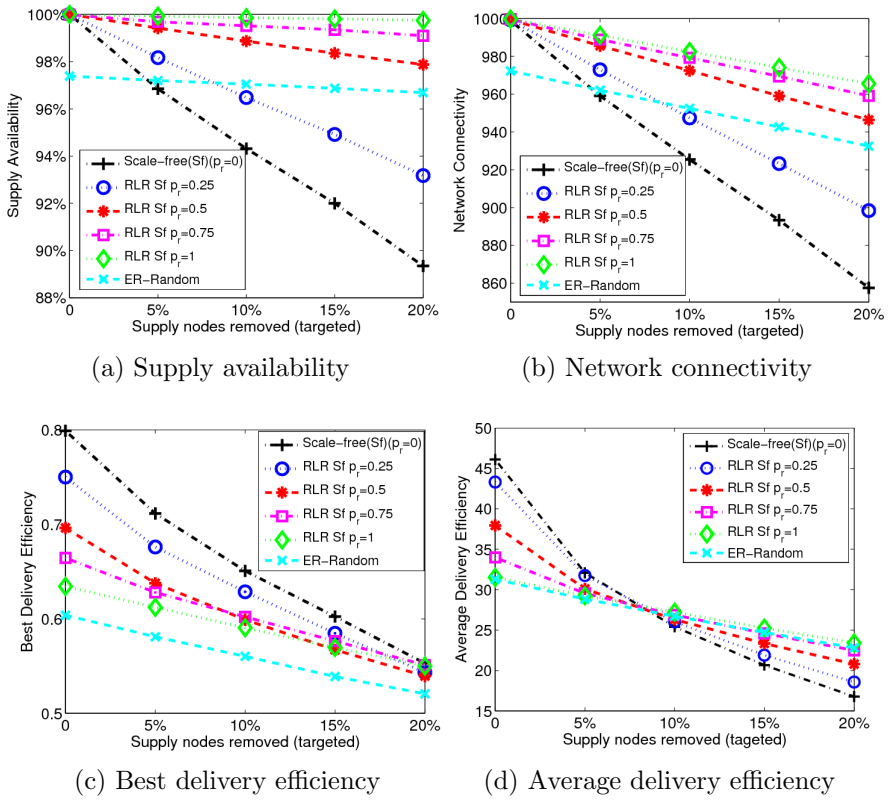
(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.13: Various military logistic networks' responses to random supply disruptions. Average of 20 runs.

Surprisingly, very good performance is found from the totally rewired scale-free network, i.e. RLR scale-free with $p_r = 1$, which incorporates some level of random attachment in all its edges. Figure 2.13a and 2.13b reveal that random disruptions to 20% of its supply nodes have negligible impact on its supply availability, which hardly decreases even 1%. Network connectivity is also well preserved as almost all the remaining nodes are still connected in one functional sub-network.

2.4.3.3 Simulation results for targeted supply disruptions

Figure 2.14 shows the responses of the six network topologies to targeted disruptions. Similar to Figure 2.13, the horizontal axes denote the percentage of supply node removal, while the vertical axes are the topology-level robustness metrics. As expected, robustness of all the four supply-chain networks suffers different levels of deterioration when compared with the case of random disruptions.



(a) Supply availability (b) Network connectivity

(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.14: Various military logistic networks' responses to targeted supply disruptions. Average of 20 runs.

For the scale-free network, its supply availability and network connectivity deteriorate very rapidly. Although the scale-free network still maintains the highest efficiency at the early stage of disruptions, the deterioration rates on the two efficiency metrics are higher than for the other topologies. For example, scale-free networks have the highest average delivery efficiency when no disruption occurs, 47% higher than that of the ER-random network. However, as supply nodes are removed, the average delivery efficiency falls the fastest for scale-free, which is a major drawback. When 10% of the supply nodes are removed, scale-free already has lower average delivery efficiency than ER-random. An additional 10% removal will make the average delivery efficiency of scale-free 27% lower than that of ER-random. Generally, the performance of scale-free seems to confirm previous research that it is very fragile to targeted disruptions [41].

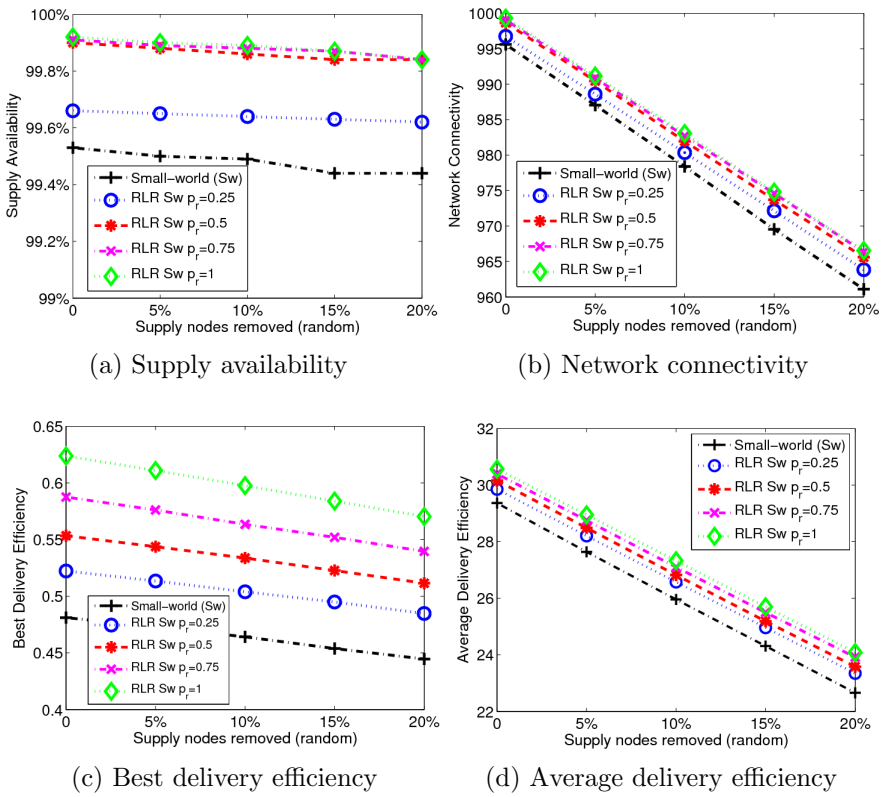
At the other end of the spectrum, the ER-random network deteriorates slowly

with increasing failure rate on all the four metrics, but it still suffers from poor initial values. For instance, even though its rate of decrease in best delivery efficiency is much lower than for the scale-free and several RLR scale-free networks, 20% supply node removal is still not enough for the ER-random to catch up with the other topologies on the delivery efficiency metrics.

As for the RLR scale-free supply-chain network, it still offers better availability and connectivity than scale-free. Its availability and connectivity also increase when using higher rewiring probability p_r. The supply availability and network connectivity of the RLR scale-free network with $p_r \geq 0.5$ are also higher than ER-random when 0%-20% of the supply nodes are removed. In terms of the efficiency metrics, its performance generally lies in between the scale-free and the ER-random, although the RLR scale-free with $p_r = 1$ slightly outperforms the ER-random in average delivery efficiency. Higher p_r generally leads to a lower initial delivery efficiency values, but also a slower deterioration rates. Take average delivery efficiency as an example. At the very beginning, the scale-free ($p_r = 0$) has the highest average delivery efficiency and the ER-random the least. Most RLR scale-free networks fall between the two and rewired scale-free with higher p_r have better initial efficiency than those with lower p_r. As supply nodes are removed, the gaps in average delivery efficiency get smaller. After 10% of supply-node removal, the RLR scale-free with $p_r = 1$ actually has the best average delivery efficiency, better than the scale-free, the ER-random and other rewired scale-free with lower p_r.

2.4.4 Rewiring Small-world Networks with RLR

Besides the scale-free topology, another popular complex network topology is the small-world [9], which features high clustering coefficients and low characteristic path length. The degree distribution of a small-world network is also more uniform than the highly skewed power law distribution. Adopting the simulation settings in Section 2.3.3., this research generates a small-world military logistic network using the model proposed in [9], applies RLR to the small-world network, and compares the robustness of the original small-world logistic network with that of the rewired ones (referred to as *RLR small-world networks*).



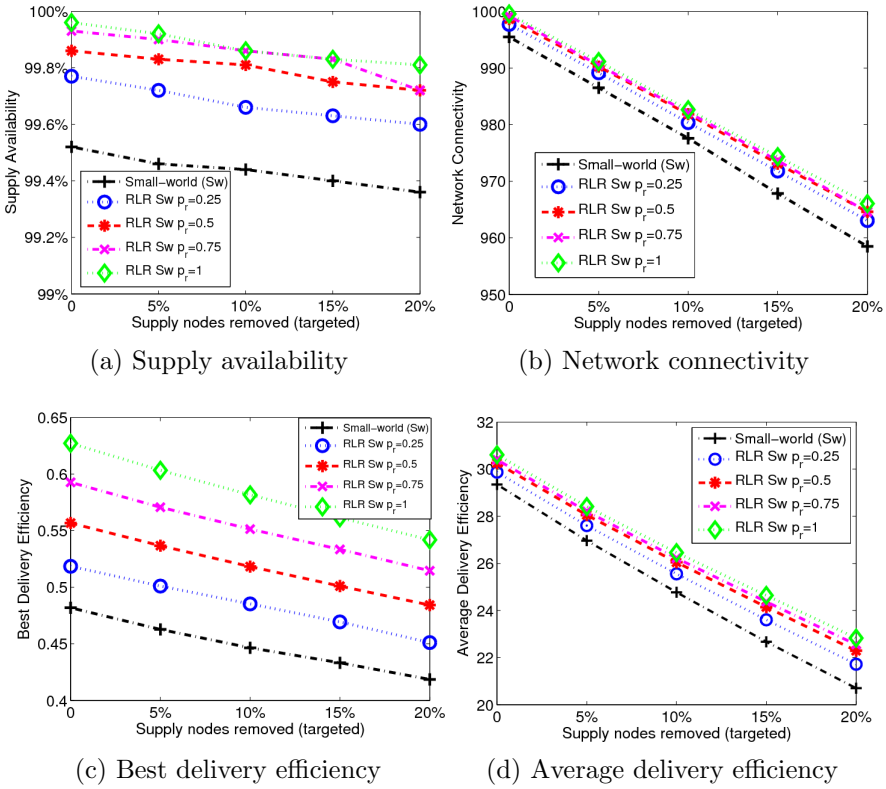
(a) Supply availability (b) Network connectivity

(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.15: Various military logistic networks' responses to random supply disruptions. Average of 20 runs.

Figure 2.15 illustrates the robustness of the military logistic network with the small-world topology and RLR small-world networks with various p_r when random disruptions occur. RLR small-world networks dominate the original small-world networks on all four metrics. Higher rewiring probability p_r generally leads to better performance. The advantage of RLR small-world networks mainly lies in their good initial performance, especially on the metric of best delivery efficiency. In terms of the rates of performance deterioration when supply nodes are removed, RLR small-world networks are only slightly slower than the small-world network on the metrics of network connectivity and average delivery efficiency.

The five networks' robustness against targeted disruptions (shown in Figure 2.16) is similar to that against random disruptions. Although the performance of all the supply-chain networks deteriorates slightly faster than in random disruptions, more rewiring still helps RLR small-world to become more robust than small-world.



(a) Supply availability (b) Network connectivity

(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.16: Various military logistic networks' responses to targeted supply disruptions. Average of 20 runs.

2.4.5 Discussion

According to the simulation study, rewiring a supply-chain network with RLR will affect its robustness in different ways. I first discuss the case when RLR is applied to a military logistic network with the scale-free topology. Comparing the scale-free supply-chain network with rewired RLR scale-free networks, it is found that *there is no optimal network topology* that dominates all others on every metric in both supply disruption scenarios. Some previous conclusions about the robustness of scale-free networks still hold in the heterogeneous military logistic networks. For example, while a scale-free network is able to maintain good average delivery efficiency in random disruptions, its robustness against targeted disruptions is unacceptable with low supply availability and network connectivity. Although its initial delivery efficiency values are better than other topologies, they deteriorate rapidly when high-degree supply nodes are disrupted.

Then, I turn to RLR scale-free networks. Recall that the rewiring probability p_r is tunable: a lower p_r will retain more aspects of the original topology; while higher p_r will add greater randomness. One advantage of RLR scale-free with high p_r, say 0.75 or 1, is that very high supply availability and network connectivity are maintained in both random and targeted disruptions (even as high as 20%!). However, its delivery efficiency is often not as good as for the other topologies, even with a low rewiring probability, in both disruption scenarios. While topologies such as scale-free and ER-random are usually robust against one type of disruption but fragile to another type, *RLR scale-free networks have the unique property that their behaviors on all four metrics are relatively consistent in both types of disruptions*, which gives rewired scale-free balanced robustness against both type of supply disruption.

In addition, for RLR scale-free networks, better performance on availability and connectivity is often gained at the cost of delivery efficiency. This trade-off, together with the tunable rewiring probability, provides another way to balance a supply-chain network's robustness. If availability or connectivity is more important than delivery efficiency for a supply-chain network, an RLR scale-free supply-chain network with high rewiring probability will be preferred. Conversely, for networks in which delivery efficiency is critical, an RLR scale-free supply-chain network with low rewiring probability may offer a better choice.

Another interesting issue pertains to the RLR scale-free network with $p_r = 1$. Since it randomly rewires every edge in the scale-free network, one might expect it to approach a random network. However, the ER-random network, which is constructed in a purely random manner, does not demonstrate its well-known robustness against targeted disruptions in heterogeneous supply-chain networks. Although the ER-random network generally features low performance deterioration rates when supply nodes are removed (in both random and targeted disruptions), it is outperformed by the RLR scale-free network with $p_r = 1$.

Then, what are the differences between the two supply-chain network topologies? The experimental results suggest that the advantage of RLR scale-free with $p_r = 1$ mainly lies in its initial performance. In an ER-random network, edges between any pair of nodes are established based on a pre-determined probability. I fear this growth model may leave some nodes with no connection at all.

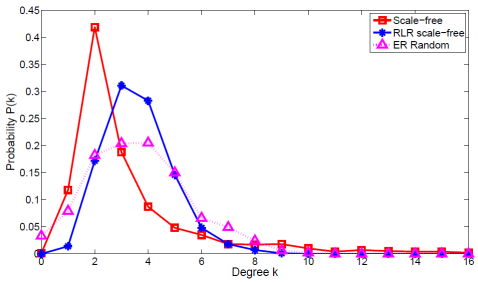


Figure 2.17: Degree distributions of simulated supply-chain networks ($p_r = 1$ for the RLR scale-free supply-chain network. All networks have 1000 nodes and 1815 edges).

To confirm the conjecture, I draw the degree distributions of simulated ER-random, scale-free and RLR scale-free with $p_r = 1$ networks in Figure 2.17. All three simulated networks have 1000 nodes and 1815 edges. The horizontal axes denote the degree of nodes, while the vertical axes reflect the probability of finding a node with a given degree. One might observe in Figure 10 that some nodes have degree 0 and are thus isolated from other nodes in the ER-random network. Such isolation directly affects its initial values on the metrics. As a comparison, the simulated scale-free network has almost no zero-degree node but features few nodes with very high degrees. The RLR scale-free network takes some advantage of the scale-free network's good initial performance by using a scale-free network as the basis for rewiring. As a result, the RLR scale-free has few, or none, disconnected nodes.

Moreover, the degree distributions of simulated RLR scale-free networks are in accord with my theoretical analysis in Appendix B. Compared with the simplified RLR scale-free network with $p_r = 1$ in Appendix B, the simulated rewired one with $p_r = 1$ in Figure 2.17 has a much shorter right tail and exhibits a near-Poisson degree distribution. The degree distribution of the simulated RLR scale-free supply-chain network and its performance in the simulations also validate my previous hypothesis regarding its networks' topology and robustness. In addition, some may notice that the degree distribution of the scale-free network in Figure 2.17 features fewer 1-degree nodes than 2-degree nodes. This is different from the monotonically decreasing distribution of the scale-free network in Appendix B. The low probability for 1-degree nodes can be explained by the simulation setting:

each new node initiates an average of 1.8 edges in the simulation and thus very few nodes have only one edge in the resulting network.

In contrast with the trade-offs between scale-free and RLR scale-free, when RLR rewiring is applied to a military logistic network with small-world topology, its impact is very positive. RLR small-world networks can help to improve the original network's robustness against both random and targeted disruptions, reflected by their very good performance on all four robustness metrics. Meanwhile, similar to RLR scale-free, the performance of RLR small-world supply-chain networks is also consistent in random and targeted disruptions.

In sum, when applied to scale-free networks, RLR rewiring is able to represent both the preferential and random attachment networks at the extreme rewiring probabilities $p_r = 0$ *and* $p_r = 1$ *respectively. Varying* p_r *between 0 and 1 generates intermediate supply-chain networks that can balance the robustness between scale-free and random supply-chain networks. When applied to small-world supply-chain networks, RLR can improve the robustness against both random and targeted disruptions, with higher* p_r *leading to better performance.*

2.4.6 An Experiment for a Retailer's Distribution Network

To further evaluate the performance of the RLR approach, another experiment is conducted to apply the model to a commercial supply-chain network–a leading retailer's distribution network in the state of California. Different from the synthesized scale-free or small-world military logistic networks, the retailer's distribution network does not conform to any established complex network topology and thus provides a more realistic test bed for RLR.

2.4.6.1 Simulation setting

The distribution network consists of 185 nodes, including 2 warehouses (both are located in southern California), 7 distribution centers (DCs), and 176 stores. Because the address of each node is available, it is possible to calculate the physical distance between any two nodes. As this three level structure is also quite similar to the military logistic network in the previous experiment, warehouses and DCs are considered as supply nodes and stores as demand nodes.

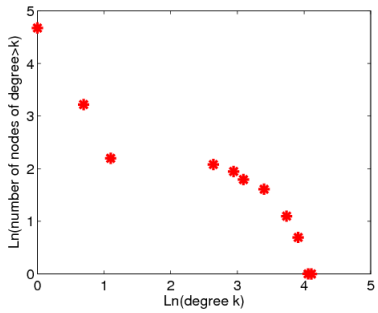


Figure 2.18: The ln-ln degree distribution of the retailer's distribution network in California.

On the basis of the retailer's rule of building a distribution network, a distribution network with the following realistic configuration is generated: (1) the two warehouses are connected; (2) each DC randomly connects to 1 or 2 warehouses and 2 other DCs; (3) each store uses preferential attachment to connect to a 1 or 2 DCs within a radius of 300 miles. 10% of all stores are able to connect directly to a warehouse.

The degree distribution of the retailer's distribution network (shown in Figure 2.18) is a combination of two power-laws. This is because the roles of nodes in the somewhat hierarchical distribution network polarize the node degrees. Lower in the hierarchy, stores' degrees are in the range of 1 to 3, and follow the power-law distribution (on the left); while DCs and warehouses generally have degrees higher than 12, and their node degrees are reflected in the power-law distribution on the right. Even though this network is not a scale-free or small-world network, I can still apply the RLR approach and rewire it.

As California is very large and the retailer's facilities are located all across the state, the RLR approach in this experiment will consider the radius limit when rewiring edges. While robustness is important, the cost to operate the distribution network is also a major concern for the retailer. It is expensive to rewire an edge and connect a store in southern California, an area with several DCs, to a DC in northern California 600 miles away. In this experiment, the maximum rewiring radius is set to $d_{max} = 300$ miles, so that an edge's rewiring destination must be within a physical distance of 300 miles from the node that the edge currently attaches to.

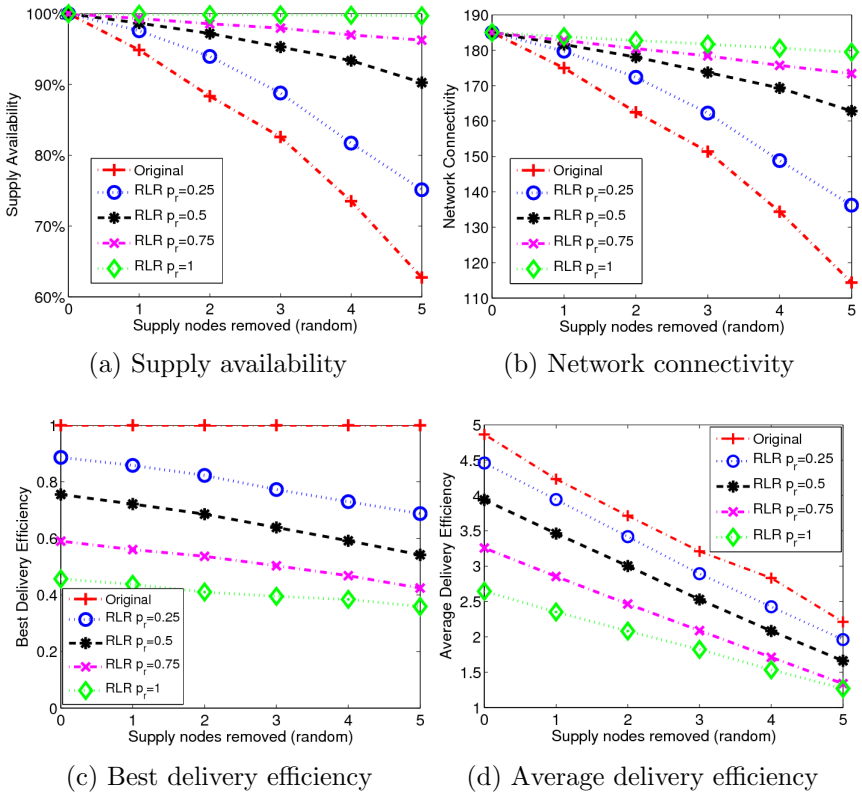
In the experiment, both random and targeted disruptions to the distribution network are considered. While random disruptions can happen in any distribution network, this retailer's distribution network may also face targeted disruptions. As a major retailer with hundreds of billions of annual sale around the world, the company is a strategic target for terrorist attacks and cyber-attacks [61], which tend to aim at warehouses and DCs that play more important roles in the network. Also, some of the retailer's highly connected warehouses and DCs are located near the coast of southern California, an area that is more subject to earthquakes and tsunamis than other areas in the state.

In my simulation, 5 supply nodes (out of 9) are removed – one at a time between successive observations. During the process of node removal, the robustness metrics for each network topology are tracked. While evaluating the supply-chain networks, I also set $f(j) = 1$ in Equation 2.12 for average delivery efficiency. I apply the RLR approach with four different rewiring probabilities ($p_r = 0.25, 0.5, 0.75$, and 1) to the original network ($p_r = 0$) and compare the robustness of rewired networks (referred to as *RLR retail networks*) with the original one.

2.4.6.2 Experimental results

Figures 2.19 and 2.20 show the performance of the five distribution networks against random and targeted disruptions respectively. The horizontal axes denote the number of supply nodes that have been removed, and the vertical axes are the topology-level supply-network robustness metrics. As one can see, the results are similar to those for the military logistic network.

The original distribution network is able to maintain very good delivery efficiency in both types of disruptions. Its delivery efficiency does not change when supply nodes are removed because, as noted earlier, nodes in the network generally follow a hierarchical structure. Stores only connect to DCs or warehouses, but do not connect to other stores. Thus if a store can still access a DC or a warehouse after disruptions, the closest DC or warehouse is always the store's immediate neighbor. However, when disruptions happen, the network is fragmented easily and many stores lose access to DCs or warehouses, as shown by its very fast deterioration in network connectivity and supply availability, especially in targeted disruptions.



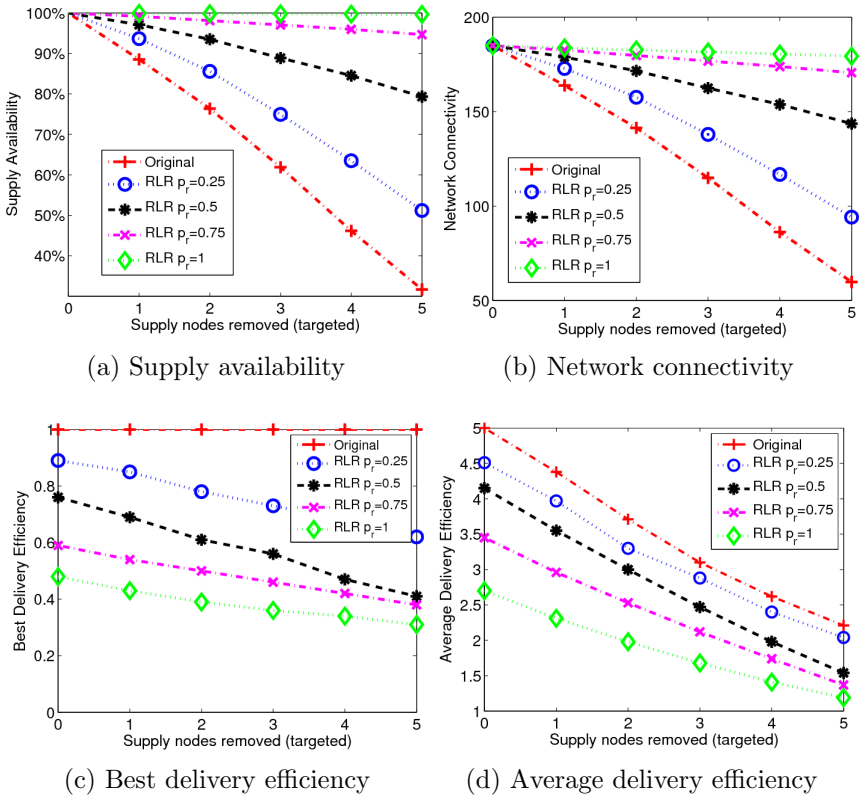
(a) Supply availability (b) Network connectivity

(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.19: Various retail distribution networks' responses to random supply disruptions. Average of 30 runs.

In summary, the experiment on the retailer's distribution network also reveals that no network topology can dominate all others on every metric, but it is important to understand the tradeoffs that are involved. The experiment also validates the advantage of RLR retail networks on supply availability and network connectivity, as well as its stable performance in both random and targeted disruptions.

As mentioned earlier, some retailers may adjust or rebuild their distribution networks to adapt to the changing market conditions. The RLR approach offers a simple yet effective heuristic strategy to improve the distribution network's robustness on supply availability and network connectivity. A manager of the distribution network can add some controlled randomness into the existing network by rewiring edges. The manager can select new destination of a rewired edge among nodes that are not too far away from the remained node of the edge. The tunable re-wring probability and the maximum rewiring distance also allow the retailer to balance



(a) Supply availability (b) Network connectivity

(c) Best delivery efficiency (d) Average delivery efficiency

Figure 2.20: Various retail distribution networks' responses to targeted supply disruptions. Average of 30 runs.

the distribution network's availability/connectivity and delivery efficiency on the basis of robustness requirements and the operation budget.

2.5 Summary and future work

The objective for the research in this chapter is to study how topological properties affect supply flows and the robustness of supply-chain networks against disruptions. The consideration of nodes play heterogeneous roles and supply flows among these nodes in a supply-chain network, which is not the case for many other networks, leads to novel robustness metrics. Then new supply-chain topologies that perform well when their robustness was evaluated with the new metrics are also needed?

Therefore, this study first proposes the taxonomy of metrics for supply-chain network robustness to capture the characteristics of supply flows and reflect the

node heterogeneity. The taxonomy consists of system-level metrics, including supply availability, network connectivity, accessibility, and delivery efficiency, and corresponding topology-level metrics. Corresponding to the changes in robustness metrics, two strategies are proposed to design or adjust the topology of supply-chain networks.

The first strategy, DLA, is a new general and hybrid growth model to build or design supply-chain networks. Attachments in this model are based on both degree and locality. Compared with other growth models, DLA represents a different strategy for distributed connections: nodes consider both efficiency and distance when deciding which nodes to connect to. Using simulations, I compared the robustness of supply-chain networks generated with various network growth models. The results reveal that DLA networks have desirable robustness properties in both random and targeted disruptions. Even though DLA is not the dominant model in terms of robustness against all types of disruptions, it offers an excellent *compromise strategy* to establish connections in supply-chain networks when it is not possible to predict whether a random or a targeted attack will occur, which is often the case in the real world. It has been shown that by adjusting the parameters of the DLA model, one is able to tune DLA networks' relative performance against the two types of disruptions. Recall that the multiple robustness metrics can be combined to generate a single objective function for a specific supply chain. Consequently, for supply-chain network managers or individual decision-makers, the DLA model represents a simple yet effective heuristic strategy to build a Pareto-optimal supply-chain network with balanced robustness in the supply-chain network's operational context.

The second strategy, RLR, is based on probabilistic and localized rewiring of a supply-chain network and can be used to adjust the topology of existing supply-chain networks. Supply-chain networks rewired with the RLR approach with various rewiring probabilities have some nice robustness properties under both random and targeted supply disruptions. In the experiments, I apply the model to military logistic networks with scale-free and small-world topologies and a retailer's distribution networks to compare the robustness of original and rewired networks in detail using computational simulations. The results suggest that the proposed RLR approach can improve the robustness of a supply-chain network in

supply availability and network connectivity, and is a good option in situations where both random and targeted disruptions can occur. I also compare the effect of varying the rewiring probability p_r on the performance. Increasing p_r, i.e., bringing more controlled randomness into the network, generally leads to better supply availability and network connectivity, but at the cost of delivery efficiency if the original network has a non-uniform or skewed degree distribution. This is intuitively meaningful. To decide the optimal value for p_r, a designer or manager should consider the trade-off between the higher costs of accessing supply nodes because of longer distances when p_r is high versus the cost of stock outs if some nodes cannot obtain supplies when p_r is low.

Although a military logistic network and a retailer's distribution network are used as case studies from different industries, the taxonomy of robustness metrics and the DLA and RLR approaches may also provide insights to the study and design of robust supply-chain networks in other domains or industries. Specifically, my approach provides simple strategies for supply chain designers or managers to build new or fine-tune existing supply-chain networks and get balanced robustness on the new metrics when both random and targeted disruptions are likely to happen. By shedding new light on the robustness of supply-chain networks with heterogeneous nodes, this research could help to understand how an existing supply-chain network evolved over time and how different microscopic paradigms to design or adjust a supply-chain network will affect its robustness at a macroscopic level. With these topological considerations in mind, managers are also better informed in the future expansion or adjustment of existing supply-chain networks.

The simulation platform also provides the basis for the future development of a decision support tool. Such a tool will help managers to: (1) evaluate the robustness of an existing supply network, (2) assess how possible rebuilding strategies affect the resulting network's robustness, and (3) quickly make rewiring decisions when disruptions occur. In addition, the research may also be applicable in other complex networks whose operations rely on the delivery of people, information, goods, or services between entities with heterogeneous roles. Example networks with similar features may include communication networks such as the Internet (with servers and clients), and infrastructure networks such as power grids.

There are several areas that I would like to address in future. First, in my simulation of disruptions, only the removal of nodes are considered. This corresponds to the failures of entities in a supply network. Actually, a real-world supply network may also face disruptions of connections, e.g. a road block may force a manufacturer to find an alternative path to deliver the goods. Thus, it would be useful to study the removal of edges from the supply network. Second, my analysis of robustness in this paper is only on the topological level. Nevertheless, the design of supply networks in the real world needs to consider more operational level factors and constraints, such as the capacities of supply nodes, the needs of demand nodes, and the cost of transportation. Thus I also plan to apply network optimization techniques to robustness analysis, so that this research can provide better decision support. Third, as suggested in previous research [62], I would also like to study the effect of adaptive behavior, such as increased capacity and new delivery routes, on the distribution network robustness in the presence of disruptions. Fourth, this research classifies nodes in a distribution network into two roles: supply and demand nodes. While this classification is simple and straightforward, it may overlook the diversity of node roles in a supply network. By incorporating more heterogeneous roles into this research, I will be able to improve the validity of my conclusion. Other possible research directions include a more analytical study of the robustness of networks rewired with RLR, and the cascading failures of nodes caused by load redistribution [63].

Chapter 3

Information Flows in Online Health Communities–An Analysis at the Individual Level

While supply or information flows in a supply-chain network may be planned or managed, the flows of information among individuals in social networks are often driven by distributed inter-personal interactions. Such flows may change people's behaviors, attitudes, or emotions. Focusing on network flows at the micro level, the research in this chapter leverages the large-scale data of individual users' interactions and utilizes the information flow to identify influential users from social networks in online health communities (OHCs). This chapter starts with introduction to this research and related work. Then two approaches are used, compared, and eventually integrated to identify influential users in a popular OHC among more than 27,000 cancer survivors. The outcome of this research has been published in two conference papers [64][65].

3.1 Introduction

In recent years, more and more people turn to Internet for health-related needs [66]. A recent study by the Pew Research Center found that seeking health-related information becomes the third most popular online activity, with 83% of Amer-

ican adult Internet users looking for health information online, with more than 25% of them seeking support through joining and participating in an online health community [67]. Compared with traditional health websites that only allow users to retrieve information, online health communities (OHC), which focus more on social networking among users, can better meet such diverse needs of users. Patients use OHCs for a wide range of information and support needs, such as "find second opinions, seek support and experiential information from other patients, interpret symptoms, seek information about tests and treatments, help interpret consultations, identify questions for doctors, make anonymous private inquiries" [68]. Social networking through an OHC can help users to obtain more social support, which is beneficial to patients in adjusting to the stress of the disease [69] and is a consistent indicator of survival [70]. An OHC can also serve as an outlet for users' emotional needs and improves users' offline life [71].

The effectiveness and proper functioning of these OHCs depends greatly on influential users (IUs) [72]. Identifying IUs has utilities for community building and management, marketing, and information retrieval and dissemination. For an OHC specifically, finding IUs may have additional implications in advocating new treatments, guiding the proper use of drugs, and encouraging healthy lifestyles and positive attitude [73][74][75][76].

Then who are these influential users? How can they be identified? This research shows how IUs can be identified automatically from the online forum of a popular OHC among cancer survivors. Two approaches are adopted–the first one is a data-driven approach and the second one utilizes the sentimental impact of information flows among OHC members.

3.2 Related work

Social network theories on IUs fall into four schools, all of which are based on a network representation of the community. The first school asserts that people who are on the paths between pairs of individuals in the network can influence the flow of information or resources. Hence, betweenness centrality and its variants have been introduced [77][6][78]. The second school addresses the importance of high-degree nodes (i.e., hubs) and demonstrates that the removal of hubs, espe-

cially from a scale-free network, can break up a network into multiple disconnected components [41]. The third school emphasizes on the distance between nodes and argues that those who have shorter path lengths to other nodes in the network are more influential [79]. The fourth school distinguishes between a person's in-degrees (e.g., followers) and out-degrees (e.g., following). The rank of an individual is determined by the rank of those with a link pointing to that individual [80]. These theories and their extensions have been widely adopted in the analysis of real-world social networks.

The emergence of online communities, where users often interact through open discussions, provides important opportunities for using novel computational social science approaches [81] to identify influential users from large-scale social networks. In addition to the structure of users' social networks, online communities also capture detailed information of micro-level online interactions (e.g., the amount, content and time of interactions) that are typically not available in other types of social networks. Research that tries to identify IUs in these communities considers not only network-level centralities, but incorporates individual users' behaviors and contributions. For example, in online communities that feature significant contagion or diffusion phenomena, as seen in Twitter or more generally with Internet viral marketing, one can use contagion maximization to find influential users [82]. In an online Q&A community, the difference in knowledge between question-askers and answerers has been used to find experts [83], a type of influential user. Analyses of the blogosphere have utilized blogger contributions (e.g. the number and length of posts) and reader activities (e.g., comments to a posts) to assess the influence of a blogger [84]. Analyses of dark web forums have used content similarity among users' posts to identify IUs [85].

What is an accurate metric to identify IUs in an OHC? An OHC provides multiple types of social support to users, including information support (e.g. asking and answering questions about a specific treatment), emotional support (e.g. sharing experience and seeking prayers), and companionship (e.g. initiating discussions about weekend plans or playing online Scrabble) [86]. In such a multi-purpose and interactive community, existing metrics in the literatures cannot be directly used. For example, an IU may actively start and follow threaded discussions, possess high network centrality, and post content that brings a positive spirit to the

community. Consequently, this research proposes two ways to find a metric from the large-scale data of users' online activities and distributed interaction. It is also worth noting that both approaches in this study try to find IUs for the community, not those who are influential for a specific OHC member.

3.3 The dataset

This research uses de-identified data from the online forum in the American Cancer Society Cancer Survivors Network (CSN, http://csn.cancer.org). Cancer accounts for 1.8 million deaths in China, 0.57 million in the U.S. (both in 2010), and 7.6 million deaths worldwide in 2008 [87]. As of 2007, 11.7 million Americans have been diagnosed with cancer and are either living free of cancer or still have evidence of the disease. Social support can help cancer survivors cope better with the disease and improve their life [69]. Moreover, the information and support from online peers can be quite valuable, since cancer survivors' experiences are unique and may not be well understood by their family members, friends or care providers.

CSN is an OHC with more than 146,000 registered members created by and for cancer survivors and caregivers. CSN provides a peer support service and psychosocial intervention that utilizes group dynamics to facilitate therapeutic factors such as instillation of hope, universality, catharsis, existentialism, altruism, interpersonal learning and group cohesiveness. The conceptual framework for CSN is based on the work of Irvin D. Yalom, a foremost group work theoretician and practitioner [88].

While CSN provides several services to its members, such as newsletters, online chat rooms, and personal blogs, the most commonly used features is its online forum. The forum consists of 38 discussion boards–25 of them are cancer-specific ones, such as the popular breast cancer forum and colorectal cancer forum. The dataset used in this research consists of forum posts from July 2000 to October 2010, comprised of 48,779 threaded discussions with more than 468,000 posts from 27,173 users. A nationwide ACS marketing campaign featuring CSN, conducted in spring 2008, resulted in greater than 300% growth in membership; 70% of all forum posts in the dataset were after the end of 2008. There was relatively stable traffic in the forum from May 2009 to October 2010, with an average of 16,550

posts per month. The number of posts (both initial posts and responding replis) each user published in the forum follows a power-law distribution (Figure 3.1).

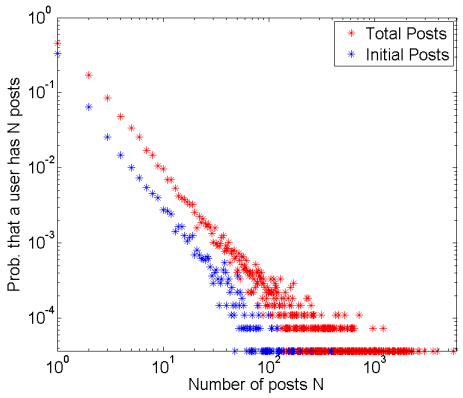


Figure 3.1: Distributions of number of posts each user published.

Each threaded discussion could consist of three types of posts: it starts with an *initial post*, which is published by the thread originator and followed by two types of replies. Replies from other users (respondents) are called *responding replies*. In many cases, the thread will contain additional replies from the originator (*self-replies*). Figure 3.2 shows an example of threaded discussions.

The number of replies (both responding and self replies) in a thread approximately follows a power-law distribution (see Figure 3.3). The life span of a thread, defined as the time between the initial post and the last reply (either a self-reply or a responding reply), follows the distribution pattern in Figure 3.4. Note that while the mean life span is as long as 1,725 hours (or about 72 days), the median is only 58 hours and 72% of threads have life spans shorter than one week. Interestingly, there is not a strong correlation between the number of replies and the life span of a thread. The Pearson correlation coefficient is 0.04, suggesting that a long thread with many replies does not necessarily have a long life span. Table 3.1 summarizes the basic statistics of the forum dataset used in the analysis.

3.4 A classification approach

As mentioned before, user's activities and contributions can hardly be captured by a single feature and a combination of multiple features may be necessary. Thus a

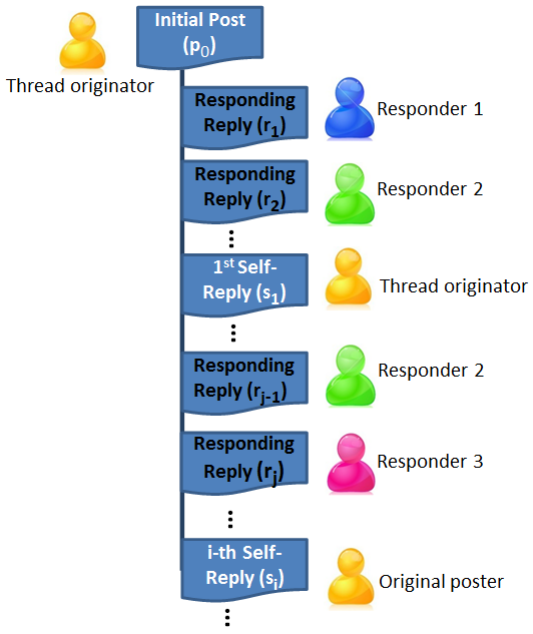


Figure 3.2: An example of threaded discussions.

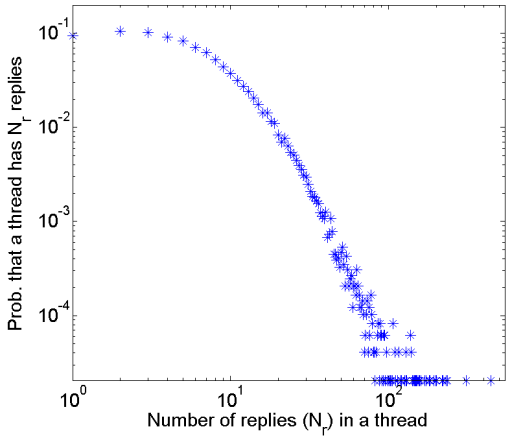


Figure 3.3: The distribution for the number of responding replies in threads

classification approach is used to find proper ways to combine multiple features of users' activities.

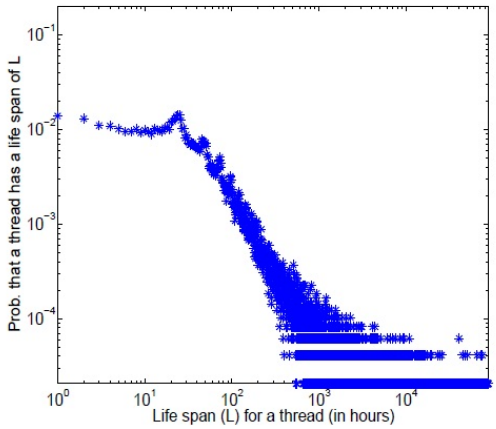


Figure 3.4: The distribution for the life span of threads

Table 3.1: Summary of statistics for the CSN forum dataset used in this research.

	Mean	Median	Maximum
Number of posts by a user	17.25	2	5,607
Number of replies per thread	8.7	6	442
Life span of threads	1,725 hrs	58 hrs	87,846 hrs

3.4.1 Features and initial results

Three groups of features are extracted from the forum for IU classification: contribution, network, and semantic features (see Table 3.2). Contribution features, as the name implies, measure a user's direct contribution to the forum, such as number of discussion threads initiated and replies posted, number of days the user has actively posted, length of posts, etc. Network features reflect users' centrality (e.g. in/out-degree and betweenness) in a post-reply network, where there is an edge pointing from user A to user B if A replied a thread started by B. Derived from the content of users' posts, semantic features reflect the positive/negative sentiment, strength of emotion, diversity of topical coverage, etc. The list of sentimentally positive and negative word is based on the work of Hu and Liu [89], and the positive and negative emoticon lists are collected from Internet. The topical diversity is calculated using Latent-Dirichlet Allocation [90] (with 50 clusters).

Classification techniques are utilized to distinguish between IUs and regular users. To train a classifier, one first needs to label a set of users as either an IU (a positive record, +1) or non-IU (a negative record, -1). While it is easy to

Table 3.2: Summary of basic user features

Group	Features
Contribution features	Number of one's initial posts (i.e., posts that start threads) Number of one's responding replies to threads (i.e., following posts) Number of threads that one contributed responding replies to Number of other users' responding replies published after one's responding replies in the same thread Avg. response delay between one's responding replies and the next responding reply by others in the same thread (in minutes)
	Total length of one's post (in bytes) Avg. length of one's post (in bytes) Avg. content length of one's top 30 longest posts (in bytes) Number of one's active days (a user needs to publish at least 1 post in an active day) Time span of one's activity (from first active day to the last) Avg. number of posts per active day Avg. number of posts per day during one's time span of activity
Network features	One's in-degree and out-degree in the post-reply network One's betweenness centrality in the post-reply network One's PageRank in the post-reply network
Semantic features	Avg. percentage of words with positive sentiment in one's posts Avg. percentage of words with negative sentiment in one's posts Avg. percentage of Internet slang/emoticons in one's posts Avg. percentage of words with strong emotion in one's posts Ratio between the numbers of words with positive and negative words in one's posts Topical diversity (Shannon entropy and log energy of topic distribution in a user's posts)

label users whose activity levels are very low as non-IUs, finding an IU requires good knowledge of the user's activity history and other users' reactions to the user's posts over an extended period of time. As this task is very difficult even for someone who knows the forum well, domain experts are consulted. The CSN community manager and two staff members who monitor the forum content on a full-time basis are asked to nominate IUs. 41 members were nominated as IUs (referred to as *IU List-1*) . Although a subjective and imperfect labeling, the non-exclusive list provides a good starting point and help to understand what type of behaviors contributes to the status as an IU and how to improve classifiers. It is

worth noting that ranking of List-1 is not performed because these experts do not think that it is feasible to do so reliably for such a high number of users.

Identifying IUs posed a challenge for regular classifiers–an unbalanced dataset. As only a small number of users are IUs, the dataset contains many more negative than positive records. With such an unbalanced dataset, a regular classifier is biased towards simply classifying all users as non-IUs yet achieving a very high accuracy rate. To tackle this problem, this research processes user features in two steps. First, assuming that it is almost impossible for one to become an IU with little contribution to the forum, this research rules out obvious non-IUs using two empirical rules based on users' activities: (1) users whose total number of posts is below average (17 posts) or (2) users whose total number of active days is below average (7.7 days) are non-IUs. In fact, all 41 nominated IUs made a considerable contribution in terms of total posts and active days. With the two rules, 91.4% of all users in the dataset are eliminated from consideration as IUs, leaving 2,336 users as unclassified. However, the much smaller dataset is still unbalanced. Second, positive records (IUs) are over-sampled 20 times, a common practice in dealing with unbalanced datasets, to raise the ratio between positive and negative records to about 0.36. This step provides a more balanced dataset (hereafter referred to as Dataset-1) with which to train classifiers.

Five classifiers are applied to Dataset-1 using 10-fold cross-validation: Naíve Bayesian, Logistic Regression, Random Forest, one-class SVM, and two-class SVM. A proper metric is also needed to evaluate the performance of these classifiers. Traditional classification metrics, such as F-measure and area under the ROC, place equal importance on precision and recall. However, in this case, the correct identification of the 41 nominated IUs (i.e., recall) is more important. Because the 41 nominated users are not the only IUs in the forum, identifying IUs who are not nominated is a helpful undertaking. Although recall is emphasized, one certainly does not want to get a perfect recall of 1 by classifying all users as being IUs. Instead, top-K recall (also known as recall@K) is used to evaluate classifiers' performance. Basically, classifiers are asked to provide a probability of being an IU to each user. Then K users whose probabilities are among the top K of all users are identified (note that one-class SVM only provides binary decisions. Thus "outlier" ratio in the classification process is specified, so that the classifier identifies K IUs).

Table 3.3: Compare the top-150 recall of 5 individual classifiers on 3 datasets (using IU List-1).

	Naíve Bayesian	**Logistic Regression**	**Random Forest**	**One-class SVM**	**Two-class SVM**
Dataset-1	0.789	0.685	0.738	0.557	0.732
Dataset-2	0.796	0.681	0.779	0.561	0.724
Dataset-3	0.781	0.706	0.731	0.781	0.739

The top-K recall metric is the fraction of the 41 nominated IUs that are among the top K users. For example, if 30 out of the 41 nominated IUs can be ranked within top 150, then the top-150 recall is $30/41 = 0.732$. The top-150 recall obtained from the 5 classifiers range from 0.557 to 0.789 (see Table 3.3 for details).

3.4.2 Improved results with new features and ensemble methods

It is observed that identifying IUs in boards that are not very active is challenging. There are 25 cancer-specific boards, such as "Brain cancer", and 13 non-cancer-specific ones, such as "Caregivers". About 90% of all users contributed to only one discussion board (see Figure 3.5), so users may be fragmented into multiple sub-communities, each of which could have different IUs. Because sizes of sub-communities vary, IUs in these sub-communities may exhibit different online behaviors. For example, one nominated IU who participated in only one of the most active boards published 5 times more posts and had 3 times higher out-degree than another nominated IU who participated in a much less active board. This is most likely due to the fact that the board with less activity attracts fewer users and posts than highly active ones. While it may take more contribution for one to stand out from the crowd and become an IU in a larger and more active sub-community, it can also be difficult for even an IU to have high in/out-degrees in a sub-community with fewer users.

Then how can one identify IUs in both large and small sub-communities at the same time? Intuitively, a board-based division can be used: each discussion forum can be considered as a sub-community, from which users' features are gathered and

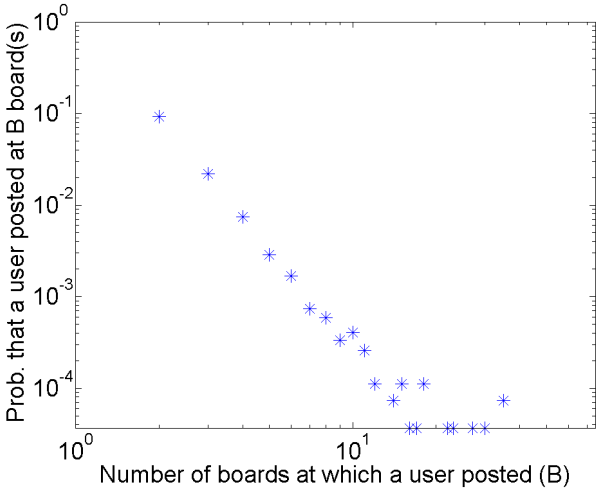


Figure 3.5: Distribution of the number of discussion boards to which a user contributed

normalized. Then there are 38 datasets, one for each sub-community (i.e., a board). Each sub-community may have its own IUs. However, the one-to-one mapping between boards and sub-communities is arbitrary and may be disadvantageous to those contributing to multiple boards. The division of user sub-communities should be based on users' interactions, which are not necessarily separated by the boundaries of boards. When several discussion boards share a similar user base, these users have closer relationship with each other and should be viewed as one sub-community that spans these boards. For instance, a group of more than 3,000 users contributed to both the colorectal and the anal cancer boards, while there were no common users between the stomach and the uterine cancer boards. With the board-based division, if an IU contributes almost evenly to multiple boards that share a similar user base and gains reputation among this group of users, her contribution will be divided into each board. As a result, her contribution to each individual board may not be high enough for her to be recognized by classifiers as an IU. By contrast, classifiers may favor another user who has made less overall contribution but participates in one board only. Also, while one can get a user's contribution and semantic features in a board by examining the user's posts in this board only, the board-based division of sub-communities may lead to inaccurate board-based network features, as it will arbitrarily cut many inter-user ties that are formed through their interactions in other boards. Thus this approach may

fail to find IUs who act as a bridge or broker between two sub-communities that are not well connected.

Instead of the arbitrary board-based sub-community division, my approach emphasizes on users' interactions and proposes two new groups of features to facilitate IU identification in sub-communities, especially smaller or less active ones.

The first group, *neighborhood-based features*, takes an ego-centric approach and focuses on local neighborhood. The idea is that a user whose contribution and centrality are much higher than her network neighbors is more likely to be an IU. I look at who a user interacts with and measure how the user stands out in her neighborhood. In other words, I measure how a user's contribution and network centrality differ from those of her neighbors. For each user in Dataset-1, I generate a new neighborhood-based feature from each of the user's contribution and network features using Equation 3.1, where N_i is the set of user i's neighbors. As Equation 3.1 shows, the neighborhood-based feature j for user i ($F'_{i,j}$) is the difference between user i's value on feature j ($F_{i,j}$) and the average of user i's neighbors values on feature j. Adding neighborhood-based features to the original Dataset-1, another dataset (Dataset-2) is created for classifiers. As Table 3.3 suggests, adding neighborhood-based features has improved the top-150 recall of 2 of the 5 classifiers.

$$F'_{i,j} = F_{i,j} - \frac{\sum_{k \in N_i} F_{k,j}}{|N_i|} \tag{3.1}$$

While neighborhood-based features are simple to calculate and help some classifiers find more IUs, disadvantages arise when examining users who connect to other IUs with high contributions. Consider Figure 3.6 that depicts two sub-communities in the user network (users' values on feature f are listed below their IDs). Users G and I are IUs and are connected to each other. As a result, user I has a neighborhood-based feature value of only $7 - 24/4 = 1$ on feature f. Meanwhile, user F, who connects only to the low-contribution user B, gets a neighborhood-based value of 4. As users I and F have the same original value 7 on feature f, the classifier will tend to consider user F, who has higher neighborhood-based feature value, as an IU instead of user I, who is actually a "big fish in a small pond" that I want to find.

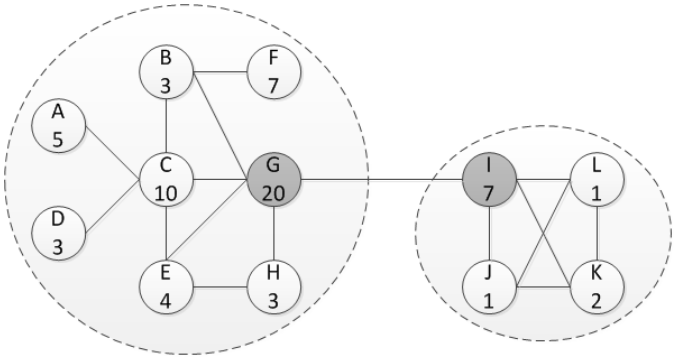


Figure 3.6: An example of a user network with two sub-communities (IUs are in gray. Users' values on feature f are below their IDs).

To complement neighborhood-based features, a second group of *cluster-based features* are added. Cluster-based features go beyond the local neighborhood and emphasize on the sub-community structure in the user network. The modularity-maximization algorithm [11] is used to cluster the post-reply network among users into 16 sub-communities: the largest has about 5,000 users and the smallest has only 90 users. Similar to neighborhood-based features, the cluster-based contribution or network feature j for each user i in Dataset-1 is calculated using Equation 3.2, where C_i is the network cluster to which i belongs. Back to the example in Figure 3.6, cluster-based features of user I do not depend on user G's, but on those of users J, L, and K, who belong to the same cluster as user I. Adding new cluster-based features to Dataset-2 produced Dataset-3. It turns out the new dataset complements Dataset-2 very well, raising top-150 recalls on 3 classifiers that Dataset-2 fails to improve (see Table 3.3). In summary, the new neighborhood-based and cluster-based features can help all the 5 classifiers improve classification performance.

$$F''_{i,j} = F_{i,j} - \frac{\sum_{k \in C_i} F_{k,j}}{|C_i|} \tag{3.2}$$

To further improve the classification of IUs, an ensemble approach [91] is adopted. The basic idea is to combine the outcome from various classifiers. For each user, a classifier gives a classification result, either as a probability or a binary value to denote whether the user is considered an IU. Then each user's five classification results from the five individual classifiers are fed to another classi-

Table 3.4: Compare top-150 recalls of various approaches.

Approach	**Top-150 recall**
The ensemble classifier	0.854
Ranking by the number of a user's total posts	0.732
Ranking by the number of a user's total active days	0.756
Ranking by a user's total degree in the post-reply network	0.707
Ranking by a user's PageRank in the post-reply network	0.731
Ranking by a user's betweenness in the post-reply network	0.463

fier. For each individual classifier, I pick the dataset that enables the classifier to achieve the highest top-150 recall, e.g., Dataset-2 for Naíve Bayesian, Dataset-3 for Logistic Regression, and so on. Among many ensemble methods, the ensemble classifier based on Random Forest achieves the best performance: an average top-150 recall of 0.854, which is higher than any of the individual classifiers in Table 3.3. In addition, when compared to user ranking based on traditional contribution and centrality metrics, the ensemble classifier performs better in terms of top-150 recalls (see Table 3.4).

3.5 A sentiment approach

In the classification approach, various user features are extracted to approximate a user's influence from various perspectives. However, the classification has two limitations: *First*, it lacks transparency, especially when an ensemble approach is used to boost its performance. It is difficult to give intuitive explanation on what an IU is like. *Second*, these features do not directly measure influence defined as "the power or capacity of causing an effect in indirect or intangible ways" [92]. Social contacts are known to influence health-related behaviors and emotions in individuals [93][94]. Provision of emotional support is a key component of OHCs, especially OHCs that cater to individuals with serious medical conditions, such as cancer. Individuals with serious medical conditions often experience stress and anxiety especially around the time of first diagnosis and during treatment [95]. However, this important function of OHCs has not been reflected in the literature and the classification approach.

The sentiment approach proposes a novel way to IU identification. It is based on

Table 3.5: Examples of posts in CSN and their sentiment classes.

Sentiment class	**Content of the post**
Negative	My mom became resistant to chemo after 7 treatments and now the trial drug is no longer working :(, ...
Positive	..., I love the way you think, ..., hope is crucial and no one can deny that a cure may be right around the corner!!!

the assumption that influential OHC users affect the emotion of other community members through online interactions. Hence, by measuring the effect of interpersonal information flow and influence, the approach should be able to identify IUs in an OHC. The proposed approach utilizes individual OHC users' sentiment dynamics and develops a new metric based on sentiment influence.

3.5.1 Sentiment analysis of posts

In an OHC, user emotions cannot be directly observed, but the sentiment of their posts can reflect their emotions at the time of posting. Manually labeling sentiment for a large number of posts is a labor-intensive task that is often infeasible. Instead, text mining techniques are used to detect the sentiment of posts automatically [96] and classify texts into *positive* or *negative* sentiment classes. Calibration of the classification algorithm relies on training data consisting, in this case, of 300 randomly selected posts that are manually labeled into positive or negative sentiment classes. An example of a negative and of a positive post is shown in Table 3.5. Next, lexical and stylish features features are extracted from each post, including the numbers of words with positive and negative sentiment, the number of Internet slangs, the numbers of question marks and exclamation marks, etc [64]. Finally, machine learning classifiers are trained (fitted or calibrated) with these data. The ultimate goal is having a fitted classifier that best classifies posts back to their originally labeled sentiment class. Figure 3.7 illustrates how the automatic sentiment analysis is conducted on posts in the forum.

Of the several classifiers that are tried, AdaBoost [97], which incorporates regression trees as weak learners, has the best sentiment classification performance [64]. Its classification accuracy of 79.2% is reasonable and in line with other senti-

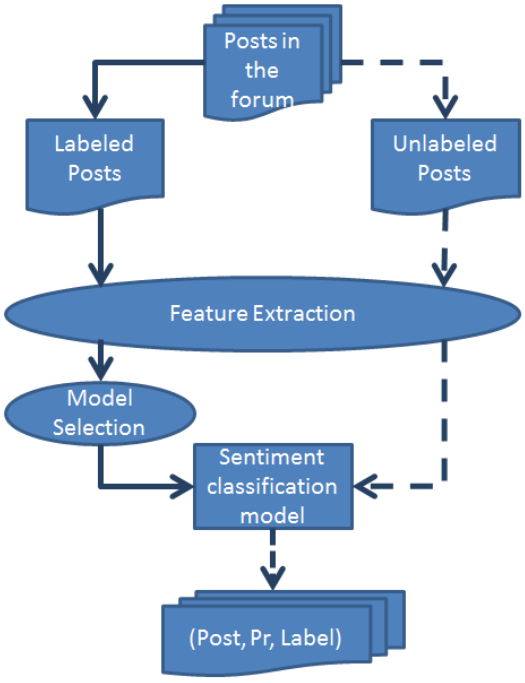


Figure 3.7: A flow chart of the automatic sentiment analysis using classification.

ment analysis of various domains that have re-ported accuracy rates ranging from 66% to 84%. This sentiment classification model is applied to all unlabeled posts, producing a sentiment label for each. Specifically, for each post p_i, the sentiment classification model estimates a sentiment posterior probability, $P_r(c = pos|p_i)$, which measures how likely the post belongs to the positive sentiment class (class "pos") given its post characteristics. If $P_r(c = pos|p_i) > 0.5$, post p_i is labeled as positive; otherwise, it is labeled as negative.

3.5.2 The new metric based on sentiment dynamics

Given the assigned sentiment of each post, this research utilizes the sentiment dynamics caused by information flows within threads to develop a metric that reflects each user's ability to influence others' sentiment. Thread originators are often seeking some form of support from the community. It is expected that information in replies from other users will exert some level of influence on the originator, so that sentiment of the originator's subsequent self-replies may change. From such sentiment change, a metric can be derived to measure how influential

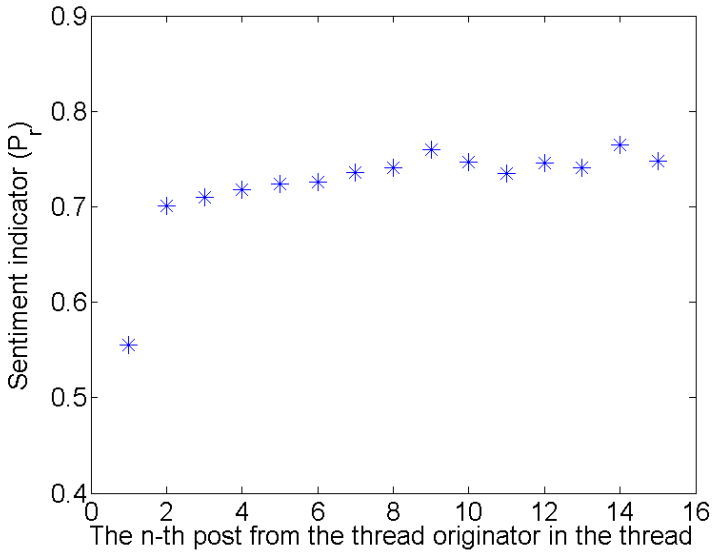


Figure 3.8: Sentiment change of thread originators by number of posts. A point represents the average sentiment of thread originators' n-th posts in threads they initiated. As the 2nd post from the originator is the 1st self-reply, the 2nd data point from the left-hand side denotes the average sentiment of originators' first self-replies.

responding users are.

If a thread does not receive any responding reply, presumably the originator does not receive information and support from the community. If the thread originator does not publish any self-reply to responding replies, their sentiment change cannot be gauged. From the 48,779 threads in the forum, only 23,000 threads satisfy two essential conditions for examining sentiment change: (1) a thread must have at least one responding reply, and (2) it must contain at least one self-reply from the thread originator. By comparing sentiment in the initial post with the sentiment in subsequent self-replies, it is possible to measure the impact of respondents on originators.

As Figure 3.8 illustrates, the sentiment of thread originators do change within threads they initiated. Significant change often occurs between their initial posts and their first self-replies. Additional self-replies they published within the thread do not reflect as much sentiment change. For this analysis, the sentiment of originators' self-replies is simply averaged as Equation 3.3.

$$S_F = \sum_{i=1}^{N} P_r(c = pos|s_i)/N \tag{3.3}$$

where s_i refers to one of the N self-replies from the thread originator. Similarly, the sentiment of responding replies is averaged as Equation 3.4.

$$S_R = \sum_{j=1}^{M} P_r(c = pos|r_j)/M \tag{3.4}$$

where r_j is one of the M responding replies in the thread. Then the *sentiment change indicator* for a thread originator is computed as $\Delta P_r = S_F - S_0$, where $S_0 = P_r(c = pos|p_0)$ is the sentiment of the thread's initial post.

Plotting ΔP_r against S_R (Figure 3.9) demonstrates that ΔP_r tends to have higher values as S_R increases. While the curve does demonstrate some non-linearity, the Pearson correlation coefficient between ΔP_r and S_R is 0.96 ($P-val \leq 0.0001$). This suggests that the more positive the sentiment of responding replies, the greater the positive change in originator sentiment.

While this finding only establishes association, it does lend support to the assumption that the sentiment of thread respondents can impact that of originators. The analysis also establishes that social support from respondents generally influences thread originators in a positive way. After at least one responding reply is received, about 75% of all the thread originators who started with negative sentiment expressed positive subsequent sentiment; among those who started with positive sentiment, 85% stayed positive in their subsequent sentiment [64]. Figure 3.10 shows an example where the thread originator's sentiment is positively influenced by a responding reply.

Having demonstrated that the sentiment of respondents has an impact on the change in sentiment of the thread initiator, this section returns to the issue of identifying IUs. Positing that influential users post greater numbers of influential responses, this study proposes to use the number of *influential responding replies (IRR)* as a metric of influence. An IRR is a responding reply that is able to affect the sentiment of thread originators. While all responding replies in a thread may alter the sentiment of originators' self-replies, only responding replies that are published before originators' first self-reply within the thread are considered. The

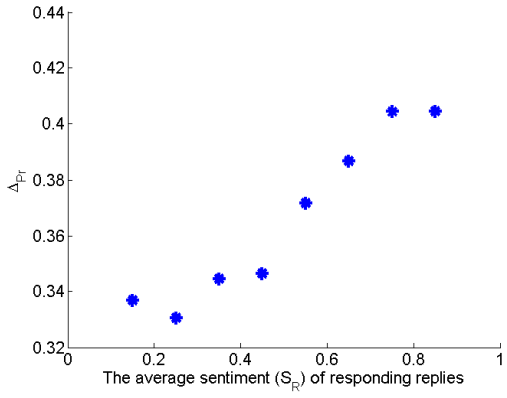


Figure 3.9: Change in originators' sentiment as a function of the average sentiment of responding replies.

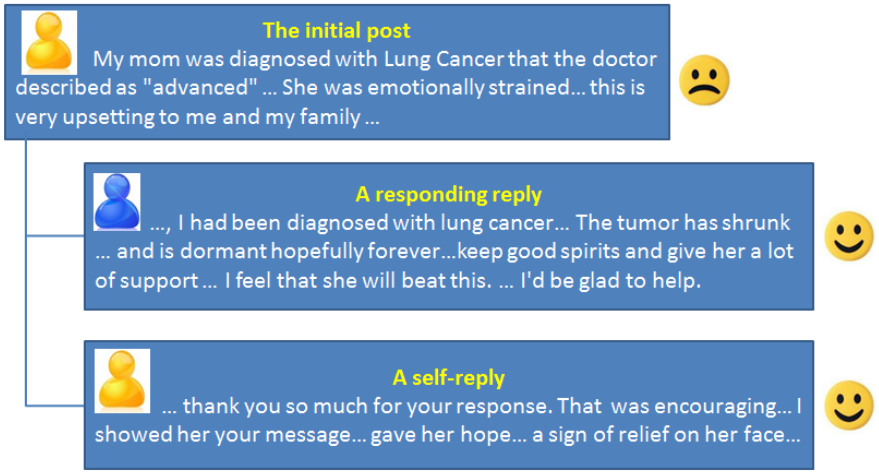


Figure 3.10: An example of how a responding reply influences the thread originator's sentiment (happy and sad faces illustrate the sentiment of a post).

rationale for this is twofold.

First, a thread originator's sentiment often changes significantly between the initial post and the first self-reply, but it only changes a little afterwards (See Figure 3.8). Hence, when responding replies to a thread are published before the thread originator posts again, they are more likely to be influential.

Second, the temporal intervals between initial posts and first self-replies have a median value of 17 hours, with two-thirds of them below 24 hours. The cumulative distributions of temporal intervals between initial posts and the first/last self-reply (Figure 3.11) shows that threads can sometimes last quite a long time

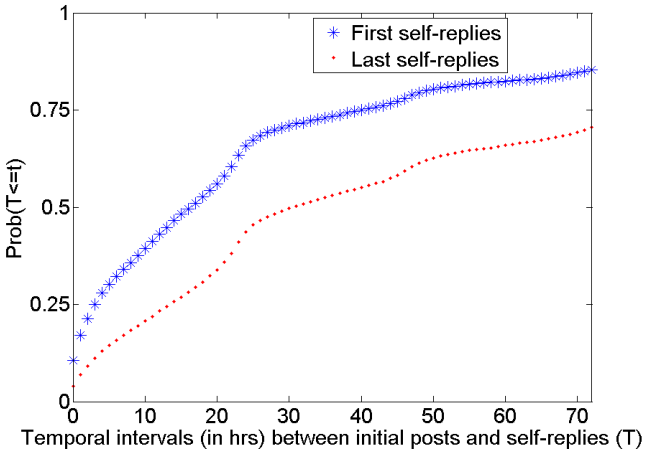


Figure 3.11: Cumulative distribution of the intervals between initial posts and their first/last self-replies.

with the originator often keeping the conversation going with multiple self-replies. Thus by focusing on those early responding replies, this approach can further increase the probability that an originator's sentiment change is actually due to the responding replies and not due to other events happening offline, such as changes in their physical conditions, personal support from friends and acquaintances, or approaching holidays.

Operationally, an IRR moves the thread originator's sentiment in the direction of the reply's sentiment (positive or negative). If an IRR r_j brings in positive sentiment to a thread (i.e., the sentiment classifier assigns $P_r(c = pos|r_j) > 0.5$ for post r_j), the originator's sentiment in the first self-reply (s_1) should become more positive compared to that in the initial post (p_0), i.e., $P_r(c = pos|s_1) > P_r(c = pos|p_0)$; if r_j contains negative sentiment (i.e., $P_r(c = pos|r_j) \leq< 0.5$), the originator's sentiment in the first self-reply should become more negative than it was in the initial post, i.e., $P_r(c = pos|s_1) < P_r(c = pos|p_0)$. If there are multiple IRRs between the initial post and the first self-reply, all are considered to be contributory to the originator's sentiment change. This is analogous to the aggregation of influence from multiple actors in the threshold model [98][99], even though thresholds that represent individual differences on the ease of being influenced are not explicit in the definition of IRR. Figure 3.12 shows three examples on how to identify IRRs from threads.

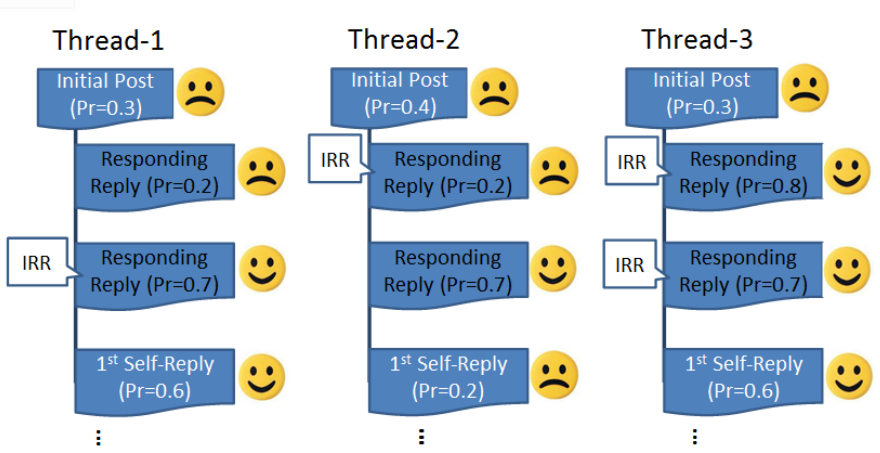


Figure 3.12: Examples of how to identify IRRs from a thread. Note that $P_r > 0.5$ means positive sentiment (denoted with happy faces); $P_r \leq 0.5$ indicates negative sentiment (denoted with sad faces).

Formally, a responding reply r_j in the thread started by initial post p_0 is an IRR if and only if the following two conditions are met:

Condition 1: $T(p_0) < T(r_j) < T(s_1)$, where $T(p)$ is the publishing time of post p. Condition 2: $P_r(c = pos|r_j) > 0.5$, if $P_r(c = pos|s_1) > P_r(c = pos|p_0)$ or $P_r(c = pos|r_j) \leq> 0.5$, if $P_r(c = pos|s_1) < P_r(c = pos|p_0)$.

Please note that the metric considers responding replies that bring both positive and negative sentiment to thread originators. There are at least two reasons to consider negative sentiment influence. First, in some contexts, it is not always appropriate to publish responding replies with positive sentiment. For example, if a thread reports the death of a community member, influential users may show their sympathy in the thread that may deepen the sadness of the originator (Threads where IRRs had negative sentiment impacts on thread originators are examined. A preliminary lexical search finds that 13% of the initial posts contain words or expressions related to death. See Appendix C for the list of words and expressions). Responding replies like this can also be supportive in this special context and do not necessarily have a negative impact on the community. Second, a small number of the influential users may feel so passionate about an opinion that they, while contributing to the emotion support of some members, could annoy those who do not agree with their opinions. Distinguishing different types of negative sentiment influence and finding IUs whose activities may have negative impacts on the whole

Table 3.6: Compare the Top-K recall from various single-metric user rankings (using IU List-1).

Ranking Metric	**K=50**	**K=100**	**K=150**
Total number of threads initiated	0.342	0.439	0.585
Total number of posts	0.415	0.707	0.781
In-degree in the post-reply network	0.317	0.512	0.610
Out-degree in the post-reply network	0.390	0.659	0.780
Betweenness in the post-reply network	0.293	0.366	0.488
PageRank in the post-reply network	0.390	0.561	0.732
Total number of IRRs	0.511	0.732	0.805

community is beyond the scope of this research.

In general, the number of a user's IRRs is a reflection of the individual's engagement in and contribution to the community, promptness in providing support, and more importantly, the level of influence that this user can exert on others. These are all important characteristics of influential users in the OHC.

One is able to use this metric to rank users according to their numbers of IRRs–the higher the number of IRRs is, the more likely they are influential users. Similar to the classification approach, Top-K recall is still used to evaluate the performance of this new metric and the 41 nominated IUs (i.e., IU List-1) are also used.

Table 3.6 lists the Top-K recalls of various user rankings by traditional metrics, including contribution metrics and network centralities in the post-reply network. Intuitively, higher K values lead to higher Top-K recalls. More importantly, the performance of the new IRR metric is better than that of all other metrics for top-50 recall, top-100 recall, and top-150 recall.

3.6 Comparison and integration of the two approaches

Both the classification approach and the IRR metric outperform some traditional metrics for IU identification. Then how does their the performance compare to each other? Will the single-metric based IRR ranking prevail? Or will the combined power of various traditional metrics from the classifier triumph?

Table 3.7: Compare the recalls and precisions of the IRR ranking and an ensemble classifier (using IU List-2).

	K=50		**K=100**		**K=150**	
	Recall	Prec.	Recall	Prec.	Recall	Prec.
IRR Ranking	0.349	0.880	0.627	0.790	0.762	0.640
The original ensemble classifier	0.278	0.700	0.532	0.670	0.698	0.587
The new ensemble classifier with IRR	0.373	0.940	0.579	0.730	1.000	0.840
Note: Because there are 126 IUs and $126 > 100 > 50$, the maximum possible values for Top-50 and Top-100 recalls are not 1. For instance, even though the top 50 users are all influential ones, the maximum possible Top-50 recall is $50/126 = 0.397$. Similarly, the maximum Top-100 recall is $100/126 = 0.794$; the maximum Top-150 precision is $126/150 = 0.84$.						

In the previous two sections, top-K recalls have been used to evaluate the performance of both approaches. To assess their precision, domain experts are asked to review a new list of users that are ranked within the top 150 by the classifier but were not included in IU List-1. Using criteria similar to those that generated IU List-1, domain experts endorse an additional 85 influential users, yielding a total of 126 IUs acknowledged by domain experts (referred to as *IU List-2*). It enables better comparison between the IRR ranking and the classifier.

Surprisingly, it turns out that the IRR metric is not only intuitive, but also really powerful. As Table 3.7 shows, the performance of the IRR ranking is better than that of the classifier in both Top-K recalls and precisions, even though the IRR ranking only uses one metric and the other classifier uses 60 features. In addition to recall, the high precision of the IRR metric is also very useful. For example, the Top-50 precision is 0.880, meaning that 88% of the top 50 users are indeed influential ones.

To further improve the identification of influential users and illustrate the synergistic benefit of the two approaches, the IRR metric is incorporated as a new feature into the original ensemble classifier. The individual classifiers and the ensemble classifier are re-trained. The IRR-enhanced classifier performs much better than the previous one (See the last row in Table 3.7). Its strong performance in Top-50 recall and precision is especially desirable for accurately finding commu-

nity members with very high influence in the OHC. In addition, the new classifier's perfect Top-150 recall means that it could find all of the nominated and endorsed influential users within top 150. The imperfect Top-150 precision is also acceptable because the 126 IUs in List-2 still may not include all influential users in the OHC.

3.7 Summary and future work

In summary, this research develops two approaches to identify influential users in OHCs and uses the large-scale data from a cancer survivor's online community as a case study.

The first approach is based on classification. The classifiers can only utilize a partial list of subjectively identified leaders from domain experts. In addition to traditional metrics for users' contributions, network centralities, and contents of their posts, the structure of the network among users is utilized to extract neighborhood-based and cluster-based features. Combining these metrics using individual classifiers and ensemble methods, this research achieves reasonable performance that is better than those obtained using traditional user ranking methods.

The second approach focuses on information flows among individuals and the subsequent sentimental effect of such flows. The proposed new metric–the number of influential responding replies–directly measures an OHC member's ability to influence others' sentiment. The new metric is more intuitive than the classification approach. More importantly, the metric outperforms not only many traditional metrics, but also the classifier that combines the power of 60 user features, in identifying IUs from the OHC.

The proposed IRR metric for influential users is significant not only because it has been shown to be effective for identifying IUs from a large community with a long history, but also because it provides fundamentally new insights into understanding the nature of social influence at multiple scales. The concept of "influential post" introduces, for the first time, a fundamental element of social influence at the inter-personal level. The concept is intrinsically multi-scale, because it is based on the alignment of a responding reply's sentiment (inter-personal level) with the direction of the sentiment change of the thread originator at the individual level. The concept of "influential post" also compliments the previous emphasis

on analyzing "relationship" networks with a fundamentally new perspective that analyzes the conversation of actual social interactions, in which information flows and influence take place. This new perspective is especially suitable for analyzing information flows and influence in online communities, where interactions emerge and evolve among people previously unconnected. This new vantage point will provide an important basis for advancing our understandings about influence, information flows, human behaviors, and the dynamics of online communities. For instance, longitudinal studies about influential posts can be useful for studying the dynamic patterns of user engagement in online communities.

Admittedly, our approach only examined inter-personal sentiment influence through online interactions in an OHC forum. This leads to two limitations. First, the sentiment in a user's post may also be influenced by offline issues. This has been a issue for many other studies of online social influence. This study has tried to eliminate as much offline influence as possible by focusing on the sentiment effect of prompt replies to thread originators. Second, only sentiment influence has been analyzed. Although providing emotional support is a major goal of many OHCs, other types of influence may exist, such as the influence on OHC members' decisions on medications and treatments. Such influence may also be conveyed through other interaction channels, such as offline communications and online chats/messages. To address these limitations and achieve a more comprehensive understanding of influence among OHC members, researchers need to capture and analyze members' activities both online and offline. This can then be combined with a more fine-grained text analysis of the content of their interactions.

Another limitation of this work is that it does not distinguish healthy negative sentiment influences (e.g., sadness due to the death of a community member) from those that are not healthy for the community (e.g., opinions that are so strong that can annoy certain community members). Such a distinction, which requires a more fine-grained analysis of the content of the threads, can contribute to the identification of IUs who exhibit disruptive behaviors and have negative impacts on the community. This is also an important area for future research.

In general, as the first large-scale computational study of sentiment influence in OHCs, this automatic identification of IUs in an OHC not only reduces labor-intensive monitoring over a long period of time to establish behavior patterns, but

also has implications for building an active, supportive, and sustainable OHC. Such an OHC can play important roles in providing personalized support, disseminating medical knowledge, encouraging healthy lifestyles, advocating new treatments, etc. Specifically, the practical contributions of this research are three-fold: *First*, the early and proactive identification of IUs provides community managers an opportunity to publicly recognize their contributions by awarding them prestigious status (e.g., presenting virtual badges of honor) and to encourage other members' participation in the OHC. The status as an IU may also be helpful when a user faces conflicting information in the forum or in resolving relationship conflicts. *Second*, the identification of IUs helps OHC managers to assure consistent and strong peer leadership in the community. This is especially valuable to OHCs, because influential users are sometimes lost as a result of health-related factors that limit or preclude their continued involvement in the community. Such loss of IUs may lead to a less supportive and more pessimistic community. When this happens, community managers can guide emerging community IUs to assume greater leadership roles and help the community recovery from the loss. *Last*, but not least, the application of the sentiment approach is not limited to OHCs only. It is also possible to adopt similar approaches to identify influential users and reveal sentiment dynamics in other online communities, such as those for political opinions and product reviews.

Chapter 4

Information Flows in Inter-organizational Networks–Connect Individual Interactions and Network Topologies

The studies of supply-chain networks and online social networks have illustrated that supply or information flows are interrelated with network topologies and individual interactions respectively. In this chapter, my research on networks among humanitarian organizations takes a step further and tries to connect micro-level interactions among individuals and macro-level network topologies using the flow of information. Taking a multi-relational perspective, the research tries to answer "How does the flow of information in one network affects the evolution and structure of another network among the same group of individuals?"

This research develops an agent-based model to simulate how a collaboration network among organizations emerges from organizations' interactions and the flow of information through another network-the inter-organizational communication network. The proposed approach also models the competitive yet non-exclusive dissemination of information among organizations, organizations' dynamic prioritization of candidate projects, and network-based influence. The model adds links (or edges) into the collaboration network on the basis of events, which correspond

to organizations' formation of collaborative teams for joint projects. Applying the model to a case study of the humanitarian sector, this research configures and validates the agent-based simulation, and use it to analyze how to promote inter-organizational humanitarian collaboration by encouraging communication. The simulation results suggest that encouraging communication between peripheral organizations can better promote collaboration than other strategies. The outcome of this research has been published in 3 conference papers [100][101][102] and 2 journal papers [103][104].

The chapter starts with an introduction, followed by the background of this research and how our approach differs from existing research. Then I will introduce an agent-based model for the emergence and evolution of inter-organizational collaboration network. The implementation and validation of the model is then presented, along with a study of how to promote humanitarian collaboration. The chapter concludes with discussions of future research directions.

4.1 Introduction

In recent years, inter-organizational collaboration has increased in both for-profit and non-for-profit domains [105]. Inter-organizational collaboration takes place when organizations share authority and responsibility for planning and implementing an action to solve a problem [106]. According to [107], inter-organizational collaboration occurs when different organizations work together to address problems through joint effort, resources, decision making and share ownership of the final product or service. Inter-organizational collaboration could benefit individual organizations in a community, clients of organizations in a community, and the community as a whole [107][108]. Research also identifies potential gains that nonprofit organizations could ripe from collaborating with others, including economic efficiencies, more effective response to collective problems, improvements in the quality of services, the spreading of risks, and increased access to resources [109]. As an important topic in organizational research, collaboration has drawn the attention of many researchers. A better understanding of inter-organizational collaboration may reveal ways to facilitate and improve collaboration activities.

The collaboration relationship among organizations can be denoted by an inter-

organizational collaboration network. In such a collaboration network, a node represents an organization and an edge that connects two nodes means that the two organizations are collaborators. At the same time, collaboration is only one of the many types of relationships that could exist among organizations, such as information sharing, business transaction, etc. Also, the existence of other types of relationships among organizations may affect or influence the formation of collaboration relationship and hence the inter-organizational collaboration network.

The goal of this research is to model the evolution of collaboration networks among organizations. Specifically, it models how the flow of information through organizations' interactions in a communication network lead to the event of team formation and the subsequent emergence of collaboration networks. An inter-organizational network in the humanitarian sector is used as a case study.

In the past few years, the world has suffered from several major natural disasters. Humanitarian efforts after these tragedies have highlighted the need for greater levels of inter-organizational collaboration among humanitarian agencies. Such collaboration can improve humanitarian response so that the response meets the needs of the affected population to the maximum extent possible [110]. In addition to better understanding of how inter-organizational a collaboration network emerges, I hope this study can also help to provide recommendations on how to promote humanitarian collaboration, which will eventually benefit disaster victims.

4.2 Background

To promote inter-organizational collaboration, one approach taken by humanitarian agencies has been to organize "coordination bodies," one of whose goals is to improve disaster relief efforts through collaboration among its member organizations. These coordination bodies may be temporary, special initiatives, or permanent incorporated non-profit organizations, and provide a non-hierarchical environment for organizations to interact with each other and form collaborative teams for joint projects. As participation in collaborative teams is undertaken on a purely voluntary basis, mutually beneficial joint projects and corresponding teams "emerge" from the collective behaviors of individual organizations in this non-hierarchical setting.

In interviews and surveys conducted by my colleagues and I, many humanitarian organizations acknowledged that communication plays a very important role in the forming of collaboration relationship [111]. First, communication often antecedes collaboration and serves as the basis for establishing future collaboration relationship. This is because communications enable the flow of information among organizations. In other words, organizations need to communicate with acquaintance to get information about different joint project initiatives, so that they can identify interesting projects and collaborate on them. Second, an organization's decision on whether to collaborate with others on a joint project is mainly based on its own evaluation of the project. However, through communication, organizations are often able to exert various levels of influence on others' decisions on collaboration. In other words, the collaboration network emerges from individual organizations' interaction through a communication network.

Besides, my previous analysis of the inter-organizational communication and collaboration networks also revealed the connection between communication and collaboration [101]. In our analysis of assortative patterns, which describes the tendency of nodes in a network being connected with nodes with similar degrees, inter-organizational communication network and collaboration networks have similar dis-assortative patterns. In other words, in both networks, high-degree nodes tend to connect to low-degree nodes. In fact, topologies of the two networks are positively correlated.

Therefore, to model the collaboration network among organizations, one has to adopt a multi-relational perspective and incorporate the impact of the communication network. From the perspective of network modeling, the research needs to simulate how organizations create links (or edges) in the collaboration network. Meanwhile, the decisions to create links are facilitated and affected by the flow of information and influence, which are transmitted in another network–the communication network. Hence our work is related to existing literatures in two areas (1) link formation in networks, and (2) the dissemination of information and influence. I will briefly review the two areas in this section and discuss how our model is different from previous approaches. Before that I first introduce what a multi-relational perspective means.

4.2.1 The multi-relational perspective

Connecting a group of nodes with edges (or links), networks are able to represent relationships among entities, such as people, organizations, locations and so on. If one looks at a network from an aggregated perspective, an edge connecting two nodes in a network means that the two nodes are somehow related. However, scrutinizing how connected nodes are related to each other in the network, one will find the heterogeneous nature of these edges. For example, in a social network, the most general relationships is "know," i.e. a person knows another person if they have any relationship. A social network can then be specialized by categorizing the relationship as one of business, family, friend, and so on, and each of these can be further sub-classified. For example, Tom and Jack may be brothers with a link between them representing a family relationship; while Jack and David may be graduates of the same college thus their link reflects an alumni relationship. Similarly, edges between two organizations in an inter-organizational network may denote different types of relationships, such as collaboration, information sharing, trading, etc.

In other words, for many real-world networks, edges in the same network could mean different types of relationships, each of which spans a network of its own. Thus a network is multi-relational with multiple relationships and heterogeneous edges. The heterogeneity of edges can be further illustrated by representing such a multi-relational network with multiple uni-relational networks. Each type of relationship can be represented by its own uni-relational network. Given a network $G(V, E)$ with node set V and edge set E (note that as multiple types of relationships could exist between two individuals, more than one edge between two nodes in G is allowed), N types of relationships are represented by edges $e_i \in E$ in the network. Then one can divide all edges in set E into N disjoint sub-sets, which satisfy $E_1, E_2, ..., E_N \subset E$, $\forall i, j \in [1, N](i \neq j) : E_i \cap E_j = \Phi$, and $E_1 \cup E_2 \cup ... \cup E_N = E$. Then $G(V, E)$ can be divided into N sub-networks that share the same set of nodes but have different sets of edges: $G_1(V, E_1), G_2(V, E_2), ..., G_N(V, E_N)$. Each of the sub-network is a uni-relational network that represents only one type of relationship in the aggregated multi-relational network $G(V, E)$.

Many studies on social and organizational networks took a uni-relational perspective. On one hand, a lot of research focused on networks based on a specific

type of relationship. Examples include the email communication among employees within an organization [112]; the collaboration among organizations [113]; the overlapping of board of directors among large corporations [114], and so on. On the other hand, many studies did not make distinctions between different types of relationships. Instead, they often took a coarse-grained approach and aggregated multiple types of relationships into one network with homogeneous edges. For instance, in studies of online social networks, a link between two users in a social networking website, such as LinkedIn or Twitter, was often only considered to show that the two are somehow related but it was often disregarded whether the link is one of family, co-worker, classmate, etc [115][116].

While uni-relational approaches are simple and intuitive, they inevitably lose valuable information. First, one uni-relational network may affect or influence another, as catalysts or constraints. For example, the classmate or roommate network among college students may affect their email communication network; the collaboration network among organizations may depend on the communication network; the transportation network may constrain a retailer's distribution network. Second, an individual may exhibit multi-faceted behaviors and possess different structural positions in different uni-relational networks. For instance, in an online social network like Facebook, a user may have many online "friends" (the friendship network) but his/her status or shared links seldom draw comments from others (a comment network); in an inter-organizational network, one could be a hub in the trading network but at the peripheral of the collaboration network.

By contrast, a multi-relational perspective can shed new lights on the study of networks and help to understand real-world networks in a more systematic way. For example, the structural difference between the sibling network and the farm work assistance network among villagers helped to explain several sociological phenomena [117]. Analyzing various types of networks among online gamers has validated the social balance theory in a large scale [118]. [119] inferred the trust network among online gamers from their mentoring network. The incorporation of multiple networks that social media users are involved in also helps to predict users' collective behaviors [120]. Nevertheless, previous multi-relational analysis focused on relationship between different networks' static topologies.

This research adopts the multi-relational network perspective and study the

evolution of an inter-organizational collaboration network. Thus our approach needs to emphasize on how organizations' interaction and the information flow through the communication network affects the dynamic growth, i.e., the formation of links, of the collaboration network.

4.2.2 Formation of links in networks

Representing relationship among a group of organizations, inter-organizational networks have drawn the attentions of many scholars in organization and management science [121][15]. Specifically, many studied what drive organizations to form ties and identified important factors such as influence from other organizations, communication, history of relationship, reputation, and level of inter-dependency [122][123][124][125]. However, these works stayed at the organizational level and relied mainly on organizational characteristics.

At the network level, the link prediction problem has attracted researchers from different disciplines. On one hand, many researchers studied various statistical attachment rules that guide a node's connections to another node. These models are usually based on connection heuristics and calculate the probability of connections between two nodes using the two nodes' topological attributes. Examples include the preferential attachment model[10], Jaccard index [126], Adamic/Adar index [127], and Katz index [128]. Limited research used this approach to study inter-organizational networks [113][129]. On the other hand, some research used classification or maximum likelihood approaches to predict links. The basic idea is to consider connected node pairs as positive samples, and other node pairs as negative samples. Then a classifier is trained [130][131][132] or parameters that maximize the likelihood of data is found [133] using a set of positive and negative samples.

However, most of the aforementioned approaches are rooted in building dyadic or binary relationship between two nodes. As a result, they may not capture two important aspects in humanitarian inter-organizational collaboration.

First, most existing approaches are limited to link formation within the one network. In other words, they take a uni-relational approach and do not consider the interaction between different types of relationships. Thus they are unable to

capture how one network affects the formation of links in another, with the recent exception of [119]. As mentioned before, inter-organizational collaboration relies heavily on their communication relationship. Thus our model has to incorporate the multi-relational perspective.

Second, for humanitarian agencies, collaboration is often project-based activities. The joint project they collaborated on could be joint training of staff members, coordinated data collection, shared database, and so forth. A joint project may start by only one or two organizations but could have more than two collaborators. According to our survey, more than 80% of the collaborative projects had more than 3 collaborating organizations [111]. The Information and Communication Technology (ICT) Skills Building Program of the ReliefTechNet[1], a major coordination body, is an example of such projects. The goal of this project was to provide training on latest ICTs to staff members of humanitarian agencies, so that their response to emergency and enhance their organizational effectiveness can be improved. This project was initially proposed by one organization in the coordination body ReliefTechNet, but was then developed with inputs and contributions from a team of more than ten different member organizations.

From a network perspective, the formation of inter-organizational collaboration network is based on events of collaborations. An event here is the formation of a collaborative team for a joint project. This type of event-based collaboration relationship can be described as a bipartite graph with two types of nodes: projects and organizations. Organizations are connected to a project if they form a team to collaborate on the project. Figure 4.1 shows an example bipartite graph with 2 joint projects and 8 organizations: 4 organizations collaborate on Project 1; Project 2 has 5 collaborators. The inter-organizational collaboration network can be constructed by converting the bipartite collaboration graph into a network that directly represents collaboration relationship among organizations. In other words, organizations in the a team for the same joint project will be connected to each other (as shown in Figure 4.2).

The important implication of the event-based collaboration is that collaboration relationship is *n-ary*, instead of binary. In other words, several edges in the

[1]In this dissertation, pseudonyms of organizations are used to protect the confidentiality of humanitarian organizations.

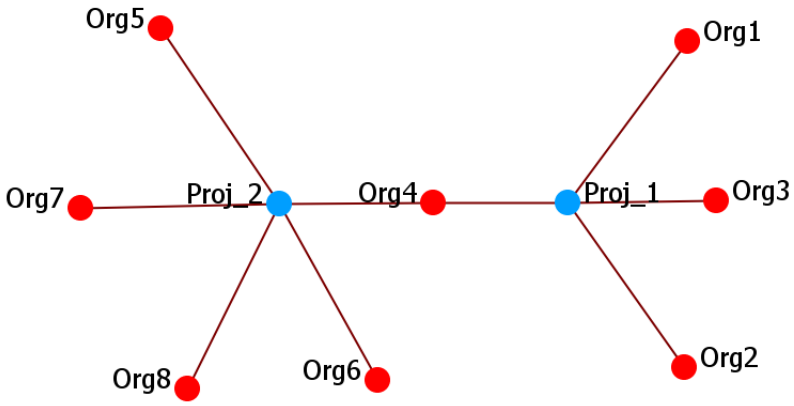


Figure 4.1: An example bipartite graph for collaboration.

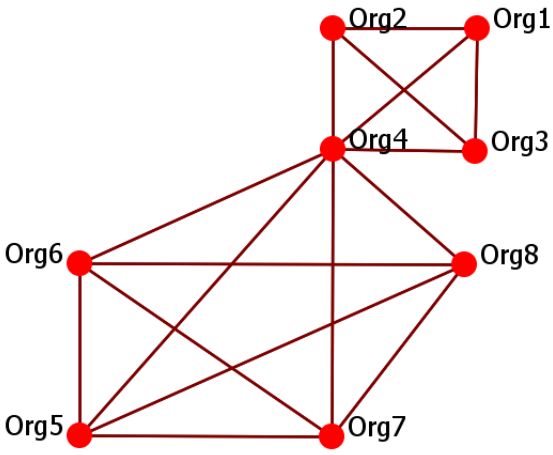


Figure 4.2: An example inter-organizational collaboration network.

corresponding inter-organizational network may co-occur at the same time and are not independent from each other. For instance, if organizations in Figure 4.1 fail to form a team for Project 1, there might be no collaboration relationship among the participation organizations of Project 1. In other words, organizations 1, 2, 3, and 4 may not be connected in the inter-organizational collaboration network. Current statistical approaches focus on binary or dyadic relationship, but cannot capture this type of concurrent formation of multiple links and the dependency between links.

4.2.3 Dissemination in networks

A network often plays a key role for epidemic or social contagions to disseminate across multiple individuals. Epidemic contagion concerns the spread of infectious diseases, such as flu. Examples include the Susceptible-Infected-Recovered (SIR) model [134] and its variants. Social contagion refers to the diffusion of informaiton, products, fashion, behaviors, and so on in the human society.

Most network dissemination models consider the disseminative flow as a process based influence. The basic idea is that dissemination happens between network neighbors as one node is able to influence its neighboring node to change status or adopt new behaviors. In terms of how to model such influence among neighboring nodes, there are two types of approaches. One approach is based on independent cascade between neighboring nodes [135]. When a node changes its state, it has one chance to "infect" its neighboring nodes in other states with a probability. Another type of models is based on the idea of threshold [98]. In such models, a node will change its state or adopt a new behavior when a certain fraction or number of other nodes have changed or adopted. In this case, the dissemination does not have to happen between neighboring nodes. The threshold of each individual may also vary.

However, existing approaches cannot be directly used in our model for two reasons.

First, few dissemination models consider individuals' endogenous factors, which are especially important in social contagion. While exogenous influence from one's peers in a network can often affect whether she/he is "infected" by a behavior, information or product, such a decision is also based on one's individual characteristics or her/his independent evaluation of the "infectant". In fact, recent research claimed that the impact of peer influence is overestimated [136]. Instead, homophily [137] plays an important role in the dissemination of the service. Homophily, which represents the similarity among individuals, reflects the endogenous factors that may affect dissemination. Similarly, in the context of inter-organizational collaboration, our survey revealed that when organizations evaluates a candidate joint project, they consider both endogenous and exogenous factors, such as whether the goal of the project aligns with the organizational mission, whether similar projects have been on the organizational agenda, which and

how many organizations have decided to collaborate on the project, etc. Therefore, I need a model that incorporate both endogenous and exogenous factors in the dissemination process.

Second, most studies only considered the dissemination of a single "infectant" or "contagion" but paid little attention to the competitive yet non-exclusive dissemination of multiple "infectants" or "contagions". When organizations collaborate, multiple projects may be proposed as candidate joint projects. However, an organization cannot work on all candidate projects, because it has limited resources. Put it another way, various candidate joint projects, whose information is disseminated through the communication network, are competing for organizations' resources. The interplay between these candidate projects will eventually affect the outcome of collaboration. A recent study [138] proposed a model for competitive diffusion of political standings in networks. Also, political standings are mutually exclusive of each other: one cannot support both Democrats and Republicans at then same time. However, this type of exclusiveness does not hold in the dissemination of projects information, as an organization can work on multiple joint projects if resources permit. Therefore, our model needs to capture both the competition and the non-exclusiveness of multiple candidate projects in the dissemination process.

4.3 Proposed approach

Based on existing literatures and the needs of the research, agent-based models (and simulations) is chosen for this study. Computational models and simulations, especially agent-based ones, have been widely used to study a variety of social, organizational, and natural phenomenon [139]. Agent-based models are capable of simulating macro-level structures or patterns resulting from micro-level interactions and decisions of heterogeneous agents within complex systems [140]. An agent-based simulation is especially helpful for decision-makers and policy-makers in organizations, because it is often very difficult to manipulate organizations to evaluate the impact of a policy or a decision. A computational simulation for inter-organizational collaboration network not only enables one to study the outcome of different policies, but also helps to gain insights into the patterns and characteristics of inter-organizational collaboration at both micro and macro levels.

This section will propose an agent-based model for information flows among organizations and the evolution of inter-organizational collaboration network. The model adopts the multi-relational perspective and simulates how agents' interactions and the flow of information through one uni-relational network (the communication network) lead to the emergence of another uni-relational network (the collaboration network). Representing organizations as agents and embedding them in a communication network, the model incorporates *the disseminative flow of project information* through the communication network, *organizations' decision-making* on whether to work on a candidate project in the context of network influence, and how the inter-organizational collaboration network emerges through *the event of forming multi-agent teams for joint projects.*

In the agent-based model, organizations are represented as agents and are embedded in a communication network $G_c(V, E_c)$. $V = \{a_1, a_2, \ldots, a_N\}$ is a set of N agents. The edge $e_{i,j} \in E_c$ denotes that agents a_i and a_j communicate with each other through the network. The weight of an edge $w_{i,j}$ denotes the strength of ties between the two agents. $P = \{P_1, P_2, \ldots, P_M\}$ are a set of M candidate projects. Each agent a_i maintains a prioritized to-do list of projects: $L_i \subseteq P$. A project in agent a_i's to-do list $P_k \in L_i$ is the project that agent a_i would like to work on or collaborate with others. Agent a_i is called a supporter of project P_k if $P_k \in L_i$. Agent a_i also assigns project $P_k \in L_i$ a priority score $C_{i,k}$ that reflects how important this project is to the agent. Projects with higher priority scores are ranked higher in the list. The maximum number of projects S_i in a to-do list may vary from agent to agent. The limited size of to-do lists reflects the resource constraints, which are important to modeling the competitiveness and non-exclusiveness of different candidate projects.

After each agent initializes its to-do list, agent interaction starts as an iterative process with 3 steps in each iteration. Figure 4.3 shows the pseudo-code of the model.

The first step is the dissemination of information about candidate projects through the communication network. To find collaborators, an agent first needs to disseminate information about projects in its to-do list, so that other organizations are aware of these candidate projects. An agent a_i spreads project information by proposing the top-ranked project P_t in its list to its neighbors in the communication

While (the simulation tick $t < T$) {
*/**Inter-agent communication and dissemination of project information**/*
For each agent $a_i \in V$\{
Initialize a to-do list L_i, where $|L_i| = S_i$;
Calculate priority scores for $P_k \in L_i$;
Identify $P_t \in L_i$, so that $\forall P_k \in L_i, C_{i,t} \geq C_{i,k}$;
Send information of top project P_t to all a_i's neighbors B_i in $G_c(V, E_c)$;
}
*/**Inter-agent influence and projects prioritization**/*
For each agent $a_i \in V$\{
For every $P_k \in L_i$, update the priority score using Equation 4.2;
For every newly received project $P_k \notin L_i$, assign a priority score using Equation 4.1;
Adjust L_i and only keep S_i projects with top priority scores;
}
*/**Possible occurrence of team formation events**/*
Check the supporters $SP_k \subseteq V$ for $P_k \in P$;
If $|SP_k| \geq TH_k$\{ */**The team-formation event occurs**/*
All agents $Ai \in SP_k$ form a team for project P_k ;
$\forall a_i, a_j \in SP_k$, add edges $e'_{i,j}$ to $G_b(V, E_b)$;
}
}

Figure 4.3: Pseudo code for the agent-based model.

network. a_i's neighbors in $G_c(V, E_c)$ are defined as $B_i \subseteq V$, so that $\forall a_j \in B_i, \exists e_{i,j} \in E_c$.

The second step is the evaluation of candidate projects. Upon receiving a new candidate project $P_k \notin L_i$ proposed by its neighbors, agent a_i will evaluate the project using various criteria of its own and assign the project an initial priority score. As shown in Equation 4.1, the initial and independent evaluation score $C'_{i,k}$ is determined by function *Score*, which is based on the characteristics of the project and agent a_i's evaluation criteria R_i. Function *Score* can be configured by the modeler to cater different scenarios.

$$C'_{i,k} = \text{Score}(P_k, R_i) \tag{4.1}$$

Moreover, other agents also exert various levels of influence on an agent through

the communication network. The priority score that an agent assigns to a project is influenced by other agents' evaluation of the same project. This research uses a network influence model, which extends the social influence model in [141], to handle the exogenous influence from the network. Similar to most influence models, this model also assumes that an agent knows priority scores that other agents assign to candidate projects. Although this assumption may not hold in all situations for collaboration, it is a reasonable one in a collaborative environment, such as a coordination body. As coordination bodies would like to foster communication and collaboration among their member organizations, they often host regular meetings and group discussions, which provide an open venue for member organizations to exchange information openly and learn about projects others would like to work on. Equation 4.2 describes how an agent's project evaluation, which is reflected by the priority score assigned to the project, is iteratively influenced by other agents' evaluations of the same project. The right hand side of the equation consists of two parts.

$$C_{i,k}(t) = E_i \times F_i \times SC_k(t-1) + (1 - E_i) \times C'_{i,k} \tag{4.2}$$

where $C_{i,k}(t)$ is the priority score of candidate project P_k assigned by agent a_i at time t ($t > 0$). Please note that before agent a_i receives project P_k, the agent does not assign any score to the project. In other words, $C_{i,k}(t) = 0$ for $t < t^r_{i,j}$, where $t^r_{i,j}$ is the time stamp when a_i receives the P_k for the first time.

The first part describes the exogenous influence. F_i is a $1 \times N$ vector that represents influences on agent a_i from all the N agents in the communication network, including agent a_i itself. Elements in F_i are called influence indexes, with $F_i[j]$ representing the influence index of agent a_j over agent a_i. The sum of all influence indexes in F_i is 1, as shown in Equation 4.3.

$$\sum_{j=1}^{N} F_i[j] = 1 \tag{4.3}$$

$SC_k(t)$ is an $N \times 1$ vector that stores project P_k's priority scores assigned by all the N agents at time t. Namely, $SC_k(t) = [C_{1,k}(t), C_{2,k}(t), ..., C_{i,k}(t), ..., C_{N,k}(t)]^T$. Thus the product of F_i and $SC_k(t-1)$ is a score that reflects agent a_i's combined

consideration both its independent evaluation and exogenous influence from all other agents' evaluations of the same project P_k at time $t-1$.

The second part is agent a_i's initial and independent evaluation of project P_k. The initial evaluation score is kept because it is made independently by the agent without exogenous influence. This will generally serve as the basis for possible deviations of priority scores even though network influence exists.

The two parts are connected and balanced with the influence coefficient $E_i (0 \leq E_i \leq 1)$, which denotes how likely agent a_i's project evaluation is influenced by others'. A higher influence coefficient means an organization is more subject to exogenous influence, while agents with lower influence coefficient are more independent when evaluating projects and making decisions on collaboration.

The last step in each iteration is the adjustment of to-do lists. With priority scores for candidate projects from the previous step, an agent may add new projects with higher priority scores to its to-do list, re-evaluate and re-rank existing projects, or remove projects with lower priority scores from the list, as the list is limited in size. In the next round of the iterative process, an agent will again advocate for its top-ranked project and disseminate information about this project to its neighbors, even though this agent receives information about the project from another agent. This type of advocation further facilitates the flow of project information through the communication network.

As mentioned before, this study needs to model the competitive and non-exclusive dissemination of project information. The design of to-do lists and the adjustment of project prioritization models such dissemination, as projects compete for positions and rankings in a to-do list that can accommodates multiple projects.

As you can see, agents' interactions and the flow of information are mainly through the communication network. Then how will such interactions through the communication network affect the collaboration network? After the three steps in each iteration, the model checks whether the event of team formation occurs. With the help of inter-agent interactions through the communication network $G_c(V, E_c)$, a candidate project may disseminate to many agents and appear on the to-do lists of some agents. As each agent's its to-do list is openly available to all the other agents, when the number of a candidate project P_k's support reaches a project-

specific threshold TH_k, supporters of this project will form a team to work on the project together. Such a team-formation event achieved through the communication network also leads to the establishment of collaboration relationship among agents in the same team. Consequently, edges are added to connect all the team members to each other in the collaboration network $G_b(V, E_b)$, where $e'_{i,j} \in E_b$ denotes the collaboration relationship between Agents a_i and a_j. Take one of the teams in Figure 4.1 for example. Organizations (agents) 1, 2, 3, and 4 form a team for Project 1. Thus 6 edges connect the four team members in the collaboration network in Figure 4.2.

4.4 A case study of the humanitarian sector

In this section, the proposed agent-based model is used to simulate the emergence of collaboration networks among organizations in the humanitarian sector. The simulation is implemented with the Repast toolkit [142], a Java library for agent-based simulations. Then I configure and validate the simulation with empirical data. An experiment is also conducted to evaluate the effectiveness of different strategies that aim at facilitating collaborations.

4.4.1 Configuration of the simulation

In order to implement a trustworthy simulation, one needs to configure our simulation properly. The simulation is first configured using empirical data of humanitarian organizations' demographics and project preference. Then I validate the configured simulation by comparing the simulated inter-organizational collaboration network with the real-world one. If the two networks are not similar, the configuration is tweaker and the simulation is run again, till the simulation gets a configuration that leads to satisfactory results. Figure 4.4 illustrates the configuration process.

The configuration has to be backed by empirical data. Two surveys and numerous interviews were conducted among member organizations of GlobalSympNet, a major coordination body with 119 member organizations. My colleagues and I collected these organizations' demographic data, including missions, focus regions,

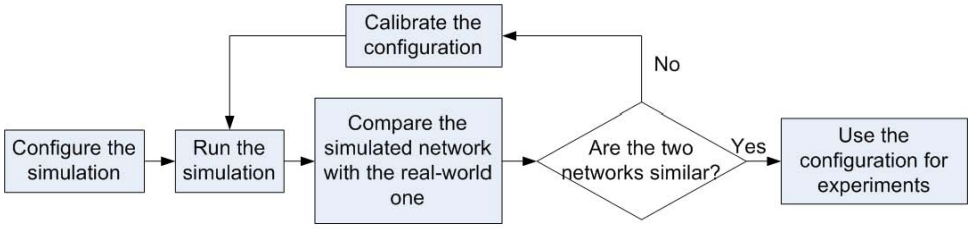


Figure 4.4: The process of simulation configuration and calibration.

numbers of full-time employees. Table 4.1 lists the 9 major missions and 7 focus regions of member organizations in GlobalSympNet. Note that an organization may have multiple missions and more than one focus region. In addition, data about 30 collaborative projects that humanitarian organizations worked on was gathered. The data includes where the project was implemented, the goal of the project, and how many organizations got involved, etc. The data of the 30 projects was augmented to generate 150 synthesized candidates projects.

The surveys and interviews also help to understand how humanitarian organizations evaluate candidate collaborative projects and how their decisions are influenced by others. It is worth noting that, although collaborative projects are expected to be implemented immediately after disasters, most of the projects in the dataset were pre-disaster projects, whose goals were to improve humanitarian organizations' capabilities in disaster response and relief. Therefore, time pressure, which is very important in forming team for poster-disaster projects, does not seem to be a key issue when organizations evaluate pre-disaster projects.

Further, this research needs to find different types of relationships among organizations. Two uni-relational networks among organizations are identified: communication and collaboration. Note that, the communication network is based on the inter-organizational relationship of advice seeking and giving. Research has shown that such advice exchange behavior often plays a major role in the exchange of information among humanitarian organizations [143]. More importantly, the advice exchanged among humanitarian organizations through the network is mostly about humanitarian projects, this network can be considered as a more focused communication network with stronger ties. In the collaboration network, two organizations are connected by an edge if they used to collaborate on humanitarian

Table 4.1: List of missions and focus regions for organizations in GlobalSympNet

Mission	**Focus Region**
1. Provide food	1. Sub-Saharan Africa
2. Provide shelter	2. Middle East & North Africa
3. Provide water	3. Europe & Central Asia
4. Provide sanitation	4. South Asia
5. Provide medical care	5. South East Asia
6. Provide funding	6. North America
7. Provide information services	7. Latin America & Caribbean
8. Provide training and advice	
9. Provide IT infrastructure and/or applications	

information management projects.

To find a proper configuration for the simulation, the collaboration network among 30 member organizations of the GlobalSympNet is simulated. The simulation takes as inputs the 30 organizations' demographic data, synthesized data of candidate projects, and the 30 organizations' communication network, which was identified in May 2008 (see Figure 4.5). I run the simulation for 40 ticks, allowing 8 rounds of inter-agent interactions. After the simulation stops, I get a collaboration network among the 30 organizations.

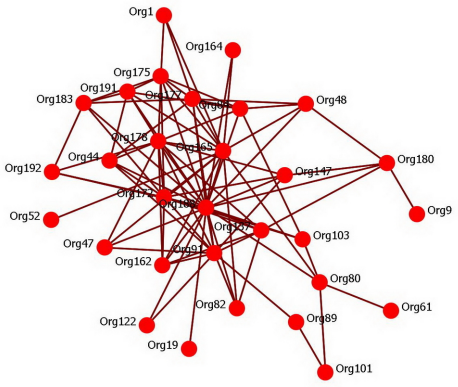


Figure 4.5: The communication network among 30 member organizations of GlobalSympNet as of May 2008.

Then the simulated collaboration network is compared with the actual collaboration network, which was gathered in a follow-up survey in October 2009 (see

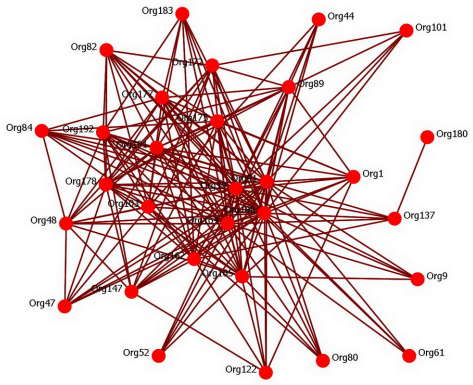


Figure 4.6: The collaboration networks among 30 organizations in GlobalSympNet as of October 2009.

Figure 4.6). To compare the two networks, the research evaluates how close the simulated one is to the actual one using several metrics, including the number of total edges, the clustering coefficients, the average path length, and the accuracy of link prediction. On the basis of evaluation outcomes, I adjust and calibrate the configuration of the simulation and re-run the simulation till a satisfactory configuration is found.

Here one such configuration is illustrated. The survey results suggest that when evaluating a candidate project, an organization considers two key factors: whether the purpose of the project matches the mission of the organization, and whether the beneficial (geographical) area of the project is within the organization's focus regions. Thus agent a_i calculates the initial priority score $C'_{i,k}$ for project P_k using Equation 4.4. $\mathcal{M}_i(m)$ specifies whether organization a_i has mission m. If organization a_i has mission m, $\mathcal{M}_i(m) = 1$, otherwise, $\mathcal{M}_i(m) = 0$. Similarly, $\mathcal{M}_k(m)$ denotes whether m matches one of project P_k's purposes. Likewise, $\mathcal{R}_i(r)$ and $\mathcal{R}_k(r)$ refer to whether region r is within organization a_i's focus region and whether project P_k will be implemented in or bring benefit to region r, respectively. Equation 4.4 calculates the number of overlapping missions/purposes and the number of overlapping regions between organization a_i and project P_k. Coefficients α_m and α_r denotes the relative importance of the two numbers. On the basis of organizations' average rankings of the two factors, I choose $\alpha_m = 5$ and $\alpha_r = 10$. Also, for the purpose of simplicity, the two coefficients are applied to all the agents. The initial priority score $C'_{i,k}$ that agent a_i assigns to project P_k

is the weighted sum of the numbers of mission and area matches. Intuitively, the better the project's purposes and beneficiary regions match the organization's missions and focus regions, the higher the score is. This initial score also reflects homophily-based project selection, because organizations with similar missions and regions may prefer similar candidate projects.

$$C'_{i,k} = \alpha_m \sum_{m=1}^{9} \mathcal{H}_{match}(\mathcal{M}_i(m), \mathcal{M}_k(m)) + \alpha_r \sum_{r=1}^{7} \mathcal{H}_{match}(\mathcal{R}_i(r), \mathcal{R}_k(r)), \quad (4.4)$$

$$\text{where } \mathcal{H}_{match}(x, y) = \begin{cases} 1 & \text{if } x = 1 \text{ and } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

In terms of exogenous influence, the survey results indicate that organization a_i is more likely to be influenced by organization a_j in the following scenarios: (1) the two organizations directly communicate with each other; (2) a_j is widely considered a leader in the community; and (3) a_j is a larger organization (and thus generally has more resources). Therefore, in Equation 4.5, the configuration incorporates these factors into the calculation of $F_i[j]$, organization a_j's influence index on organization a_i.

$$F_i[j] = \frac{f_i[j]}{\sum_{k=1}^{N} f_i[k]}, \text{ where } f_i[j] = \mathcal{D}(i, j)\mathcal{L}(j)\mathcal{S}(i, j) \quad (4.5)$$

According to Equation 4.5, $F_i[j]$ is essentially normalized $f_i[j]$, which is a product of multiple parts. Function $\mathcal{D}(i, j)$ is based on the geodesic distance $dist(i, j)$, i.e., the number of hops, between organizations a_i and a_j in the communication network. Intuitively, the longer the distance between a_i and a_j, the smaller the value of $\mathcal{D}(i, j)$. This configuration represents $\mathcal{D}(i, j)$ with an exponential function, as shown in Equation 4.6, so that the influence decades very fast as the distance increases.

$$\mathcal{D}(i, j) = e^{-[dist(i,j)-1]} \quad (4.6)$$

Function $\mathcal{L}(j)$ concerns whether organization a_j is considered a leader in this community. A leader organization a_j will have $\mathcal{L}(j) = 1.2$, while $\mathcal{L}(j) = 1$ for non-leaders. Function $\mathcal{S}(i, j)$ reflects how sizes of the two organizations affect

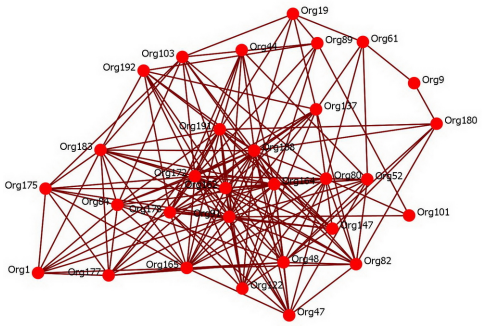


Figure 4.7: The predicted collaboration networks among 30 organizations in GlobalSympNet.

Table 4.2: Statistics of the simulated and actual collaboration network

	Simulated Network	**Actual Network**
Number of Edges	183.6 (178.9-188.3)	186
Clustering Coefficient	0.69 (0.68-0.71)	0.73
Average Path Length	1.64 (1.63-1.65)	1.63
Note: 95% CIs are reported in parentheses.		

the influence power. On the basis of the number of full-time employees in each organization, organizations are categorized into micro, small, medium, large and very large organizations. $\mathcal{S}(i,j)$, defined in Equation 4.7, should yield a higher value if a_j is larger than a_i, and a lower value if a_i is larger than a_j. In this configuration, $\phi = 0.3$ is picked.

$$\mathcal{S}(i,j) = [\frac{size(a_j)}{size(a_i)}]^{\phi} \tag{4.7}$$

The influence coefficient E_i, which connects the independent evaluation and the external influence in Equation 4.2, is based on each organization's response in a survey question about how likely their decisions will be influenced by those of other organizations.

With the above configuration, I run the simulation 30 times and get 30 simulated inter-organizational collaboration networks. Figure 4.7 shows one of them. Table 4.2 lists the basic statistics of simulated collaboration networks, along with those of the actual one. Statistics of simulated networks are the average of results from 30 different runs. The simulated results are very close to the statistics of the

actual collaboration network. In terms of link predication accuracy, simulations with this configuration gets an average accuracy rate of around 70%, with an average sensitivity of 64% and an average specificity of 75%. This means that the simulated network can predict whether two specific nodes are connected or not with a success rate of 70%. This is much higher than random prediction, which has an accuracy rate of only 21%.

Overall, taking the communication network as one of the inputs, the properly configured simulation is able to generate collaboration networks that are very similar with the actual collaboration network in the number of edges, average path length, and clustering coefficient. Although the simulation does not excel at link prediction accuracy, the main goal of this simulation is not to predict whether two specific organizations are connected or not in the collaboration network either. In fact, the validity of the configured simulation in basic statistics of the simulated collaboration network paves the way for our experiment in the following subsection, because the number of edges in the collaboration network is used as a key metric to evaluate the effectiveness of different strategies.

4.4.2 An experiment on how to facilitate collaboration among organizations

The GlobalSympNet is very interested in finding strategies that could facilitate or promote inter-organizational collaboration among its member organizations. However, as the GlobalSympNet is a coordination body without formal hierarchy and does not participate in any humanitarian project, it may not directly get involved in the process of identifying collaborative projects and forming teams. From a network perspective, it cannot directly help to build connections in the inter-organizational collaboration network. By contrast, as many organizations indicate that communication is important for and often serves as the prerequisite for collaboration, the GlobalSympNet could focus its effort on another uni-relational network, the communication network among organizations. In other words, it could try to facilitate collaboration by promoting communication and the information flow among its member organizations.

The analysis of the inter-organizational communication network (Figure 4.8)

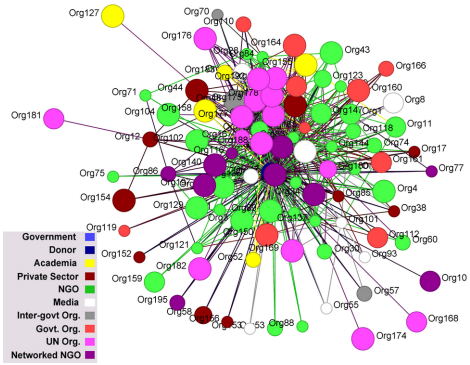


Figure 4.8: The communication network among the 95 humanitarian organizations as of October 2009. Node colors denote organization types.

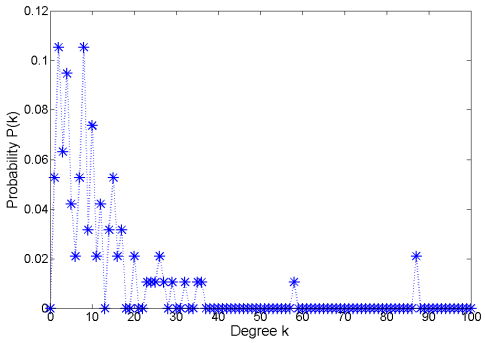


Figure 4.9: Degree distribution of the communication network among the 95 humanitarian organizations.

inside the GlobalSympNet reveals that organizations in the community are polarized in their network positions. The network has 95 nodes and 574 edges. As the degree distribution of the communication network in Figure 4.9 shows, there are some highly-active core organizations with high degrees. In other words, some organizations communicate with a lot of other organizations and are in the core of the community. Meanwhile, many organizations only talk to few other organizations and are at the periphery of the community [101]. Then the question for the GlobalSympNet is, among many organizations that have not communicated with each other before, which ones should the GlobalSympNet picks so that its staff members can try to introduce them to each other and encourage them to communicate. This provides a good scenario to use our simulation, because trying different strategies on its member organizations in the real world is often difficult, risky, or expensive.

Therefore, in this experiment, the simulation is used to explore how three strategies that enhance the communication network can facilitate collaboration: *Strategy 1* encourages core members to communicate more with other core members; *Strategy 2* encourages core members to communicate with peripheral members; *Strategy 3* encourages the communication between peripheral members.

4.4.3 Simulation setup and results

As the research tries to evaluate how different strategies to enhance the communication network will affect the simulated collaboration network, four scenarios are designed to manipulate the communication network topology: one baseline scenario with no changes to the communication network in Figure 4.8, and three scenarios with enhanced communication networks as the simulation input. To simulate the Strategy 1, 57 new edges (about 10% of the total number of existing edges) are added to the communication network. Each new edge lies randomly between two high-degree nodes, whose degrees are within the top 25% of all nodes. For the Strategy 2, the simulation also adds 57 edges to the original communicaiton network but each edge has to connect a random high-degree node, whose degrees are within the top 25%, and a random low-degree node, whose degrees are within the bottom 25%. Similarly, Strategy 3 is simulated by adding 57 edges between randomly chosen low-degree nodes, whose degrees are within the bottom 25%.

Besides the topology of the communication network, two other groups of factors or parameters will affect the outcome of our simulation: (1) agents' attributes and criteria of project evaluation; and (2) candidate projects that agents put in their initial to-do lists. In order to balance between internal and external validity, this research controls parameters in the first group and ensure that all simulations for the four scenarios will have the same agent attributes and project evaluation criteria. This is because agents' attributes and evaluation criteria, such as size, focus regions, whether it is a community leader, etc., are based on real-world data and have been validated in section 4.4.1. Meanwhile, because the simulations are using synthesized project data and there is no empirical data about which project an agent will pick at the very beginning, I also introduce some level of randomness in the second group of factors. In different runs of the simulation, an agent is

allowed to pick projects randomly from the same pool of 300 synthesized candidates projects, and put them into their initial to-do lists, as long as these projects match one of the agent's missions or focus regions. Thus repeating multiple runs of a simulation scenario will help to improve the validity of the results.

After running simulations 30 times (each run lasts for 40 ticks, as in Section 4.4.1.), for each of the four scenarios, the effectiveness of the four strategies are is evaluated to see which one promotes more collaboration. An effective strategy is one that can facilitate or promote more collaboration, but how do I decide which strategy is more effective?

This research first considers the density of the collaboration network as an important and intuitive indicator for how well collaboration is promoted, because more edges in a collaboration network often mean more collaboration among organizations and a more collaborative environment in a coordination body. The increase in collaboration helps organizations to get more collaborators and access more resources that may be unavailable internally [144]. More edges in a network will general decrease the distance (such as the average shortest path length) between nodes and make the community more close-knit. Network density is also among the commonly used metrics to evaluate an inter-organizational network [145] in organizational research. Admittedly, the metric of density emphasizes more on the quantity of collaboration. Although the quality of collaboration is also important, it is often out of the control of coordination bodies and thus is beyond the scope of this research.

Then, the strategy, whose corresponding simulation scenario can generate a collaboration network with more edges in our experiment, is considered more effective at facilitating inter-organizational collaboration. Figure 4.10 shows the number of edges in simulated collaboration networks after implementing different strategies on the communication network. Each data point is the average of 30 runs. Vertical bars at data points indicate the 95% confidence interval.

The comparison first suggests a surprising result: Strategy 1, adding edges between core members in the communication network, performs worse than adding no edges to the communication network. In other words, although it is expected that adding edges to the communication network will always have positive impact on the collaboration network, adding edges only between high-degree nodes

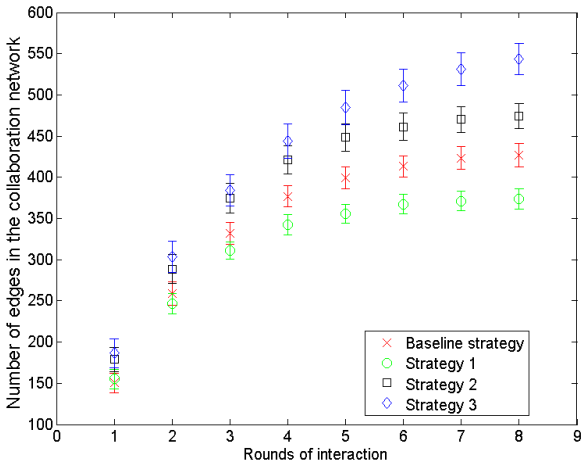


Figure 4.10: Effectiveness of different strategies to promote collaboration.

does not help. In the context of the GlobalSympNet, focusing only on promoting communication among its core members may not facilitate collaboration.

Meanwhile, Strategies 2 and 3, especially 3, can increase the density of the resulting collaboration network, compared to the baseline strategy. In other words, if GlobalSympNet can encourage peripheral members to get more involved in the community by introducing them to other organizations, especially other peripheral members, collaborations among humanitarian organizations will be facilitated.

4.4.4 Discussions

Why does Strategy 1 fail to work? Why does Strategy 3 outperform Strategy 2? From a multi-relational network perspective, it is hypothesized that the information flow, i.e., dissemination of candidate project information, through the communication network may have contributed to the difference in the densities of simulated collaboration networks. Figure 4.11 shows the total number of unique candidate projects that are evaluated by all agents. This number serves as a good metric of how well project information disseminates through the communication network. If more projects get evaluated by organizations, there is a higher chance that more collaboration can happen and thus more edges are formed in the collaboration network. As you can see from the figure, the curves of the number of evaluated projects are very similar to the curves for the number of collaboration

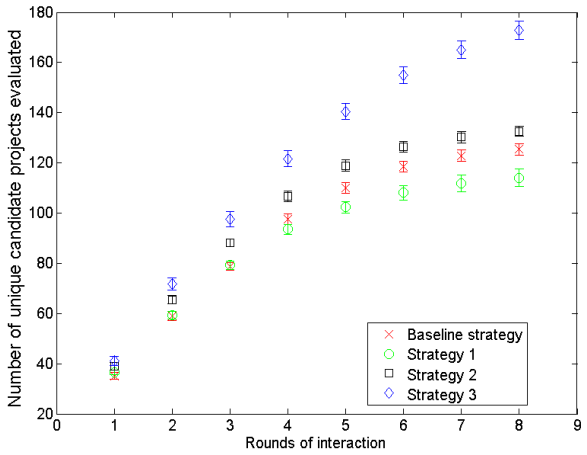


Figure 4.11: The total number of unique candidate projects that agents evaluates.

Table 4.3: Comparison of four communication networks.

Networks	**Avg. Path Length**	**Clustering Coefficient**
The original network	1.9559	0.6600
The network enhanced with Strategy 1	1.9350 (1.9342-1.9359)	0.7031 (0.7008-0.7054)
The network enhanced with Strategy 2	1.9098 (1.9068-1.9129)	0.6345 (0.6309-0.6381)
The network enhanced with Strategy 3	1.9263 (1.9254-1.9272)	0.5620 (0.5590-0.5649)
Note: 95% CIs are reported in parentheses if available.		

edges in Figure 4.10. Strategy 1 has the lowest number of evaluated projects, while Strategy 3 generates up to 30% more evaluted projects than Strategy 2 does.

Now it is clear that different levels of information flow in the communication network lead to different collaboration networks. Then what characteristics of the communication network could reflect such difference in information flow and the structure of the collaboration network? As all the three strategies add the same number of edges to the communication network, the densities of the three enhanced communication networks are the same. Then average path length and clustering coefficient are left for considerations. Table 4.3 compares the two measures for inter-organizational communication networks that are enhanced by the three strategies.

Although previous research argued that dissemination is easier in networks with shorter path lengths [146], Table 4.3 reveals that the average path length does not correlate well with network dissemination in this study. Intuitively all three strategies can reduce the average path length because they add new edges to the original network. Despite of the shorter average path length, the network enhanced by Strategy 1 still hinders the flow of project information compared to the original network. In addition, Strategy 2 is able to generate networks with the lowest average path length by connecting peripheral nodes with hub nodes. However, Strategy 2 is still outperformed by Strategy 3 in terms of facilitating dissemination and collaboration. Thus average path length of the communication network does not seem to be the key factor in affecting the outcome.

By contrast, the clustering coefficient sheds light on the problem. As Table 4.3 shows, Strategy 1 increases the clustering coefficient. Strategies 2 and 3, on the other hand, are able to decrease the clustering coefficient. The order of networks (by descending clustering coefficients) correlates well with the order of strategies (by descending effectiveness in facilitating inter-organizational collaboration). In other words, the clustering coefficient of the inter-organizational network based on communication relationship has important impact on the density of the inter-organizational collaboration network.

From a multi-relational network perspective, the simulation data confirms the close relationship between communication and collaboration networks, as the flow of project information through the communication network plays an important role in the formation of the collaboration network. It also suggests that simply decreasing the average path length in the communication network may not necessarily increase the density of the collaboration network. Instead, adding more edges for low-degree nodes in the communication network can better facilitate the flow of project information and foster a better connected inter-organizational collaboration network. In addition, adding more edges between high-degree nodes in the communication network may have negative impacts on the collaboration network. A possible reason is that, in a highly-clustered communication network, paths for the information flow across the network is often controlled by high-degree nodes. Adding edges between high-degree nodes reinforced their key roles in the network. By contrast, new edges for low-degree nodes add shortcuts that may

connect some poorly connected nodes or loose clusters a little bit. Consequently, project information does not have to go through the few high-degree nodes and have more alternative paths to be disseminated to another nodes or clusters.

From a organizational management perspective, if GlobalSympNet would like to facilitate collaboration among its members, it may want to get peripheral members more involved in the community by encouraging them to communicate with others, especially other peripheral organizations. Empirical studies found that core organizations in the communication network are often larger or general-purpose humanitarian organizations, such as the Red Cross or the Food and Agriculture Organization of the United Nations. These organizations are often well funded and have less need from others. Meanwhile, peripheral organizations are often smaller ones. They may also specialize in a specific humanitarian area, such as land mine detection, or a geographic region, such as North America. They may have with limited information, resources or expertise. Thus peripheral organizations have greater need for external resources and information and are generally more motivated to collaborate.

If core organizations are more densely connected to each other, project information is often exchanged among this highly-connected group. If an organization at the periphery of the communication network would like to send out information of a candidate project, for which it wants to solicit collaborators, or to get information of other candidate projects from peer organizations, it has to rely on its one or two points of contact among core organizations. If a peripheral organization has a candidate project that fails to get support from core organizations, the project will have little chance to be received and evaluated by other organizations, who may be very interested in collaborating on the project.

Conversely, if communication between core and peripheral organizations is encouraged, information about candidate projects can flow more easily between information- and resource-rich organizations and organizations who desperately need more information and resources. Further, if a peripheral organization can talk to more peripheral ones, those who are highly motivated to collaborate are directly connected, which will likely to lead to more dissemination and collaboration.

4.5 Summary and future work

This research models the evolution of inter-organizational collaboration networks. The model is based on the interaction and information flow among organizations in a non-hierarchical context. The model adopts a multi-relational perspective and investigate how organizations' communication network affects the growth of the collaboration network. It also uses an event-based approach and generates collaboration networks from the formation of teams for joint projects. When modeling organizations' decision-makings, this research captures both endogenous and exogenous factors that affect organizations' decisions on whether to collaborate on a project and to disseminate project information. To my best knowledge, this study represents the first attempt to use agent-based models for the evolution of inter-organizational networks.

The model is implemented as an agent-based simulation for a case study of the humanitarian sector. After configuring and validating the simulation, the simulation is used to study how to promote collaboration among humanitarian organizations by enhancing the communication network. The simulation helps to compare the effectiveness of three different strategies that enhance the communication network. From the perspective of multi-relational network analysis, the simulation results suggest that adding edges for low-degree nodes, which will leads to a lower clustering coefficient, in the communication network may help to facilitate the information flow on the communication network and the connectivity of the collaboration network. The organizational implication for humanitarian coordination bodies is that encouraging peripheral organizations to communicate more with others, especially with other peripheral ones, can help them reach out and consequently facilitate more inter-organizational collaboration in the humanitarian sector. The case study also helps to provide recommendations on how to promote humanitarian collaboration, which will eventually benefit disaster victims.

In addition, the implication of this research is not limited to this case study only. *First*, this research illustrates that micro-level interactions and the flow in one network (the communication network) could impact on the growth and macro-level structure of another network (the collaboration network). This research should have implications to the study of complex networks that involve

multiple types of relationships, such as social networks and supply-chain networks. *Second*, this study models the formation of edges in a collaboration network using an event-based approach, which can better reflect the n-ary nature of collaboration relationship in the real world. This approach can also be applied to the modeling of other networks with event-based n-ary relationships. For example, in a co-authorship network, an event refers to scholars' collaboration on a paper; in an online social network, an event could be several users' participation in the same threaded discussion. *Third*, the work proposes a novel way to model the flow of competitive yet non-exclusive information through networks. Incorporating individuals' endogenous prioritizations and exogenous influence from the network, the new approach can be used to model the network flow of fashion, behaviors, products, and so on. For instance, different digital gadgets, such as iPads and netbooks, are competing in the sense that they all need users' investment, but they can co-exist too, because a user may not be limited to only one gadget. *Last*, when properly configured, the model may also be used to simulate emerging collaboration networks in other domains, where individuals or agencies interact and influence each other in an environment that has no formal hierarchy and welcomes collaboration. Example domains include social services, environmental protection, open-source software development, education, academic research, and so on.

For future research, I would like to configure and calibrate the simulation configuration further, and improve the accuracy of edge-by-edge prediction, so that this simulation can be used for more experiments on other inter-organizational collaboration issues. One possible way to improve the simulation is to introduce more heterogeneity among agents. For example, in Equation 4.4, one can assign different coefficients values (α_m and α_r) to agents, which means they evaluate candidate projects in different ways. This will bring the simulation closer to real-world situations. Another possible way is through the use of more data. The current research is limited by the amount of data that was collected through traditional data collection methods (surveys and interviews). More effective methods are needed to collect more data, especially on how networks among humanitarian organizations evolve over time. More data will also help to further improve and validate our simulation.

Moreover, this research studies how the communication network affects the

collaboration network and the interaction between the two networks is one-way only. I also hope to explore how the collaboration network in turn affects the influence and information dissemination among nodes in the communication network. Another possible research direction is to incorporate other relationships in this multi-relational network, such as the business transaction network and the funding network, into the agent-based model.

Chapter 5

Summary

This dissertation focuses on network flows, an important topic that spans network research at both macro and micro levels. Specifically, it studies study information and supply flows in social and business networks, including supply-chain networks, online health communities, and inter-organizational networks.

The research on supply-chain networks examines supply flows at the macro level. A topological analysis, the study first develops new performance metrics at the topological level to capture the characteristic of supply flows and reflect different roles that multiple types of nodes play in a supply-chain network (e.g., providers, distributors, and consumers, etc.). Then two customizable heuristic strategies are proposed to improve or balance the robustness of supply-chain networks against disruptions. The first strategy is suitable for building or designing supply-chain networks. The second strategy could adjust topologies of existing supply-chain networks that may be difficult or expensive to re-build. Simulation experiments show that both strategies lead to supply-chain network topologies that can preserve more flows than traditional topologies after random and targeted disruptions, especially when it is not possible to predict which type of disruption will occur.

The study on online health communities emphasizes on information flows at the micro level. An individual-level analysis of large-scale online social networks, it tried to identify influential individuals. A data-drive approach based on classification is first adopted. The second approach leverages the sentimental impact of online interactions and information flows in online health communities. To find out

the sentimental impact of information flow in OHCs, only looking at the topology of users' social networks does not suffice. Fortunately, the large-scale data from OHCs not only enables one to build a online social network, but also provides the content of users' online interactions. Thus the content of individual users' posts is analyzed using text mining techniques, and users' sentiment dynamics in their posts are revealed. It is found that the information flow among users does influence one's sentiment: a thread originator's change of sentiment within the thread correlates with the sentiment of thread participants. Then a new metric based on the sentiment influence–the number of users' influential responding replies–is proposed. The influential user ranking by the new metric outperforms rankings by many traditional metrics, and even the classifier that utilizes more than 60 user features. The accuracy of the influential user identification is further improved by incorporating the new metric into the classifier.

The investigation on inter-organizational networks tries to connect micro-level individual interactions to macro-level network structure. The research proposes three models: a network influence model for how organizations influence others' decisions, an information dissemination model for how the information about candidate collaborative projects disseminates through organizations' interactions, and an event-based network evolution model for the evolution of inter-organizational collaboration networks. Agent-based simulations are used to analyze how changes to the communication network affect the flow of information and topology of the resulting collaboration network. Simulation results also suggest that encouraging interactions among organizations at the peripheral of the communication network can better facilitate the flow of information and collaboration.

A multi-level study of network flows in three types of networks, this dissertation can contribute to the better understanding, utilization, and management of various social and business networks. The outcome of the dissertation also has the potential to make a difference in the real world. The research in Chapter 2 provides two heuristic strategies to balance or improve supply-chain networks' robustness against disruptions. More robust supply-chain networks (such as power grids and water supply network) are not only valuable for business operations, but also important for people's daily life. The online social network research in Chapter 3 is the first large-scale computational study of sentiment influence in OHCs.

It has implications for the design and management of an active, supportive, and sustainable OHC, which can play important roles in providing personalized support, disseminating medical knowledge, encouraging healthy lifestyles, advocating new treatments, etc. In Chapter 4, the study on inter-organizational networks provides insights into the evolution of collaboration networks and gives suggestions on how to promote humanitarian collaboration, which will eventually benefit disaster victims.

Further, the three pillar research of this dissertation could have general implications for the area of networks analysis. Through the research on supply-chain network, it is illustrated that incorporating the heterogeneous roles of nodes into the analysis of network flows could provide a new perspective to the performance evaluation of many networks, such as power grids, Internet, and so on. The exploration on online health communities shows that analyzing the content of individuals' interactions and conducting sentiment analysis on large-scale data can help to understand individuals' behavioral dynamics and the impact of social influence in online settings. The study on inter-organizational networks proposes a new dissemination model and an event-based network evolution model. It also demonstrates the potential to use network flows to connect micro-level and macro-level network phenomena, and the power of the multi-relational perspective of network analysis.

Appendix A

Properties of a Good Metric for Average Delivery Efficiency

The *average delivery efficiency* metric must integrate in the evaluation *both the number of accessible supply nodes* and *the length of supply paths to those supply nodes*. For demand node D_i, its average delivery efficiency $AVG\_DEF_{D_i}$ must satisfy the following three requirements (under the reasonable assumption that higher average delivery efficiency is better):

1. If two demand nodes can access the same number of supply nodes, the metric should favor the one with shorter supply-path length. For example, given two demand nodes $D_p, D_q \in V_D$, both nodes can access m supply nodes $S_1, S_2, ..., S_m \in V_S$, with shortest supply-path lengths $l_{p,1}, l_{p,2}, ..., l_{p,m}$ and $l_{q,1}, l_{q,2}, ..., l_{q,m}$ respectively. Then, the average delivery efficiency value of the two demand nodes must satisfy Equation A.1:

$$If\ \forall i \in [1, m], l_{p,i} \leq l_{q,i},\ then\ AVG\_DEF_{D_p} \leq AVG\_DEF_{D_q} \tag{A.1}$$

2. The metric should provide some reward to a demand node that can access more supply nodes. For instance, demand node $D_p \in V_D$ can only access supply nodes $S_1, S_2, ..., S_m \in V_S$, with shortest supply-path lengths $l_{p,1}, l_{p,2}, ..., l_{p,m}$. Demand node $D_q \in V_D$ can also access these m supply nodes, with shortest supply-path lengths $l_{q,1}, l_{q,2}, ..., l_{q,m}$ respectively. However, D_q can also access k, where $k \geq 1$, additional supply nodes $S_{m+1}, ..., S_{m+k} \in V_S$, with shortest supply-path

lengths $l_{p,m+1}, ..., l_{p,m+k}$. Then Equation A.2 must hold:

$$If\ \forall i \in [1,m] : l_{p,i} = l_{q,i},\ then\ AVG\_DEF_{D_p} < AVG\_DEF_{D_q}. \tag{A.2}$$

3. The metric must be able to handle the situation when a demand node is not connected to any supply node, i.e. its shortest supply-path length to any supply node is infinity. If demand node $D_p \in V_D$ cannot access any supply node, then $AVG\_DEF_{D_p} = 0$.

Appendix B

Theoretical Analysis of the Degree Distribution for Simplified RLR Scale-free Networks

In networks with homogeneous nodes, degree distribution is often closely related to the network's robustness [41]. Now let us briefly look at the degree distribution $P_{RL}(k)$ of a simplified RLR scale-free network with n nodes and m edges. Assume $P(k_0) = {k_0}^{-r}$ is the degree distribution of the scale-free network generated with the preferential attachment model [10]. The maximum node degree is D and the rewiring probability is p_r. Then we have Equation B.1, where $P_{discon}(x, k_0)$ is the probability that x edges are disconnected from a node with degree k_0; $P_{new}(k-k_0+x, mp_r)$ is the probability that among all the mp_r edges to be rewired, $k - k_0 + x$ connect to the same node. Both probabilities are based on binomial distributions.

$$P_{RL}(k) = \sum_{k_0=0}^{D} P(k_0)[\sum_{x=0}^{k_0} P_{discon}(x, k_0) P_{new}(k - k_0 + x, mp_r)] \tag{B.1}$$

For the purpose of simplicity, we assume that when rewiring an edge, we randomly choose one end to disconnect and set $d_{max} = \infty$. Using this simplified rewiring, we have

$$P_{discon}(x, k_0) = \binom{k_0}{x} (\frac{p_r}{2})^x (1 - \frac{p_r}{2})^{k_0 - x} \tag{B.2}$$

$$P_{new}(k-k_0+x, mp_r) = \begin{cases} \binom{mp_r}{k-k_0+x}(\frac{1}{n-2})^{k-k_0+x}(1-\frac{1}{n-2})^{mp_r-k+k_o-x}, \\ \text{if } k-k_0+x \geq 0; \\ 0, \qquad \text{otherwise.} \end{cases} \tag{B.3}$$

The expected value of $P_{discon}(x, k_0)$ is $E(P_{discon}(x, k_0)) = k_0 p_r/2$. This means that a node with degree k_0 in the scale-free network will on average lose $k_0 p_r/2$ edges in the rewiring process. The expected value of $P_{new}(k-k_0+x, mp_r)$ is $E(P_{new}(k-k_0+x, mp_r)) = mp_r/(n-2)$. As $n >> 2$, $E(P_{new}(k-k_0+x, mp_r)) \approx mp_r/n$, where $m/n = \langle k_0 \rangle$ is the average degree in the network. We can then infer that a node in the original scale-free network will, on average, get $\langle k_0 \rangle p_r$ new edges in the rewiring process. As a result, a node with original degree k_0 will on average have degree $k_0 + \langle k_0 \rangle p_r - k_0 p_r/2$ after rewiring. In the rewired network, a node with original degree $k_0 > 2\langle k_0 \rangle$ will most likely have lower degree, while a node with original degree $k_0 < 2\langle k_0 \rangle$ will generally get higher degree.

Using Equations B.1, B.2 and B.3, we can infer the degree distributions of simplified rewired networks. In Figure B.1, we draw the inferred degree distributions of three simplified rewired networks, each with different rewiring probabilities. The horizontal axes denote the degree of nodes; the vertical axes reflect the probability that a node has a given degree. As a comparison, we also include the scale-free network with degree distribution $P(k) = k^{-2.9}$ [10]. Each network has 1000 nodes and 1815 edges. The maximum node degree D is set to 70, but Figure B.1 only shows probabilities for node degrees up to 12, because the probabilities for still higher degrees nodes is very small. As the figure shows, a higher rewiring probability for the simplified rewired network will lead to fewer high-degree nodes and more nodes with low and medium degrees. The degree distributions of rewired networks generally become less skewed than that of the scale-free network. As the rewiring probability increases, the degree distribution of a simplified rewired scale-free network gradually approaches a Poisson distribution.

Recall that in the RLR approach, instead of randomly choosing which node to disconnect, we actually disconnect an edge from the node with higher degree, which means high-degree nodes would lose and low-degree nodes would gain more edges

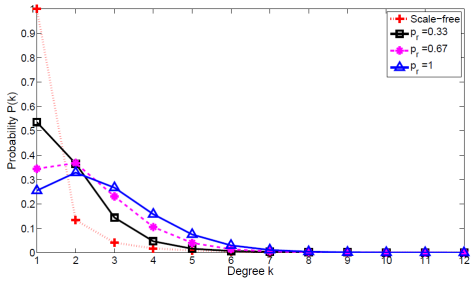


Figure B.1: Inferred degree distributions of scale-free and three simplified RLR rewired scale-free networks.

in the rewiring process. As a result, the degree distribution of a RLR scale-free network should be more homogeneous than that of a simplified rewired network with the same rewiring probability, which means a RLR scale-free network will have many nodes with medium degrees and few nodes with high or low degrees. Previous research on networks revealed that robustness in the presence of targeted failures increases when nodes have similar degrees [41]. Thus we hypothesize that RLR scale-free should have better robustness than scale-free networks in targeted supply disruptions. We will validate this hypothesis with our simulations of specific distribution networks.

Appendix C

The list of words and expressions related to death

This list is picked from two lists: (1) "List of expressions related to death" at Wikipedia (`http://en.wikipedia.org/wiki/List_of_expressions_related_to_death`) and (2) "Death and general words relating to death" at the MacMillan Dictionary Thesaurus (`http://www.macmillandictionary.com/thesaurus-category/american/Death-and-general-words-relating-to-death`).

The list only contains the original form of these words and expressions. Their variants in different tense or plural forms are also used in the search.

- pass away
- funeral
- die
- death
- memorial
- is gone
- at rest
- final summons

- assume room temperature
- at peace
- in peace
- beyond the grave
- beyond the veil
- over the big ridge
- the last roundup
- the great majority
- the ultimate sacrifice
- a last bow
- last breath
- bereavement
- demise
- obituary

Bibliography

[1] Borner, K., S. Sanyal, and A. Vespignani (2007) "Network science," *Annual Review of Information Science and Technology*, **41**(1), pp. 537–607.

[2] Alexanderson, G. (2006) "About the cover: Euler and Konigsberg's Bridges: A historical view," *The Bulletin of the American Mathematical Society*, **43**(4), pp. 567–573.

[3] Freeman, L. (1979) "Centrality in social networks conceptual clarification," *Social Networks*, **1**(3), pp. 215–239.

[4] Dijkstra, E. (1959) "A note on two problems in connexion with graphs," *Numerische mathematik*, **1**(1), pp. 269–271.

[5] Erdos, P. and A. Renyi (1959) "On random graphs," *Publicationes Mathematicae.*, **6**, pp. 290–297.

[6] Borgatti, S. (2005) "Centrality and network flow," *Social Networks*, **27**(1), pp. 55–71.

[7] Garlaschelli, D. and M. I. Loffredo (2004) "Patterns of Link Reciprocity in Directed Networks," *Physical Review Letters*, **93**(26), p. 268701, pRL.

[8] Liben-Nowell, D. and J. Kleinberg (2007) "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, **58**(7), pp. 1019–1031.

[9] Watts, D. J. and S. H. Strogatz (1998) "Collective dynamics of 'small-world' networks," *Nature*, **393**(6684), pp. 440–442.

[10] Barabasi, A. L. and R. Albert (1999) "Emergence of scaling in random networks," *Science*, **286**(5439), pp. 509–512.

[11] Newman, M. E. J. (2006) "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, **103**(23), pp. 8577–8582.

[12] Bertsekas, D. (1998) *Network optimization: continuous and discrete models*, Athena Scientific Belmont, Massachusetts.

[13] Leskovec, J., L. Backstrom, and J. Kleinberg (2009) "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1557077, pp. 497–506.

[14] Pathak, S. D., J. M. Day, A. Nair, W. J. Sawaya, and M. M. Kristal (2007) "Complexity and adaptivity in supply networks: Building supply network theory using a complex adaptive systems perspective," *Decision Sciences*, **38**(4), pp. 547–580.

[15] Powell, W. W. and L. Smith-Doerr (1994) *Networks and economic life*, Princeton University Press, Princeton, N.J., pp. 368–402.

[16] Zhao, K., A. Kumar, and J. Yen (2009) "Effect of topology on robustness of a supply network-metrics and results," in *Proceedings of the 19th Annual Workshop on Information Technologies and Systems (WITS'09)*, pp. 97–102.

[17] Zhao, K., A. Kumar, T. P. Harrison, and J. Yen (2011) "Analyzing the Resilience of Complex Supply Network Topologies Against Random and Targeted Disruptions," *IEEE Systems Journal*, **5**(1), pp. 28–39.

[18] Zhao, K., A. Kumar, and J. Yen (2011) "Achieving High Robustness in Supply Distribution Networks by Rewiring," *IEEE Transactions on Engineering Management*, **58**(2), pp. 347–362.

[19] Surana, A., S. Kumara, M. Greaves, and U. N. Raghavan (2005) "Supply-chain networks: a complex adaptive systems perspective," *International Journal of Production Research*, **43**(20), pp. 4235–4265.

[20] Lee, H. L., V. Padmanabhan, and S. Whang (1997) "The bullwhip effect in supply chains," *Sloan Management Review*, **38**(3), pp. 93–102.

[21] Rice, J. and F. Caniato (2003) "Building a Secure and Resilient Supply Network," *Supply Chain Management Review*, **7**(5), pp. 22–30.

[22] Chopra, S. and M. S. Sodhi (2004) "Managing risk to avoid supply-chain breakdown," *MIT Sloan Management Review*, **46**(1), pp. 53–61.

[23] Hendricks, K. B. and V. R. Singhal (2005) "An empirical analysis of the effect of supply chain disruptions on long-run stock price performance and equity risk of the firm," *Production and Operations Management*, **14**(1), pp. 35–52.

[24] Kleindorfer, P. R. and G. H. Saad (2005) "Managing disruption risks in supply chains," *Production and Operations Management*, **14**(1), pp. 53–68.

[25] Wu, T., J. Blackhurst, and P. O'Grady (2007) "Methodology for supply chain disruption analysis," *International Journal of Production Research*, **45**(7), pp. 1665–1682.

[26] Thadakamalla, H. P., U. N. Raghavan, S. Kumara, and R. Albert (2004) "Survivability of Multiagent-Based Supply Networks: A Topological Perspective," *IEEE Intelligent Systems*, **19**(5), pp. 24–31, 1032412.

[27] Lee, S. H. (1980) "RELIABILITY EVALUATION OF A FLOW NETWORK," *IEEE Transactions on Reliability*, **29**(1), pp. 24–26.

[28] Lin, Y. K. (2001) "A simple algorithm for reliability evaluation of a stochastic-flow network with node failure," *Computers and Operations Research*, **28**(13), pp. 1277–1285.

[29] Wang, D. and W. H. Ip (2009) "Evaluation and Analysis of Logistic Network Resilience with Application to Aircraft Servicing," *IEEE Systems Journal*, **3**(2), pp. 166–173.

[30] Ball, M. O. (1980) "Complexity of network reliability computations," *Networks*, **10**(2), pp. 153–165.

[31] Jayaraman, V. and A. Ross (2003) "A simulated annealing methodology to distribution network design and management," *Euro. Journal of Operational Research*, **144**(3), pp. 629–645.

[32] Altiparmak, F., M. Gen, L. Lin, and T. Paksoy (2006) "A genetic algorithm approach for multi-objective optimization of supply chain networks," *Computers and Industrial Engineering*, **51**(1), pp. 196–215.

[33] Yeh, W. (2005) "A hybrid heuristic algorithm for the multistage supply chain network problem," *The International Journal of Advanced Manufacturing Technology*, **26**(5), pp. 675–685.

[34] Choi, T. Y. and Y. Hong (2002) "Unveiling the structure of supply networks: case studies in Honda, Acura, and DaimlerChrysler," *Journal of Operations Management*, **20**(5), pp. 469–493.

[35] Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang (2006) "Complex networks: Structure and dynamics," *Physics Reports*, **424**(4-5), pp. 175–308.

[36] Anderson, R. M. and R. M. May (2002) *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press.

[37] Sun, H. and J. Wu (2005) "Scale-Free Characteristics of Supply Chain Distribution Networks," *Modern Physics Letters B*, **19**(17), pp. 841–848.

[38] Wang, K., Z. Zeng, and D. Sun (2008) "Structure Analysis of Supply Chain Networks Based on Complex Network Theory," in *Proceedings of the 2008 4th International Conference on Semantics, Knowledge and Grid*, IEEE Computer Society, pp. 493–494, 1495576.

[39] Albert, R. and A. L. Barabasi (2002) "Statistical mechanics of complex networks," *Reviews of Modern Physics*, **74**(1), pp. 47–97.

[40] Pathak, S. D., D. M. Dilts, and G. Biswas (2007) "On the Evolutionary Dynamics of Supply Network Topologies," *IEEE Trans. on Engineering Management*, **54**(4), pp. 662–672.

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

[42] Holme, P., B. J. Kim, C. N. Yoon, and S. K. Han (2002) "Attack vulnerability of complex networks," *Physical Review E*, **65**(5), p. 056109.

[43] Grubesic, T. H., T. C. Matisziw, A. T. Murray, and D. Snediker (2008) "Comparative Approaches for Assessing Network Vulnerability," *International Regional Science Review*, **31**(1), pp. 88–112.

[44] Klau, G. W. and R. Weiskircher (2005) *Robustness and Resilience*, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 417–437.

[45] Costa, L. d. F., F. A. Rodrigues, G. Travieso, and P. R. V. Boas (2007) "Characterization of complex networks: A survey of measurements," *Advances in Physics*, **56**(1), pp. 167 – 242.

[46] Albert, R., I. Albert, and G. L. Nakarado (2004) "Structural vulnerability of the North American power grid," *Physical Review E*, **69**(2), p. 4.

[47] Neely, A., M. Gregory, and K. Platts (1995) "Performance measurement system design: A literature review and research agenda," *International Journal of Operations and Production Management*, **15**(4), pp. 80–166.

[48] Feigenbaum, A. (1961) *Total Quality Control*, McGraw-Hill, New York, NY.

[49] Vickery, S., R. Calantone, and C. Drge (1999) "Supply chain flexibility: an empirical study," *Journal of Supply Chain Management*, **35**(3), pp. 16–24.

[50] Beamon, B. M. (2004) "Humanitarian relief chains: issues and challenges," in *Proceedings of the 34th International Conference on Computers and Industrial Engineering*, pp. 77–82.

[51] Martin, R. (1993) "Changing the Mind of the Corporation," *Harvard Business Review*, **71**(6 November-December), pp. 81–94.

[52] Fassoula, E. D. (2006) "Transforming the supply chain," *Journal of Manufacturing Technology Management*, **17**(6), pp. 848 – 860.

[53] Subramanian, G. H. and A. C. Iyigungor (2006) "Information systems in supply chain management: a comparative case study of three organisations," *Int. J. Bus. Inf. Syst.*, **1**(4), pp. 370–386, 1356420.

[54] Lao, G. and L. Xing (2008) *Supply Chain System Integration in Retailing: A Case Study of LianHua*, vol. 254, Springer Boston, pp. 519–528.

[55] BaselineMagazine (2006) "The Limited Designs Supply Chain to Begin and End With the Customer," *Baseline*, (April 3).

[56] Zhang, H., B. Qiu, K. Ivanova, C. L. Giles, H. C. Foley, and J. Yen (2010) "Locality and Attachedness-based Temporal Social Network Growth Dynamics Analysis," *Journal of American Society of Information Science and Technology (JASIST)*, **61**(5), pp. 964–977.

[57] Law, A. and W. Kelton (2000) *Simulation modeling and analysis*, 3rd ed., McGraw-Hill, New York.

[58] Latora, V. and M. Marchiori (2005) "Vulnerability and protection of infrastructure networks," *Physical Review E*, **71**(1), p. 4.

[59] Newman, M. E. J. (2008) *Mathematics of networks*, 2nd ed., Palgrave Macmillan, Basingstoke, UK.

[60] Latora, V. and M. Marchiori (2001) "Efficient Behavior of Small-World Networks," *Physical Review Letters*, **87**(19), p. 198701.

[61] Warren, M. and W. Hutchinson (2000) "Cyber attacks against supply chain management systems: a short note," *International Journal of Physical Distribution and Logistics Management*, **30**(7-8), pp. 710–716.

[62] Choi, T. Y., K. J. Dooley, and M. Rungtusanatham (2001) "Supply networks and complex adaptive systems: control versus emergence," *Journal of Operations Management*, **19**(3), pp. 351–366.

[63] Crucitti, P., V. Latora, and M. Marchiori (2004) "Model for cascading failures in complex networks," *Physical Review E*, **69**(4), p. 045104.

[64] Qiu, B., K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G. E. Greer, and K. Portier (2011) "Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community," in *Proceedings of the Third IEEE Third International Confernece on Social Computing (SocialCom'11)*, pp. 274–281.

[65] Zhao, K., B. Qiu, C. Caragea, D. Wu, P. Mitra, J. Yen, G. E. Greer, and K. Portier (2011) "Identifying Leaders in an Online Cancer Survivor Community," in *Proceedings of the 21st Annual Workshop on Information Technologies and Systems (WITS'11)*, pp. 115–120.

[66] Eaton, L. (2002) "Europeans and Americans turn to internet for health information," *British Medical Journal*, **325**(7371), pp. 989–989.

[67] Zickuhr, K. (2010) *Generations 2010*, *Tech. rep.*, Pew Internet.

[68] Ziebland, S., A. Chapple, C. Dumelow, J. Evans, S. Prinjha, and L. Rozmovits (2004) "How the internet affects patients' experience of cancer: a qualitative study," *British Medical Journal*, **328**(7439), p. 564.

[69] Dunkel-Schetter, C. (1984) "Social Support and Cancer: Findings Based on Patient Interviews and Their Implications," *Journal of Social Issues*, **40**(4), pp. 77–98.

[70] McClellan, W., D. Stanwyck, and C. Anson (1993) "Social support and subsequent mortality among patients with end-stage renal disease," *Journal of the American Society of Nephrology*, **4**(4), pp. 1028–1034.

[71] Maloney-Krichmar, D. and J. Preece (2005) "A multilevel analysis of sociability, usability, and community dynamics in an online health community," *ACM Trans. Comput.-Hum. Interact.*, **12**(2), pp. 201–232.

[72] Mehra, A., A. L. Dixon, D. J. Brass, and B. Robertson (2006) "The Social Network Ties of Group Leaders: Implications for Group Performance and Leader Reputation," *Organization Science*, **17**(1), pp. 64–79.

[73] Anonymous (2008) "Calling all patients," *Nature Biotech*, **26**(9), pp. 953–953.

[74] Brownstein, C. A., J. S. Brownstein, D. S. Williams, P. Wicks, and J. A. Heywood (2009) "The power of social networking in medicine," *Nature Biotech*, **27**(10), pp. 888–890.

[75] Ma, X., G. Chen, and J. Xiao (2010) "Analysis of an online health social network," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ACM, pp. 297–306.

[76] Valente, T. (2010) *Social networks and health: models, methods, and applications*, Oxford University Press.

[77] Freeman, L. (1977) "A set of measures of centrality based on betweenness," *Sociometry*, **40**, pp. 35–41.

[78] Newman, M. (2005) "A measure of betweenness centrality based on random walks," *Social Networks*, **27**(1), pp. 39–54.

[79] Cobb, N. K., A. L. Graham, and D. B. Abrams (2010) "Social Network Structure of a Large Online Community for Smoking Cessation," *American Journal of Public Health*, **100**(7), pp. 1282–1289.

[80] Brin, S. and L. Page (1998) "The anatomy of a large-scale hypertextual Web search engine," *Computer networks and ISDN systems*, **30**(1-7), pp. 107–117.

[81] Lazer, D., A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann (2009) "Life in the network: the coming age of computational social science," *Science*, **323**(5915), pp. 721–723.

[82] Kempe, D., J. Kleinberg, and E. Tardos (2003) "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 137–146.

[83] Zhang, J., M. S. Ackerman, and L. Adamic (2007) "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*, ACM, pp. 221–230.

[84] Agarwal, N., H. Liu, L. Tang, and P. S. Yu (2008) "Identifying the influential bloggers in a community," in *Proceedings of the International Conference on Web Search and Web Data Mining*, ACM, pp. 207–218.

[85] Yang, C. C., X. Tang, and B. M. Thuraisingham (2010) "An Analysis of User Influence Ranking Algorithms on Dark Web Forums," in *ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010)*, ACM.

[86] Bambina, A. (2007) *Online social support: the interplay of social networks and computer-mediated communication*, Cambria Press, Youngstown, N.Y.

[87] WHO (2009) *Cancer*, *Tech. rep.*, World Health Organization.

[88] Yalom, I. (1970) *The Theory and Practice of Group Psychotherapy*, New York: Basic Books.

[89] Hu, M. and B. Liu (2004) "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*, ACM, 1014073, pp. 168–177.

[90] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003) "Latent dirichlet allocation," *J. Mach. Learn. Res.*, **3**, pp. 993–1022.

[91] Seni, G. and J. Elder (2010) "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery*, **2**(1), pp. 1–126.

[92] Merriam-Webster.com (2011), "influnce," .

[93] Centola, D. (2010) "The Spread of Behavior in an Online Social Network Experiment," *Science*, **329**(5996), pp. 1194–1197.

[94] Fowler, J. and N. Christakis (2008) "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study," *BMJ*, **337**(a2338), pp. 1–9.

[95] Barraclough, J. (1999) *Cancer and Emotion: A Practical Guide to Psycho-oncology*, 3rd ed., Wiley.

[96] Pang, B. and L. Lee (2008) "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, **2**(1-2), pp. 1–135.

[97] Freund, Y. and R. Schapire (1995) *A desicion-theoretic generalization of on-line learning and an application to boosting*, vol. 904, Springer Berlin / Heidelberg, pp. 23–37.

[98] Granovetter, M. (1978) "Threshold Models of Collective Behavior," *The American Journal of Sociology*, **83**(6), pp. 1420–1443.

[99] Valente, T. W. (1996) "Network models of the diffusion of innovations," *Computational and Mathematical Organization Theory*, **2**(2), pp. 163–164.

[100] Zhao, K., J. Yen, C. Maitland, A. Tapia, and L.-M. N. Tchouakeu (2009) "A formal model for emerging coalitions under network influence in humanitarian relief coordination," in *Proceedings of the 2009 Spring Simulation Multiconference*, vol. Agent-Directed Simulation, Society for Computer Simulation International, p. Article No.10, 1639820 1-7.

[101] Zhao, K., L.-M. Ngamassi, J. Yen, C. Maitland, and A. Tapia (2010) "Assortativity Patterns in Multi-dimensional Inter-organizational Networks: A Case Study of the Humanitarian Relief Sector," in *Advances in Social Computing–Proceedings of the 2010 International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP10)* (S.-K. Chai, J. J. Salerno, and P. L. Mabry, eds.), vol. 6007/2010 of *LNCS*, Springer, pp. 265–272.

[102] Zhao, K., J. Yen, L.-M. Ngamassi, C. Maitland, and A. Tapia (2010) "From Communication to Collaboration: Simulating the Emergence of Inter-organizational Collaboration Network," in *Proceedings of 2010 IEEE International Conference on Social Computing (SocialCom'10)*, IEEE Press, pp. 413–418.

[103] Zhao, K., J. Yen, L.-M. Ngamassi, C. Maitland, and A. H. Tapia (2009) "Modeling Emerging Coalitions in the context of Inter-organizational Networks: A Case Study of Humanitarian Coordination," *International Journal of Intelligent Control and Systems*, **14**(1), pp. 97–103.

[104] Zhao, K., J. Yen, L.-M. Ngamassi, C. Maitland, and A. Tapia (2012 (Accepted for publication)) "Simulating Inter-organizational Collaboration Network: a Multi-relational and Event-based Approach," *Simulation: Transactions of the Society for Modeling and Simulation International*.

[105] Powell, W. W. (1998) "Learning from collaboration: Knowledge and networks in the biotechnology and pharmaceutical industries," *California Management Review*, **40**(3), pp. 228–240.

[106] Wood, D. and B. Gray (1991) "Toward a comprehensive theory of collaboration," *Journal of Applied Behavioral Science*, **27**(2), pp. 139–162.

[107] Guo, C. and M. Acar (2005) "Understanding Collaboration Among Nonprofit Organizations: Combining Resource Dependency, Institutional, and Network Perspectives," *Nonprofit and Voluntary Sector Quarterly*, **34**(3), pp. 340–361.

[108] Jang, H. S. and R. C. Feiock (2007) "Public Versus Private Funding of Nonprofit Organizations: Implications for Collaboration," *Public Performance and Management Review*, **31**(2), pp. 174–190.

[109] Gazley, B. (2008) "Beyond the contract: The scope and nature of informal governmentnonprofit partnerships," *Public Administration Review*, **68**(1), pp. 141–154.

[110] Bennett, J. (1994) *NGO coordination at field level : a handbook*, INTRAC, Oxford.

[111] Maitland, C., A. Tapia, L. Ngamassi, K. Zhao, and E. Maldonado (2010) *Global Symposium +5 Information for Humanitarian Action Second Follow-up Survey Report*, *Tech. rep.*, Penn State University.

[112] Diesner, J., T. Frantz, and K. Carley (2005) "Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different"," *Computational and Mathematical Organization Theory*, **11**(3), pp. 201–228.

[113] Powell, W. W., D. R. White, K. W. Koput, and J. Owen-Smith (2005) "Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences," *The American Journal of Sociology*, **110**(4), pp. 1132–1205.

[114] Davis, G., M. Yoo, and W. Baker (2003) "The small world of the American corporate elite, 1982-2001," *Strategic organization*, **1**(3), pp. 301–326.

[115] Leskovec, J., L. Backstrom, R. Kumar, and A. Tomkins (2008) "Microscopic evolution of social networks," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 462–470.

[116] Huberman, B., D. Romero, and F. Wu (2009) "Social networks that matter: Twitter under the microscope," *First Monday*, **14**(1), p. 9.

[117] Entwisle, B., K. Faust, R. R. Rindfuss, and T. Kaneda (2007) "Networks and contexts: Variation in the structure of social ties," *American Journal of Sociology*, **112**(5), pp. 1495–1533.

[118] Szell, M., R. Lambiotte, and S. Thurner (2010) "Multirelational organization of large-scale social networks in an online world," *Proceedings of the National Academy of Sciences*, **107**(31), pp. 13636–13641.

[119] Ahmad, M. A., Z. Borbora, J. Srivastava, and N. Contractor (2010) "Link Prediction Across Multiple Social Networks," in *Domain Driven Data Mining Workshop (DDDM2010) at ICDM 2010*.

[120] Tang, L. and H. Liu (2010) "Toward Predicting Collective Behavior via Social Dimension Extraction," *IEEE Intelligent Systems*, **25**(4), pp. 19–25.

[121] Oliver, A. and M. Ebers (1998) "Networking network studies: an analysis of conceptual configurations in the study of inter-organizational relationships," *Organization Studies*, **19**(4), p. 549.

[122] Gulati, R. and M. Gargiulo (1999) "Where Do Interorganizational Networks Come From?" *The American Journal of Sociology*, **104**(5), pp. 1439–1493.

[123] Oliver, C. (1990) "Determinants of interorganizational relationships: Integration and future directions," *Academy of Management Review*, **15**(2), pp. 241–265.

[124] Ostrom, E. (1990) *Governing the commons: The evolution of institutions for collective action*, Cambridge Univ Pr.

[125] Van de Ven, A. and G. Walker (1984) "The dynamics of interorganizational coordination," *Administrative Science Quarterly*, **29**(4), pp. 598–621.

[126] Jaccard, P. (1901) "tude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin del la Socit Vaudoise des Sciences Naturelles*, **37**, pp. 547–579.

[127] Adamic, L. and E. Adar (2003) "Friends and neighbors on the web," *Social Networks*, **25**(3), pp. 211–230.

[128] Katz, L. (1953) "A new status index derived from sociometric analysis," *Psychometrika*, **18**(1), pp. 39–43.

[129] Rosenkopf, L. and G. Padula (2008) "Investigating the microstructure of network evolution: alliance formation in the mobile communications industry," *Organization Science*, **19**(5), pp. 669–687.

[130] Benchettara, N., R. Kanawati, and C. Rouveirol (2010) "A supervised machine learning link prediction approach for academic collaboration recommendation," in *Proceedings of the 4th ACM conference on Recommender systems*, ACM, pp. 253–256.

[131] Hasan, M. A., V. Chaoji, S. Salem, and M. Zaki (2006) "Link prediction using supervised learning," in *Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

[132] Wang, C., V. Satuluri, and S. Parthasarathy (2008) "Local probabilistic models for link prediction," in *7th IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 322–331.

[133] Clauset, A., C. Moore, and M. E. J. Newman (2008) "Hierarchical structure and the prediction of missing links in networks," *Nature*, **453**(7191), pp. 98–101.

[134] Bailey, N. T. J. (1975) *The Mathematical Theory of Infectious Diseases and its applications*, Griffin, London.

[135] Goldenberg, J., B. Libai, and E. Muller (2001) "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, **12**(3), pp. 211–223.

[136] Aral, S., L. Muchnik, and A. Sundararajan (2009) "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, pp. 21544–21549.

[137] McPherson, M., L. Smith-Lovin, and J. M. Cook (2001) "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, **27**(1), pp. 415–444.

[138] Broecheler, M., P. Shakarian, and V. S. Subrahmanian (2010) "A Scalable Framework for Modeling Competitive Diffusion in Social Networks," in *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pp. 295–302.

[139] Carley, K. and L. Gasser (1999) *Computational organization theory*, chap. 7, MIT Press, pp. 299–330.

[140] Berry, B. J. L., L. D. Kiel, and E. Elliott (2002) "Adaptive agents, intelligence, and emergent human organization: Capturing complexity through agent-based modeling," *Proceedings of the National Academy of Sciences*, **99**(3), pp. 7187–7188.

[141] Friedkin, N. E. and E. C. Johnsen (1997) "Social positions in influence networks," *Social Networks*, **19**(3), pp. 209–222.

[142] North, M., T. Howe, N. Collier, and R. Vos (2005) "The Repast Simphony Development Environment," in *Agent 2005 Conference on Generative Social Processes, Models, and Mechanisms* (C. Macal, M. North, and D. Sallach, eds.), pp. 159–166.

[143] Leonardi, P. M. (2007) "Activating the Informational Capabilities of Information Technology for Organizational Change," *Organization Science*, **18**(5), pp. 813–831.

[144] Boyd, B. (1990) "Corporate linkages and organizational environment: A test of the resource dependence model," *Strategic Management Journal*, **11**(6), pp. 419–430.

[145] Provan, K. and J. Sebastian (1998) "Networks within networks: Service link overlap, organizational cliques, and network effectiveness," *The Academy of Management Journal*, **41**(4), pp. 453–463.

[146] Kleinberg, J. (2007) *Cascading Behavior in Networks: Algorithmic and Economic Issues*, chap. 24, Cambridge University Press, pp. 613–632.

Vita

Kang Zhao

Kang Zhao joined the Ph.D. program at the College of Information Sciences and Technology, Penn State University in 2007 as a recipient of the Graham Endowed Graduate Fellowship. He holds a Bachelor's degree in Information Engineering from Beijing Institute of Technology and a Master's degree in Computer Science from Eastern Michigan University.

His research focuses on social computing and business analytics, including the analysis, modeling, mining, and simulation of social/business networks and social media. His research has been published in seven peer-reviewed journal articles (four as the lead author) and presented at sixteen academic conferences or workshops. He was also the recipient of the Network Science Exploration Research Grant by the Penn State University.