STATISTICAL MODELING OF MULTI-DIMENSIONAL KNOWLEDGE DIFFUSION NETWORKS: AN ERGM-BASED FRAMEWORK

by

Shan Jiang

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2015
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Shan Jiang, titled Statistical Modeling of Multi-dimensional Knowledge Diffusion Networks: An ERGM-based Framework and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_______________________________________________________________________
Hsinchun Chen
Date: (04/06/2015)

_______________________________________________________________________
Paulo B. Goes
Date: (04/06/2015)

_______________________________________________________________________
Susan A. Brown
Date: (04/06/2015)

_______________________________________________________________________
Mingfeng Lin
Date: (04/06/2015)

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

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

_______________________________________________________________________
Dissertation Director: Hsinchun Chen
Date: (04/06/2015)
STATEMENT BY AUTHOR

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

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Shan Jiang
ACKNOWLEDGMENTS

I am grateful to my advisor, Dr. Hsinchun Chen, for his guidance and encouragement throughout my doctoral study at the University of Arizona. He has played an integral role in my scholarly development. My experience in Artificial Intelligence Lab is invaluable to me. He has taught me not only how to succeed in academia, but also how to grow as a person. I am also thankful to my committee members Dr. Paulo B. Goes, Dr. Susan A. Brown, and Dr. Mingfeng Lin, for their guidance and direction over the years. Additionally, I would like to thank Cathy Larson for her help with everything in the lab and for her constant encouragement.

My dissertation has been partially supported by National Science Foundation (NSF) programs: "EAGER: Long-term View on Nanotechnology R&D as Reflected in Scientific Papers, Patents, and NSF Awards," (CMMI-1057624), "EAGER: Two Decades of Nanotechnology Development: Global Competitive Landscape and Knowledge Diffusion via ERGM and SIR Analysis," (CMMI-1249210), and "EAGER: A Systems Approach for Identification and Evaluation of Nanoscience and Nanomanufacturing Opportunities and Risks," (CMI-1442116). I would like to thank my friends and colleagues at the Artificial Intelligence Lab: Yukai Lin, Qiang Gao, Victor Benjamin, David Zimbra, Xiao Liu, Hongyi Zhu, Shuo Yu, Shuo Zeng, Jaebon Son, Andy Pressman, Ximiing Yu, Chun-Neng Huang, Yulei Zhang, and Yan Dang for their support.

Lastly, I would like to thank my parents Chunbo Jiang and Lihui Du, my sister Meijia Jiang, my other family members Renee Yang and Coconut for their constant support and encouragement. This dissertation would not have been possible without them.
DEDICATION

The dissertation is dedicated to my parents
for their unconditional love and support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>12</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>12</td>
</tr>
<tr>
<td>1.2 A General Research Framework</td>
<td>16</td>
</tr>
<tr>
<td>1.3 Areas of Study</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER 2. KNOWLEDGE DIFFUSION NETWORKS IN COMMERCIAL TECHNOLOGY: THE INFLUENCE OF NETWORK STRUCTURE ON KNOWLEDGE DIFFUSION</td>
<td>26</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Related Work</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1 Knowledge Diffusion in Commercial Technology</td>
<td>28</td>
</tr>
<tr>
<td>2.2.2 Application of ERGM</td>
<td>35</td>
</tr>
<tr>
<td>2.3 Research Gaps, Research Questions, and Research Hypotheses</td>
<td>37</td>
</tr>
<tr>
<td>2.3.1 Research Gaps</td>
<td>37</td>
</tr>
<tr>
<td>2.3.2 Research Questions</td>
<td>38</td>
</tr>
<tr>
<td>2.3.3 Research Hypotheses</td>
<td>38</td>
</tr>
<tr>
<td>2.3.3.1 Interactions between Knowledge Co-creation and Knowledge Transfer</td>
<td>38</td>
</tr>
<tr>
<td>2.3.3.2 The Roles of Inventors in Key Positions</td>
<td>40</td>
</tr>
<tr>
<td>2.3.3.3 The Impact of Public Funding</td>
<td>41</td>
</tr>
<tr>
<td>2.4 Research Framework</td>
<td>42</td>
</tr>
<tr>
<td>2.4.1 Patent Data Collection</td>
<td>42</td>
</tr>
<tr>
<td>2.4.2 Inventor Network Extraction</td>
<td>43</td>
</tr>
<tr>
<td>2.4.3 ERGM Analysis</td>
<td>44</td>
</tr>
<tr>
<td>2.5 Results and Discussions</td>
<td>48</td>
</tr>
<tr>
<td>2.5.1 ERGM Parameter Estimation</td>
<td>48</td>
</tr>
<tr>
<td>2.5.2 Goodness-of-fit Tests</td>
<td>53</td>
</tr>
<tr>
<td>2.6 Conclusions</td>
<td>54</td>
</tr>
<tr>
<td>CHAPTER 3. MULTI-DIMENSIONAL KNOWLEDGE DIFFUSION NETWORKS FOR SCIENTIFIC INNOVATION: THE EFFECT OF INDIVIDUAL ATTRIBUTES</td>
<td>58</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>58</td>
</tr>
<tr>
<td>3.2 Related Work</td>
<td>59</td>
</tr>
<tr>
<td>3.2.1 Research Objectives</td>
<td>60</td>
</tr>
<tr>
<td>3.2.2 Attributes of Knowledge Artifacts</td>
<td>61</td>
</tr>
<tr>
<td>3.2.3 Attributes of Knowledge Entities</td>
<td>63</td>
</tr>
<tr>
<td>3.2.4 Types of Knowledge Flows</td>
<td>65</td>
</tr>
<tr>
<td>3.2.5 Research Gaps</td>
<td>66</td>
</tr>
<tr>
<td>3.3 Theory and Research Hypotheses</td>
<td>67</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1.1 Typical Configurations in ERGM................................................................. 19
Table 2.1 Selected Recent Studies on Knowledge Diffusion Networks of Patent Inventors
........................................................................................................................................... 29
Table 2.2 Selected Recent Knowledge Diffusion Studies using ERGM .................... 35
Table 2.3 1991-2000 ERGM Analysis Results............................................................ 48
Table 2.4 2001-2010 ERGM Analysis Results............................................................ 49
Table 2.5 Goodness-of-fit Test Results ........................................................................ 54
Table 2.6 A Summary of Hypothesis testing Results.................................................... 56
Table 3.1 Selected Recent Individual-level Knowledge Diffusion Studies ............... 60
Table 3.2 Statistics of Scientific Publication Data from WoS...................................... 77
Table 3.3 Research Hypotheses, ERGM Configurations, and Illustrations.............. 81
Table 3.4 ERGM Analysis: Entire Knowledge Diffusion Network............................ 85
Table 3.5 ERGM Analysis: Models with Different Levels of Knowledge Impact ....... 88
Table 3.6 ERGM Analysis: Models with Different Levels of Knowledge Age .......... 90
Table 4.1 Selected Recent Studies on Social Media Networks .................................... 97
Table 4.2 NATERGM Temporal Patterns for Directed Networks: Reciprocity ....... 109
Table 4.3 NATERGM Temporal Patterns for Directed Networks: K-out-star ........... 109
Table 4.4 NATERGM Temporal Patterns for Directed Networks: K-in-star ............ 109
Table 4.5 NATERGM Temporal Patterns for Directed Networks: Transitivity ......... 110
Table 4.6 NATERGM Temporal Patterns for Directed Networks: Cyclicity .......... 111
Table 4.7 Specification of NATERGM Terms .............................................................. 113
Table 4.8 Data Collection From WikiAnswer.com .................................................... 118
Table 4.9 Parameter Estimates for TERGM and NATERGM ................................. 123
Table 4.10 Wilcoxon Signed Rank Tests for Prediction Errors ................................. 126
LIST OF FIGURES

Figure 1.1 Comparison between Uni-dimensional and Multi-dimensional Networks ..... 15
Figure 1.2 A General Framework for Analyzing Multi-dimensional Networks .......... 16
Figure 1.3 Areas of Study .................................................................................. 22
Figure 2.1 Analytical Framework of Knowledge Diffusion Networks among Inventors 42
Figure 2.2 Research Hypotheses and Corresponding ERGM Configurations ........ 47
Figure 2.3 Geographic Mapping of Nanotechnology Knowledge Co-creation-Transfer Co-occurrences ............................................................ 51
Figure 3.1 A Comparison Between Knowledge Transfer and Knowledge Co-creation... 66
Figure 3.2 Research Model for Scientific Knowledge Diffusion ........................... 69
Figure 3.3 Knowledge Flow Identification from Scientific Publication Data .......... 80
Figure 4.1 Different Processes Leading to the Same Network Pattern .................. 94
Figure 4.2 Three Different Temporal Patterns Derived From Transitivity ............ 102
Figure 4.3 Network Temporal Randomization .................................................. 104
Figure 4.4 NATERGM Framework .................................................................. 106
ABSTRACT

Knowledge diffusion networks consist of individuals who exchange knowledge and knowledge flows connecting the individuals. By studying knowledge diffusion in a network perspective, it helps us understand how the connections between individuals affect the knowledge diffusion processes.

Existing research on knowledge diffusion networks mostly adopts a uni-dimensional perspective, where all the individuals in the networks are assumed to be of the same type. It also assumes that there is only one type of knowledge flow in the network. This dissertation proposes a multi-dimensional perspective of knowledge diffusion networks and examines the patterns of knowledge diffusion with Exponential Random Graph Model (ERGM) based approaches. The objective of this dissertation is to propose a framework that effectively addresses the multi-dimensionality of knowledge diffusion networks, to enable researchers and practitioners to conceptualize the multi-dimensional knowledge diffusion networks in various domains, and to provide implications on how to stimulate and control the knowledge diffusion process.

The dissertation consists of three essays, all of which examine the multi-dimensional knowledge diffusion networks in a specific context, but each focuses on a different aspect of knowledge diffusion. Chapter 2 focuses on how structural properties of networks affect various types of knowledge diffusion processes in the domain of commercial technology. The study uses ERGM to simultaneously model multiple types of knowledge flows and examine their interactions. The objective is to understand the impacts of network structures on knowledge diffusion processes. Chapter 3 focuses on
examining the impact of individual attributes and the attributes of knowledge on knowledge diffusion in the context of scientific innovation. Based on social capital theory, the study also utilizes ERGM to examine how knowledge transfer and knowledge co-creation can be affected by the attributes of individual researchers and the attributes of scientific knowledge. Chapter 4 considers the dynamic aspect of knowledge diffusion and proposes a novel network model extending ERGM to identify dynamic patterns of knowledge diffusion in social media. In the proposed model, dynamic patterns in social media networks are modeled based on the nodal attributes of individuals and the temporal information of network ties.
1.1 Background

“Scientia potestas est (Knowledge itself is power).” Since the famous philosopher, Francis Bacon, expressed the importance of knowledge in his writing in 1597, history has demonstrated the crucial roles of knowledge in changing peoples’ life and shaping our society. However, it is essential to recognize that the values of knowledge are not embodied without diffusing to the wide public. In past literature, researchers have summarized three major functions of knowledge diffusion. First, knowledge diffusion helps preserve the knowledge (Poorna et al. 2014; Wahl et al. 2009). Through diffusion, knowledge is acquired and learned by more individuals and passed from generation to generation. Second, knowledge can be potentially converted into practice when it is diffused to practitioners. It is then that knowledge is “moved into use” (Gagnon 2011). Third, by passing to others, knowledge can be improved over time and evolve into new innovation (Arthur 2011; Corredoira and Banerjee 2014). Therefore, knowledge diffusion plays a critical role in the modern society. Researchers in various fields have shown great interest in understanding the mechanisms and outcomes of knowledge diffusion (Hassan and Haddawy 2013; Liu et al. 2014; Majchrzak et al. 2013; Wasko and Faraj 2005).

In recent years, knowledge diffusion is also receiving growing attention from information systems (IS) researchers because information technology has given birth to cyber-infrastructure (CI) for knowledge management (Brown et al. 2010b; Hey and Trefethen 2005; Wright and Wang 2011). The knowledge CI is a collection of technology, machines, systems, and people that are interconnected for better knowledge processing.
For example, in the science domain, online scientific knowledge databases have greatly enhanced researchers’ access to scientific knowledge and enabled better management of scientific literature. For another example, numerous online wiki-based knowledge management platforms have emerged recently, where people can seek, share, organize, and continuously improve various kinds of knowledge. The knowledge CI positively affects knowledge diffusion in two ways. First, it facilitates the diffusion and reuse of knowledge in organizations, across organizational boundaries, and in virtual communities (Zhang and Venkatesh 2013). Second, through analyzing the electronic records, it enables the tracking of knowledge flows that are otherwise though to be “invisible” (Alcacer and Gittelman 2006).

Among the plethora of knowledge diffusion studies, many researchers have adopted a network perspective of knowledge diffusion and have used social network analysis (SNA) to understand the mechanisms of knowledge diffusion. In this perspective, the process of knowledge diffusion is modeled by network graphs that consist of nodes which represent individuals who exchange knowledge with one another, and network ties/links which represent the flows of knowledge between the nodes. The major advantage of studying knowledge diffusion in a network perspective is that it reveals how the relationships between individuals affect the roles they play in diffusing the knowledge. The SNA approaches can also take into account the interdependencies between different social relations. As a result, an increasing number of researchers have used SNA to investigate knowledge diffusion between individuals, within organizations, and across

One of the limitations that stands out in the past knowledge diffusion literature is that few studies have realized and addressed the multi-dimensionality of knowledge diffusion networks. The concept of network multi-dimensionality maintains that network nodes can be people as well as “nonhuman agents” such as documents, datasets, analytic tools, and concepts, whereas the network ties can also be various types of relationships between these heterogeneous nodes (Contractor 2009). Multi-dimensional networks are any networks that consist of more than one representation of nodes and/or ties (Contractor et al. 2011; Jiang et al. 2014b; Li et al. 2014). Figure 1.1 illustrates this idea by comparing a conventional mono-dimensional network with a multi-dimensional network. In a mono-dimensional network, all nodes and ties are of the same type and interchangeable. By contrast, in a multi-dimensional network, nodes can have different types (represented by different shapes in Figure 1.1) and the same type of nodes can vary in attributes (represented by colors). In addition, nodes can be connected by different types of ties (represented by line thickness). By this definition, knowledge diffusion networks are also multi-dimensional because the individuals who exchange knowledge can vary in a wide array of attributes, such as their positions in the network and personal characteristics. Also, there can be more than one type of knowledge flows connecting these individuals. However, most knowledge diffusion studies did not take into account the network multi-dimensionality. As a result, the impact of these varying factors on knowledge diffusion processes is not fully understood.
The objective of this research is to propose a framework that effectively addresses the multi-dimensionality of knowledge diffusion networks. The framework conceptualizes the multi-dimensional knowledge diffusion networks in various domains, suggests an effective approach to examine the mechanisms of knowledge diffusion, and provides implications on how to stimulate and control the knowledge diffusion process. In light of these objectives, the following overarching research questions highlight the essence of this research:

- How should we conceptualize and model the multi-dimensionality of knowledge diffusion networks in various domains?
- What aspects of network multi-dimensionality affect the patterns of knowledge diffusion, and how do they affect?
• How can we effectively analyze the patterns of knowledge diffusion in multi-dimensional networks and explain underlying mechanisms?

1.2 A General Research Framework

Figure 1.2 illustrates the general framework used in this research to analyze multi-dimensional knowledge diffusion networks. Each component will be discussed next.

<table>
<thead>
<tr>
<th>Multi-dimensional Network Construction</th>
<th>Network Reduction</th>
<th>ERGM Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Node Attributes</td>
<td>▶ Intensity Modeling</td>
<td></td>
</tr>
<tr>
<td>▶ Attributes of Knowledge</td>
<td>▶ Network dichotomization</td>
<td></td>
</tr>
<tr>
<td>▶ Knowledge Flows Types</td>
<td>▶ Network Segmentation</td>
<td></td>
</tr>
<tr>
<td>▶ Hypotheses Transformation</td>
<td>▶ Parameter Estimation</td>
<td></td>
</tr>
<tr>
<td>▶ Parameter Estimation</td>
<td>▶ Goodness-of-fit Test</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.2 A General Framework for Analyzing Multi-dimensional Networks

In multi-dimensional network construction, knowledge diffusion networks are extracted from various knowledge CIs and the multi-dimensionality is modeled. Examples of the knowledge CIs can be patent repositories such as United States Patent and Trademark Office (USPTO), scientific publication databases (Chen et al. 2013b), and online platforms for knowledge communities (Xu and Zhang 2013). Different types of relations between individuals are extracted, which serve as the basis for modeling various types of knowledge flows. For example, a citation (or quotation) in an article can be used to suggest a potential flow of knowledge between two individuals, while collaboration in creating knowledge may imply a process through which knowledge is co-created. At the same time, various types of individual attributes are evaluated. For instance, the number
of articles written by an individual in the community could be used to reflect the level of activeness. Furthermore, different attributes of knowledge can be modeled, such as its quality and impact based on the number of times it is cited or viewed by others. The output of this module is the multi-dimensional representation of knowledge diffusion networks.

In network reduction, the major purpose is to provide an effective way to model different flows of knowledge diffusion. It also lessens the computational burden of ERGM applied in later steps. Relevant approaches include intensity modeling, network dichotomization, and network segmentation. The intensity modeling and dichotomization approaches examine the frequencies that individuals build relations with each other, and identify only the intense links as network ties that represent knowledge flows. The underlying assumption of this approach is that not all the relations are equivalent to the flows of knowledge (Alcacer and Gittelman 2006; Roach and Cohen 2013). For example, a paper citation may not represent the actual knowledge flow from the cited researcher to the citing researcher, because the citing individual may have limited participation in the publication, and be even unaware of the citation. To better model the knowledge flows, the intensity of relations must be taken into account. A network tie between individual nodes is removed if its intensity is below a predetermined threshold value. The threshold values for dichotomization can be determined based on mean tie intensities plus several standard deviations, following prior work (Lomi and Pallotti 2012). Another way to reduce the network is to segment the network into multiple densely connected and self-contained subpopulations (Zhang et al. 2013). The knowledge diffusion patterns in each
subpopulation can be examined individually, and the observations can be combined to infer the general patterns in the overall network.

In ERGM analysis, we apply ERGM or its extensions to the reduced networks to identify knowledge diffusion patterns. ERGMs are statistical models that can be used to test whether the observed networks exhibit theoretically hypothesized structural tendencies (Brashears 2104; Robins et al. 2007a; Wasserman and Pattison 1996). These structural tendencies, or network patterns, are called configurations in ERGM. Technically, a configuration is a subset of nodes and ties in the network, reflecting a certain type of network sub-structure. Table 1.1 lists some examples of typical configurations in ERGM, including “reciprocity” (a pair of nodes connected by a pair of directed ties that have opposite directions), “triangle” (three nodes connected to each other), and “k-star” (a central node is connected to k peripheral nodes). In addition, nodal attributes can be incorporated in a configuration, such as “homophily” (two nodes of the same type are connected), “sender” (a directed tie originating from a node with a particular attribute), and “receiver” (a directed tie ending to a node with a particular attribute), thereby making ERGM particularly suitable for analyzing multi-dimensional networks. Given an observed network, the primary task of ERGM is to examine which configurations appeared statistically more than by chance. The statistical significance will lead to support of the underlying hypotheses that explain the formation of corresponding network patterns. ERGM has been widely used in management, healthcare, and social science studies (Contractor et al. 2006; Ellwardt et al. 2012; Gondal 2011). It has recently
received attention in IS research as well (Faraj and Johnson 2011; Jiang et al. 2013; Jiang et al. 2014b; Su and Contractor 2011).

Table 1.1 Typical Configurations in ERGM

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Illustration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge</td>
<td></td>
<td>Density or asymmetrical relationship</td>
</tr>
<tr>
<td>reciprocity</td>
<td></td>
<td>A tendency for reciprocated ties</td>
</tr>
<tr>
<td>k-in-star</td>
<td></td>
<td>In-degree distribution of a node and a tendency for popularity</td>
</tr>
<tr>
<td>k-out-star</td>
<td></td>
<td>Out-degree distribution of a node which reflects social activity or expansiveness</td>
</tr>
<tr>
<td>Alternating k-triangle</td>
<td></td>
<td>Propensity for ties to form transitive triad or multiple transitive triads</td>
</tr>
<tr>
<td>Alternating independent k-path</td>
<td></td>
<td>Propensity to involve multiple short paths between actors</td>
</tr>
<tr>
<td>sender</td>
<td></td>
<td>A tendency for ties to be sent from nodes with a specific attribute (darker node) to any node (lighter node)</td>
</tr>
<tr>
<td>receiver</td>
<td></td>
<td>A tendency for ties to be sent from any node (lighter node) to nodes with a specific attribute (darker node)</td>
</tr>
<tr>
<td>homophily</td>
<td></td>
<td>A tendency for nodes with a specific attribute to form ties among themselves</td>
</tr>
</tbody>
</table>
Given a network \( Y \) of \( N \) nodes, let \( Y_{ij} \) be a tie variable with \( Y_{ij}=1 \) if there is a network tie from node \( i \) to node \( j \), and \( Y_{ij}=0 \) otherwise. The network \( Y \) can be represented by a matrix with \( Y_{ij} \) as its elements. The matrix for an undirected network is symmetric and a directed one is asymmetric. A configuration is a subset of nodes and ties among them, reflecting a certain type of structural tendency of the network, which can be represented by sub-matrices of \( Y \). Given an observed network \( y \), the mathematical definition of ERGM can be specified by the following equation:

\[
\Pr(Y = y) = \left( \frac{1}{\kappa} \right) \exp \left\{ \sum_{A} \eta_{A} g_{A}(y) \right\} 
\]

(1.1)

In (1.1):

- The summation is over all configurations that need to be tested;

- \( \eta_{A} \) is a parameter corresponding to the configuration \( A \), positively related to its tendency to occur;

- \( g_{A}(y) \) is network statistics corresponding to configuration \( A \), present in the observed network \( y \);

- \( \kappa \) is a normalizing constant ensuring that \( \Pr(Y) \) is a probabilistic distribution.

The primary task is to estimate the parameters \( \eta_{A} \) given the observed network \( y \). If \( \eta_{A} \) is positive, it indicates that the pattern \( A \) is more likely to occur than by chance, while a negative parameter estimate indicates that the pattern \( A \) rarely occurs. For
estimation technique, Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE) (Snijders 2002; Wasserman and Robins 2005) has been mostly adopted in prior ERGM research.

To evaluate whether the estimated ERGM models fit to the characteristics of the observed network, goodness-of-fit (GOF) tests can be conducted. In GOF tests, a number of sample networks are generated based on the estimated ERGM parameters. The network statistics including configurations and other basic network measures (e.g., standard deviation and skewness of the in/out-degree centrality for nodes) are compared. Small differences would indicate that the ERGM model fits the observed data well (Huffaker et al. 2009; Wang et al. 2009b).

1.3 Areas of Study

In order to address the research questions about multi-dimensional knowledge diffusion networks, we have conducted several studies to explore knowledge diffusion networks in various domains. This dissertation presents these studies in three essays, all of which attempting to answer the aforementioned questions in a specific context, but each focusing on a different aspect of the knowledge diffusion mechanisms, as shown in Figure 1.3.
In Chapter 2, we focus on how structural properties of networks affect various
types of knowledge diffusion processes in the domain of commercial technology. Patent
data were used to model knowledge diffusion networks of inventors. Understanding the
knowledge diffusion networks of patent inventors can help governments and businesses
effectively use their investment to stimulate commercial science and technology
development. This chapter contends that two types of knowledge flows, namely
knowledge transfer and knowledge co-creation, interact to form knowledge diffusion
networks. We also observed that individuals in key network positions play critical roles
in knowledge diffusion. The study utilized ERGM to simultaneously model knowledge
transfer and knowledge co-creation processes, examined their interactions, and evaluated
the impacts of network structures and public funding on knowledge diffusion networks
among inventors. Experiments were conducted on a longitudinal dataset that covered two
decades (1991-2010) of nanotechnology-related USPTO patents. The results showed that
knowledge co-creation and knowledge transfer were closely inter-related. High degree
centrality and boundary spanning inventors played significant roles in the knowledge
diffusion networks. In addition, NSF funding positively affected knowledge co-creation
despite its small fraction in overall funding and upstream research topics.

In Chapter 3, the focus is on the impact of individual attributes and the attributes
of knowledge on knowledge diffusion. An empirical study was conducted in the context
of scientific innovation, using scientific publication data from Thomson Reuter’s Web of
Science (WoS) databases. The scientific knowledge diffusion networks are multi-
dimensional, and various factors interact in a complex way to affect the diffusion process.
The study explains the complex knowledge diffusion mechanisms based on social capital
theory. Specifically, the study utilized ERGM to examine how knowledge transfer and
knowledge co-creation can be affected by the attributes of individual researchers and the
attributes of scientific knowledge. Results showed that social capital factors such as
collaboration experience, credibility, and similarity in career length positively affect
knowledge diffusion. Researchers’ activity level affects knowledge co-creation positively,
but not knowledge transfer. The relationships between the activity level and knowledge
diffusion can be moderated by the impact and age of scientific knowledge: when the
knowledge is of relatively low impact, researchers are less likely to participate in
knowledge co-creation processes even if they have high activity levels; for new knowledge, active researchers have a strong tendency to seek the knowledge through knowledge transfer processes. This study provides a conceptualization of multi-dimensional knowledge diffusion networks in scientific innovation domain, and provides insight about how attributes of knowledge affect knowledge diffusion processes.

In Chapter 4, we consider the dynamic aspect of knowledge diffusion, and propose a novel network model extending ERGM to identify dynamic patterns of knowledge diffusion. A new model, Nodal Attribute-based Temporal Exponential Random Graph Model (NATERGM) was developed and tested in social media context. Social media networks are dynamic and the sequence of network ties is an important aspect of the network dynamics. More specifically, knowledge flows between online users develop in a sequence. The proposed model focuses on how the attributes of online users affect the sequence in which the knowledge flows develop. Dynamic patterns in social media networks were modeled based on the nodal attributes of individuals and the time information of network ties. Some of the nodal attributes were evaluated using text mining techniques. Using social media data collected from an online knowledge sharing community, empirical tests were conducted to evaluate the performance of the NATERGM on identifying the dynamic patterns and predicting the characteristics of the future networks. Results showed that the NATERGM demonstrated an enhanced pattern testing capability and an increased prediction accuracy of network characteristics compared to benchmark models. The proposed NATERGM model helps explain the roles
of individual attributes in the dynamic formation process of online knowledge diffusion networks.

Chapter 5 summarizes the major research contributions, discusses relevance to MIS research, and highlights some interesting future directions.
CHAPTER 2. KNOWLEDGE DIFFUSION NETWORKS IN COMMERCIAL TECHNOLOGY: THE INFLUENCE OF NETWORK STRUCTURE ON KNOWLEDGE DIFFUSION

2.1 Introduction

Knowledge diffusion in the domain of commercial technology is important for modern society. In particular, the nanotechnology field may be most representative of the status in the modern commercial technology diffusion. Nanotechnology has been recognized as a key foundation technology for economy and society (Wang and Shapira 2011). Nanotechnology has also been recognized as an indicator of a country’s technology competence and has become a national priority for many industrialized countries since 2000 when the United States announced the National Nanotechnology Initiative (NNI), which led to a significant increase of public funding for nanotechnology from $464M in 2001 and to about $1,800M in 2012 (Roco et al. 2011). Nanotechnology-related patents are one of the tangible outputs of nanotechnology innovation. Private sector has exceeded government R&D investment in nanotechnology in 2006 and has, together with government, interest in making their investment more effective. Understanding the formation and evolution of knowledge diffusion networks among nanotechnology-related patent inventors will improve understanding of how to use funding effectively and stimulate the nanotechnology development.

The advancement of information technology has enabled access to a huge amount of electronic patent data online, such as those in USPTO. As a result, studying knowledge
diffusion through patent data has been increasingly appreciated by researchers (Chen et al. 2013a; Chen et al. 2013b; Dang et al. 2009; Huang et al. 2006; Huang et al. 2005; Li et al. 2007). However, such knowledge diffusion networks usually consist of thousands of inventors and their complex relationships. The complexity of the problem is further increased by the “multi-dimensionality” of the networks: different types of inventors exist in the networks, such as publicly funded inventors and unfunded ones; and multiple types of knowledge diffusion flows exist between these inventors, including knowledge co-creation which occurs among inventors through collaboration, and knowledge transfer where knowledge flows start from cited inventors to citing ones. Therefore, an analytical approach that is capable of examining the complex interaction between network elements has been called for.

In this essay we used ERGM to characterize and understand how inventors in the nanotechnology field form different types of knowledge flows. We systematically evaluated how different types of knowledge flows interact with each other, and how public funding affected those knowledge diffusion processes. The aim is to model the knowledge diffusion networks of nanotechnology inventors, to understand how different types of knowledge flows interact, to understand how structural properties of network affect various types of knowledge diffusion processes, and to provide suggestions to funding agencies for their investment decisions.

The remainder of this chapter is organized as follows. Section 2.2 provides a review of relevant literature on commercial technology knowledge diffusion and application of ERGM to knowledge diffusion networks. Section 2.3 presents research
gaps, research questions, and research hypotheses. Section 2.4 presents the research framework and experiment design in this study, followed by results and discussions in Section 2.5. Section 2.6 makes conclusions and discusses implications.

2.2 Related Work

2.2.1 Knowledge Diffusion in Commercial Technology

Knowledge diffusion has been studied extensively in prior informatics research (Abbasi et al. 2012; Liu et al. 2011; Nerkar and Paruchuri 2005; Reinholt et al. 2011; Su and Contractor 2011; Tang and Hu 2013; Yan and Ding 2012). One of the most common approaches is to model the knowledge diffusion process via knowledge diffusion networks. In such networks, nodes represent individuals that create or own the knowledge, and ties represent knowledge flows between the individuals. The network-based approach to modeling and studying the knowledge diffusion process provides a comprehensive understanding of the overall structure of a knowledge domain and enriches the potential utility of analyzing the knowledge diffusion processes.

Knowledge diffusion networks in the commercial technology field have also been extensively studied. Table 2.1 summarizes these studies based on the types of inventor networks, measures used for network characterization, and analytical techniques applied to study the networks. We will discuss each of these points in detail in next.
Two types of knowledge diffusion networks have been investigated in prior studies. In *knowledge co-creation networks*, a network tie between a pair of nodes...
represents that two individuals are sharing common knowledge. The knowledge is usually co-created through collaboration, based on which two inventors establish a co-invention or co-patenting relationship (Breschi and Lissoni 2004; Gao et al. 2011; Liu et al. 2011; Nerkar and Paruchuri 2005). During the co-invention process, they exchange important knowledge relevant to the content of their publications (Breschi and Lissoni 2004). There is no positional difference between the inventors and thus the network is undirected. Evidence has shown that collaborative work often has a stronger impact than individual work (Hoekman et al. 2009). Consequently, the average number of inventors that contribute to a patent has increased over the past 20 years (Fleming and Frenken 2007), and knowledge co-creation networks among inventors have been increasing in size and density.

In knowledge transfer networks, a network tie represents a directed knowledge flow from one inventor to another. There is a temporal precedence between when two nodes obtain the knowledge. By citing references in a patent, an inventor acknowledges that a part of his/her knowledge is built based on others’ work. Therefore, patent citations have been used to model knowledge transfer from the cited inventors to citing inventors (Autant-Bernard et al. 2013; Breschi and Lissoni 2004; Cho and Shih 2011; Hsueh and Wang 2009; Huang et al. 2006; Huang et al. 2005; Li et al. 2007; Qiao et al. 2014). A question typically raised is whether patent citations accurately represent knowledge transfer processes. As many citations present in a patent are determined by patent examiners rather than inventors themselves, using patent citations as proxies of knowledge flows between inventors has been questioned (Alcacer and Gittelman 2006;
However, it has been argued that patent citations still account for significant knowledge transfer. As reported in Bresch and Lissoni (2004), more than 40% of citing inventors suggested that they knew about the technologies underlying cited patents before or while working on their own invention, about 25% answered that they learnt them after the invention, and only 25% were totally unaware of the cited work. Such recognition was clearer when the cited works are in the same technology field. Moreover, examiner-inserted citations can be differentiated from citations added by inventors in USPTO since 2001 (Alcacer and Gittelman 2006). Therefore, patent citations are still useful for identifying knowledge transfer.

Prior research has shown that both structural properties of the knowledge diffusion networks and individual attributes of network nodes can affect the knowledge diffusion processes. Structural properties focus on the topological characteristics of the network and the relationships between inventors. The most common metrics are centrality and structural holes. Degree centrality has been often used in previous studies (Cho and Shih 2011; Gao et al. 2011; Huang et al. 2006; Huang et al. 2005; Li et al. 2007; Liu et al. 2011). The degree centrality of an inventor node is the number of other inventor nodes directly connected to it. Based on centrality measures, influences of knowledge co-creation on knowledge transfer processes have been studied. For example, in knowledge co-creation networks, inventors with high degree centrality generally have high collaboration activities with others and their knowledge is more likely to transfer to others (Liu et al. 2011). Structural holes measure the extent to which a node is
concatenating sub-components of a network that are not inter-connected (Burt 1995).

Based on structural holes, inventors that play boundary-spanning roles can be identified. When an inventor is situated between inventors that are not directly inter-connected, the inventor controls the knowledge flows in the network with greater efficiency. Consequently, the extent of structural holes spanned by an inventor was found to be positively associated with the likelihood of the inventor’s knowledge being cited by others (Liu et al. 2011; Nerkar and Paruchuri 2005).

Individual attributes of network nodes focus on inherent characteristics of the nodes regardless of their positions in the network. The individual attributes have been found to affect the knowledge flows handled by the node. Examples include regional areas of inventors (Gao et al. 2011), institutions through which the patents are filed (Hsueh and Wang 2009), patents’ relevant technology fields (Cho and Shih 2011; Li et al. 2007), and availability of public funding to inventors (Huang et al. 2006; Huang et al. 2005). Often, these individual attributes have been combined with structural properties of the networks to discover various knowledge diffusion patterns. For example, Hsueh and Wang (2009) analyzed patents in the liquid crystal display (LCD) field and identified key institutions and countries that were associated with nodes of high degree centrality. Gao et al. (2011) combined the degree centrality and geographic information of patents and concluded that the most active knowledge exchanges occurred among the most advanced provinces in China from 1985 to 2007.

Different from these attributes, the availability of public funding is a factor that can be controlled by funding agencies to influence knowledge diffusion processes.
Therefore we focus on the influences of public funding on nanotechnology. The fact that 73% of all papers cited by the US industry patents have origins in publicly funded projects indicates the critical role of public funding on commercial technology development (Narin 1998). It is shown that public funding stimulates research activities. For example, Adams and Griliches (1998) identified a strong correlation between research output and public funding. Federal funding for universities has resulted in an increased number of papers and patents (Payne and Siow 2003). Although these studies indicated positive effects of public funding on the knowledge generation process, little has been done to examine the influences of public funding on knowledge diffusion.

Various techniques have been used to analyze patent knowledge diffusion networks. The most common one has been to characterize the networks with descriptive measures. Relevant studies usually ranked the nodes of networks based on some metrics. For example, based on US Patent Classification Codes and citation counts, important technology fields were identified in knowledge diffusion networks (Cho and Shih 2011; Hsueh and Wang 2009; Li et al. 2007). A similar approach was used to identify key countries and institutions (Hsueh and Wang 2009; Li et al. 2007) or geographical areas (Gao et al. 2011). In addition to citation counts, several other network characterization techniques have been used to evaluate the importance of nodes. For instance, the HITS algorithm was applied to calculate hub and authority scores that accounted for network nodes’ roles in information flows (Hsueh and Wang 2009; Huang et al. 2005). Blockmodel analysis was used to group structurally equivalent nodes in the knowledge diffusion network (Gao et al. 2011).
Another stream of research has used statistical analysis to study more general patterns of knowledge diffusion. For example, the Analysis of Variance (ANOVA) method was used to compare the impacts of publicly funded inventors and other groups of inventors (e.g., inventors from top universities) in knowledge diffusion networks (Huang et al. 2006; Huang et al. 2005). Logistic regression was used to analyze the impact of social distances on the occurrence of patent knowledge transfer ties (Breschi and Lissoni 2004). Moreover, a class of survival analysis techniques in statistics has been used to take time-varying effects into account. For instances, the Cox Proportional Hazards model was used to examine how patent citation rates are influenced by network structures over time (Liu et al. 2011; Nerkar and Paruchuri 2005).

Although these techniques have successfully revealed interesting findings, none of them is suitable for modeling the complete knowledge diffusion processes in a multi-dimensional network, because they do not take into account the interdependency between various knowledge diffusion processes. To analyze the complex interaction patterns of knowledge co-creation and knowledge transfer processes in such networks, a systematic and theory-grounded technique that is able to simultaneously model multiple types of nodes and edges in a network is needed. ERGM is one such technique (Robins et al. 2007a). ERGM assumes interdependencies between network ties, and tests whether the occurrence of observed network components follows some patterns or is random. By using ERGM analysis, different types of network components with interdependencies can be simultaneously modeled, leading to a parsimonious representation and a systematic exploration of the global network. In the next section, I review relevant ERGM studies.
2.2.2 Application of ERGM

Recent studies have used ERGM to study knowledge diffusion in various domains, such as citation networks (Gondal 2011), multi-cultural research collaborations (Sayogo et al. 2011), organizational information seeking (Johnson et al. 2012; Su and Contractor 2011), and medical innovation (Zappa and Mariani 2011). Table 2.2 summarizes these works based on the network dimensionality and parameter estimation methods.

Table 2.2 Selected Recent Knowledge Diffusion Studies using ERGM

<table>
<thead>
<tr>
<th>Study</th>
<th>Network Dimensionality (# of node types, # of tie types)</th>
<th>Estimation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huffaker et al. (2009)</td>
<td>&gt;2,1</td>
<td>MCMCMLE⁺</td>
</tr>
<tr>
<td>Sayogo et al. (2011)</td>
<td>&gt;3,1</td>
<td>not explained</td>
</tr>
<tr>
<td>Gondal (2011)</td>
<td>&gt;2,1</td>
<td>MCMCMLE</td>
</tr>
<tr>
<td>Su and Contractor (2011)</td>
<td>&gt;2,4</td>
<td>MCMCMLE</td>
</tr>
<tr>
<td>Zappa and Mariani (2011)</td>
<td>&gt;3,2</td>
<td>MCMCMLE</td>
</tr>
<tr>
<td>Johnson et al. (2012)</td>
<td>&gt;5,2</td>
<td>not explained</td>
</tr>
<tr>
<td>Keegan et al. (2012)</td>
<td>&gt;2,1</td>
<td>not explained</td>
</tr>
<tr>
<td>Jiang et al. (2013)</td>
<td>&gt;2,2</td>
<td>MCMCMLE</td>
</tr>
</tbody>
</table>

* Only distinct types of nodes are counted. The greater signs indicate that in these studies, the variety of nodes can be greater considering the nodal attributes, which can be binary, nominal, and continuous.

⁺ MCMCMLE=Markov Chain Monte Carlo Maximum Likelihood Estimation

ERGM is able to model multi-dimensional networks that consist of multiple types of nodes and ties. Most recent research has incorporated more than one type of nodes in the network. Some of these networks are called two-mode networks, where nodes represent different entity types. For example, Gondal (2011) built a citation network consisting of paper and author nodes, and examined citation patterns between papers and
authors. Another research studied a Wikipedia network consisting of editor and article nodes, and examined how editor experience and article quality affected the Wikipedia knowledge network (Keegan et al. 2012). In one-mode networks, node types can also be differentiated by considering their attributes. For instance, Zappa and Mariani (2011) considered physicians with different attributes in marketing pressure, receptivity to change, and absorptive capacity. They examined how these characteristics affected information-seeking processes in the medical innovation network when new drugs were launched. Huffaker et al. (2009) studied patterns of information seeking between expert online game players and non-experts. Jiang et al. (2013) examined the roles of senior researchers and productive researchers in knowledge diffusion networks. However, most prior research involved only a single type of tie between nodes. Some modeled multiple types of ties, but used separate networks without considering their interactions (Jiang et al. 2013; Johnson et al. 2012; Zappa and Mariani 2011). An exception is Su and Contractor (2011)’s work, which modeled two types of nodes (consultant and digital knowledge source) and examined how they interrelate with four types of relational ties (information seeking, expertise recognition, easy access, and social communication).

According to Robins et al. (2007a), there are two primary approaches for estimating ERGM parameters: pseudo-likelihood estimation (PLE) (Besag 1975; Strauss and Ikeda 1990) and Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) (Snijders 2002; Wasserman and Robins 2005). PLE is easy to fit to complicated models, but the estimations are not always accurate. MCMCMLE simulates new networks based on seed parameters and iteratively adjusts the parameter values
based on the difference between the generated networks and the observed network. When the differences decrease and meet certain stopping criteria, the model converges and the resulting parameters are used for the estimates. MCMCMLE deals with model degeneracy better and thus is suggested whenever possible (Robins et al. 2007a). Consequently, most recent studies have adopted the MCMCMLE method.

2.3 Research Gaps, Research Questions, and Research Hypotheses

2.3.1 Research Gaps

Only a few studies have simultaneously examined the processes of knowledge transfer and knowledge co-creation within the same knowledge diffusion network. In Bresch and Lissoni (2004), the probability of patent citation was found to be significantly affected by the social distance between two inventors, which was represented by the shortest path between them in their knowledge co-creation network. This provides a strong implication that knowledge co-creation and knowledge transfer processes are closely interrelated. Without modeling these two types of knowledge flows at the same time, we would have a limited understanding of how these processes interact to form the knowledge diffusion network among inventors in the nanotechnology field. ERGM enables simultaneous modeling of multiple types of ties in knowledge diffusion networks. In spite of this, little has been done in this direction. Moreover, prior research examining structural properties of knowledge diffusion networks did not take into account how interactions between knowledge co-creation and transfer processes could be addressed. For example, structural holes spanned by an inventor based on knowledge co-creation
ties were found to be associated with high citation rates for the inventor (Liu et al. 2011; Nerkar and Paruchuri 2005). However, the citation rate in their study was a nodal attribute and did not reflect a flow of knowledge transfer from one inventor to another in the network. By using ERGM, configurations can be used to model the interactions between structural properties, knowledge co-creation, and knowledge transfer. In addition, although public funding has been found to positively influence knowledge generation, relatively little has been done to examine how public funding affects the processes of knowledge co-creation and transfer in nanotechnology development.

2.3.2 Research Questions

The research gaps above lead to the following research questions:

- How do knowledge co-creation and knowledge transfer processes interact to form the knowledge diffusion networks in the field of nanotechnology innovation?
- What roles do inventors in the key positions in the network play and how are they associated with knowledge co-creation and knowledge transfer processes?
- What impact does public funding have on knowledge co-creation and knowledge transfer processes in the nanotechnology field?

2.3.3 Research Hypotheses

2.3.3.1 Interactions between Knowledge Co-creation and Knowledge Transfer
Prior research has found that credibility of the knowledge source is one of the antecedents of knowledge transfer (Inkpen and Tsang 2005; Joshi et al. 2007). Knowledge co-creation experiences through collaboration are expected to increase the inventors’ mutual credibility, and thus increase the likelihood that one of them transfers knowledge to the other in future. In Bresch and Lissoni (2004)’s work, they found that the more closely and directly two inventors collaborated, the more likely patent citations would occur between them. Similarly, intensive institutional collaboration has been found to increase institutional citation behaviors (Yan and Sugimoto 2011). These findings imply that knowledge co-creation may entail knowledge transfer. Therefore, we propose the following research hypothesis:

**H1**: Knowledge co-creation and knowledge transfer are likely to co-occur between a pair of inventors.

Knowledge co-creation and knowledge transfer processes can also interact in other ways. On one hand, if two inventors have cited a common work by a third inventor, it may indicate that the two inventors have a similar scope of knowledge and areas of expertise, and thus they have high cognitive proximity (Criscuolo and Verspagen 2008), which increases the likelihood that they will collaborate and share knowledge in the future. On the other hand, if two inventors have collaborated before, they share common knowledge about key inventors and work in the field and might cite from the same inventor in their individual work.
**H2**: Knowledge co-creation between two inventors and knowledge transfers from a common inventor to these two inventors are likely to co-occur.

### 2.3.3.2 The Roles of Inventors in Key Positions

Degree centrality and structural hole measures can be used to identify inventors in key positions in the knowledge diffusion network. Previous research has found that intrinsic motivation to exchange knowledge is another antecedent of knowledge flows (Ko et al. 2005; Lin 2007). When the network ties represent knowledge transfer processes from cited inventors to citing ones, inventors with high out-degree centrality are associated with high citation counts, and hence a high citation authority (Hsueh and Wang 2009; Huang et al. 2006; Huang et al. 2005). These high authority inventors should have high intrinsic motivations to collaborate with each other to increase the chance of their innovation to be patented. Moreover, studies have shown that inventors with high degree centrality had more collaboration activities (Liu et al. 2011). Therefore, we propose the following hypothesis:

**H3**: Knowledge co-creation is likely to occur between two inventors with high degree centrality.

Inventors in boundary-spanning positions control the knowledge flows in the network with greater efficiency, with more opportunities to brokerage knowledge flows,
to bring together different knowledge streams, and to possess richer information content (Hargadon and Sutton 1997). Consequently, the extent of structural holes spanned by an inventor can positively affect the likelihood that her knowledge being selected by others (Liu et al. 2011; Nerkar and Paruchuri 2005). Therefore, we propose the following research hypothesis:

**H4**: Knowledge transfer is likely to occur from boundary-spanning inventors to others.

### 2.3.3.3 The Impact of Public Funding

Previous research has found that public funding increased research output of the funded individuals or organizations (Adams and Griliches 1998; Payne and Siow 2003). A few studies have also suggested that public funding positively affects knowledge diffusion. First, public funding facilitates the resource-seeking processes between inventors. When funding is available, individuals would have stronger intrinsic motivation to communicate and exchange information with other experts, which may further lead to intense knowledge co-creation (Ko et al. 2005; Lin 2007). Using Hyperlink-Induced Topic Search (HITS) algorithm on the citation network of nanotechnology patent inventors, Huang et al. (2005; 2006) found that inventors with NSF funding were associated with high authority scores (receiving more citations). One possible explanation is that receiving funding is a positive signal for the innovativeness and average quality of the inventor’s work. In this way the signal contributes to the
source credibility of funded inventor, which then increases the likelihood that knowledge is transferred from the inventor (Inkpen and Tsang 2005; Ko et al. 2005). Based on the discussion, two additional hypotheses are developed:

**H5:** Knowledge co-creation is likely to occur between publicly funded inventors and other inventors.

**H6:** Knowledge transfer is likely to occur from publicly funded inventors to other inventors.

2.4 Research Framework

Figure 2.1 presents our analytical framework of using ERGM to test the hypotheses proposed in this study. The major components are described in detail below.

![Diagram of Research Framework]

2.4.1 Patent Data Collection
Patent data was obtained from USPTO by using an automated spider program. The spider collected nanotechnology-related patents by using a set of keywords provided by domain experts. Searched data fields included title, abstract, and full-text of the patents. The same list of keywords and the search procedure has been used in prior nanotechnology patent studies (Huang et al. 2006; Huang et al. 2005; Jiang et al. 2013; Li et al. 2007). Twenty years of patents were collected based on their issued dates, covering 165,095 patents and 253,682 inventors in total. The data was separated into two time periods, 1991-2000 and 2001-2010, to examine how knowledge diffusion networks of nanotechnology inventors evolved over the past two decades.

2.4.2 Inventor Network Extraction

Inventor network extraction consists of two steps. First, we identified knowledge co-creation ties and knowledge transfer ties between inventors based on co-patenting and citation relationships between them. A pair of inventors was linked with a knowledge co-creation tie if they co-authored a patent. A knowledge transfer tie was established from one inventor to another if the former’s work was cited by the latter, representing the flow of knowledge. For knowledge transfer ties, self-citations and examiner-inserted citations were excluded. The numbers of knowledge co-creation and knowledge transfer ties were counted to account for intensity.

Next, we dichotomized the network ties based on the tie intensities. A knowledge co-creation tie or a knowledge transfer tie was removed if its intensity was below a threshold value. This approach provides a more effective way to model the flows of
knowledge co-creation and transfer. In a patent, contributions and levels of participation for inventors can be different. Some may have marginally participated in the invention process. Knowledge co-creation between two inventors is not warranted if they co-patented only a few times. Through intensive collaborations, knowledge is exchanged more frequently between two inventors and the validity of their knowledge co-creation process is increased. Similarly, by retaining only strong knowledge transfer ties (multiple citations) between two inventors, we believe knowledge transfer paths were modeled with better validity.

To determine the threshold values for dichotomization, we followed previous research and used mean tie intensities plus one standard deviation (Lomi and Pallotti 2012). In our dataset, the mean collaboration intensity was 1.9 with a standard deviation of 5.4 during 1991-2000, and was 1.6 with a standard deviation of 6.2 during 2001-2010. The mean citation intensity was 1.9 with a standard deviation of 7.8 during 1991-2000, and was 7.3 with a standard deviation of 3.7 during 2001-2010. Based on these statistics, we set threshold values for network dichotomization and constructed the knowledge diffusion network in each time period. Each resulting network was represented by a pair of matrices $Y_S = \{y_{ij}^S\}$, and $Y_T = \{y_{ij}^T\}$, where $y_{ij}^S$ (or $y_{ij}^T$) is the dichotomized knowledge co-creation (transfer) tie variable from inventor i to j, and takes value ‘1’ if the tie intensity was greater than the threshold value, or ‘0’ otherwise.

2.4.3 ERGM Analysis

First, based on the networks extracted in the previous step, the following measures that characterize the network were computed.
1) **Degree Centrality**: The degree centrality of each inventor was calculated based on the number of knowledge transfer ties originating from the inventor, and thus is out-degree centrality. These outward links represent knowledge transfer flows from the focal inventor. Formally, with the matrix notation introduced previously, the degree centrality of inventor i was calculated using Equation (2.1):

\[
\text{Degree}_i = \sum_j y_{ij}^T
\]  

(2.1)

Inventors with high degree centrality are cited frequently by other inventors, and tend to have higher authority (Hsueh and Wang 2009; Huang et al. 2006; Huang et al. 2005). To determine inventors with high degree centrality, we first computed the mean and standard deviation of degree centrality of all inventors. Then inventors with degree centrality being greater than the mean plus one standard deviation were coded as “high degree centrality inventors.” The percentage of high degree centrality inventors was 18.6% in 1991-2000, and 15.7% in 2001-2010.

2) **Structural Holes.** Following previous patent studies, the structural hole measure of each inventor was calculated based upon knowledge co-creation ties in the network (Liu et al. 2011; Nerkar and Paruchuri 2005). In these studies, effects of structural holes on citation the effects of structural holes on citation were similar using redundancy-based and constraint-based measurement approaches. Therefore, for simplicity we measured the structural holes only with the redundancy-based approach from Burt (1995) in Equation (2.2):

\[
\text{structural}_i = \frac{\sum_q \left[1 - \sum_{q} p_{iq} m_{jq}\right]}{C_i}, \quad q \neq i \neq j
\]  

(2.2)
where $p_{iq} = \frac{(y_{iq}^S + e_{qiq})}{\sum_{j \neq i} y_{ij}^S + y_{ji}^S}$ is the proportion of inventor $i$’s ties connected to inventor $q$, $m_{jq} = \frac{(y_{jq}^S + y_{qj}^S)}{\max_{k \neq j} y_{jk}^S + y_{kj}^S}$ is the marginal strength of the connection from inventor $j$ to inventor $q$, and $C_i = \sum_{j \neq i} e_{ij}^S$ is the number of ties connected to inventor $i$.

Inventors with higher structure hole measures connect other less connected inventors and hence play boundary-spanning roles. To determine those inventors, those with structural hole measures being greater than the mean plus one standard deviation were coded as “boundary-spanning inventors.” The percentage of boundary-spanning inventors was $22.2\%$ of all inventors in 1991-2000, and $19.4\%$ in 2001-2010.

3) Availability of Public Funding. To identify inventors supported by public funding, we used National Science Foundation (NSF) nanotechnology-related grant records from 1991-2010. The grant data was provided by the NSF and retrieved by searching with the same keywords as used for patent data. The searched fields included titles and abstracts of the grant records. In each time period, an inventor was coded as “funded” if the inventor’s name was present in the grant records of the corresponding time window. The percentage of NSF-funded inventors was $2.9\%$ of all inventors in 1991-2000, and $3.8\%$ in 2001-2010.

Next, to test the research hypotheses with ERGM, the hypotheses were transformed into ERGM configurations represented by inventor nodes, edges (knowledge co-creation ties), arcs (knowledge transfer ties), and network measures associated with the nodes. Figure 2.2 summarizes the research hypotheses and their corresponding ERGM configurations.
After the extracted networks and configurations were fed into the ERGM model, we used the MCMCMLE to estimate the parameters associated with each configuration. The values of these parameters reflect the tendencies for corresponding network configurations to occur. A positive and significant parameter indicates that the corresponding configuration occurs significantly more frequently than by chance in the network. For implementation, XPnet was used because it is designed specifically for multi-dimensional network analysis (Wang et al. 2009a).

To evaluate whether the estimated ERGM models fit to the characteristics of the observed network, goodness-of-fit tests were conducted. In goodness-of-fit tests, a number of sample networks were generated based on the estimated ERGM model and compared to the observed network. Network statistics chosen to be compared included standard deviation and skewness of the in/out-degree centrality for nodes, and the counts
of six configurations associated with the hypotheses. Small differences would indicate that the estimated model fits the observed data well.

2.5 Results and Discussions

2.5.1 ERGM Parameter Estimation

Tables 2.3 and 2.4 show the ERGM analysis results during the time periods 1991-2000 and 2001-2010 respectively. The parameter estimates, standard errors, and t-statistics when MCMC MLE process converged are presented. The ‘*’ symbol indicates that the value of the parameter estimate was at least twice the standard error, meaning that the configuration appears in the network significantly more frequently than by chance and thus the corresponding hypothesis is supported. Similar criteria have been used in prior ERGM research (Pahor et al. 2008; Su and Contractor 2011). The t-statistics for ERGM parameters, defined as the difference of observation and sample means for the corresponding configuration count, divided by standard deviation, measure the degree of model convergence (Snijders 2002). In his work, Snijders (2002) suggested that the convergence of an ERGM model is excellent if |t|<0.1, and good if |t|< 0.2. In our experiment, the absolute values of t were all less than 0.1.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Estimates</th>
<th>Std. error</th>
<th>t-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>4.69*</td>
<td>0.13</td>
<td>-0.03</td>
</tr>
<tr>
<td>H2</td>
<td>7.25*</td>
<td>0.24</td>
<td>-0.02</td>
</tr>
<tr>
<td>H3</td>
<td>0.38*</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>H4</td>
<td>1.03*</td>
<td>0.11</td>
<td>0.00</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>Estimates</td>
<td>Std. error</td>
<td>t-statistics</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>H1</td>
<td>4.25*</td>
<td>0.12</td>
<td>-0.07</td>
</tr>
<tr>
<td>H2</td>
<td>7.23*</td>
<td>0.21</td>
<td>-0.01</td>
</tr>
<tr>
<td>H3</td>
<td>0.22*</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>H4</td>
<td>1.22*</td>
<td>0.11</td>
<td>-0.07</td>
</tr>
<tr>
<td>H5</td>
<td>0.67*</td>
<td>0.19</td>
<td>-0.02</td>
</tr>
<tr>
<td>H6</td>
<td>-0.10</td>
<td>0.24</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 2.4 2001-2010 ERGM Analysis Results

Hypothesis 1 was supported in both periods, suggesting that knowledge co-creation and knowledge transfer were likely to co-occur between two nanotechnology inventors. This result is consistent with previous findings in Bresch and Lissoni (2004) that inventors in close and direct collaborative relationships are likely to cite each other’s work. In addition, our results also indicate that when the collaboration is repeating, citations are likely to repeat as well (because the network was dichotomized based on tie intensities). It suggests that repeating knowledge co-creation activities can lead to enhanced credibility of the knowledge sources, which in turn entails intense knowledge transfer activities. This observation adds a new dimension to the previous finding that source credibility antecedes knowledge transfer (Inkpen and Tsang 2005; Joshi et al. 2007).
The co-occurrence of knowledge co-creation and knowledge transfer also reflects a key aspect of nanotechnology development because it represents the part of the network where knowledge diffusion is most intense. By mapping these co-occurrences to geographic locations, we can identify nanotechnology clusters--the geographic collocations of entities (inventors, institutions, etc.) that are working prominently to drive nanotechnology development (Robinson et al. 2007). To this end, we further extracted all the patents associated with knowledge co-creation and transfer co-occurrence (i.e., co-authored patents of the two inventors; citing and cited patents for the same pair of inventors), and identified the top 100 assignees based on the number of patents in this set. We then marked the cities of these assignees on the US map. Figure 2.3 shows the geographical mapping of knowledge co-creation-transfer co-occurrences. The size of the marker is proportional to the number of patents associated with each city. We also show the top 30 assignees on the map for better visualization. We observed that during 1991-2010, the California region, the Texas region, the Great Lakes region, and the upper East Coast region are the four major nanotechnology clusters based on knowledge co-creation-transfer co-occurrences. In addition, the Florida region is becoming another cluster in recent years.
Figure 2.3 Geographic Mapping of Nanotechnology Knowledge Co-creation-Transfer Co-occurrences

Hypothesis 2 was supported in both periods, confirming that two nanotechnology inventors with a common source of knowledge transfer tended to share knowledge with each other. This result supports previous findings that cognitive proximity leads to a higher chance of knowledge exchange (Criscuolo and Verspagen 2008). Note that during network dichotomization process, we cut the ties whose intensities were below the threshold. As a result, the configuration corresponding to Hypothesis 2 only occurs when
two inventors co-patented many times, and many of their patents cited the work from a third inventor. Our results suggest that such type of triadic relationship occurs more frequently than by chance in nanotechnology knowledge diffusion networks.

Hypothesis 3 was supported in both periods, indicating that nanotechnology inventors in high degree centrality positions were likely to share knowledge with each other. The works of these inventors are repeatedly cited by many other inventors, indicating that inventors in high degree centrality positions have high authority in the network. Previous studies have shown that senior or productive academic scientist is likely to collaborate with each other for better efficiency (Jiang et al. 2013). Our results complement prior work by discovering that high authority inventors in nanotechnology field are also likely to collaborate and share knowledge with each other.

Hypothesis 4 was also supported in both periods, showing that knowledge transfer was more likely to originate from boundary-spanning nanotechnology inventors, supporting previous findings that the extent of structural holes spanned by an inventor can positively affect the likelihood that the inventor’s knowledge being selected by others (Liu et al. 2011; Nerkar and Paruchuri 2005). These inventors bring different streams of knowledge together and possess richer information content than other players in the network (Hargadon and Sutton 1997). This should be why knowledge transfer is more likely to occur from these inventors.

Hypothesis 5 was supported in 2001-2010, but not in 1991-2000. Before 2000, not many inventors were supported by NSF funding and the amount of funding was also relatively low (Roco et al. 2011). Such limitations might result that inventors not having
enough intrinsic motivation to collaborate and share knowledge with others. However, after the announcement of NNI in 2001, both the number of nanotechnology scientists receiving NSF funding and the amount of funding have increased (Roco et al. 2011). Also, the number of nanotechnology-related technology topics has increased as well (Chen et al. 2013a). With the increased level of funding and greater variety of nanotechnology topics, inventors had strong intrinsic motivation to make interdisciplinary collaboration so that they could produce high impact innovations.

Hypothesis 6 was supported in neither 1991-2000 nor 2001-2010, suggesting that NSF funding did not stimulate knowledge transfer among nanotechnology inventors. One possible explanation is that the technological relevance in patented innovation rather than the source credibility could be the primary factor of citation behaviors of inventors. Source credibility contributes to knowledge transfer mainly in academics where researchers learn from valid past literature and tend to cite the works from authorities in the field. Instead, when filing patents, inventors cite because they are legally required to provide citations to relevant works in the in prior art list. To prevent their patents to be invalidated, inventors are inclined to cite any prior arts that are related to their innovation, and hence the source credibility may play a less important role here.

2.5.2 Goodness-of-fit Tests

Table 2.5 shows the results of goodness-of-fit tests for the ERGM models. The numbers are t-statistics for the network parameters, which was calculated as the difference of observation and sample mean, divided by standard deviation. All t-statistics
were less than 2, indicating that the estimated ERGM models fit to the data well (Huffaker et al. 2009; Wang et al. 2009b).

Table 2.5 Goodness-of-fit Test Results

<table>
<thead>
<tr>
<th>Network Parameters</th>
<th>1991-2000</th>
<th>2001-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 configuration</td>
<td>-0.025</td>
<td>-0.023</td>
</tr>
<tr>
<td>H2 configuration</td>
<td>0.051</td>
<td>1.946</td>
</tr>
<tr>
<td>H3 configuration</td>
<td>0.062</td>
<td>0.004</td>
</tr>
<tr>
<td>H4 configuration</td>
<td>-0.057</td>
<td>-0.129</td>
</tr>
<tr>
<td>H5 configuration</td>
<td>0.044</td>
<td>0.043</td>
</tr>
<tr>
<td>H6 configuration</td>
<td>-0.005</td>
<td>-0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tie Types</th>
<th>Knowledge co-creation</th>
<th>Knowledge Transfer</th>
<th>Knowledge co-creation</th>
<th>Knowledge Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. dev in-degree dist.</td>
<td>1.311</td>
<td>-0.681</td>
<td>1.925</td>
<td>1.399</td>
</tr>
<tr>
<td>Skew in-degree dist.</td>
<td>1.516</td>
<td>1.41</td>
<td>0.019</td>
<td>1.656</td>
</tr>
<tr>
<td>Std. dev. out-degree dist.</td>
<td>0.694</td>
<td>1.665</td>
<td>0.782</td>
<td>1.459</td>
</tr>
<tr>
<td>Skew out-degree dist.</td>
<td>1.963</td>
<td>1.771</td>
<td>1.204</td>
<td>1.873</td>
</tr>
</tbody>
</table>

2.6 Conclusions

In this study, the knowledge diffusion networks of nanotechnology inventors were been modeled with a multi-dimensional framework that accounted for both knowledge co-creation and knowledge transfer processes. Table 2.6 summarizes the hypothesis testing results for 1991-2010. To sum up, Hypotheses 1, 2, 3, and 4 were always supported, Hypothesis 5 was supported only during 2001-2010, and Hypothesis 6 was not supported. Based on these results, several conclusions can be made.
First, we identified two types of interactions between knowledge co-creation and knowledge transfer in nanotechnology inventor networks, as described in Hypotheses 1 and 2. Understanding how they interact may provide guidance for stimulating knowledge diffusion. For example, the supported Hypothesis 1 implies that by stimulating knowledge co-creation among inventors, knowledge transfer processes can also be stimulated (and vice versa), leading to a wider diffusion of the knowledge. Second, we identified the roles of inventors who are in key positions in the nanotechnology knowledge diffusion networks. We found that inventors with high degree centrality were more likely to share knowledge by collaborating with each other. Also, knowledge transfer was likely to originate from boundary-spanning inventors. These two findings suggest that inventors in key positions in the knowledge diffusion networks greatly enhance the levels of knowledge co-creation and knowledge transfer. Therefore, funding these key inventors could be more effective than funding others. Third, results of hypothesis 5 show that the effects of NSF funding on nanotechnology knowledge co-creation have increased since the NNI in 2001. This suggests that the NNI has successfully stimulated nanotechnology knowledge co-creation, and more funding can be used to support nanotechnology researchers to increase their levels of knowledge co-creation.
Table 2.6 A Summary of Hypothesis Testing Results

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>1991-2000</th>
<th>2001-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H1</strong>: Knowledge co-creation and knowledge transfer are likely to co-occur between two inventors</td>
<td>supported</td>
<td>supported</td>
</tr>
<tr>
<td><strong>H2</strong>: Knowledge co-creation between two inventors and knowledge transfers from a common inventor to these two inventors are likely to co-occur</td>
<td>supported</td>
<td>supported</td>
</tr>
<tr>
<td><strong>H3</strong>: Knowledge co-creation is likely to occur between two inventors of high degree centralities</td>
<td>supported</td>
<td>supported</td>
</tr>
<tr>
<td><strong>H4</strong>: Knowledge transfer is likely to occur from boundary spanning inventors and other inventors</td>
<td>supported</td>
<td>supported</td>
</tr>
<tr>
<td><strong>H5</strong>: Knowledge co-creation is likely to occur between publicly funded inventors and other inventors</td>
<td>NOT supported</td>
<td>supported</td>
</tr>
<tr>
<td><strong>H6</strong>: Knowledge transfer is likely to occur from publicly funded inventors to other inventors</td>
<td>NOT supported</td>
<td>NOT supported</td>
</tr>
</tbody>
</table>

Our approach that models knowledge co-creation and knowledge transfer processes in the same multi-dimensional network using ERGM provides a more comprehensive understanding of the structures of knowledge diffusion networks. By taking the interactions of knowledge co-creation and knowledge transfer into account, many of our findings confirmed or complemented prior research findings. Also, our study provides tools to improve nanotechnology-related commercial technology development. By understanding the mechanisms of knowledge diffusion among inventors in the nanotechnology field and how public funding affects knowledge diffusion, government
funding agencies and private sector can use resources more effectively. This is especially important today with nanotechnology serving as a country-level competency indicator.
CHAPTER 3. MULTI-DIMENSIONAL KNOWLEDGE DIFFUSION NETWORKS FOR SCIENTIFIC INNOVATION: THE EFFECT OF INDIVIDUAL ATTRIBUTES

3.1 Introduction

Over the past few decades, scientific knowledge has become more important than ever to global economies since many economies have become knowledge-based (Hassan and Haddawy 2013). Since 2010, the total expenditure on Research and Development (R&D) has exceeded $1 trillion worldwide and $0.4 trillion in the US, and the number of research scientists in the world has exceeded 4.2 million. With more and more resources being invested in science, it is worth noting that the diffusion of scientific knowledge has played an important role in scientific development. Understanding how scientific knowledge is diffused among individual researchers can help policy makers in making informed decisions on stimulating knowledge diffusion for scientific development.

Recently, emerging research has adopted or suggested a network perspective of knowledge diffusion, and has used social network analysis (SNA) to examine the mechanisms of knowledge diffusion (Agarwal et al. 2008; Hinz and Spann 2008; Jiang et al. 2014b; Robert et al. 2008; Su and Contractor 2011). In this perspective, the process of scientific knowledge diffusion can be modeled by network graphs which consist of nodes that represent researchers/scientists who exchange knowledge with one another, and network ties that represent flows of knowledge between them. The major advantage of studying knowledge diffusion from a network perspective is that it reveals how the connections between individuals affect the way knowledge is diffused. A network
perspective also takes into account interdependencies between various social relations. However, limited studies have addressed the multi-dimensionality of scientific knowledge diffusion networks. In such networks, the attributes of individuals may vary. Furthermore, the knowledge itself can have attributes. These factors are likely to affect the knowledge diffusion processes. The objective of this study is to conceptualize and understand the multi-dimensionality of scientific knowledge diffusion networks, to examine the mechanisms of scientific knowledge diffusion, and to provide implications for policy makers about how to influence the diffusion of scientific knowledge based on the attributes of researchers and knowledge.

The remainder of this chapter is organized as follows. Section 3.2 reviews recent research that focuses on individual-level knowledge diffusion to provide the background for this study and identify research gaps. Based on the literature and social capital theory, Section 3.3 develops the research model for scientific knowledge diffusion. Section 3.4 describes the dataset and research framework that tests the research model. Empirical results along with potential implications are discussed in Section 3.5. Finally, Section 3.6 concludes this study by summarizing its key contributions.

3.2 Related Work

Table 3.1 summarizes selected recent individual-level knowledge diffusion studies in a taxonomy based on research objective, attributes of knowledge artifact, attributes of individuals, and types of knowledge flows.
Table 3.1 Selected Recent Individual-level Knowledge Diffusion Studies in Major Management and MIS Journals

<table>
<thead>
<tr>
<th>Authors</th>
<th>Objective</th>
<th>Attributes of Knowledge Artifact</th>
<th>Attributes of Individuals</th>
<th>Types of Knowledge Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlo et al. (2012)</td>
<td>Outcome</td>
<td>-</td>
<td>Absorptive capacity</td>
<td>KT</td>
</tr>
<tr>
<td>Chang and Gurbaxani (2012)</td>
<td>Outcome</td>
<td>-</td>
<td></td>
<td>KT</td>
</tr>
<tr>
<td>Griffith et al. (2003)</td>
<td>Mechanism</td>
<td>Codifiability</td>
<td>Virtualness</td>
<td>KT; KC</td>
</tr>
<tr>
<td>Ko et al. (2005)</td>
<td>Mechanism</td>
<td>-</td>
<td>Source credibility; Motivation; Absorptive capacity</td>
<td>KT</td>
</tr>
<tr>
<td>Roach and Cohen (2013)</td>
<td>Methodology</td>
<td>-</td>
<td></td>
<td>KT</td>
</tr>
<tr>
<td>Liu et al. (2011)</td>
<td>Mechanism</td>
<td>Age</td>
<td>DC; BC; Structural holes</td>
<td>KT</td>
</tr>
<tr>
<td>Majchrzak et al. (2013)</td>
<td>Mechanism</td>
<td>Impact</td>
<td>Knowledge depth; Knowledge breadth</td>
<td>KT; KC</td>
</tr>
<tr>
<td>Nerkar and Paruchuri (2005)</td>
<td>Mechanism</td>
<td>Age</td>
<td>DC; BC; Structural holes</td>
<td>KT</td>
</tr>
<tr>
<td>Olivera et al. (2009)</td>
<td>Mechanism</td>
<td>-</td>
<td>Motivation</td>
<td>KC</td>
</tr>
<tr>
<td>Sasidharan et al. (2012)</td>
<td>Outcome</td>
<td>-</td>
<td>DC; BC</td>
<td>KT</td>
</tr>
<tr>
<td>Singh (2005)</td>
<td>Mechanism</td>
<td>-</td>
<td>Social network</td>
<td>KT</td>
</tr>
<tr>
<td>Singh and Marx (2013)</td>
<td>Mechanism</td>
<td>-</td>
<td>Geographical proximity; Technological relatedness</td>
<td>KT</td>
</tr>
<tr>
<td>Su and Contractor (2011)</td>
<td>Mechanism</td>
<td>Codifiability</td>
<td>Expertise recognition; Easy accessibility</td>
<td>KT</td>
</tr>
<tr>
<td>Susarla et al. (2012)</td>
<td>Mechanism</td>
<td>Age</td>
<td>DC</td>
<td>KT</td>
</tr>
<tr>
<td>Wasko and Faraj (2005)</td>
<td>Mechanism</td>
<td>-</td>
<td></td>
<td>KT</td>
</tr>
</tbody>
</table>

*KF=Knowledge Flow; KT=Knowledge Transfer; KC=Knowledge Co-creation; ER=Electronic Records; SEM=Structural Equation Modeling; PLS=Partial Least Square; SNA=Social Network Analysis; WESML=Weighted Exogenous Sampling Maximum Likelihood; DC=Degree Centrality; BC=Betweenness Centrality

3.2.1 Research Objectives

Knowledge diffusion has been studied for different objectives. Some studies have focused on the outcomes of knowledge diffusion. Knowledge diffusion has been
consistently found to bring positive impacts to individuals and organizations. For example, IT-related outsourcing was found to lead to the transfer of accumulated knowledge from service firms to client companies, which in turn increased the productivity of client companies (Chang and Gurbaxani 2012). Successful use of enterprise systems was more likely to happen when employees were connected in a social network where diffusion of knowledge could take place efficiently (Sasidharan et al. 2012). Other studies have tried to explain the mechanism of knowledge diffusion by examining under what conditions knowledge diffusion occurs, and what factors drive, stimulate, or facilitate knowledge diffusion activities (Griffith et al. 2003; Majchrzak et al. 2013). Overall, knowledge artifacts, knowledge entities, and knowledge flows are the three essential elements in knowledge diffusion processes. Knowledge artifacts are objects that convey or hold usable representations of knowledge (Holsapple et al. 1994), such as texts, speech, and diagrams. Knowledge entities usually represent individuals who generate, handle, and assimilate the knowledge. Knowledge flows represent the movement of knowledge between knowledge entities. Prior research has suggested that these elements vary in types or attributes, which may further influence knowledge diffusion processes. These studies will be discussed in detail next.

3.2.2 Attributes of Knowledge Artifacts

Previous literature has suggested that various attributes exist for knowledge artifacts. Impact, age, and codifiability are the most commonly studied ones (Griffith et al. 2003; Majchrzak et al. 2013; Newman 2009). First, knowledge can have different levels of impact. Knowledge with high impact is usually associated with high quality and more
substantial work. For example, knowledge adding and knowledge shaping have been
differentiated as two different knowledge contribution activities in organizational wiki
systems (Majchrzak et al. 2013). The knowledge resulting from knowledge adding was
found to be associated with higher perceived contribution for organizational improvement.
The impact of scientific knowledge is usually measured by the citation counts of research
publications (Chen et al. 2013b; Liu et al. 2011); the higher the citation counts, the
stronger the potential impact of the scientific knowledge.

Second, knowledge can be associated with age, which refers to the amount of
time that has elapsed since the creation of knowledge. Age is a critical dimension of
knowledge because knowledge can receive different types or different levels of attention
over time. For example, in YouTube communities, knowledge about new videos diffuses
to the most active users in early stages, while in later stages knowledge diffusion
occurred mainly through the homophily effects of users (Susarla et al. 2012). Age is
particularly important for scientific knowledge because of the “first-mover” advantage:
the first paper in a specific topic will generally receive the greatest attention in the field
(Newman 2009). Therefore, taking into account the age of scientific knowledge can help
better understand the patterns of knowledge diffusion in different time periods (Liu et al.
2011; Nerkar and Paruchuri 2005).

Knowledge can also be classified as explicit or tacit based on its codifiability. The
codifiability of knowledge refers to the extent to which it can be articulated or
documented. Explicit knowledge can be articulated and is thus accessible to others
(Leonard and Sensiper 1998). Tacit knowledge cannot be articulated, or at least is not
done so yet (Polanyi 1997). Some researchers have maintained that there is no hard distinction between explicit and tacit knowledge, and individual knowledge has been conceptualized as a continuum from tacit to explicit (Griffith et al. 2003). Codifiability of knowledge may affect its diffusion patterns. For instance, based on theoretical discussions, researchers have proposed that tacit knowledge is less likely to be disseminated among team members of a virtual team, compared to individuals who physically work together (Griffith et al. 2003).

3.2.3 Attributes of Knowledge Entities

Prior works have suggested that various attributes related to knowledge entities can also affect knowledge diffusion patterns. These attributes can be actor-based, dyadic, or social network-based.

Actor-based attributes focus on the characteristics associated with individuals, groups, or organizations without considering the relationships between these entities. Some typical attributes include motivation, absorptive capacity, and credibility (Carlo et al. 2012; Ko et al. 2005; Grewal et al. 1994). For example, motivation to learn has been identified as a key driver that triggers knowledge flows from knowledge senders (Ko et al. 2005; Lin 2007; Olivera et al. 2008). Absorptive capacity refers to the ability to understand and assimilate knowledge and it has been viewed as of great importance for acquiring knowledge (Carlo et al. 2012). From the perspective of knowledge source, research has also shown that credibility plays an important role when knowledge can be obtained from several candidates (Ko et al. 2005). Source credibility is the extent to which knowledge recipients perceive a source to be trustworthy (Grewal et al. 1994).
When source credibility is high, the knowledge owned by the knowledge source is perceived to be useful, which may further facilitate knowledge transfer.

Dyadic attributes refer to the characteristics of relationships between two knowledge entities. They are often used to suggest under what conditions two individuals exchange knowledge. For example, expertise recognition and easy access between two individuals have been found to positively predict knowledge seeking activities between them (Su and Contractor 2011). Homophily is another important dyadic factor that facilitates knowledge diffusion (Breschi and Lissoni 2009; Singh 2005; Sorenson et al. 2006; Susarla et al. 2012). If two knowledge entities share something in common, knowledge exchange is more likely to occur between them. For example, in the context of commercial technology, knowledge transfer was found to be more likely to occur within the same geographical area or within the same technological field (Singh and Marx 2013). Likewise, it has been shown that knowledge is likely to be exchanged between friends who share similar interests (Susarla et al. 2012).

Social network-based attributes focus on the positional characteristics of knowledge entities in a knowledge network. Previous research has shown that the structure in which people are connected affect their knowledge diffusion patterns (Sasidharan et al. 2012). Degree centrality and structural hole are the most commonly studied social network measures in previous knowledge diffusion studies. The degree centrality of an individual is the number of other individuals directly connected to the individual. It affects the likelihood that knowledge is gathered by or selected from the central knowledge entity (Liu et al. 2011; Nerkar and Paruchuri 2005; Sasidharan et al.
2012). The structural hole measures of individuals refer to the extent to which the individual is concatenating others that are not inter-connected (Burt 1995). Individuals with high structural hole measures usually have better control of knowledge flows in knowledge diffusion networks (Jiang et al. 2014b; Liu et al. 2011; Nerkar and Paruchuri 2005).

3.2.4 Types of Knowledge Flows

Previous studies have suggested that multiple types of knowledge flows exist in knowledge diffusion networks (Ding 2011; Jiang et al. 2013; Jiang et al. 2014b). Alavi and Leidner (2001) defined four different knowledge processes in organizations: creation, storage, transfer, and application. The creation and transfer processes can be adapted to the context of interpersonal knowledge diffusion. Figure 3.1 illustrates these two processes. Knowledge transfer is a directional knowledge flow from a sender to a recipient, often representing a learning process. The knowledge sender possesses the knowledge before the receiver does. In a knowledge co-creation process, individuals work together to create new knowledge based on their existing knowledge base. This process involves a continual, spiraling flow of knowledge as knowledge moves between individuals (Alavi and Leidner 2001; Nonaka 1994). For scientific knowledge, paper citations and research collaboration are the major channels through which knowledge transfer and knowledge co-creation occur (Lavie and Drori 2012; Roach and Cohen 2013; Singh and Marx 2013).
3.2.5 Research Gaps

First, most studies have examined how knowledge diffusion can be affected by the attributes of knowledge entities, or the structures of knowledge networks, independently. However, findings from prior literature have suggested that these factors are interrelated in some way (Jiang et al. 2014b; Liu et al. 2011). Combining these two types of effects will lead to a multi-dimensional view of knowledge diffusion networks, based upon which we can evaluate how knowledge entities varying in attributes are connected by multiple types of knowledge flows and thereby examine their complex interactions.

Second, ERGM is an effective emerging technique for studying the network multi-dimensionality, and has been used in social science, healthcare, and management literature to explain the formation of friendship, adoption, and communication networks (Contractor et al. 2006; Ellwardt et al. 2012; Gondal 2011; Zappa and Mariani 2011). However, relatively few studies have used ERGM to examine the multi-dimensionality of knowledge diffusion networks, especially in the context of scientific innovation.

Third, although prior literature has suggested that knowledge artifacts can have attributes that may affect knowledge diffusion patterns (Majchrzak et al. 2013; Nerkar...
and Paruchuri 2005), the effect of these attributes on knowledge diffusion has not been examined.

In order to address these research gaps, we adopt a multi-dimensional view of knowledge diffusion networks in scientific innovation by taking into account various attributes of individual researchers and multiple types of knowledge flows between them. We further examine how impact and age of scientific knowledge affect its diffusion patterns. Codifiability of knowledge is not explicitly considered in this study because scientific knowledge is usually documented as publications, and thus is considered as explicit knowledge. Since the process of knowledge exchange involves social interactions, such as learning and collaboration, between researchers, the social relationships between researchers are expected to play an important role in knowledge diffusion. Social capital theory helps explain how and why people establish social relations and it has been used in prior literature to explain why individuals exchange knowledge (Inkpen and Tsang 2005; Wasko and Faraj 2005). Therefore, we turn to social capital theory to suggest possible relationships between the attributes of researchers and knowledge flows.

3.3 Theory and Research Hypotheses

3.3.1 Social Capital Theory

The central proposition of social capital theory is that network relationships are valuable resources for members in the network (Inkpen and Tsang 2005). Social capital has been defined in different ways. One perspective has focused on personal benefits individuals can gain from social relations, and considered social capital as private good
(Belliveau et al. 1996; Burt 1997). For example, a researcher who is respected and trusted by others is likely to be exposed to opportunities such as invited research talks or career advancement. In this case, the relationship of trust is the social capital that leads to individual benefits. Another major perspective has conceptualized social capital as public good owned by a social collectivity (Bourdieu 1986; Coleman 1989; Putnam 1993). For example, collaboration within a group of researchers sharing some common background should be more efficient than a group of researchers without that common background. In this case, the shared context is the social capital for the collective that facilitates further interactions.

In this essay, we adopt both of these perspectives and define social capital for researchers as the aggregate of resources embedded in the relationships between researchers, either possessed by individual researchers or the entire research community. This broader definition is in line with previous knowledge management research that used social capital theory as the theoretical basis (Inkpen and Tsang 2005; Nahapiet and Ghoshal 1998; Newell et al. 2004; Robert et al. 2008; Wasko and Faraj 2005). In these works, four major conditions have been suggested as important facilitators for knowledge exchange and recombination. First, social connections exist between individuals. Second, the relationships between individuals are positive. Third, individuals are capable of assimilating knowledge. Finally, individuals have motivations for exchanging knowledge. The first three conditions can be mapped to the three dimensions of social capital proposed by Nahapiet and Ghoshal (1998): the structural, relational, and cognitive dimensions. In this study, we examine how the attributes of knowledge entities (i.e.,
researchers) reflected by these three dimensions of social capital and individual motivations affect knowledge flows between them, and how these relationships can be influenced by the attributes of knowledge artifacts. The research model is shown in Figure 3.2. We will discuss each construct and link in the next subsections.

![Figure 3.2 Research Model for Scientific Knowledge Diffusion](image)

**3.3.2 Knowledge Flows: Knowledge Transfer and Knowledge Co-creation**

In this study, we adapt the definitions of knowledge processes in previous work (Alavi and Leidner 2001), and focus on two types of knowledge flows that manifest in scientific innovation communities: knowledge transfer and knowledge co-creation (Jiang et al. 2014b). As previously illustrated in Figure 1, knowledge transfer is defined as a directional flow of knowledge from a sender to a recipient, often representing a learning
process. For example, researchers can learn new knowledge from others by reading papers, listening to research talks, and attending seminars. Often, these processes need to be repeated or complemented by each other to ensure that the knowledge recipient fully assimilates the transferred knowledge. In knowledge co-creation, individuals repeatedly exchange, share, and develop new knowledge based on their existing knowledge bases. This process involves a continual and spiraling flow of knowledge as knowledge moves between individuals (Alavi and Leidner 2001; Nonaka 1994). For example, research collaboration is a typical channel through which new knowledge is co-created (Katz and Martin 1997). The major difference between the two knowledge flows is that knowledge co-creation requires the active participation of both parties, whereas knowledge transfer is mainly triggered by the knowledge recipient and the knowledge sender may often be unaware of the transfer process.

3.3.3 Structural Capital

The structural dimension of social capital maintains that the existence of network ties is the basis for other social interactions to occur (Wasserman 1994). If two individuals are connected by some social ties, they have better access to each other, which is one of the conditions for knowledge exchange and recombination (Nahapiet and Ghoshal 1998). For researchers, research collaboration is one of the most important opportunities for establishing initial social relations. Different from other social connections such as graduating from the same program or working in the same institution, research collaboration provides researchers with more opportunities to access each other’s knowledge, and hence we focus on this collaboration-related structural capital in
this study. Through collaboration, researchers may become familiar with each other’s work and the knowledge embedded in these works. Such accessibility and awareness may create opportunities for further knowledge transfer processes to occur (Breschi and Lissoni 2004; Yan and Sugimoto 2011). Therefore, we propose the following research hypothesis:

**H1:** Existence of collaboration experience is positively related to the likelihood of knowledge transfer between two researchers.

3.3.4 Relational Capital

The relational dimension of social capital highlights that individuals develop personal relationships, such as friendship, respect, and trust, through a history of interactions (Nahapiet and Ghoshal 1998). The developed relationships can further influence individuals’ behaviors. One of the most important facets of the relational dimension in social capital is trust, or credibility of knowledge sources, which is particularly relevant to knowledge exchange (Inkpen and Tsang 2005; Ko et al. 2005; Zhou et al. 2012). Since knowledge transfer is often a process that is triggered by knowledge recipients, a trust relationship must exist to ensure that the incoming knowledge is believable and useful. In order for trust to develop, the knowledge sender must signal his or her credibility of knowledge in some way. Similarly, knowledge co-creation involves repeated knowledge exchanges between two researchers, and thus they
should be credible to each other. Therefore, we propose the following research hypotheses:

**H2a**: Credibility of knowledge source is positively related to the likelihood of knowledge transfer between two researchers.

**H2b**: The aggregated credibility of two researchers is positively related to the likelihood of knowledge co-creation between them.

### 3.3.5 Cognitive Capital

The cognitive dimension of social capital focuses on the shared context a social collectivity owns. In the context of scientific knowledge, the shared context mainly refers to areas of expertise and research experiences reflected by career length. Prior research has shown that knowledge exchange is easier between scientists with expertise in similar technology areas (Singh and Marx 2013). Therefore, in this study we focus on a less explored aspect of the shared context: similarity in career length of researchers. For knowledge co-creation, although the combination of different knowledge and experiences from individuals is recognized as a key factor for creating new innovation (Nahapiet and Ghoshal 1998), such a process will become more efficient when some shared norm exists between researchers, such as a common knowledge base, experiences in the field, and the way of conducting research. Therefore, we propose the following hypothesis:
**H3**: Similarity in career length between researchers is positively related to the likelihood that they co-create knowledge with each other.

3.3.6 Individual Motivations

Prior research has also shown that individual motivations are important antecedents of knowledge exchange (Lin 2007; Wasko and Faraj 2005). When co-creating knowledge, researchers may benefit intellectually from intensive discussion with each other, receiving feedback, and improving knowledge, especially when the collaborators are very active. Also, based on social exchange theory (Blau 1964), individuals engage in social interactions in an expectation of social rewards. For instance, individuals may anticipate approval, social recognition, and respect from interacting with others. When the research collaborators are active in the field, the process of co-creating knowledge is likely to lead to wider exposure of their knowledge to the entire research community, which results in more attention and recognition. In addition, researchers active in the field have great motivation to learn from others in order to improve the knowledge base. Therefore, we propose the following hypotheses:

**H4a**: The aggregated activity level of researchers is positively related to the likelihood that knowledge co-creation occurs between them.

**H4b**: Activity level of a researcher is positively related to the likelihood that knowledge transfers to the researcher.
3.3.7 Attributes of Knowledge Artifacts

We further consider how attributes of knowledge artifacts influence the patterns of scientific knowledge diffusion. Specifically, we consider knowledge impact and age which have been suggested as important attributes of scientific knowledge in prior literature.

First, we argue that the impact of knowledge should positively affect the relationships between social capital factors and knowledge diffusion. For example, knowledge transfer should occur mostly when the knowledge is of high impact and interesting to the knowledge recipient. Also, no matter how credible or active the research collaborators are, the knowledge exchange may not continue if the knowledge is perceived to be less interesting to any one of them. When co-creating knowledge, the intellectual benefit may manifest when the knowledge is of high impact because such knowledge is perceived as being more valuable (Majchrzak et al. 2013). Therefore, the following research hypotheses are proposed:

**H5a**: The impact of knowledge positively affects the relationships between social capital/individual motivation factors and knowledge transfer.

**H5b**: The impact of knowledge positively affects the relationships between social capital/individual motivation factors and knowledge co-creation.

For scientific knowledge, timeliness is also an important element (Brown et al. 2010a; Watson 1994). Over time, knowledge may take on different diffusion patterns due...
to changes in attention (Susarla et al. 2012). The attention changes mainly because the audience of scientific knowledge evolves over time. In its early stages, scientific knowledge may get some attention because it is new and receives academic curiosity. At the same time, it is less known by the entire field and individuals have the chance to access it only when proactively seeking new knowledge. As a result, previous collaboration experience and activity level of researchers should play an important role in identifying and receiving knowledge that is early in time. Moreover, new knowledge is often associated with uncertainty in application and usage (Schulz 2001). Therefore, the credibility of source may have a stronger influence when the transferred knowledge is relatively new. We propose the following hypothesis:

**H6:** The newer the knowledge, the more likely that knowledge transfer occurs based on the existence of collaboration ties, source credibility, and activity level of the knowledge recipient.

3.3.8 Untested Links

In addition to the hypothesized relationships above, we provide explanations for omitting some links in the hypotheses development. First, we omit a hypothesis about a link from collaboration experience to knowledge co-creation because based on our definition and operationalization for knowledge co-creation in later steps, knowledge co-creation requires collaboration experience as prerequisite and thus such a link is straightforward. Second, we omit a link from similarity in career length to knowledge
transfer because we expect that learning usually occurs from an experienced individual to a less experienced one. Third, we omit effects of age on the relationships between the social capital factors and the knowledge co-creation because unlike transferred knowledge which can be new or old in terms of when the knowledge is created, knowledge is always new during a co-creation process. Finally, although we expect that knowledge transfer and knowledge co-creation would have a high tendency to co-occur, this relationship has already been tested in a similar context (Jiang et al. 2014b).

3.4 Data and Methods

3.4.1 Research Test-bed

To test the aforementioned research hypotheses about knowledge diffusion, scientific publication data were collected from the Thomson Reuter Web of Science (WoS). The WoS database covers a broad range of scientific journals in more than 150 disciplines and has been used in recent knowledge management studies (Chen et al. 2013b; Liu et al. 2011). We utilized ER-based data to study knowledge diffusion networks due to the following considerations. First, scientific knowledge diffusion networks are large-scale in nature and we wanted to examine diffusion patterns based on a large-population of researchers. Second, we could demonstrate the usefulness of knowledge CI in studying knowledge diffusion.

Two cutting-edge technology fields, biotechnology and nanotechnology, were selected for analysis because of their significant development speed and active knowledge diffusion. For biotechnology data, publications under the WoS category
“Biotechnology and Applied Microbiology” were collected. Since nanotechnology is a relatively new field, in addition to articles under “Nano-science and technology” category, data was also collected by searching 29 keywords in the title and abstract areas of the articles. The keywords were provided by domain experts and have been used in prior studies that collected nanotechnology publication data (Chen et al. 2013b; Huang et al. 2006; Huang et al. 2005; Li et al. 2007; Liu et al. 2011). Table 3.2 summarizes the statistics of the data collection.

<table>
<thead>
<tr>
<th>Field</th>
<th># of Publications</th>
<th># of Researchers</th>
<th>Years Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biotechnology</td>
<td>230,060</td>
<td>194,515</td>
<td>2001-2010</td>
</tr>
<tr>
<td>Nanotechnology</td>
<td>502,535</td>
<td>352,893</td>
<td>2001-2010</td>
</tr>
</tbody>
</table>

3.4.2 Multi-dimensional Network Construction

We first extracted knowledge diffusion networks from the scientific publication data, and modeled the networks’ multi-dimensionality including attributes of researchers, attributes of knowledge, and multiple types of knowledge flows.

For the attributes of researchers, the existence of collaboration experience was evaluated based on the co-authorship information between researchers, following previous research (Ding 2011; Gao et al. 2011). It was coded as a binary dyadic variable: if two researchers co-authored at least one paper, the collaboration experience was coded as “1” for this pair of researchers, and “0” otherwise. Career length in the field was utilized to account for the credibility of researchers. Career length reflects the
experiences of a scientist in the field, and often signals the quality of the researcher’s work (Buenstorf and Geissler 2012). Also, past research has shown that career length is often connected with credibility (Chen and Starosta 1997). Career length was measured by the number of years passed since the researcher first published an article in the WoS database (which can be outside the 10-year test-bed window) until the researcher’s latest publication in the 10-year window. If a researcher published only one paper, the researcher’s career length was set to zero. Similarity in career length was measured as the difference of career length variables between two researchers. The activity level of a researcher was measured based on the number of publications in the 10-year window, because the number of publications of a researcher reflects how actively the researcher conduct research activities in the field (Laroche and Amara 2011).

For attributes of knowledge, the impact was evaluated based on the citation count of scientific publications, which has been used to represent the quality, impact, and authority of knowledge by prior studies (Chen et al. 2013b; Huang et al. 2005). Knowledge with high impact is usually associated with a high citation count. Regardless of the impact, citation count increases over time due to an accumulation effect. To control this effect, the citation count within 3 years from publication was used to calculate the impact of knowledge embedded in each article. The age of knowledge was calculated as the number of years passed since the time of publication until when it was cited, following prior research (Liu et al. 2011; Nerkar and Paruchuri 2005). Therefore, age was a variable associated with each citation, not with each publication.
The major channels of knowledge transfer and knowledge co-creation are paper citations and research collaboration (Lavie and Drori 2012; Roach and Cohen 2013). We identify these channels of knowledge flows as the basis on which to model knowledge transfer and co-creation in later steps. The channels of knowledge transfer were identified based on citation relationships between researchers, following previous studies (Ding 2011; Gondal 2011; Liu et al. 2011; Singh and Marx 2013). Citing other researchers’ works in a paper suggests that some knowledge is built upon others’ knowledge. Therefore, paper citations signal that knowledge transfer may have occurred. Self-citations were excluded from this study because they do not reflect a process of knowledge transfer. Likewise, the channels of knowledge co-creation were identified based on co-authorship relationships between researchers, following prior research (Ding 2011; Gao et al. 2011). Co-authored works are outcomes of research collaboration. During research collaboration, researchers may have intensively exchanged ideas that finally resulted in the creation of new knowledge. The new knowledge is co-owned by the researchers, representing that knowledge co-creation may have occurred between the researchers.

3.4.3 Knowledge Flow Identification

Note that these citation and collaboration channels cannot be used directly as flows of knowledge transfer or knowledge co-creation because citation/co-authorship is not necessarily equivalent to knowledge transfer/co-creation (Alcacer and Gittelman 2006; Roach and Cohen 2013). For example, a paper citation may not represent the actual knowledge flow from the cited researcher to all citing researchers, because co-authors
may have different roles in the publication process. Some may be even unaware of the citation. For the same reason, knowledge co-creation does not necessarily occur between any pair of co-authors. To provide a better modeling of knowledge flows, the intensity of citations and co-authorship relationships must be taken into account by measuring how many times they have occurred between a pair of researchers. Figure 3.3 illustrates the identification of the two types of knowledge flows. A knowledge flow between a pair of researchers was established only if the intensity was above a predetermined threshold value. The threshold values were determined based on mean intensities plus two standard deviations. Similar approaches have been used to dichotomize networks in previous research (Jiang et al. 2013; Jiang et al. 2014b; Lomi and Pallotti 2012). Finally, researchers without any knowledge flows to others (i.e., isolated nodes) resulting from this process were dropped from further analysis.

3.4.4 ERGM Analysis

In order to test the research hypotheses about scientific knowledge diffusion, we transformed H1 to H4b into network patterns as shown in Table 3.3. If some of these
patterns occurred more frequently than by chance in the constructed networks, the corresponding hypotheses would be supported.

Table 3.3 Research Hypotheses, ERGM Configurations, and Illustrations

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Configuration</th>
<th>Illustrations*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H1</strong>: Existence of collaboration experience is positively related to the likelihood of knowledge transfer between two researchers.</td>
<td>$A_1$: collaboration-arc</td>
<td><img src="image1" alt="Illustration" /></td>
</tr>
<tr>
<td><strong>H2a</strong>: Credibility of knowledge source is positively related to the likelihood of knowledge transfer between two researchers.</td>
<td>$A_2$: credibility-sender</td>
<td><img src="image2" alt="Illustration" /></td>
</tr>
<tr>
<td><strong>H2b</strong>: The aggregated credibility of two researchers is positively related to the likelihood of knowledge co-creation between them.</td>
<td>$A_3$: credibility-sum</td>
<td><img src="image3" alt="Illustration" /></td>
</tr>
<tr>
<td><strong>H3</strong>: Similarity in career length between researchers is positively related to the likelihood that they co-create knowledge with each other.</td>
<td>$A_4$: career-difference</td>
<td><img src="image4" alt="Illustration" /></td>
</tr>
<tr>
<td><strong>H4a</strong>: The aggregated activity level of researchers is positively related to the likelihood that knowledge co-creation occurs between them.</td>
<td>$A_5$: active-sum</td>
<td><img src="image5" alt="Illustration" /></td>
</tr>
<tr>
<td><strong>H4b</strong>: Activity level of a researcher is positively related to the likelihood that knowledge transfers to the researcher.</td>
<td>$A_6$: active-receiver</td>
<td><img src="image6" alt="Illustration" /></td>
</tr>
</tbody>
</table>

* The circle nodes represent researchers. Attributes of researchers such as credibility and activity level are represented by color and size of nodes.

Given these hypothesized network patterns, the ERGM model in this study can be represented as Equation (3.1). In the equation, $Y$ is a matrix of random variables representing ties in a network, and $y$ is its realization. $\eta_{Ai}$ is a parameter corresponding to configuration $Ai$, reflecting the tendency for the corresponding pattern to occur, and
\[ g_{Al}(y) \] is network statistics corresponding to \( A_i \), measured based on \( y \). In addition to \( A_1 \) to \( A_6 \), we included ‘edge’ and ‘arc’ as control configurations to account for general structural effects. The ‘edge’ configuration represents undirected ties in the network and models a tendency for knowledge co-creation to occur between any pair of researchers in general, while the ‘arc’ configuration represents directed ties and models a tendency for knowledge transfer to occur between any pair of researchers. \( \kappa \) is a normalizing constant ensuring that \( \Pr(Y) \) is a probabilistic distribution.

\[
\Pr (Y = y) = \left( \frac{1}{\kappa} \right) \exp \left\{ \sum_{s \in \text{[edge, arc]}} \eta_s g_s(y) + \sum_{i \in \{1,2,\ldots,6\}} \eta_{Ai} g_{Ai}(y) \right\} \\
(3.1)
\]

The primary task of ERGM is to estimate the parameters \( \eta_{Ai} \) (\( i=1,2,\ldots,6 \)) associated with the research hypotheses. For any hypothesis listed in Table 3 except H3, if its parameter estimate was positive and significant, it would indicate that the corresponding network pattern occurred more frequently than by chance in the constructed knowledge diffusion network, supporting the corresponding hypothesis. The configuration used for H3 (career-difference) models a tendency for two researchers with great absolute difference in career length to develop knowledge co-creation ties, and hence the parameter of H3 should be negative in order for the hypothesis to be supported. XPnet was used to run the ERGM analysis because of its capability of modeling multiple types of network ties at the same time (Wang et al. 2009a; Wang et al. 2013). Markov
Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE) (Snijders 2002; Wasserman and Robins 2005) was used as the estimation method.

In order to test H5a and H5b which take into account the impact of scientific knowledge, we reconstructed the knowledge diffusion networks based on the citation count of publications. We first divided the publications into three sets based on their 33% and 66% percentiles of citation counts. Separate knowledge diffusion networks were then constructed based on the articles falling into each interval. For example, the network constructed based on the [66%, 100%] interval accounted for the diffusion of high impact knowledge. The ERGM analysis results were then compared between the three networks to examine how the impact of scientific knowledge affected the diffusion process.

Similarly, in order to test H6 which takes into account the age of scientific knowledge, we divided the paper citations into three sets based on 1) citations that occurred within 3 years of publication; 2) citations that occurred after 3 years and within 6 years of publication, and 3) citations that occurred after 6 years from publication. Flows of knowledge transfer were then re-identified based on the intensity of citations in each interval. For example, the knowledge diffusion network constructed based on the [0, 3 years] interval accounted for the diffusion of newborn knowledge within three years of publication. Again, the ERGM parameter estimates were compared between the three networks to examine how the age of the knowledge affected the diffusion process.

Furthermore, when the MCMCMLE processes converged, goodness-of-fit (GOF) tests were conducted for each parameter to evaluate whether the estimated model fit the data well. GOF tests generate a number of sample networks based on the estimated
ERGM parameters and compare their network statistics against the original network. Small differences would indicate that the estimated ERGM model fit the data well (Huffaker et al. 2009; Wang et al. 2009b).

3.5 Results and Discussions

3.5.1 Social Capital and Individual Motivation Factors

We first performed ERGM analysis for H1 to H4b based on the knowledge diffusion network constructed from the entire dataset to examine how social capital and individual motivation factors affect knowledge diffusion. Table 3.4 shows estimated ERGM parameters and standard errors (in parentheses). The estimated parameter values are log-odds of the ties of corresponding patterns forming in the network. These values are positively related to the probabilities that corresponding patterns would occur. For example, the value of estimated parameter for the configuration ‘collaboration-arc’ is positive (1.14). Then the conditional log-odds of a directed tie adding a ‘collaboration-arc’ pattern is 

\[ -5.36 + 1.14 = -4.22 \]

(because the tie automatically adds an ‘arc’ pattern as well). The probability that such a tie would develop in the network is 

\[ \frac{\exp(-4.22)}{1 + \exp(-4.22)} = 0.014, \]

which is greater than the probability that a directed tie would develop between any pair of nodes, which is 

\[ \frac{\exp(-5.36)}{1 + \exp(-5.36)} = 0.004. \]

The ratio between the parameter estimate and the corresponding standard error is the t-statistics of the parameter. If the absolute value of this ratio is greater than 2, it indicates that the parameter is significantly non-zero with at least 95% confidence (Snijders 2001). The same criteria have been used in prior ERGM research (Pahor et al. 2008; Su and
Contractor 2011). The significant parameters are flagged by ‘*’ symbols. All GOF parameters were smaller than 2 when the model converged, indicating a good fit of the model (Huffaker et al. 2009; Wang et al. 2009b).

Table 3.4 ERGM Analysis: Entire Knowledge Diffusion Network

<table>
<thead>
<tr>
<th>Parameters (η_{Ai})</th>
<th>Nanotech.</th>
<th>Biotech.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structural Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edge</td>
<td>-4.36 (0.14)*</td>
<td>-4.27 (0.11)*</td>
</tr>
<tr>
<td>arc</td>
<td>-5.36 (0.20)*</td>
<td>-4.62 (0.16)*</td>
</tr>
<tr>
<td><strong>Hypothesized Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>collaboration-arc (H1)</td>
<td>1.14 (0.25)*</td>
<td>1.25 (0.31)*</td>
</tr>
<tr>
<td>credibility-sender (H2a)</td>
<td>0.34 (0.12)*</td>
<td>0.47 (0.10)*</td>
</tr>
<tr>
<td>credibility-sum (H2b)</td>
<td>0.42 (0.25)</td>
<td>0.55 (0.29)</td>
</tr>
<tr>
<td>career-difference (H3)</td>
<td>-0.09 (0.04)*</td>
<td>-0.07 (0.02)*</td>
</tr>
<tr>
<td>active-sum (H4a)</td>
<td>0.79 (0.35)*</td>
<td>0.84 (0.20)*</td>
</tr>
<tr>
<td>active-receiver (H4b)</td>
<td>0.07 (0.12)</td>
<td>0.04 (0.13)</td>
</tr>
</tbody>
</table>

First, a significant and positive effect was observed for the collaboration-arc configuration, indicating that the existence of collaboration experience between researchers was positively associated with the likelihood that knowledge transfer occurred between them, supporting H1. This result suggests that structural capital plays an important role in knowledge diffusion. Through collaborations, researchers begin to know their collaborators’ specialties and expertise, which can be a point of reference in future. When relevant knowledge is needed to develop new innovation, researchers may recall the collaborators’ knowledge and seek help, which in turn leads to knowledge transfer. A direct implication is that weak ties between researchers create opportunities for them to involve in more intense knowledge exchange activities. To this end, policy
makers may consider enabling more workshops and collaboration grants to gather researchers for opportunities to develop weak ties.

Second, a significant positive effect was observed for the credibility-sender configuration, indicating that credibility of a researcher was positively correlated with the likelihood of the researcher to send out flows of knowledge transfer, supporting \( H2a \). This result suggests that relational capital plays an important role in the transfer of scientific knowledge. Researchers develop a relationship of trust through a history of interactions either directly or indirectly (Nahapiet and Ghoshal 1998). The trust acts as relational capitals to facilitate knowledge transfer. For policy makers, the result implies that researchers with long careers can be selected as key disseminators when key scientific knowledge needs to be promoted. However, we did not observe a significant effect for credibility-sum configuration. Therefore, \( H2b \) was not supported. It implies that credibility may not be that important for initiating knowledge co-creation, because trust may develop during the process of co-creation. This developed credibility was not measured in our model.

Third, a significant negative effect was observed for the career-difference configuration, indicating that researchers similar in career length were likely to co-create knowledge with each other, supporting \( H3 \). This result suggests that cognitive capital in the researchers’ network affects knowledge diffusion positively. It is very likely that researchers who are similar in career length have gone through similar professional experiences such as education. The similarity has given the researchers shared ideas, language, and norms which facilitate their knowledge exchange activities. The result also
supports that homophily is an important antecedent for social interaction, which has been
demonstrated in many prior studies (Breschi and Lissoni 2009; Singh 2005; Sorenson et
al. 2006; Susarla et al. 2012). Based on the results, an implication is that researchers who
are similar in career length can be grouped together in scholarly activities to stimulate
knowledge diffusion. For instance, some international conferences provide separate
consortiums for graduate students, junior faculties, and mid-career faculties. The
grouping can be extended to provide further segmentation of participants.

Fourth, we observed that the active-sum configuration had significant and positive
correlations with knowledge co-creation, indicating that active researchers have strong
motivations to collaborate with each other to co-create knowledge. The observation is
consistent with previous work where productive researchers were found to be likely to
collaborate with each other (Ding 2011). However, we did not observe a significant
relationship between the activity level of knowledge recipients and the likelihood of
knowledge transfer. H4b was not supported.

3.5.2 Knowledge Impact

Table 3.5 shows estimated ERGM parameters for models with different levels of
knowledge impact. In each interval of knowledge impact, the parameter estimates are
presented and standard errors are shown in parentheses. As discussed earlier, the
estimated parameter values are log-odds of the ties of corresponding patterns forming in
the network, and are positively correlated with the probabilities that those patterns would
occur. The ratio between the parameter estimate and the standard error needs to be
greater than 2 in order for the parameter estimate to be considered as significant at 95%
confidence level. If an estimated parameter is not significant in intervals representing lower levels of knowledge impact, but becomes significant in intervals representing higher levels of knowledge impact, it would indicate that knowledge impact positively affected the relationships between social capital factors and knowledge transfer/co-creation. For example, under the ‘Nanotech.’ column, the parameter for active-sum configuration was not significant in [0, 33%] and (33%, 66%] intervals, but became significant in the [66%, 1] interval, implying the influence of knowledge impact on knowledge co-creation in nanotechnology field. For simplicity, structural effects are omitted in the table.

Table 3.5 ERGM Analysis: Models with Different Levels of Knowledge Impact

<table>
<thead>
<tr>
<th>Parameters (η_{Ai})</th>
<th>Nanotech.</th>
<th>Biotech.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[0, 33%]</td>
<td>(33%, 66%]</td>
</tr>
<tr>
<td>collaboration-arc (H5a)</td>
<td>1.19 (0.41)*</td>
<td>1.21 (0.33)*</td>
</tr>
<tr>
<td>credibility-sender (H5a)</td>
<td>0.25 (0.09)*</td>
<td>0.31 (0.08)*</td>
</tr>
<tr>
<td>credibility-sum (H5b)</td>
<td>0.20 (0.13)</td>
<td>0.21 (0.11)</td>
</tr>
<tr>
<td>career-difference (H5b)</td>
<td>-0.06 (0.01)*</td>
<td>-0.07 (0.01)*</td>
</tr>
<tr>
<td>active-sum (H5b)</td>
<td>0.53 (0.29)</td>
<td>0.61 (0.35)</td>
</tr>
<tr>
<td>active-receiver (H5a)</td>
<td>0.06 (0.15)</td>
<td>0.03 (0.06)</td>
</tr>
</tbody>
</table>

As shown in the table, H5a was not supported, because the significance of parameter estimates for knowledge transfer ties (i.e., collaboration-arc, credibility-sender, and active-receiver) did not change across models with different levels of knowledge impact. However, we observed that the active-sum configuration had significant and
positive correlations with knowledge co-creation only in intervals representing high impact knowledge (above 66% percentile for nanotechnology and above 33% for biotechnology), which suggests that co-creation of high impact knowledge was highly likely to occur if the researchers’ total activity level was high. Parameter estimates for other two knowledge co-creation ties (i.e., credibility-sum and career-difference) did not change across models with different levels of knowledge impact. Therefore, the results partially support H5b. This suggests that researchers may participate in knowledge co-creation in an expectation of both intellectual merits and social rewards that they gain from collaborating with active researchers (Wasko and Faraj 2005).

3.5.3 Knowledge Age

Table 3.6 shows estimated ERGM parameters for knowledge diffusion networks accounting for different levels of knowledge age. In each interval of knowledge age, the parameter estimates are presented and standard errors are shown in parentheses. If an estimated parameter is significant in intervals representing newer knowledge (e.g., 0-3 years since its creation), but becomes insignificant in intervals representing older knowledge, it would indicate that newer knowledge led to stronger relationships between social capital factors and knowledge transfer. Since only knowledge transfer ties were affected by incorporating the age of knowledge, configurations that did not involve knowledge transfer (i.e., career-difference, experience-sum, and active-sum) were excluded. For simplicity, structural effects are also omitted in the table.
Table 3.6 ERGM Analysis: Models with Different Levels of Knowledge Age

<table>
<thead>
<tr>
<th>Parameters (η_{ai})</th>
<th>Nanotech.</th>
<th>Biotech.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0, 3]</td>
<td>(3, 6]</td>
</tr>
<tr>
<td>collaboration-arc (H6)</td>
<td>0.99</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>(0.33)*</td>
<td>(0.36)*</td>
</tr>
<tr>
<td>credibility-sender (H6)</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>(0.08)*</td>
<td>(0.11)*</td>
</tr>
<tr>
<td>active-receiver (H6)</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>(0.01)*</td>
<td>(0.02)*</td>
</tr>
</tbody>
</table>

As shown in the row of activity-receiver, the activity level of a researcher was found to positively correlate with the likelihood that the researcher would receive knowledge that had been created within the first six years. After six years, the parameter estimate became insignificant. For other two configurations, the significance of parameter estimates did not change across models with different knowledge age. These observations partially support H6. The coefficients were especially high when knowledge transfer within three years was modeled, suggesting that active researchers have great motivation to monitor the latest trends in the field, and learn new knowledge as soon as possible.

3.6 Conclusions

Based on social capital theory, this study examined how the flows of knowledge transfer and knowledge co-creation could be affected by attributes of individual researchers and attributes of scientific knowledge. Our empirical results showed that social capital factors including collaboration experience, credibility, and similarity in career length positively affected knowledge diffusion flows between researchers. Also, researchers’ activity level in general affected knowledge co-creation positively, but not knowledge transfer. However, the relationship between motivations and knowledge
diffusion can be moderated by the impact and age of scientific knowledge: when the knowledge is of relatively low impact, researchers were less likely to participate in knowledge co-creation processes even if they are active; when knowledge was new, active researchers had a stronger tendency to seek knowledge through knowledge transfer processes.

We have multiple contributions to research. First, we provided conceptualization of the multiple types of knowledge flows in multi-dimensional scientific knowledge diffusion networks. Existing knowledge literatures mainly focused on only one type of knowledge diffusion process (Singh and Marx 2013; Wasko and Faraj 2005), while knowledge can diffuse in many different ways. Previous studies have recognized this problem as one limitation, and suggested that future research should consider multiple aspects of knowledge activities such as how knowledge is created, shared, and learned (Wasko and Faraj 2005). This research aligns with their suggestions and models knowledge transfer and knowledge co-creation processes in the same network. Second, this study provides insight into how attributes of knowledge artifact affect knowledge diffusion processes. Most prior research focusing on knowledge diffusion mechanisms did not explicitly consider how knowledge with different attributes diffused in different ways. Prior research has suggested considering the influence of knowledge attributes in knowledge diffusion research (Inkpen and Tsang 2005). We shed light on this aspect and provided explanations for how knowledge diffusion can be affected by the impact and age of scientific knowledge. Third, the process of modeling and analyzing multi-dimensional scientific knowledge diffusion networks used in this study can be applied to
any general multi-dimensional networks, thereby providing a useful methodological framework to conceptualize and examine multi-dimensional networks in general. Finally, we demonstrated the usefulness of knowledge cyber-infrastructure in examining large-scale knowledge diffusion networks.
4.1 Introduction

Social media networks are emerging online networks that virtually connect individuals. These networks consist of nodes that represent individual social media users and ties that represent various relationships between the users. Examples of social media networks include online friendship networks (Heatherly et al. 2013; Lin et al. 2013), following-follower networks (Grabowicz et al. 2012), and content sharing networks (Shi et al. 2014; Stieglitz and Linh 2013). The relationships between the online users are often public information, which provides opportunities for using social network analysis (SNA) to better understand how and why individuals establish social connections online (Kane et al. 2014). As a result, a growing number of studies have used SNA to examine social media networks (Oestreicher-Singer and Sundararajan 2012; Shi et al. 2014; Singh et al. 2011; Stieglitz and Linh 2013; Susarla et al. 2012).

Social media networks have two important characteristics. First, they are dynamic in nature. Network ties develop in a sequence, but not simultaneously. As such, relationships between individuals may change over time. Second, social media users differ in various attributes, such as gender, functional role in online communities, and reputation. As a result, social media networks are multi-dimensional networks (Tang et al. 2012; Wang et al. 2014) and different node types exist in the network. A consequence of these two characteristics is that the seemingly same network patterns can result from
different network formation processes, depending on the sequence in which the network ties develop (Jiang and Chen 2014). For example, Figure 4.1 illustrates two processes in forming a two-star pattern. Here, we assume that the black nodes represent highly active individuals (e.g., individuals who frequently come online and leave messages) in online communities and the numbers next to network ties indicate the sequence in which the relationships develop. The Pattern A illustrates a process where highly active individuals are prioritized over others when developing relationships, while the pattern B illustrates the opposite tendency. If the sequence of the network ties is ignored, we are unable to differentiate between these two patterns and understand how highly active individuals participate in the dynamic process of network formation.

![Figure 4.1 Different Processes Leading to the Same Network Pattern](image)

Differentiating between various temporal patterns is thus critical to understand the formation mechanisms of social media networks. However, current social network
research usually adopts a static view of networks based on the assumption that all network ties have developed concurrently upon observation. This assumption, while contributing to simplicity and being useful for identifying static patterns of networks, leads to reduced representation of real social media networks. As a result, the ability of social network analysis to identify network patterns may be negatively affected. The problem can further reduce the practical value of social network analysis to understand various network phenomena in social media contexts.

In this study, we propose a novel dynamic network model, the Nodal Attribute-based Temporal Exponential Random Graph Model (NATERGM), for dynamic network analysis. NATERGM is an extension of TERGM (Hanneke et al. 2010) and focuses on how nodal attributes of networks affect the sequence in which network ties develop. The proposed model extracts nodal attributes of individuals and time information of network ties from social media networks, based on which various temporal patterns are modeled and their likelihoods of occurrence are estimated. In this study we used NATERGM to examine dynamic knowledge diffusion processes in social media. With empirical data obtained from an online knowledge sharing community, we demonstrate that NATERGM provides an enhanced pattern testing capability compared to TERGM. Moreover, NATERGM is able to predict the characteristics of social media networks in future and we show that our approach outperforms TERGM-based prediction models. The major objective of this study is to provide a framework to explore, analyze, and explain the dynamic formation mechanisms of social media networks.
The remainder of this chapter is organized as follows. Section 4.2 reviews relevant social network literature to provide the background for this research and address the need for analyzing temporal patterns in social media networks. Relevant dynamic network models are also briefly introduced and research gaps are summarized. Section 4.3 presents our NATERGM model. Empirical tests that demonstrate the pattern testing and network prediction performance of the model are outlined in Section 4.4. Results are shown and implications are discussed in Section 4.5. Finally, we conclude the study in Section 4.6.

4.2 Related Work

In this section we first review recent studies examining social media networks. Then, we review emerging network models for dynamic network analysis.

4.2.1 Social Media Networks

Table 4.1 shows selected recent studies on social media networks. They are summarized in a taxonomy based on tie types, research objectives, analytical methods, and network dynamics.

Based on a theoretical conceptualization of network ties (Borgatti et al. 2009), four types of social media network ties have been summarized in prior research (Kane et al. 2014). Proximity ties represent that two individuals belong to the same sub-communities (e.g., Facebook Group) or locational areas. Social relation ties represent social connections between individuals, such as virtual friendships and subscription relationships in micro-blogging sites (Kwak et al. 2010; Shriver et al. 2013). Interaction
ties represent interactive behaviors between individuals, such as information exchanges via message replies (Stewart and Abidi 2011). Flow ties represent the movement of goods or information between network nodes, such as retweets (Baker 2015).

<table>
<thead>
<tr>
<th>Authors</th>
<th>Tie Types</th>
<th>Objectives</th>
<th>Analytical Methods</th>
<th>Network Dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grabowicz et al. (2012)</td>
<td>Social Relation</td>
<td>Community Detection</td>
<td>Clustering</td>
<td>N</td>
</tr>
<tr>
<td>Faraj and Johnson (2011)</td>
<td>Interaction</td>
<td>Network Mechanisms</td>
<td>ERGM</td>
<td>N</td>
</tr>
<tr>
<td>Lin et al. (2013)</td>
<td>Social Relation</td>
<td>Structural Capital</td>
<td>Regression</td>
<td>N</td>
</tr>
<tr>
<td>Qiu et al. (2013)</td>
<td>Social Relation</td>
<td>Structural Capital</td>
<td>Simulation</td>
<td>N</td>
</tr>
<tr>
<td>Shriver et al. (2013)</td>
<td>Social Relation</td>
<td>Structural Capital</td>
<td>Regression</td>
<td>Partially addressed</td>
</tr>
<tr>
<td>Stewart and Abidi (2012)</td>
<td>Interaction</td>
<td>Community Detection</td>
<td>Clustering</td>
<td>N</td>
</tr>
<tr>
<td>Traud et al. (2012)</td>
<td>Social Relation</td>
<td>Network Mechanisms; Community Detection</td>
<td>Modularity optimization; ERGM</td>
<td>N</td>
</tr>
<tr>
<td>Wang et al. (2013)</td>
<td>Flow</td>
<td>Community Detection</td>
<td>Regression</td>
<td>Y</td>
</tr>
<tr>
<td>Wang et al. (2014)</td>
<td>Interaction</td>
<td>Community Detection</td>
<td>Random Walk</td>
<td>Y</td>
</tr>
</tbody>
</table>

Some researchers have argued that these types of ties are not necessarily decoupled, but represent a continuum (Atkin 1977). For example, proximity may further lead to social relations; interactions and flows of knowledge may occur at the same time.
Social media networks have been studied for different purposes. In general, the research objectives of these studies can be classified into three categories. The first stream of research focuses on explaining network mechanisms. This type of research aims at understanding in what conditions individuals are more likely to establish social connections online. For example, demographic homophily was found to exist in online friendship networks (Traud et al. 2012). Students of the same gender, major, and residence area were more likely to establish social connections in Facebook friendship networks. Prior research has also found that direct reciprocity, indirect reciprocity, and preferential attachment occur very frequently in online web forums (Faraj and Johnson 2011). The second stream of research examines how the structure of a social media network affects the outcomes of individuals in the network. This type of research is referred to as structural capital studies (Borgatti and Foster 2003). For example, an examination of friendship networks in an online micro-lending platform led to discoveries that the chances of successful funding were significantly affected by the number of friendship ties and by the types of friendship (Lin et al. 2013). Research has found that individuals in a connected network are able to predict outcomes of a given problem more accurately, compared to the cases when they are isolated (Qiu et al. 2013). Another popular research area is to partition the network into sub-graphs and detect sub-communities. These studies usually aim at identifying key groups or players in the network and understanding the characteristics of these sub-communities. For example, based on centrality and coreness measures, core groups and key members in the core group who were most active were identified in a clinical discussion forum (Stewart and
Abidi 2011). Another study identified Twitter user clusters from following-follower networks in Twitter.com, and examined the influence of intra-group ties, inter-group ties, and intermediary ties on retweeting behaviors (Grabowicz et al. 2012).

Previous studies focusing on community detection mainly use clustering or modularity optimization algorithms (Newman 2006). In structural capital studies, regression analysis has been frequently used to examine the relationships between network structures and individual outcomes. Dependent variables are the outcomes of network nodes, such as funding success (Lin et al. 2013) and online users’ activity levels (Shriver et al. 2013). Independent variables can be various network metrics of the nodes, such as degree centrality, betweenness centrality (Freeman 1979), and structural holes (Burt 1995). To explain the mechanisms of network formation, network models can be used, such as the Latent Space Model (Hoff et al. 2002), p1 models (Holland and Leinhardt 1974), and the ERGM (Wasserman and Robins 2005). In social media network research, ERGM has received increased attention recently (Faraj and Johnson 2011; Jiang et al. 2013; Jiang et al. 2014b; Traud et al. 2012). ERGMs are statistical models that test whether observed networks show theoretically hypothesized structural tendencies (Robins and Pattison 2005; Wasserman and Pattison 1996). These structural tendencies, or configurations, are subsets of nodes and ties in the network, reflecting certain types of network sub-structures. Examples of typical configurations can be “triangle” and “k-star” (Robins et al. 2007a; Robins et al. 2007b). In addition, nodal attributes can be incorporated in a configuration.
Although various analytical methods have been used to study social media networks, studies that address the dynamics of social media networks are still scarce. Only a few studies have taken into account the time information relating to when network ties are developed. For instance, Shriver et al. (2013) considered the number of friendship ties at previous time points in their time series regressions. Another study analyzed the sequence in which retweeting links were activated in micro-blogging sites, and found that the extent to which an individual could reach other parts of the network positively affected the popularity of the content posted by that individual (Wang et al. 2013).

Overall, the dynamics of social media networks have been addressed in few prior studies. Nevertheless, dynamic network analysis is an emerging area of network research, and relevant studies have been conducted in biology, neural science, healthcare, and social science domains. We review existing dynamic network analysis approaches next.

4.2.2 Dynamic Network Analysis

Generally, two different approaches can be used for dynamic network analysis. Cross-sectional approaches analyze network data where time information is embedded within the network. Longitudinal approaches observe networks at multiple time points and track the evolution of networks based on comparisons (Tang et al. 2012). Previous research has proposed various dynamic network models, including both types of approaches, for studying the dynamic process of network formation, evolution, and dissolution. We review selected dynamic network models next.
Temporal Exponential Random Graph Model (TERGM) is an extension of the ERGM for dynamic networks (Guo et al. 2007; Hanneke et al. 2010; Kolar et al. 2010). A simple TERGM model under the first-order Markov dependency can be written as:

$$
\Pr(Y^t = y^t | Y^{t-1} = y^{t-1}) = \left( \frac{1}{\kappa} \right) \exp \left\{ \sum_A \eta_A \cdot g_A(y^t, y^{t-1}) \right\}
$$

(4.1)

In Chapter 1, Equation (1.1) specified the expression of ERGM. The Equation (4.1) uses the same notation, where $Y$ is a matrix of random variables representing network ties and $y$ is its realization; $\eta_A$ is a parameter corresponding to configuration $A$, positively related to the likelihood of configuration $A$ to occur; $g_A()$ is network statistics corresponding to $A$; $\kappa$ is a normalizing constant ensuring that $\Pr(Y)$ is a probabilistic distribution. Note that the major difference between (4.1) and (1.1) is the specification of network statistics for each temporal pattern $A$ (i.e., $g_A()$), which is now determined by network realizations in multiple observational time points (observed at $t$ and $t-1$ in this case). Given multiple observations, TERGM can be used to test whether a certain temporal pattern is more likely to occur than by chance. For example, as illustrated in Figure 4.2, three different temporal patterns can be derived from a transitivity pattern, depending on the sequence in which the three ties develop. Compared to the conventional ERGM where only a tendency for transitivity can be tested, TERGM differentiates between three different sequences of network ties which all finally lead to the same
transitivity pattern. TERGM can further test the likelihood of each temporal pattern to occur.

In addition to the transitivity in this example, TERGM can also include network configurations of many other types such as temporal stability and temporal reciprocity (Hanneke et al. 2010; Krivitsky and Handcock 2014). TERGM can also be applied to cross-sectional data if time duration information for network ties is provided. However, none of the TERGM research has considered the effects of nodal attributes in temporal network configurations.
Separable Temporal Exponential Random Graph Model (STERGM) separates TERGM into a formation model and a dissolution model, thereby modeling not only the temporal patterns of network formation, but also the temporal patterns of network dissolution (Goodreau et al. 2014; Krivitsky 2012; Krivitsky and Handcock 2014). STERGM addresses the concern that some existing network ties might disappear over time, such as a broken friendship, for example. STERGM identifies new connections and dissolved ties by comparing networks at multiple time points. A variant of STERGM for cross-sectional data is also proposed for the case when longitudinal data is unavailable (Krivitsky 2012).

Hidden Temporal Exponential Random Graph Model (HTERGM) is a model that combines TERGM with hidden Markov models (Guo et al. 2007). It assumes that (1) network structure at time $t$, $Y_t$, is dependent on the structure of the network in the previous time point $Y_{t-1}$, and (2) nodal attributes of the network, $x_t$, are dependent on the network structure $Y_t$. It further assumes that only nodal attributes are observable, while network structures are hidden states. The major aim of HTERGM is to estimate the transition probabilities $P(Y_t|Y_{t-1})$ and emission matrices $\Lambda = P(x_t|Y_t)$ so that hidden network structures can be inferred given time series of nodal attributes $x_1, x_2, \ldots, x_t$. However, HTERGM does not explain how nodal attributes affect the formation process of networks.

Temporally Randomized Reference Models (TRRM) investigates the dynamic characteristics of networks by comparing observed networks with an ensemble of temporally randomized networks (Holme 2003; Holme and Saramäki 2012; Karsai et al. 2012).
Temporal randomization generates new networks by rewiring ties in the original networks or changing time information associated with the ties. Typical randomization methods include randomized edges, randomly permutated times, random times, edge randomization, and time reversal (Holme and Saramäki 2012). Figure 4.3 shows examples of randomized edges and randomly permutated times. By comparing original networks with temporally randomized networks, key dynamic characteristics of original networks can be understood. For example, Holme (2003) compared e-mail networks with their temporally randomized samples and found that in general the average time it took to pass information between network nodes is longer in the original email networks.

(a) an original network with numbers indicating the sequence of tie activation; (b) a randomized network by iteratively rewiring network ties among four selected nodes; and (c) another randomized network by permuting the time associated with ties.

Figure 4.3 Network Temporal Randomization

Latent space models (Hoff et al. 2002) assume that each node in a network is associated with a latent position in a low dimensional space. The probabilities of tie occurrences are determined by the distances between nodes in the latent space. The latent space model estimates the parameters associated with latent positions based on the
observed networks. The estimated model can be used to visualize a spatial representation of network relationships (Hoff et al. 2002; Krivitsky et al. 2009). Dynamic Latent Space Model (DLSM) is an extension of the latent space model and allows the latent positions to change over time (Sarkar and Moore 2005; Sarkar et al. 2007).

4.2.3 Research Gaps

Based on the prior literature, several research gaps can be identified. First, social media networks are dynamic in nature. However, little research has explained the mechanisms of network formation with a dynamic perspective. Dynamic network analysis has been frequently used to detect communities from networks (Tang et al. 2012; Wang et al. 2014), but not to explain the mechanisms of network formation. Most network mechanisms studies focused on identifying static network patterns, but did not explain how these patterns developed dynamically. Second, emerging network research has given rise to various approaches for examining temporal networks and has suggested that the sequence of network ties is an important aspect of network dynamics (Hanneke et al. 2010; Holme and Saramäki 2012; Wang et al. 2013). However, none of the existing dynamic network models have focused on how nodal attributes affect the sequence of network ties. In addition, network prediction has been an under-studied research area (Goldenberg et al. 2010). Although prior research has helped identify dynamic network patterns, little has been done to predict future networks based on the identified patterns.

4.3 Nodal Attribute-based Temporal Exponential Random Graph Model

The proposed NATERGM focuses on how nodal attributes of networks affect the sequence in which network ties develop. Because the sequence of network ties needs to
be tracked accurately, NATERGM examines cross-sectional network data with time information for network ties. Figure 4.4 presents the framework of NATERGM. The major components include network extraction, temporal pattern analysis, and network prediction. In the network extraction step, social connections are identified between individuals in social media, along with the timestamps of these relationships and nodal attributes of the individuals. Temporal patterns of the networks are modeled, and the likelihood of each pattern is estimated in the temporal pattern analysis step. Based on the estimated model, new networks are simulated and compared to the original network to evaluate how effectively the model can predict future networks. We explain each component in the following subsections.

Figure 4.4 NATERGM Framework

4.3.1 Network Extraction

First, network ties are extracted from social media based on relationships between online users. Among the various types of social media network ties summarized by Kane et al. (2014), the interaction/flow and social relation ties are the ones that are the most dynamically established (i.e., these ties are often associated with timestamps). Different
types of network ties can be identified depending on specific social media contexts. For example, directed interaction/flow ties can be established if an individual sends greetings to another individual; undirected social relation ties can be established if two individuals become friends by using friending functions provided in social media platforms. After identifying network ties between all possible pairs of individuals, a network of size N is represented by a matrix $Y=[Y_{ij}]$, $(i, j =1, 2, \ldots N)$. For undirected networks, $Y_{ij}=1$ if a tie exists between nodes (i.e., individuals) i and j, and $Y_{ij}=0$ otherwise. For directed networks, $Y_{ij}=1$ if a tie starts from i and ends at j, and $Y_{ij}=0$ otherwise.

For timestamp modeling, we use $T_{ij}$ to represent the time when each network tie (i, j) is established. A matrix $T=[T_{ij}]$, $(i, j =1, 2, \ldots N)$ records the timestamps for all network ties and can be used to model the sequence of network ties. For example, if $T_{12}<T_{21}$, it would represent a process where node 1 sent out a tie to node 2 first, and then received a tie from the node 2 in return.

Nodal attributes of individuals can be evaluated using different approaches. Prior studies have characterized individual social media users based on three types of features. Platform-based features refer to individual attributes that are directly provided by social media platforms. For example, registered users are often associated with usernames while an unregistered user is represented by a "visitor" tag or an IP address in the name space. Some social media platforms also assign functional roles to users such as members or administrators. This type of information can be directly used as nodal attributes of individuals. Textual features refer to attributes that are inferred by texts posted by the individuals. Social media users typically leave many textual traces, such as private
messages and message postings. Various characteristics of social media users can be evaluated based on these texts, such as general opinions, writing proficiency, and topics of interests. Social network features refer to individual attributes that are inferred by their connections or positions in the network. Social relations between individuals in part reflect their personality, status, and roles. For example, an individual who is linked with many others is expected to have a high level of popularity compared to others who have fewer connections. Such information can thus be used as nodal attributes of individuals. After evaluating the nodal attributes of individuals, they are represented by a vector $X = (x_1, x_2, ..., x_N)$.

4.3.2 Temporal Pattern Analysis

To model temporal patterns, the nodal attributes and timestamps of network ties are used to represent various temporal patterns regarding the dynamics of network formation. By taking into account the sequence in which network ties develop, common static network patterns such as reciprocity, k-star, transitivity, and cyclicity can have different temporal variations. Tables 4.2 to 4.6 list examples of temporal patterns for directed networks. White nodes represent individuals in general and black nodes represent individuals with key nodal attributes (e.g., highly active individuals). Dashed arrows represent network ties that developed after solid ones.
Table 4.2 NATERGM Temporal Patterns for Directed Networks: Reciprocity

<table>
<thead>
<tr>
<th>Static Pattern</th>
<th>Temporal Pattern</th>
<th>Illustration</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>reciprocity</td>
<td>feedback</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to receive feedback.</td>
</tr>
<tr>
<td></td>
<td>response</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to respond to incoming ties.</td>
</tr>
</tbody>
</table>

Table 4.3 NATERGM Temporal Patterns for Directed Networks: K-out-star

<table>
<thead>
<tr>
<th>Legend: General nodes</th>
<th>Nodes with key attributes</th>
<th>Ties activated at $T_1$</th>
<th>Ties activated at $T_2 (T_1 &lt; T_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-out-star</td>
<td>prioritization</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to be prioritized when forming relationships.</td>
</tr>
<tr>
<td></td>
<td>de-prioritization</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to be de-prioritized when forming relationships.</td>
</tr>
</tbody>
</table>

Table 4.4 NATERGM Temporal Patterns for Directed Networks: K-in-star

<table>
<thead>
<tr>
<th>Legend: General nodes</th>
<th>Nodes with key attributes</th>
<th>Ties activated at $T_1$</th>
<th>Ties activated at $T_2 (T_1 &lt; T_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-in-star</td>
<td>initiative</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to take the initiative in multi-actor relationships.</td>
</tr>
<tr>
<td></td>
<td>laziness</td>
<td><img src="image" alt="Illustration" /></td>
<td>Nodes with some attribute have a high tendency to hold off in multi-actor relationships.</td>
</tr>
</tbody>
</table>
Table 4.5 NATERGM Temporal Patterns for Directed Networks: Transitivity

<table>
<thead>
<tr>
<th>Transitivity</th>
<th>Pattern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bridge</td>
<td><img src="image" alt="Diagram" /></td>
<td>Nodes with some attribute have a high tendency to bridge new relationships between others.</td>
</tr>
<tr>
<td>co-supporting</td>
<td><img src="image" alt="Diagram" /></td>
<td>If two nodes are supporting a common node with some attribute, they have a high tendency to build a new relationship.</td>
</tr>
<tr>
<td>co-supported</td>
<td><img src="image" alt="Diagram" /></td>
<td>If two nodes are supported by a common node with some attribute, they have a high tendency to build a new relationship.</td>
</tr>
<tr>
<td>remarked-supporter</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to receive attention from another node, if both co-support a common node.</td>
</tr>
<tr>
<td>remarked-supported</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to receive attention from another node if they are supported by a common node.</td>
</tr>
<tr>
<td>remarking-supporter</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to pay attention to another node, if they co-support a common node.</td>
</tr>
<tr>
<td>remarking-supported</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to pay attention to another node, if they are supported by a common node.</td>
</tr>
<tr>
<td>follow-up</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to pay attention to another node, if a third node bridges their relationship.</td>
</tr>
<tr>
<td>reference</td>
<td><img src="image" alt="Diagram" /></td>
<td>A node with some attribute has a high tendency to receive attention from another node, if a third node bridges their relationship.</td>
</tr>
</tbody>
</table>
As can be seen from the table, the temporal patterns modeled by NATERGM provide an extended hypothesis testing capability about network formation compared to static patterns. In particular, these temporal patterns can be used to examine the roles of nodal attributes in determining the sequence of network ties. For example, assuming that we are interested in the role of highly active individuals in developing message flows in social media, the static reciprocity pattern would only model a tendency for two individuals (at least one of them being highly active) to exchange messages. In comparison, if we observed many “feedback” patterns in the network, it would suggest a tendency for highly active individuals to receive returning messages after they sent out messages first; if we observed many "response" patterns, it would suggest a tendency for highly active individuals to respond to others' incoming messages. Although both
"feedback" and "response" patterns finally lead to the same "reciprocity" pattern, they model two distinct dynamic processes. In a similar way, NATERGM extends other static patterns (i.e., k-star, transitivity, and cyclicity) to their temporal variations by considering the possible sequences of network ties, which provides richer insight about the dynamic process of network formation.

Given the list of temporal patterns in Tables 4.2 to 4.6, the major objective of NATERGM is to test which of these temporal patterns are more likely to be observed than to occur by chance in a network. The NATERGM model can be written as:

$$\Pr(Y = y | \eta) = \left( \frac{1}{\kappa} \right) \exp \left\{ \sum_{a \in A} \eta_a g_a(y, T, X) \right\}$$ (4.2)

In (4.2), A is a set of temporal patterns to be tested, $\eta = [\eta_a]$ is a vector of parameters representing the strength of each temporal pattern’s effect in network formation, and $\kappa$ is a scaling parameter to ensure (3) is a probability distribution. $g_a(\bullet)$ is the network statistic of temporal pattern a, evaluated with network y, timestamp matrix T, and vector of nodal attributes X. Table 4.7 provides definition of $g_a(\bullet)$ for each temporal pattern listed in Tables 4.2 to 4.6, with the assumption that nodal attributes are binary or categorical. I() is an indication function that takes the value 1 if and only if the expression inside results in TRUE values. For cases when nodal attributes are continuous variables, I(X_j) is replaced by the value of X_j.
Table 4.7 Specification of NATERGM Terms  
(Directed Network, Binary or Categorical Attributes)

<table>
<thead>
<tr>
<th>NATERGM Term</th>
<th>Network Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>reciprocity</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i \leq T_j)$</td>
</tr>
<tr>
<td>feedback</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>response</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i \leq T_j)$</td>
</tr>
<tr>
<td>2-out-star</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>prioritization</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &lt; T_j)$</td>
</tr>
<tr>
<td>deprioritization</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>2-in-star</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>initiative</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &lt; T_j)$</td>
</tr>
<tr>
<td>laziness</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>transitivity</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &lt; T_j)$</td>
</tr>
<tr>
<td>bridge</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>cosupporting</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &lt; T_j)$</td>
</tr>
<tr>
<td>cosupported</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>remarked-supporter</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>remarked-supported</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>remarking-supporter</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>remarking-supported</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>follow-up</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>reference</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>cyclicity</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>reversed_reference</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>reversed_followup</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
<tr>
<td>reversed_bridge</td>
<td>$\gamma_{\delta}(y,T,X) = \sum_{i\neq j} y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_i &gt; T_j)$</td>
</tr>
</tbody>
</table>

The likelihood of occurrence for each temporal pattern can be assessed by estimating the parameters $\eta$. If a parameter is positive and significant, it indicates that the corresponding temporal pattern appears more frequently than by chance in the network.

For parameter estimation, the Markov Chain Monte Carlo (MCMC) method is used,
following prior ERGM literature (Snijders 2002). The procedure is modified to adapt to temporal settings. In general, the model fitting procedure iteratively generates random networks based on the given set of parameters and updates the parameters based on the difference between the generated networks and the observed network. For a given set of parameters $\eta = [\eta_a]$, the following algorithm was developed to generate random networks.

**Algorithm 4.1 NATERGM Random Network Generation**

Initialize network as $Y = Y^{(t=0)}$

**repeat** until maximum rounds of iterations are made

**for** each element $Y_{ij}$ in $Y^{(t)}$:

change the value of $Y_{ij}$ based on the conditional distribution defined by

$$
\logit \{ \Pr(Y_{ij} = 1 | Y_{kl} = y_{kl} \text{ for all } (k,l) \neq (i,j)) \} = \eta^T \left( g(y^{(ij1)}, T_{\text{Gibbs}}, X) - g(y^{(ij0)}, T_{\text{Gibbs}}, X) \right) 
$$

(4.3)

**end for**

$t \leftarrow t + 1$

return $Y(t)$

$y^{(ij1)}$ and $y^{(ij0)}$ are matrices that only differ in element $Y_{ij}$, taking 1 or 0 respectively. $g = [g_{A1}, g_{a1}, g_{a1}, \ldots, g_{aM}]$, and definitions for $g(\bullet)$ can be found in Table
6. $T_{Gibbs}$ is a timestamp matrix where $T_{ij} > T_{kl}$ for all $(k,l) \neq (i,j)$. It assumes that the stochastic process $Y(t)$ develops over time. Given the random network generation procedure, the following parameter updating procedure is used to estimate parameter values.

**Algorithm 2. NATERGM Parameter Updating**

initialize $\eta = \eta^{(0)}$

**repeat** from n=0:

generate K networks $(y_1, y_2 \ldots y_K)$ independently based on $\eta^{(n)}$ and **Algorithm 1**

define

$$\bar{g} = \left(\frac{1}{K}\right) \sum_{k=1}^{K} [p_k^{(n)} g(1 - y_k^{(n)}) + (1 - p_k^{(n)}) g(y_k^{(n)}) - g_0]$$  \hspace{1cm} (4.4)$$

and

$$D_0 = diag\left(\left(\frac{1}{K}\right) \sum_{k=1}^{K} [p_k^{(n)} g^T(1 - y_k^{(n)}) g(1 - y_k^{(n)})

+ (1 - p_k^{(n)}) g^T(y_k^{(n)}) g(y_k^{(n)}) - g^T \bar{g} \bar{g}]\right)$$  \hspace{1cm} (4.5)$$

calculate $Z^{(n)} = (Z_1^{(n)}, Z_2^{(n)}, \ldots, Z_K^{(n)}),$

$$Z_k^{(n)} = p_k^{(n)}(y) \left(1 - y_k^{(n)}\right) + (1 - p_k^{(n)}(y)) g\left( y_k^{(n)} \right) - g_0]$$  \hspace{1cm} (4.6)$$

where $p_k(y) = \frac{\exp (\eta^T y_k^{(n)} g(1 - y_k))}{\exp (\eta^T g(1 - y_k)) + \exp (\eta^T g(y_k))}$  \hspace{1cm} (4.7)$$
update $\eta^{(n+1)}$ using Robins-Monro Algorithm

$$\eta^{(n+1)} = \eta^{(n+1)} - s_n D_0^{-1} z^{(n)}, \text{ n} \leftarrow \text{n}+1$$  \hspace{1cm} (4.8)

until convergence criterion is met

$s_n$ is a sequence of positive numbers converging to 0. In this study we used $s_n=2\exp(n)/10$, as suggested in prior research (Snijders 2002). For convergence criterion, we also used the t-ratio methods (Snijders 2002).

4.3.3 Network Prediction

After estimating the parameters in NATERGM, the fitted model can be used to predict the characteristics of future networks with the following procedures.

Based on the actual network observed at time point t-1, NATERGM parameters $\eta_{t-1}$ are estimated. A number (=K) of networks at time point t are then simulated based on the parameters $\eta_{t-1}$ using Algorithm 1. However, network at the time point t-1 is used as the initial network, instead of a randomly initialized network.

Each generated network at time point t does not necessarily look exactly like the actual network at time point t. However, network statistics averaged over K generated networks should resemble those of the actual network. In order to evaluate how close the generated networks are to the actual network, we calculate the absolute difference (AD) for each network statistic $a'$ at prediction period t:

$$AD_{a'} = |g_{a'}(y_0^t) - \left(\frac{1}{K}\right) g_{a'}(y_{k}^t)|$$  \hspace{1cm} (4.9)

where $y_0^t$ is the observed network at t, and $y_{k}^t$ (k=1,2,..., K) is the k-th generated network.
based on the fitted model at t. Small difference would indicate that the estimated model predicts the network well.

4.4 Research Design

In order to evaluate the performance of NATERGM, we conducted two empirical tests. The first test focused on the pattern testing capability of NATERGM. The second test focused on how accurately our model can predict the characteristics of future networks. This section describes the research test-bed used for empirical study and outlines the two experiments.

4.4.1 Research Test-bed

Social media data were collected from WikiAnswer.com, which is a large online knowledge sharing community. Community members can ask questions about any topics, and answer others’ questions as well. Open questions go through the hands of many contributors, some of whom directly provide answers, while others edit the posted answers in terms of content, language, or format. Finally, the questions and answers are organized into Q&A entries that can be accessed by all community members including the questioners. It is a good test-bed for testing NATERGM because its wiki-based “answer history” system allows us to see how members in this community develop social connections when seeking help and answering questions. We established a directed tie from member A to member B if A answered a question from B. Therefore, network ties represent knowledge flows in this community and as a result, knowledge diffusion.
networks were extracted from social media. The test-bed also allows for identifying timestamps associated with these network ties.

Since many questions require contributors to have relatively deep knowledge in a field to answer, members in WikiAnswer.com form specialized sub-communities to handle questions that belong to similar topics. In this study, we focused on three sub-communities: diabetes, online shopping, and real estate, based on the popularity of these topics and the number of relevant Q&A entries in the community. Furthermore, we observed that a great number of members were inactive and only participated in knowledge sharing activities very limitedly (e.g., only made a change to capitalization once). Since we were interested in the most representative members in the community, we restricted analysis within members who asked questions, provided the first answers, or made significant content change (this type of contribution is separately classified in WikiAnswer.com) to answers more than once. The data collection statistics are shown in Table 4.8.

<table>
<thead>
<tr>
<th>Community</th>
<th># of Q &amp; A Entries</th>
<th>Total # of Members</th>
<th># of Members for Analysis</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>27,349</td>
<td>70,164</td>
<td>2,499</td>
<td>2008-2013</td>
</tr>
<tr>
<td>Online Shopping</td>
<td>14,111</td>
<td>39,631</td>
<td>1,058</td>
<td>2008-2013</td>
</tr>
<tr>
<td>Real Estate</td>
<td>10,725</td>
<td>36,517</td>
<td>963</td>
<td>2008-2013</td>
</tr>
</tbody>
</table>
4.4.2 Pattern Testing

In order to show the enhanced pattern testing capability of NATERGM, we compared it with TERGM and used both models to explain network formation. TERGM was chosen as the baseline model because it is also capable of modeling the sequence of network ties. However, it does not explain what roles nodal attributes play in determining the sequence of network ties.

NATERGM and TERGM used for pattern testing included different sets of model terms. Baseline model terms included arc, reciprocity, 2-out-star, 2-in-star, transitivity, and cyclicity. For all the significant terms in the baseline model, they were further extended into corresponding temporal patterns in Table 2 and tested using NATERGM. We evaluated three types of nodal attributes for pattern testing with NATERGM.

For platform-based features, we identified registered members and unregistered visitors (attr=reg). Although WikiAnswer.com does not mandate a new user to register for posting or answering questions, registered members can accumulate “trust points” and obtain honor badges based on their contributions over time. We expected that registered members might have more commitment to the community and more willingness to contribute than unregistered anonymous visitors, thereby showing different behaviors from unregistered visitors when developing network ties. The attribute was measured as a binary variable, where reg = 1 if the member was registered, and reg=0 otherwise.

For textual features, we evaluated writing proficiencies (attr=pro) for community members. Writing proficiency reflects an individual’s level of literacy, expertise, and educational background (Jiang et al. 2014a; Zheng et al. 2006). We expected that
members with high levels of writing proficiency would contribute significantly to the online knowledge sharing communities. Hence, we were interested in understanding what roles these members would play in the formation process of the dynamic knowledge diffusion networks. Prior linguistic studies have suggested that writing proficiency can be assessed based on various factors (Li 1997; Nadarajan 2011). We employed text mining techniques to evaluate the following five metrics of each member in WikiAnswer.com. First, the average length of a member’s answers was evaluated because it reflects the depth of the member’s knowledge about the problem (Larsen-Freeman 2009). Second, T-unit (a single main clause + other subordinate clauses) is an index of syntactic complexity (Hunt 1977) and reflects the member’s effectiveness in organizing words in sentences. We used the number of words per T-unit to evaluate this metric. Third, lexical richness measures the extent to which different words are used in texts. We used Hapax Legomena and Hapax Dislegomena to evaluate lexical richness, following prior studies (Jiang et al. 2014a; Zheng et al. 2006). Fourth, objectivity reflects whether answers are provided without biases. Objectivity of a member was evaluated based on the percentage of sentences that are classified as “objective” by the Opinion Finder (OF) System (Wiebe et al. 2005; Wilson et al. 2005) of all sentences posted by that member. Fifth, we used one minus percentage of misspellings (i.e., percentage of correct spellings) to measure the readability of texts for each member. Finally, these metrics were normalized and averaged to represent the member’s writing proficiency.

For social network features, we evaluated out-degree centrality (attr=odc) of each member. The out-degree centrality of a member in this knowledge sharing community
reflects how actively the member helps answering others’ questions. Individuals with high out-degree centrality should play a critical role in the community, and thus we were interested in understanding how they would impact network formation dynamically. The out-degree centrality of a member was measured based on the number of out-going ties associated with him or her, and was normalized to $[0, 1]$.

4.4.3 Network Prediction

To evaluate the performance on network prediction, we compared the prediction results of NATERGM and TERGM. For NATERGM, the prediction procedure followed the procedure described in Section 3.3. For TERGM, the model was trained using the previous two time points $t-2$ and $t-1$ in order to predict network at time point $t$. As for network statistics to be compared, we focused on comparing the degree distribution of simulated networks and the actual network following previous studies, and the set of network statistics $A'$ included standard deviation and skewness of in-degree and out-degree distributions of nodes. Prediction was conducted for each month during 2008-2013. Absolute difference vectors $(\text{AD}_a^1, \text{AD}_a^2, \ldots, \text{AD}_a^{72})$ were obtained for both the baseline model and NATERGM.

The Wilcoxon signed rank test was used to validate that the prediction errors of the NATERGM were statistically lower than that of the baseline model. The procedure first calculated the differences of absolute prediction errors between the two models ("TERGM" minus "NATERGM"), and ranked the pairs according to the absolute values of the differences (from lower to higher). It then calculated the sum of the ranks where the differences were positive, as shown in Equation (4.10).
\[ W^+ = \sum_{t=1}^{T} I(AD^{t\text{ baseline}} - AD^{t\text{ NATERGM}} > 0) \cdot R_i \]  

(4.10)

T is the total number of periods for comparison, I() is an indication function that the prediction error of baseline is greater than that of the NATERGM, and \( R_i \) is the rank of the pair.

The Z statistic was calculated as \( Z = (|W^+ - \mu| - 0.5)/\sigma \). High Z values would indicate that the differences of prediction errors between the two models were statistically significant.

4.5 Results and Discussions

4.5.1 Pattern Testing Results

For each of the selected nodal attributes including registration status (reg), writing proficiency (pro), and out-degree-centrality (odc), we tested how the attribute affected the dynamic process of network formation by fitting the NATERGM model to networks extracted from each of the three sub-communities in our test-bed. Table 4.9 shows the estimated parameters for the baseline model and NATERGM.
Table 4.9 Parameter Estimates for TERGM and NATERGM

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Terms</th>
<th>Estimate (S.E.)</th>
<th>Diabetes</th>
<th>Online Shopping</th>
<th>Real Estate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(S.E.)</td>
<td>(S.E.)</td>
<td>(S.E.)</td>
</tr>
<tr>
<td>TERGM</td>
<td>arc</td>
<td>-5.11*** (0.72)</td>
<td>-4.85*** (0.78)</td>
<td>-5.07*** (0.63)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>reciprocity</td>
<td>0.59 (1.01)</td>
<td>0.41 (1.23)</td>
<td>0.59 (0.97)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-out-star</td>
<td>2.51*** (0.40)</td>
<td>2.34*** (0.44)</td>
<td>2.74*** (0.38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-in-star</td>
<td>2.56* (0.85)</td>
<td>2.45* (0.79)</td>
<td>2.63** (0.73)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>transitivity</td>
<td>1.31 (1.05)</td>
<td>1.16 (0.72)</td>
<td>1.28 (1.10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cyclicity</td>
<td>-0.09 (0.72)</td>
<td>0.41 (0.46)</td>
<td>-0.17 (0.24)</td>
<td></td>
</tr>
<tr>
<td>NATERGM</td>
<td>arc</td>
<td>-5.26*** (0.81)</td>
<td>-5.16*** (0.62)</td>
<td>-5.33*** (0.59)</td>
<td></td>
</tr>
<tr>
<td>(NA=reg)</td>
<td>2-out-star extended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[reg]-prioritization</td>
<td>2.66*** (0.37)</td>
<td>2.35*** (0.45)</td>
<td>2.78** (0.63)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[reg]-deprioritization</td>
<td>2.23* (1.01)</td>
<td>2.39* (0.89)</td>
<td>2.26* (0.94)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-in-star extended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[reg]-initiative</td>
<td>3.43** (0.97)</td>
<td>3.70** (1.01)</td>
<td>2.83* (0.95)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[reg]-laziness</td>
<td>1.51 (1.01)</td>
<td>1.23 (1.12)</td>
<td>1.58 (1.57)</td>
<td></td>
</tr>
<tr>
<td>NATERGM</td>
<td>arc</td>
<td>-4.72*** (0.73)</td>
<td>-5.05*** (0.67)</td>
<td>-5.27*** (0.81)</td>
<td></td>
</tr>
<tr>
<td>(NA=pro)</td>
<td>2-in-star extended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[pro]-laziness</td>
<td>2.68*** (0.58)</td>
<td>2.99** (0.70)</td>
<td>3.22** (0.73)</td>
<td></td>
</tr>
<tr>
<td>NATERGM</td>
<td>arc</td>
<td>-5.94*** (1.01)</td>
<td>-5.41*** (0.79)</td>
<td>-5.72*** (0.93)</td>
<td></td>
</tr>
<tr>
<td>(NA=odc)</td>
<td>2-in-star extended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[odc]-initiative</td>
<td>3.87*** (0.61)</td>
<td>3.46*** (0.53)</td>
<td>3.81*** (0.73)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: *: p<0.05, **: p<0.01, ***: p<0.001. Significant variables are marked in boldface. NA=Nodal Attribute.

The estimated parameter values are log-odds of the ties of corresponding patterns forming in the network. These values are positively related to the probabilities that corresponding patterns would occur.

For the nodal attribute of registration status, similar results were observed in all three sub-communities. In the baseline model, we observed that “2-out-star” and “2-in-
"starterms" were significant. It indicates that members in the community were likely to help more than one peer over time by providing answers to their questions, and questioners were also likely to receive answers from multiple contributors over time. When these patterns were further extended to their corresponding temporal terms in NATERGM, additional insights could be obtained. For example, when temporal 2-out-star patterns were considered, we found that both “[reg]-prioritization” and “[reg]-deprioritization” terms were positive and significant. This suggests that when a member helped others, he or she did not prioritize registered members over unregistered visitors in terms of time. In other words, everyone in the community was treated equally in terms of receiving answers. However, for temporal 2-in-star patterns, only the “[reg]-initiative” term was significant. It suggests that when a member receives answers from multiple contributors, the answers were likely to come from registered members first. An explanation is that registered members may have more commitment and want to establish a good image in the community. As a result, registered members may try to help others as soon as possible. Using NATERGM terms, we saw that the proposed model was able to reveal how registered members affect the dynamic process of knowledge diffusion.

When writing proficiency was considered, none of the extended 2-in-star terms were significant. However, when extending the 2-out-star terms, we found that only the term “[pro]-laziness” was positive and significant. This suggests that when a member receives answers from multiple contributors, answers from members with good writing proficiency tended to arrive later. A possible reason is that members with good writing proficiency try to provide well-constructed, unbiased, and helpful answers to questioners.
They may take some time to do research before providing answers, which delays the time when they delivered answers to questioners. Similar results were observed in all-three sub-communities. Using NATERGM terms, we saw that the proposed model was able to reveal how the writing proficiency of members affect the dynamic process of knowledge diffusion.

As for out-degree centrality, none of the extended 2-out-star terms were significant. However, when temporal 2-in-star patterns were considered, we found that only the “[odc]-initiative” term was positive and significant. It suggests that when a member receives answers from multiple contributors, answers from members with high out-degree-centrality were likely to arrive early. High out-degree centrality of a member indicates that he or she frequently helps others, reflecting the member’s high level of activity and willingness to help. Consequently, members with high out-degree centrality tried to help others as early as possible. Again, using NATERGM terms, we saw that the proposed model was able to reveal how the out-degree centralities of members affect the dynamic process of knowledge diffusion.

In sum, by comparing the pattern testing results of TERGM and NATERGM, we could always gain additional insights about how nodal attributes of social media users affect the dynamic process of network formation. Therefore, NATERGM has an enhanced pattern testing capability compared to the benchmark network model.

4.5.2 Network Prediction Results

Table 4.10 shows the results of Wilcoxon signed rank tests for the prediction errors of NATERGM and TERGM. The numbers are Z-statistics and standard errors are
shown in parentheses. Results show that the prediction errors obtained by NATERGM were statistically lower than that of TERGM for all of the four selected network statistics. Therefore, NATERGM is able to predict the characteristics of networks more accurately than TERGM in terms of the networks’ degree distribution.

Table 4.10 Wilcoxon Signed Rank Tests for Prediction Errors

<table>
<thead>
<tr>
<th>Community</th>
<th>Test statistics</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>std. in-degree</td>
<td>std. out-degree</td>
<td>skewness in-degree</td>
<td>skewness out-degree</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>6.26***</td>
<td>5.73***</td>
<td>4.44**</td>
<td>4.92***</td>
<td></td>
</tr>
<tr>
<td>Online</td>
<td>5.12***</td>
<td>4.86***</td>
<td>4.49**</td>
<td>4.77**</td>
<td></td>
</tr>
<tr>
<td>Real Estate</td>
<td>5.79***</td>
<td>5.11***</td>
<td>5.44***</td>
<td>4.92***</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: *: p<0.05, **: p<0.01, ***: p<0.001.

4.6 Conclusions

Dynamic interaction between various types of individuals in social media is a complex process and the sequence of network ties is an important aspect of social media network dynamics. We represented various temporal patterns of network formation based on nodal attributes and sequences of network ties and developed NATERGM model for dynamic network analysis. We conducted empirical tests to evaluate the performance of NATERGM and results showed that NATERGM has an enhanced pattern testing capability and increased prediction accuracy of network characteristics compared to previous dynamic network models. The proposed model can be used to evaluate the impact of individuals’ attributes in the formation process of dynamic social media...
networks. By examining these attributes, social media designers can understand what factors are critical to the social network evolution and determine what functionalities to add or promote in their platforms.

The contributions of this study are manifold. First, this study provides an extended ERGM-based network model to examine temporal patterns in dynamic networks. The extended model can examine how nodal attributes of networks affect the sequence in which network ties develop. Previous models were unable to examine the network dynamics from this perspective. Second, this study provides a list of temporal terms that expands static ERGM terms and dynamic TERGM terms without nodal attributes. The list of temporal terms is designed to be adaptable to any general network. Given a new network, these temporal terms can be used to understand the impact of other nodal attributes beyond the attributes used as examples in this study. Furthermore, this study provides a network prediction framework based on temporal patterns identification, which has been an under-studied area in social network research. In our current model, each temporal pattern only considers one attribute at a time. We plan to extend from this point and consider the interactions of multiple attributes in future research.
CHAPTER 5. CONTRIBUTIONS AND FUTURE DIRECTIONS

5.1 Contributions

In Chapter 2, we modeled knowledge diffusion networks in commercial technology with a multi-dimensional framework that accounts for both knowledge co-creation and knowledge transfer processes. The findings in Chapter 2 provide guidance for practitioners to effectively control the knowledge diffusion processes. We identified two types of interactions between knowledge co-creation and knowledge transfer in nanotechnology inventor networks. Understanding how they interact provides guidance for stimulating knowledge diffusion. For example, by stimulating knowledge co-creation among inventors, knowledge transfer processes can also be stimulated, leading to a wider diffusion of the knowledge. We also identified the roles of inventors who are in key positions in the commercial technology knowledge diffusion networks. We found that inventors with high degree centrality were more likely to share knowledge by collaborating with each other, and knowledge transfer was likely to originate from boundary-spanning inventors. These two findings suggest that inventors in key positions in the knowledge diffusion networks could greatly enhance the levels of knowledge co-creation and knowledge transfer. Therefore, funding these key inventors could be more effective than funding others.

As for methodology, our approach that modeled knowledge co-creation and knowledge transfer processes in the same multi-dimensional network using ERGM provides a more comprehensive understanding of the structures of knowledge diffusion.
networks. By taking the interactions of knowledge co-creation and knowledge transfer into account, many of our findings confirmed or complemented prior research findings. In addition, the framework in Chapter 2 provides tools to improve nanotechnology-related commercial technology development. By understanding the mechanisms of knowledge diffusion among inventors in the nanotechnology field and how public funding affects knowledge diffusion, government funding agencies and private sector can use resources more effectively. This is especially important today with nanotechnology serving as a country-level competency indicator.

In Chapter 3, we further examined how the flows of knowledge transfer and knowledge co-creation could be affected by attributes of individual researchers and attributes of scientific knowledge, based on social capital theory. The findings provide implications about how social capital factors affect knowledge diffusion. Our empirical results showed that social capital factors including collaboration experience, credibility, and similarity in career length positively affected knowledge diffusion flows between researchers. Also, researchers’ activity level in general affected knowledge co-creation positively, but not knowledge transfer. However, the relationships between motivations and knowledge diffusion can be moderated by the impact and age of scientific knowledge: when the knowledge is of relatively low impact, researchers were less likely to participate in knowledge co-creation processes even if they are active; when knowledge was new, active researchers had a stronger tendency to seek knowledge through knowledge transfer processes.
Also, we provided conceptualization of multi-dimensional scientific knowledge diffusion networks. Existing knowledge literatures mainly focused on only one type of knowledge entity (i.e., individual), or only one type of knowledge diffusion process (Singh and Marx 2013; Wasko and Faraj 2005), while knowledge can diffuse in many different ways. Previous studies have recognized this problem as one limitation, and suggested that future research should consider multiple aspects of knowledge activities such as how knowledge is created, shared, and learned (Wasko and Faraj 2005). This research aligns with their suggestions and models knowledge transfer and knowledge co-creation processes in the same network. Second, this study provides insight about how attributes of knowledge artifact affect knowledge diffusion processes. Most prior research focusing on knowledge diffusion mechanisms did not explicitly consider how knowledge with different attributes diffused in different ways. Prior research has suggested considering the influence of knowledge attributes in knowledge diffusion research (Inkpen and Tsang 2005). We shed light on this aspect and provided explanations for how knowledge diffusion can be affected by the impact and age of scientific knowledge. Finally, the process of modeling and analyzing multi-dimensional scientific knowledge diffusion networks used in this study can be applied to any general multi-dimensional networks, thereby providing a useful methodological framework to conceptualize and examine multi-dimensional networks in general.

In Chapter 4, we modeled various temporal patterns of knowledge diffusion based on nodal attributes and sequences of network ties and developed NATERGM model for dynamic network analysis. The study provides a dynamic perspective of multi-
dimensional knowledge diffusion networks, while existing research primarily has focused on static networks. The proposed model can be used to evaluate the impact of individuals’ attributes in the formation process of dynamic social media networks. By examining these attributes, social media designers can understand what factors are critical to the social network evolution and determine what functionalities to add or promote in their platforms.

Chapter 4 also makes important methodological contributions. First, we provided an extended ERGM-based network model to examine temporal patterns in dynamic networks. The extended model can examine how nodal attributes of knowledge diffusion networks affect the sequence in which knowledge flows develop. Previous network models have been unable to examine the dynamics of knowledge diffusion from this perspective. Second, this study provides a list of temporal terms that expands static ERGM terms and dynamic TERGM terms without nodal attributes. The list of temporal terms is designed to be adaptable to any general network. Given a new network, these temporal terms can be used to understand the impact of other nodal attributes beyond the attributes used as examples in this study. Furthermore, this study provides a network prediction framework based on temporal patterns identification, which has been an under-studied area in social network research. Finally, we conducted empirical tests to evaluate the performance of NATERGM and results showed that NATERGM has an enhanced pattern testing capability and increased prediction accuracy of network characteristics compared to previous dynamic network models.
5.2 Relevance to MIS

This dissertation falls into the category of design science, which is one of the key paradigms in MIS. The design science paradigm focuses on doing innovative design that leads to useful, clear, and significant contribution to knowledge (Hevner et al. 2004). The output of design science research is an artifact that takes the form of constructs, models, methods, and instantiations (March and Smith 1995). Design science paradigm is ultimately a problem-solving paradigm (Hevner and Chatterjee 2010). The objective of design science research is to provide innovative and useful solutions to important business problems.

The essays in this dissertation are all pertaining to the design product aspect of design science. In Chapters 2 and 3, we not only used the existing ERGM model, but also designed a framework to extract multi-dimensional knowledge diffusion networks from Electronic record-based data, use network reduction techniques to process the extracted network data for better analysis, transform the assumptions and hypotheses to network patterns for testing, and finally test the patterns with ERGM and evaluate the results with GOF tests. The framework can be used to systematically analyze the knowledge diffusion patterns in both commercial technology and scientific innovation contexts. The framework was also design to be general so that it can be applied to other similar problem contexts.

In Chapter 4, we further extended the existing ERGM and TERGM models and developed a new dynamic network model, NATERGM. Its modules and sub-approaches including identification of network tie sequence, representation of dynamic networks
with two matrices and attribute vectors, modeling of temporal patterns by combing nodal attributes and network tie sequence, and NATERGM parameter estimation are all examples of IT artifacts resulted from this essay. The model was further rigorously evaluated using empirical data collected from an online knowledge sharing community, following one of the key guidelines in design science research. Moreover, the framework was designed to be general and can be easily applied to examine any dynamic networks in social media.

In addition, as mentioned earlier in the dissertation, the emergence of knowledge cyber-infrastructure (CI) in recent years has provided opportunities for studying knowledge diffusion networks (Brown et al. 2010a; Hey and Trefethen 2005; Wright and Wang 2011). Examples of knowledge CI include the scientific databases and online knowledge communities as used for analysis in this dissertation. In this work, we also demonstrated the usefulness of knowledge CI in examining large-scale knowledge diffusion networks.

5.3 Future Directions

The future work will continue to broaden and deepen work relating to multi-dimensional network modeling, model scalability, and network dynamics.

1) **Multi-dimensional Network Modeling**: When modeling the attributes of network nodes (individuals), our current work mainly relied on information available online. This might not be sufficient for accurately measuring the attributes. For example, in Chapter 3 we used career length of researchers to model the credibility of their
knowledge. Although past research has shown that career length is closely related to credibility (Chen and Starosta 1997), we recognize that the credibility can be influenced by factors other than career length (e.g., reputation). The measurement can be improved if we collect data from additional sources (e.g., using survey). So far we have focused mainly on electronic record-based data. Leveraging multiple data sources for analysis is a possible future direction.

2) **Model Scalability:** One typical problem in ERGM-based methods is the scalability. When applying ERGM-based models to very large networks, the models suffer from serious model degeneracy problems and it is very difficult to achieve good model convergence. Our current work circumvents this problem by using network reduction approaches, such as intensity modeling/network dichotomization. The intensity modeling and dichotomization approach examines the frequencies that individuals build relations with each other, and identifies only the intense relationships as network ties. By reducing the network size, however, a lot of information is lost. In future works, we plan to employ more efficient representation of network data (e.g., triangle motif representation) and utilize distributed computing to increase the scalability of ERGM-based approaches.

3) **More complex network dynamics:** In our current NATERGM model, each temporal pattern only considers one attribute at a time. We plan to extend from this point and consider the interactions of multiple attributes in NATERGM. For example, when an individual retrieves knowledge from multiple sources (e.g., friends, online databases, books etc), we may be interested in which source tends to be selected first and last. A
model that can concurrently incorporate more than two different types of nodal attributes will be useful in such types of problems.
REFERENCES


Li, C. 1997. *Is lexical richness an essential criterion in judging a piece of writing?*, The University of Hong Kong (Pokfulam, Hong Kong).


