Dynamics of Information Diffusion

Minkyoung Kim

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

April 2015

Except where otherwise indicated, this thesis is my own original work.

Minkyoung Kim
30 April 2015

To my parents, Youngsam and Insun

List of Publications

1. **Minkyoung Kim**, David Newth, and Peter Christen, "Macro-level Information Transfer in Social Media: Reflections of Crowd Phenomena," *Neurocomputing*, Elsevier, 2015 (Accepted).
2. **Minkyoung Kim**, David Newth, and Peter Christen, "Uncovering Diffusion in Academic Publications using Model-Driven and Model-Free Approaches," In Proceedings of the 7th IEEE International Conference on Social Computing and Networking (SocialCom'14), Sydney, Australia, December 2014, Pages 564–571 (Full Paper, Acceptance rate 28%).
3. **Minkyoung Kim**, David Newth, and Peter Christen, "Macro-Level Information Transfer across Social Networks," In Proceedings of the 23rd International World Wide Web Conference (WWW'14), Seoul, Korea, April 2014, Pages 321–322 (Poster Paper, Invited from research track, Invitation rate 16%).
4. **Minkyoung Kim**, David Newth, and Peter Christen, "Trends of News Diffusion in Social Media based on Crowd Phenomena," In Proceedings of the Workshop on Social News on the Web (SNOW) at the 23rd International World Wide Web Conference (WWW'14), Seoul, Korea, April 2014, Pages 753–758 (Full Paper).
5. **Minkyoung Kim**, David Newth, and Peter Christen, "Modeling Dynamics of Diffusion across Heterogeneous Social Networks: News Diffusion in Social Media," *Entropy*, MDPI, Volume 15, Issue 10, October 2013, Pages 4215–4242.
6. **Minkyoung Kim**, David Newth, and Peter Christen, "Modeling Dynamics of Meta-Populations with a Probabilistic Approach: Global Diffusion in Social Media," In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13), San Francisco, United States, October 2013, Pages 489–498 (Full Paper, Acceptance rate 16.86%).
7. **Minkyoung Kim**, David Newth, and Peter Christen, "Modeling Direct and Indirect Influence across Heterogeneous Social Networks," In Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNA-KDD) at the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13), Chicago, United States, August 2013, Article Number 9 (Full Paper, Acceptance rate 30%).
8. **Minkyoung Kim**, Lexing Xie, and Peter Christen, "Event Diffusion Patterns in Social Media," In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12), Dublin, Ireland, June 2012, Pages 178–185 (Full Paper, Acceptance rate 20%).

Acknowledgments

I am thrilled to fill in this page. This is personally the most valuable page out of this thesis. Without people mentioned in this page, this dissertation work would not have been possible. I am privileged to take this opportunity to express my appreciation for their efforts, considerations, respect, and unfailing trust.

First and foremost, I would like to thank my supervisors, Peter Christen (Research School of Computer Science, Australian National University), David Newth (Center for Complex Systems Science, CSIRO), Lexing Xie (Research School of Computer Science, Australian National University), and Robert Ackland (Australian Demographic and Social Research Institute, Australian National University). Thanks to Peter's continuous encouragement, unsparing guidance, and valuable feedback never even late, my PhD study could be on the right track. I have learned a lot from his devoted supervision, the way of sharing research ideas with group members, and friendly atmosphere he creates. I am also deeply grateful to David who has given confidence in my research, has shared interesting research ideas, and has shown wise generosity. My research has benefited from his insightful and analytical views on complex systems. I would also like to thank Lexing who motivated my first year PhD study. Thanks to her valuable advice and deep insight, I could begin my PhD study at a good starting point. I am full of gratitude to Robert who, as a social scientist, has given different and interesting point of views on my research. His very constructive and helpful comments have guided my study to be more consistent and coherent. No words can express my greatest appreciation to my supervisory panel.

I would like to thank ANU, Research School of Computer Science, and CSIRO for their financial supports for my PhD study including PhD scholarships and travel funding for international trips for my paper presentations.

I am thankful to my fellow PhD students in our data mining group who have encouraged, have shown friendships, and have shared, sometimes tough and sometimes also tough, PhD lives with each other without losing sense of humor. Thanks to their careful considerations, my PhD life has been full of wonderful memories.

On the personal front, I thank my parents, Youngsam and Insun, and parents-in-law, Donghan and Namhee, who have shown unfailing trust, love, and sacrifices. Without their moral supports, I could not even start my PhD study.

My enormous gratitude goes to my soulmate for sharing difficulties from my PhD study as well as happiness from our lives. Thanks to his supports, encouragements, and respect, my PhD experience has been pleasant and joyful. The way he has shown his absolute faith to me is the most invaluable thing I have learned during my PhD.

Abstract

Real diffusion networks are complex and dynamic, since underlying social structures are not only far-reaching beyond a single homogeneous system but also frequently changing with the context of diffusion. Thus, studying topic-related diffusion across multiple social systems is important for a better understanding of such realistic situations. Accordingly, this thesis focuses on uncovering topic-related diffusion dynamics across heterogeneous social networks in both model-driven and model-free ways.

We first conduct empirical studies for analyzing diffusion phenomena in real world systems, such as new diffusion in social media and knowledge transfer in academic publications. We observe that large diffusion is more likely attributed to interactions between heterogeneous social networks as if they were in the same networks. Thus, external influences from out-of-the-network sources, as observed in previous work, need to be explained with the context of interactions between heterogeneous social networks. This observation motivates our new conceptual framework for cross-population diffusion, which extends the traditional diffusion mechanism to a more flexible and general one.

Second, we propose both *model-driven* and *model-free* approaches to estimate global trends of information diffusion. Based on our conceptual framework, we propose a model-driven approach which allows internal influence to reach heterogeneous populations in a probabilistic way. This approach extends a simple and robust mass action diffusion model by incorporating the structural connectivity and heterogeneity of real-world networks. We then propose a model-free approach using information-theoretic measures with the consideration of both time-delay and memory effects on diffusion. In contrast to the model-driven approach, this model-free approach does not require any assumptions on dynamic social interactions in the real world, providing the benefits of quantifying nonlinear dynamics of complex systems.

Finally, we compare our model-driven and model-free approaches in accordance with different context of diffusion. This helps us to obtain a more comprehensive understanding of topic-related diffusion patterns. Both approaches provide a coherent macroscopic view of global diffusion in terms of the strength and directionality of influences among heterogeneous social networks. We find that the two approaches provide similar results but with different perspectives, which in conjunction can help better explain diffusion than either approach alone. They also suggest alternative options as either or both of the approaches can be used appropriate to the real situations of different application domains.

We expect that our proposed approaches provide ways to quantify and understand cross-population diffusion trends at a macro level. Also, they can be applied to a wide range of research areas such as social science, marketing, and even neuroscience, for estimating dynamic influences among target regions or systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the motivation and general outline of our research methodology, and then presents a detailed overview of the thesis including its main scientific contributions.

1.1 Motivation

Users of online services are able to access increasingly large volumes of information with much less time and efforts than ever before. For instance, online users are being kept frequently updated on their interested content through varieties of feed readers (aggregating news feeds such as Flipboard and Google Reader), or social media aggreators (integrating social networks such as TweetDeck and Hootsuite), or through news agency's active participations to social networking platforms. Such open web environment enables unlimited digital content to be navigated without the need to hop multiple social media platforms. In addition, the ubiquitous use of mobile devices delivers new information in real time. Such easy and high accessibility to a wide range of information not only makes underlying social networks far-reaching across multiple social platforms, but also leads their connectivity patterns to change dynamically in accordance with the context of information.

These dynamic and complex networks can be observed in various application domains such as hyperlink networks on the Web [43, 105, 136], academic citation networks across multiple disciplines [38, 92, 140], and even cortical networks across different brain regions [55, 150, 164]. That is, studying significant patterns of (social) interactions in the real world is meaningful as they are possibly applicable to diverse research fields due to universal properties of real-world networks [15, 37, 127, 182].

However, previous diffusion studies have mostly focused on diffusion within a single homogeneous system such as Twitter or Facebook, and thus influences from out-of-the-network sources have been regarded as only external influence. In addition, prior work covering heterogeneous populations has not assessed dynamic influences formed among multiple (social) systems and has often been without the diverse context of diffusion. In this regard, the research presented in this thesis is about **uncovering dynamics of topic-related global diffusion across heterogeneous social networks**, beyond a single homogeneous social system.

For studying information diffusion, we aim to address a number of key research questions including: (1) How does information spread across multiple social systems? (2) How do we categorize heterogeneous social networks into a meaningful set and understand its intra- and inter-relationships? (3) How do diffusion patterns vary according to different information topics? (4) How do we estimate diffusion trends across multiple populations?

We expect that the ways of answering these questions can help to study dynamics of complex systems in other research areas.

1.2 Approach of Research

In order to understand such cross-population diffusion phenomena in the real world, our research is conducted based on four scientific methods throughout the thesis.

1.2.1 Step 1 – Empirical Analysis

We first analyze real-world diffusion phenomena in two different application domains: news diffusion in social media and knowledge diffusion in academic publications. Findings from these empirical analyses become the basis for developing our new approaches in the next step. In this step, we construct explicit citation networks through hyperlinks in the main content of web documents in social media, and through the bibliography in academic publications, and analyze diffusion patterns across heterogeneous social systems.

1.2.2 Step 2 – Proposal of Approach

Based on the empirical findings from Step 1, we design a new conceptual framework for diffusion processes. Following this conceptual design, we propose an approach. In this step, we propose two model-driven and model-free approaches with different aspects. That is, a model-driven approach is based on the assumption of underlying social interactions in a probabilistic way, while a model-free approach does not require any assumptions on the dynamic and complex interactions in the real world.

1.2.3 Step 3 – Estimation with Proposed Approach

We evaluate our approaches from Step 2, and estimate diffusion trends with each proposed approach, using the preprocessed real data from the two different application domains from Step 1. Applying a proposed approach to different domains and evaluating the outcomes help to validate its feasibility and to strengthen and generalize the proposed approach. Steps 2 and 3 are repeated twice in our study, i.e. one for a model-driven approach and the other is for a model-free approach.

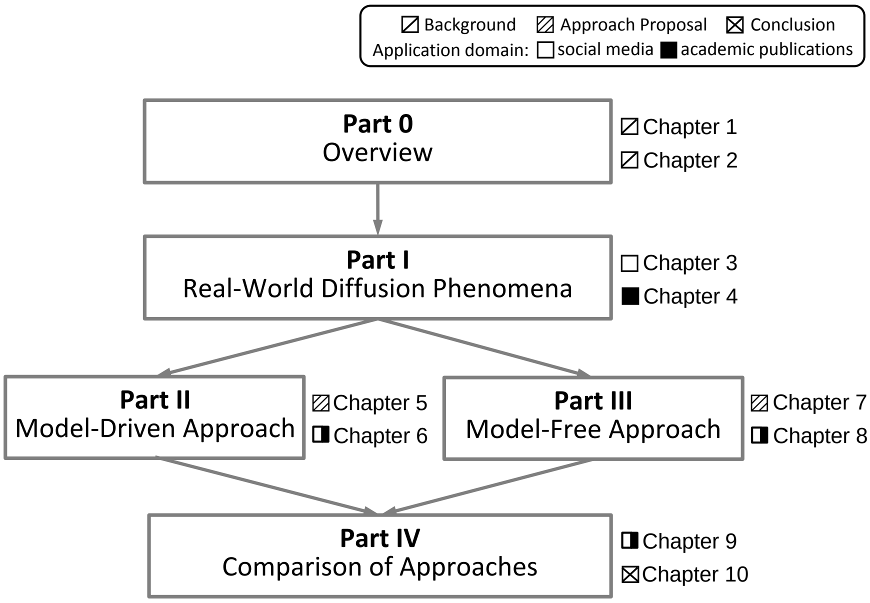


Figure 1.1: Thesis outline.

1.2.4 Step 4 – Comparison and Interpretation

We finally compare our approaches, the model-driven and model-free approaches, from Step 2 based on the estimation results from Step 3. We examine how they are different or similar and interpret what they provide for obtaining meaningful insights into cross-population diffusion phenomena in the real world. This comparison of the two different approaches aims to provide a comprehensive macroscopic view of topic-related global diffusion.

1.3 Thesis Overview

This dissertation focuses on dynamics of information diffusion at a macro level based on the research steps discussed in the previous section. Figure 1.1 shows the overall thesis outline by mapping the thesis chapters to corresponding thesis parts as well as application domains. As the figure illustrates, this thesis consists of four main parts and ten chapters. The first two chapters give an overview of this dissertation work. Chapter 1 provides introduction of the thesis including motivation, outline, and contributions. In Chapter 2, we define the concepts used in this thesis and discuss related work. The rest of the thesis outline is described in detail in the following subsections.

1.3.1 Part I – Real-world Diffusion Phenomena

Analyzing real-world diffusion patterns from interconnected user-contributed content provides clues for a new model design. In this regard, Chapter 3 and 4 analyze cross-population diffusion in social media and in academic publication domains, respectively. In common, identifying trending topics is important to trace global diffusion across different social systems. Additionally, categorizing information topics enables us to find common or distinct diffusion patterns between different categories, and to obtain higher level views from an aggregation of individual topic diffusion within a same category.

In Chapter 3, we analyze news diffusion patterns in social media by constructing different levels of underlying networks, such as document, authorship, and information networks [91]. We target trending news items by referring to Wikipedia event profiles. We observe that news spreads across different types of social media, such as mainstream news, social networking sites, and blogs, beyond a single social platform, and the majority of users (55%) cite to a small fraction of prolific users (4%) from different types of social media platforms, a notable departure from the balanced traditional bow-tie model of web content.

In Chapter 4, we focus on knowledge diffusion patterns in computer science [89]. Based on the constructed citation networks, we analyze linkage patterns between top-ranked conferences in computer science as well as between its different subdomains. We observe that premier conferences lead computer science research and that they are not only interconnected but also clustered along their common research interests. The majority of conferences (51%) cite to a small proportion of other conferences (10%) belonging to different subdomains, showing a similar pattern of a skewed bow-tie as occurred in Chapter 3.

In common across the two application domains, the underlying network spans multiple social systems. The majority of individuals in the network obtain new information from a smaller proportion of other individuals from different social systems, i.e. external sources, and the connectivity of the network is far-reaching across the systems. That is, external influence is not ignorable, and underlying networks are heterogeneous in real-world diffusion.

1.3.2 Part II – Model-Driven Approach

From the empirical analyses of real-world diffusion in Part I, we found that large diffusion is attributed to interactions between heterogeneous populations. However, most prior work has focused on diffusion within a homogeneous population such as local contact networks in a single social platform (e.g., Twitter or Facebook). In this respect, our research aims to uncover macro-level diffusion mechanisms across heterogeneous social networks.

In Chapter 5, we design a new conceptual framework for cross-population diffusion, which allows direct interactions between different social networks as if they were in the same networks. Accordingly, we propose a macro-level diffusion model, named *Dynamic Influence Model*, with a probabilistic approach by incorporating two main features of real-world networks, *heterogeneity* and *structural connectivity* [85, 86], which has not been studied in previous research.

In Chapter 6, we evaluate our proposed model using both synthetic and real data. As real data, we take cases from two different domains, social media and academic publication, from Part I. Based on the evaluation of our model, we estimate real-world diffusion at a macro level in terms of the strength and directionality of influence [84, 85, 86, 89]. We find that influences between heterogeneous social networks vary with the topics of information, leading to different diffusion patterns.

1.3.3 Part III – Model-Free Approach

As mentioned in the previous subsection, most studies have focused on micro-level diffusion within a single homogeneous system, and their models are based on assumptions that social interactions are from sampled snapshots of current social networks. In this regard, in Part II we proposed a macro-level diffusion model which estimates influence of diffusion across multiple social networks without the need of knowledge of local network structures in detail. However, this model-driven approach is based on the assumption of a real-world network property, i.e. a power-law degree distribution, and thus it still needs to estimate the value of the scaling exponent for the power-law tail. In this context, in Part III we aim to discover an approach of estimating diffusion without any assumptions on dynamically changing social interactions in the real world.

In Chapter 7, we propose a model-free approach which estimates macro-level information transfer across heterogeneous social systems by using information-theoretic measures [87, 89, 90]. This is enabled by considering a stochastic process at a system level and by defining the system's signal at a population level. Time-delay and memory effects are also considered so that we can understand how recent and long adoption histories of information in a source system have effect on diffusion in a destination system.

In Chapter 8, we estimate macro-level information transfer in the social media and academic publication domains as we did in Chapter 6. This enables us to compare the proposed approaches based on these estimation results in Part IV. We analyze topic-related information pathways across different social systems using the transfer entropy with time-delay and memory effects. We find that the estimated topic-related diffusion trends from this model-free approach exhibit similar inter-relationships to those from the model-driven approach in terms of the strength and directionality of influence. Inter-relationships among the systems are varied with news categories or research keywords in our target application domains, leading to different diffusion patterns [87, 88, 89, 90].

1.3.4 Part IV – Comparison of Approaches

Finally, we compare our proposed model-driven (Part II) and model-free (Part III) approaches with respect to common and different characteristics of the approaches [87, 88, 89, 90]. Applying both approaches to different real-world domains helps us to compare and generalize the methods in a more consistent way.

In Chapter 9, we first extract several distinct diffusion patterns from time-series cumulative adoption rates in social media and academic publications for each. We examine how the two approaches estimate and distinguish the patterns. In common across the domains, heterogeneous social systems in transitive relations have more opportunities to bring about synchronous diffusion patterns across them, but the strength as well as balance of intra- and inter-relationships within and between the systems are also important to lead to cross-population diffusion. That is, synchronous diffusion needs strong intra-relationships as momentum to make individuals to move to a certain state, i.e. critical mass, within each system, which is triggered by strong inter-relationships between different social systems.

Finally, in Chapter 10, we conclude this thesis by summarizing all the main findings and contributions, and discuss future directions for this research.

1.4 Scientific Contributions

In this section, we summarize the main contributions of our study for each part of the thesis. All citations provided in the previous and current sections refer to our publications from this dissertation work.

1.4.1 Part I – Real-world Diffusion Phenomena

We provide major characteristics of structural diffusion patterns across heterogeneous social networks in social media [91] and academic publications [89]. These application domains have drawn large attention from diverse research areas, and thus our empirical findings can provide the research communities with fundamental statistics as a background. Most studies have focused on site-specific diffusion of particular keywords or quotes, but our study attempts to discover cross-population diffusion of non-site specific topics for obtaining more universal diffusion patterns in the real world. Thus, identifying trending topics is important to trace global diffusion across heterogeneous populations. Accordingly, we propose a novel method to identify note-worthy news by presenting each news item with named entities that embody the "5W1H", i.e. *Who, What, Where, When, Why, and How*, of journalistic practice. Categorizing trending topics provides higher level views of diffusion patterns by aggregating individual topic diffusion within a same information category. By constructing different levels of underlying networks, such as document, authorship, and information linkage patterns, we provide interpretations of structural connectivity, social interactions, and relationships between different information topics.

1.4.2 Part II – Model-Driven Approach

We provide a new conceptual design for diffusion across heterogeneous social networks [85, 86], which shifts the viewpoint of a traditional diffusion framework (external and internal influences on a homogeneous population) to the dynamics of heterogeneous populations (external and internal influences on heterogeneous populations). That is, our framework allows that internal influence reaches beyond a single social system into multiple systems as if the underlying network were homogeneous, and thus it extends the traditional diffusion mechanism to a more general and flexible one. Based on our conceptual framework, we propose a macro-level diffusion model, called *Dynamic Influence Model*, by considering both heterogeneity and structural connectivity of real-world networks in a probabilistic way. Our proposed model can improve the accuracy of diffusion models dealing with a homogeneous population within a single social system only, since it does not neglect the effects of interactions between different populations on diffusion and differentiate the strength of each population's influence on others. That is, this model provides both intra- and inter-relationships among heterogeneous populations in terms of the strength and directionality of influence. By applying this approach to our target application domains, we provide topic-related diffusion patterns across heterogeneous social networks, which provides a way of understanding cross-population diffusion phenomena in the real world [84, 85, 86, 89].

1.4.3 Part III – Model-Free Approach

We propose a model-free approach of estimating macro-level diffusion without any assumptions on social interactions [87, 88, 89, 90], which not only provides benefits of studying nonlinear dynamics of complex systems but also gives an abstract view of cross-population diffusion. We define a stochastic process at a system level and the systems' signals transmitted to other systems at a population level, which can be applicable to a variety of real-world scenarios. This conceptual framework enables us to estimate macro-level information transfer across heterogeneous social systems in terms of the strength and directionality of influence. To the best of our knowledge, this is the first attempt to apply this macro-level model-free approach to social media and academic publications. Estimation results can be comparable with the outcomes from the model-driven approach, which provides opportunities to enhance the performance of both approaches. When estimating information transfer from a source to a destination system, the effects of time-delay and memory (the length of adoption histories) are all considered. That is, behavioral characteristics of different social systems can be obtained, which helps to better understand the context of diffusion. By applying this approach to our target application domains, we provide different ways of understanding topic-related diffusion across heterogeneous social networks.

1.4.4 Part IV – Comparison of Approaches

Our model-driven approach can provide both external and internal (intra- and inter-relationships) influences among heterogeneous social systems but with assumptions on the underlying social interactions. On the other hand, our model-free approach can provide only inter-relationships among the systems but with no assumptions on the interactions, and it also suggests behavioral characteristics of each system. Both our approaches provide an abstract view of cross-population diffusion in terms of the strength and directionality of influence. We find that the two approaches show similar results but with different perspectives, which in conjunction can help to obtain a more coherent overall picture of diffusion dynamics than either approach alone [88, 89, 90]. They also suggest alternative options to choose a more appropriate approach according to the conditions of different application domains. Finally, interpreting crowd phenomena with respect to both approaches suggests diverse aspects of diffusion analytics and further provides a more comprehensive understanding of topic-related diffusion.

We expect that our proposed approaches can help to quantify and understand cross-populations diffusion in a wide range of research areas including neuroscience, social science, and marketing for estimating functional connectivity of the human brain, and dynamics of influence among social organizations or countries.

In the next chapter, we describe common and fundamental concepts used in this thesis and discuss related work before starting the four main parts of the thesis.

Chapter 2

Overview and Survey

In this chapter, we first describe the common and fundamental concepts used in this thesis as preliminaries and define relevant terminology and notation. Then, we discuss related work on analyzing real-world diffusion phenomena, modeling diffusion processes, and predicting future behavior at individual or group levels with information-theoretic measures.

2.1 Concepts and Definitions

We review the basics of graph-theoretic terms, concepts of network structures, and information-theoretic measures, which are all used for studying diffusion dynamics in this thesis.

2.1.1 Fundamental Graph-Theoretic Concepts

Graph: a graph G consists of two sets V and E as $G = (V, E)$, where V represents a set of *vertices* or *nodes* and E denotes a set of *edges* or *links*. The order of G is the number of vertices $|V|$, and the size of G is the number of edges $|E|$. Each edge is associated to two vertices (not necessarily distinct), called *endpoints*. An edge is called a *self-loop* or *loop* if it joins a single endpoint to itself. If an edge's two endpoints are designated as the *tail* and the *head* for each, then it is called a *directed edge*. Otherwise, it is called a *undirected edge*. A directed edge is also called an *arc*.

Adjacent and **neighbors:** if two vertices v and u are two endpoints of an edge e, then a vetex v is *adjacent* to vertex u. The two adjacent verices are called *neighbors*.

Degree: the *degree* of a vertex v, denoted by $d(v)$, is the number of edges incident with the vertex or equivalently to the number of its neighbors. If a vertex v has no neighbors, i.e. $d(v) = 0$, then the vertex is called an *isolate*. The *Euler's handshaking lemma* states that the sum of the vertex degrees in any graphs is equal to twice the number of edges as:

$$\sum_{v \in V} d(v) = 2\,|E|$$

Indegree and **outdegree:** the *indegree* of a vertex v, denoted by $d_{in}(v)$, is the number of arcs whose endpoints are designated toward the vertex v (the number of incoming arcs). The *outdegree* of a vertex v, denoted by $d_{out}(v)$, is the number of arcs whose endpoints are designated toward neighbors of the vertex v (the number of outgoing arcs).

Directed graph: if all of a graph's edges are all directed, then it is called a *directed graph* or *digraph*.

Undirected graph: if all of a graph's edges have no designations of head and tail (unordered pairs of endpoints), then it is called a *undirected graph*.

Simple graph: if a graph has no self-loops or multi-edges (slings), then it is called a *simple graph*.

Trivial graph: if a graph consists of only one vertex without any edges, then it is called a *trivial graph*.

Clique: if a subset of vertices in an undirected graph are all connected between every pair of vertices, then they are called a *clique*.

Complete graph: if all pairs of vertices are adjacent, then a graph is said to be *complete*, and this graph is its own maximal clique.

Subgraph: a *subgraph* G_s of a graph $G = (V, E)$ consists of a subset of E and their endpoints as $G_s = (V_s, E_s)$, where $E_s \subseteq E$ and $V_s = \{v, u : (v, u) \in E_s\}$.

2.1.2 Structure of Networks

The structural properties of complex networks in the real world have drawn large attention across diverse research areas [37, 53, 127, 128, 181]. We review fundamental structural properties representing networks.

Connected component: if there is a path between two vertices in an undirected graph, the two vertices are called *connected*. A *connected component* is a maximal set of nodes whose every pair of vertices is connected.

Connectivity: when the removal of a set of vertices makes a connected graph G disconnected, we call this set of vertices as a *cut* or *vertex cut*. The *connectivity* of G is the size of a minimal cut, and the graph G is called k-connected when its connectivity is more than k.

Density: the *density* measures the proportion of the current edges to the maximum possible number of edges in a graph $G = (V, E)$. An undirected graph has no di-

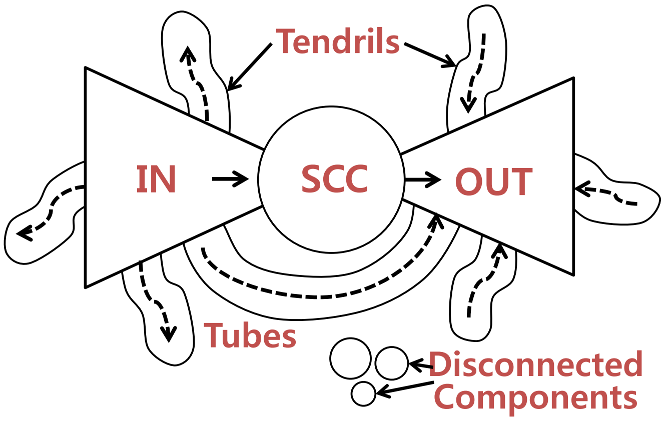


Figure 2.1: Overview of the bow-tie model consisting of six components: SCC, IN, OUT, Tendrils, Tubes, and Disconnected.

rectionality between vertices, and thus its maximum possible number of edges is $\frac{1}{2}|V|(|V|-1)$, and its density is $\frac{2|E|}{|V|(|V|-1)}$. For a directed graph, its maximum number of arcs is $|V|(|V|-1)$, and its density is accordingly measured as $\frac{|E|}{|V|(|V|-1)}$.

Weakly connected component: If all arcs are replaced with undirected edges in a directed graph, and every pair of vertices v and u in the graph has a undirected path between v and u, then this directed graph is called *weakly connected*. A *weakly connected component* is a maximal (largest) weakly connected subgraph.

Strongly connected component: A directed graph is called *strongly connected* if every pair of vertices v and u in the graph has a bidirectional path, i.e. a directed path from v to u and a directed path from u to v. A *strongly connected component* is a maximal (largest) strongly connected subgraph.

Triad: a triple of connected vertices is called a *triad* or a *triangle*.

k-core: a subgraph is called *k-core* if every vertex in the subgraph is adjacent to at least k other vertices in the subgraph.

Bow-tie model: the *bow-tie model*, introduced by Broder [40], visualizes the reachability of a network by partitioning its structure into six components as a bow-tie structure, which is illustrated in Figure 2.1. As the figure shows, the bow-tie structure consists of (1) "SCC" (strongly connected component) as a main building block in the center, (2) "IN" as a component outgoing to the SCC in the left wing, (3) "OUT" as a component incoming from the SCC in the right wing, (4) "Tendrils" as peripherals of IN and OUT, not reachable to/from the SCC, (5) "Tubes" as a component connecting from IN to OUT but without passing through the SCC, and (6) "Disconnected" as isolated components from the bow-tie (see [40] for details).

Degree distribution: there are variations of vertices' degrees in a graph. The *degree distribution* means the probability distribution of the degrees in the graph. Accordingly, the degree distribution, denoted by $P(k)$, is obtained by counting the number of vertices with degree k. If a graph has n vertices in total and n_k vertices with degree k, then the graph has a degree distribution of degree k as:

$$p(k) = \frac{n_k}{n} \ .$$

Binomial distribution: if the total number of vertices of a graph is n, where each vertex has independent probability p to be connected and $1 - p$ to be not connected, then the graph has a *binomial distribution* of degree k as:

$$p(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \ . \tag{2.1}$$

Multinomial distribution: the binomial distribution allows only two possible outcomes, success (p) and failure ($1 - p$), of each trial, but the *multinomial distribution* generalizes the binomial distribution by extending a Bernoulli trial to m possible outcomes as:

$$p(k_1, ..., k_m) = \begin{cases} \dfrac{n!}{k_1! \ ... \ k_m!} p_1^{k_1} ... p_m^{k_m} \ , & \text{when } \sum_i^m k_i = n \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

Power law degree distribution: if a graph's degree distribution follows a power law, then the graph has a *power-law distribution* as:

$$p(k) \propto k^{-\alpha} \ , \tag{2.3}$$

where α is a parameter that scales its power-law tail. Networks whose degree distributions follow a power law are called *scale-free networks* [127].

Riemann zeta function: Equation (2.3) can be rewritten as:

$$p(k) = Ck^{-\alpha} \ , \tag{2.4}$$

where C is a constant, and α is an exponent scaling the power-law tail. For all possible degrees k, the degree distribution $p(k)$ adds up to 1 as $\sum_{k=1}^{\infty} p(k) = 1$. Note that k starts at 1, since $p(0)$ would be infinite. Thus, C in Equation (2.4) can be obtained as:

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)} \ , \tag{2.5}$$

where $\zeta(\alpha)$ is the *Riemann zeta function* [172]. $\zeta(s)$ is a function of a complex variable s that analytically adds up infinite series as $\sum_{n=1}^{\infty} \frac{1}{n^s}$ and converges when $\Re(s) > 1$. Real-world networks are scale-free networks, and they exhibit a power-law degree distribution whose power law exponent α is typically in the range $2 < \alpha < 3$ [48, 127]. Thus, this function is needed for normalization of a power-law distribution, and consequently Equation (2.3) can be rewritten as:

$$p(k) = \frac{k^{-\alpha}}{\zeta(\alpha)} \ , \tag{2.6}$$

where $k > 0$, $\alpha > 1$, and $p(0) = 0$.

2.1.3 Information-Theoretic Measures

The main purpose of information theory, as introduced by Shannon [154], is to reproduce original massages which are initially generated by an information source, pass through a noisy communication channel after being encoded, and are decoded at a receiver. Regardless of its initial objectives, information-theoretic concepts have been applied and extended across diverse research areas, such as computer science, physics and neuroscience, as nonparametric statistics.

An information source produces messages x, and they are considered as possible values of a random variable X. Information-theoretic quantities are obtained based on a probability distribution function $p(x)$ of the random variable X. We review the fundamental information-theoretic measures used in this thesis as follows.

Entropy: in information theory, information is considered as the amount of uncertainty. That is, no uncertainty provides no additional information. If $p(x)$ is the probability that a message x is observed, then information $I(x)$ of observing the message is represented as:

$$I(x) = \log \frac{1}{p(x)} \ . \tag{2.7}$$

Note that we express all entropies as bits, so all logarithms are taken to the base 2 throughout the thesis. An information source produces messages each of which has a fixed probability to be observed. Thus, the average amount of information obtained by a message is:

$$\langle I \rangle = \sum_x p(x) I(x) \ . \tag{2.8}$$

This is the definition of *entropy* exhibiting the average amount of information. Thus, entropy measures the degree of uncertainty of predicting the messages produced by the information source, and is represented as:

$$H(X) = -\sum_x p(x) \log p(x) \ . \tag{2.9}$$

Joint entropy: the *joint entropy* of two random variables X and Y measures the uncertainty of the joint distribution of X and Y as:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \ . \tag{2.10}$$

Conditional entropy: the *conditional entropy* of X given Y is the average uncertainty about X when we have knowledge of Y. Specifically, the entropy of X given a message $Y = y$ is:

$$H(X|Y=y) = -\sum_{x} p(x|y) \log p(x|y) \ . \tag{2.11}$$

Thus, the conditional entropy of X given Y becomes the average uncertainly over all the possible outcomes of Y as:

$$\begin{aligned} H(X|Y) &= \sum_{y} p(y) H(X|Y=y) \\ &= -\sum_{y}\sum_{x} p(y)p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x,y) \log p(x|y) \ . \end{aligned} \tag{2.12}$$

Mutual Information: the *mutual information* is information between two random variables X and Y, and it measures the reduction in X's uncertainty due to the knowledge of Y, or vice versa as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ . \end{aligned} \tag{2.13}$$

Note that the mutual information $I(X;Y)$ is the *Kullback-Leibler* divergence [109] of the joint distribution $p(x,y)$ from the product distribution $p(x)p(y)$. The Kullback-Leibler divergence between two probability distribution $p(x)$ and $q(x)$ is defined as $D_{KL}(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.

Transfer Entropy: the mutual information can be extended to a collection of n random variables, i.e. a random process, X^n and Y^n as $I(X^n;Y^n)$. This is a symmetric measure such that $I(X^n;Y^n) = I(Y^n;X^n)$. The *transfer entropy* (TE), introduced by Schreiber [153], enables us to obtain the directionality between two stochastic processes X and Y as:

$$TE_{Y \to X} = H(X^t|X^{t-1:t-k}) - H(X^t|X^{t-1:t-k}, Y^{t-1:t-l}) \ , \tag{2.14}$$

where k and l denote Markov orders for the previous states of X and Y respectively from the time step $t-1$. Thus, the transfer entropy measures the reduction of uncertainly in X's state at time t due to the knowledge of Y's recent l states.

2.1.4 Definitions of Terminology

Diffusion network: a *diffusion network* consists of individuals who are involved in spreading a new idea or information through contact neighbors in the network.

Information cascade: chain reactions of adopting new information are called *information cascade* in this thesis.

Social system: we follow the definition of a *social system* by Talcott Parsons [137], which consists of individuals who interact and influence each other's behavior. We consider a population belonging to a single social system as a *homogeneous population* exhibiting the same behavioral characteristics within the system's boundary. We call a homogeneous population's interaction network a *homogeneous social network* as its narrower concept. As we discussed in Chapter 1, social interactions are far-reaching across different social media platforms in social media, and thus the underlying networks span multiple social systems. In this context, we call populations from different social systems *heterogeneous populations*, and their interaction networks are accordingly considered as *heterogeneous social networks*.

Meta-populations: we consider a *meta-population scheme* [29] which categorizes heterogeneous populations into a small number of meta-populations according to the characteristics of target meta-populations in diverse application domains. This helps to obtain a more simple and clearer macroscopic picture of global diffusion.

Social media: the term "social media" has been widely used, but with different and limited understanding [76]. In general, social media has been understood as user generated contents which are easily accessible by other users [12, 35, 185]. In more detail, Kaplan and Haenlein [76] defined the concept of social media as a group of internet-based applications allowing the creation and exchange of user generated contents on the ideological and technological foundations of Web 2.0. In this thesis, we follow this detailed and comprehensive definition of Kaplan and Haenlein.

The authors of [12] divided the social media domain into blogs, social networking platforms such as Twitter and Facebook, social bookmarking sites such as Delicious and CiteULike, and photo/video sharing communities such as Flickr and YouTube. Meanwhile, Kaplan and Haenlein [76] classified social media into six types according to the two dimensions, "Social Presence" and "Self-Presentation": namely, blogs (date-stamped entries managed by one person), social networking sites (applications enabling users to connect to social networks such as Facebook and MySpace), collaborative projects (outcomes by joint effort of many actors such as Wikipedia and Delicious), content communities (to share diverse types of media contents such as Flickr, YouTube, and Slideshare), virtual game worlds (three dimensional virtual environments with restricted rules of interactions), and virtual social worlds (three dimensional virtual environments with no rules of interactions).

2.2 Survey of Real-World Diffusion Studies

Collective behavior often emerges from sporadic interactions between individuals in a contact network. Such dynamic behavior consequently forms large complex diffusion networks in the real world such as the Internet [45, 56, 139, 176, 187], online social networks [11, 66, 94, 105, 115, 122], email networks [77, 116, 129, 148, 160], mobile phone networks [18, 54, 71, 96, 131], collaboration networks [27, 34, 67, 101, 125, 184], cortical networks [16, 55, 112, 150, 164], metabolic and protein networks [28, 36, 58, 73, 74], and so on. Thus, analyzing the linkage patterns of real-world complex networks is a meaningful first step towards a systematic understanding of the underlying diffusion mechanisms.

2.2.1 Inference of Diffusion

Online diffusion tracking has been conducted from different perspectives. For instance, information diffusion can be defined on shared keywords or similar text between documents [11, 66], on shared quotes called *meme phrases* [103], on recommendations [102, 106], on hyperlinks [43, 105, 123], or on site-specific actions such as retweets, mentions, or hashtags in Twitter [75, 98, 147]. However, inferring diffusion with the shared meme is implicit, and even inference with explicit hyperlinks can be noisy because there are an increasing number of spam-hyperlinks on web pages. In order to avoid a noisy tracking of diffusion in social media, we try to collect hyperlinks in the main text of a document to generate accurate and explicit citation networks, and extract key memes to observe the information evolving patterns [91], which is discussed in detail in Chapter 3.

However, it is still challenging to unveil the exact sources that have brought the spread (e.g., adopting new behaviors or ideas, purchasing a product, etc.). That is, an *influencer* is not explicitly acknowledged by its individual *influenced*, and thus it is hard to collect or define real diffusion networks such as explicit consumer networks [72]. In this regard, more recently there has been an attempt to model diffusion without explicit knowledge of underlying network structures [186]. Nevertheless, diffusion processes are not independent on diffusion contents [13, 123, 147], and thus it is meaningful to discover the interplay between them. Accordingly, one of our research goals is to understand how diffusion mechanisms across heterogeneous populations varied with the topics of information [84, 85, 86, 89, 90], which will be discussed in Chapter 6, 8, and 9 from the aspects of our model-driven and model-free approaches.

2.2.2 Diffusion Space

Diverse information sources have become increasingly accessible on the Web. This is partly because of news outlet's active participation in social networking platforms such as Twitter and Facebook and the help of web technologies such as news feeds, feed readers, and social media aggregators. That is, the underlying social interactions have become more dynamic and far-reaching across multiple social media platforms, which leads to heterogeneous social networks becoming more interconnected than ever before [85]. Such emergent phenomena of complex social networks have been investigated in a wide spectrum of diffusion spaces.

When it comes to social media, the diffusion boundary, covered by previous studies, has been mostly limited to a single social networking platform such as Twitter [75, 98, 123, 147] or Facebook [24, 167]. The interactions between different social media platforms have also been common interest, such as blogosphere [11, 105], blogosphere-to-news [62, 63, 103], or blogosphere-to-YouTube [43]. In order to obtain a global view of diffusion in social media, our research extends the diffusion space from the blogosphere to mainstream news to social networking sites (SNS), i.e. blogosphere-news-SNS [91], which is discussed in Chapter 3, 6, 8, and 9.

Similarly, in academic publications, knowledge spreads through complex pathways within a local research community [190], within a single research domain such as computer science [19, 156], or across diverse disciplines [38, 92, 140]. In our study, we also investigate academic publications in computer science, and focus on global diffusion trends across different subdomains viewed as heterogeneous populations, which is presented in Chapter 4, 6, 8, and 9.

2.2.3 Identification of Target Information

The operational definition for identifying target information in diffusion has included a number of text clusters as approximated events [68], Twitter hashtags [49, 51, 98, 183], which are manually grouped into broad topical themes [147], or meme phrases which are commonly observed across diverse media and are thus considered as reflecting real-world events [103]. The authors of [33] designed a two-step approach to first cluster the input Twitter stream and then perform event versus non-event classification on the clusters based on their statistical features. However, these identified streams can be platform-specific topics. For instance, the authors of [98] found that only 3.6% of Twitter's trending topics exist in the hot search keywords from Google.

Our research aims at understanding cross-population diffusion, and thus identifying trending topics covering multiple populations is essential. This enables us to trace beyond the bounds of site-specific (local) diffusion. Finally, categorizing the topics of information enables us to obtain higher-level interpretations, such as common or distinct diffusion patterns between different categories. Detailed steps of targeting and identifying information topics are described in Chapter 3 and 4.

2.2.4 Structural and Temporal Patterns of Diffusion Networks

Diffusion has been studied widely in computer science, marketing, social science, statistical physics, epidemiology, and neuroscience. A common theme across the study of diffusion processes is that the structure of the network, its density, and patterns of connections govern emergent phenomena, such as information cascades on the Web [11, 43, 91, 103, 105, 147], frequent subgraph patterns of a recommendation network [106], online diffusion networks [61], or diverse real-world networks [119], firm innovation [14, 141, 151], contagion of obesity [46], happiness [57] or disease [138, 144, 179], and brain functionality [10, 32, 41, 164]. Accordingly, there has been increasing attention to the universal structural properties of large complex networks such as relatively small characteristic path lengths, high clustering coefficients, power-law degree distributions, and the presence of community structures [15, 26, 37, 182].

Human activity such as blogging and tweeting often shows periodicities (weekly and yearly regularity). Traffic in the blogosphere leads to a seven-day periodicity due to a *weekend effect* of sharply dropping off at weekends [105]. Such periodicity is also found in our study of news diffusion in social media covering mainstream news, the blogosphere, and social networking sites [91]. In addition, heavy-tailed dynamics has been observed in the popularity of blogs, the distribution of blog sizes, bursty patterns in blog activity [43, 64, 105], and information cascades in social media [91].

These universal properties [37, 124, 128] of user behaviors can serve as one of the parameters in modeling spreading processes on networks. Our model-driven approach [85] assumes one of these real-world network properties, a power-law degree distribution, which is discussed in detail in Chapter 5.

2.2.5 Diffusion Factors

Studies on diffusion models have mainly focused on three factors affecting information propagation such as social structure (e.g., peer influence or connectivity), node characteristics (e.g., individual interests or receptivity), and the nature of target information (e.g., endogenous or exogenous information). Individual or combined parameters can be one of the essential factors in successful information diffusion [13, 20, 22, 165].

There have been attempts to separate peer influence and peer similarity (homophily) effects on information propagation so that more detail and accurate level of interpretations is achieved [20, 22]. For instance, user behavior contagion (e.g., the purchase of a product and adoption of a new idea) is possibly due to the direct recommendation or exposure to information from socially connected neighbors, or due to the similar preferences between users. Meanwhile, the nature of information items needs to be considered as an additional important factor of information spread. The authors of [13] assign every information item two parameters, *endogeneity* and *exogeneity*. The endogeneity of the item quantifies its tendency to spread through the underlying connections between nodes, and the exogeneity quantifies its likelihood to be acquired by a node, independently of the underlying network.

In this regard, our proposed model-driven approach [85] considers all these factors in that (1) influence by contact networks (internal influence) is separated from confounding factors (external influence), (2) our model quantifies the probabilities of internal and external influences, providing likelihood of endogeneity and exogeneity of an information item, and (3) we consider a meta-population scheme [29] which categorizes heterogeneous social networks into meta-populations, exhibiting different node characteristics between meta-populations. Note that our model-driven approach estimates macro-level diffusion across heterogeneous meta-populations, and thus it considers node characteristics at a meta-population level rather than an individual node level. Details are discussed in Chapter 5, 6 and 9.

2.2.6 Fundamental Frameworks of Diffusion Processes

One of the best ways to understand effects of network properties from empirical studies is to build mathematical models. There have been attempts to model a diffusion process on contact networks for understanding its underlying mechanisms from different research fields such as social science, marketing, medicine, physics, and computer science.

In the early 1960s, adopting behaviors were classified in the social sciences into five categories in terms of the timing of adoption, such as innovators, early adopters, early majority, late majority and laggards [146]. In the marketing literature, this idea was mathematically represented with a conditional likelihood of adoption by the Bass Model [30]. This model consists of likelihoods of "innovation" and "imitation", which correspond to the concept of external and internal influences in computer science literature [85, 107, 123], respectively.

The mathematical modeling of epidemics has also been an active field of research across different disciplines such as epidemiology [17, 39, 79, 157], physics [37, 120, 126, 138], mathematics [78, 149], and computer science [44, 83, 118, 189] for the description of the diffusion of pathogens, knowledge, and innovation. Depending on the stage of the disease, such as *susceptible* (denoted by S, those who can contract the infection), *infected* (I, those who have contracted the infection and are contagious), and *recovered* (R, those who have recovered from the disease), there have been several commonly used baseline models to describe the general properties of epidemic spreading processes; namely, the *susceptible-infected* (SI) model, the *susceptible-infected-susceptible* (SIS) model, and the *susceptible-infected-recovered* (SIR) model [21, 23, 70].

In this thesis, the Bass Model is considered as a fundamental framework for our model-driven approach in Chapter 5 due to its simplicity and robustness [31], which are important to model the dynamics of heterogeneous populations. Compared to the Bass Model, epidemic models are too complicated to incorporate the heterogeneity and structural connectivity of social networks at the same time. That is because we need to consider distinct stages of disease for each population such as S, I, and R, the transmission rate between different stages, and the diffusion factors discussed in the previous subsection. However, this epidemic approach can be one of future research topics.

2.2.7 Effects of Structure and Heterogeneity of Populations on Diffusion

The Bass Model has provided realistic and robust estimation of new product growth patterns [31], and thus, it has been one of the influential diffusion models across diverse areas, such as marketing, computer science, economics and operations research

[31, 95, 107, 110, 130, 188]. Its fundamental assumption is that each individual meets completely at random and has an equal chance of coming into contact with every other individuals (*fully mixed*, or *mass-action approximation*) [127]. This simplicity has enabled intuitive interpretation and has led to a wide range of extensions of the model [31]. Also, the traditional approach of epidemic models [23] disregards underlying contact networks and assumes that a population is homogeneous and fully connected in the same way as traditional macro-level diffusion models [30, 127]. This is a strong assumption, but it allows for models in the form of systems of ordinary differential equations for the fraction of individuals in the various compartments (*Susceptible*, *Infected*, and *Recovered*) [37].

However, many real-world networks exhibit non-uniform but heavy-tailed degree distributions [175], as discussed earlier. In terms of network structures, there has been interest in the effect of network topologies on the diffusion, such as cluster density and reachability [97, 151], and degree distributions [107]. The authors of [25] make an assumption of two levels of mixing based on the SIR model. Each node belongs both to the network as a whole, and to one of a specified set of subgroups within the network such as families. Accordingly, there has been interest in the effect of clustering on the propagation of diseases. For instance, disease transmission much more likely takes place within subgroups than transmission between subgroups [25, 97]. The authors of [107] incorporated degree distributions into the Bass Model, but their assumption of a linear influence of the number of neighboring adopters does not guarantee the probabilistic constraint. Details are discussed in Chapter 5.

All these studies are still limited to a single social platform or a homogeneous population. One extension of the Bass Model allows heterogeneous populations such as multinational diffusion of a product. For example, the adoption rate of a consumer product in one country indirectly influences that in other countries [95, 142]. However, this extension disregards the effect of network topologies on diffusion. The authors of [132] integrated different layers of single social networks into a weighted composite network scheme, such as Bluetooth proximity networks, call log networks, affiliation networks, and friendship networks, for 55 university students. They focused on homogeneous social networks represented by multi-layered networks with different levels of importance, but our study covers heterogeneous social networks with intra- and inter-network interactions.

In contrast to the previous studies, our research aims at considering the effects of structures and heterogeneity of populations on diffusion at the same time. Accordingly, we propose the Dynamic Influence Model in Chapter 5 and apply the proposed model to real-world diffusion in Chapter 6.

2.2.8 Information-Theoretic Measures of Information Flow

Diffusion models assume social interactions based on sampled snapshots of current social networks [43, 105], site-specific actions (e.g., mentions, retweets, and hashtags in Twitter) [75, 123, 147], or real-world network properties [85, 86, 107]. However, real-word network structures are hard to collect or define particularly in the case where heterogeneous social networks are interconnected. In this regard, there have been attempts to estimate diffusion trends without any assumptions on complex and dynamic social interactions in the real world by using information-theoretic measures such as the mutual information and transfer entropy.

Extensive research on such information-theoretic measures has been conducted in diverse areas including computer science [60, 177, 178, 180], neuroscience [133, 134, 135, 143, 173], and economics [99, 108, 113]. When it comes to social media, the predictability of user behavioral patterns has drawn attention, such as individual or group level future interactions [180], information pathways between Twitter users without explicit follower-followee relationships [177, 178], and classifications of Twitter user behaviors [60].

These studies focus on homogeneous population within a single social platform alone. However, as discussed earlier, heterogeneous social networks are increasingly interconnected across multiple social media platforms (e.g., CNN, BBC, Facebook, Twitter, WordPress, and Tumblr). Our research proposes a model-free approach in Chapter 7 which shows a way of estimating inter-relationships between heterogeneous populations at a macro level. Also, we apply the proposed model-free approach to real-world diffusion in Chapter 8.

2.2.9 Topic-Related Diffusion

These modeling approaches can be applied to technological and commercial domains where rapid and efficient spread of information or service is often desired such as marketing campaigns [102, 121]. One of the crucial studies has focused on identifying influential nodes [81, 82, 100, 104]. From the perspective of recommendation performance, the authors of [163] predict the destination of information flow, and ranks users based on the estimated time of how quickly information will reach to them. The authors of [114] predict the size of diffusion of a specific item. In common, these researches assume a peer influence-based influence propagation on the underlying networks and no distinctions between different information items. Accordingly, learning approaches to the probabilities of influence between nodes has been proposed by harnessing the observation of the spread of a particular information item [62, 65]. Also, they regard identical propagation of diverse information items.

Significant variations in diffusion patterns have been observed between different topics [123, 147]. For instance, diffusion of political issues is considerably driven by external influence, while entertainment topics spread through internal communications [123]. The authors of [147] showed that political topics are relatively persistent compared to non-controversial subjects. However, these studies have focused on a single social platform, such as Twitter, so the dynamics of external and internal influences is limited to a single homogeneous social system. In this respect, our study examines global diffusion across different types of social media, i.e. heterogeneous social systems, with comprehensive topics using the Wikipedia Current Events [9] and is not limited to site-specific trending topics.

Table 2.1: Table of common symbols used in this thesis.

Symbol	**Description**
$a(t)$	The number of new adopters at time t
$A(t)$	The number of cumulative adopters until time t
$f(t)$	The proportion of new adopters at time t
$F(t)$	The proportion of cumulative adopters until time t
p_i	Probability of external influence on a meta-population i
c_{ji}	Probability of influence of a meta-population j on a meta-population i
α	Power law coefficient
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Stochastic processes (an ordered set of random variables)
$\mathrm{TE}_{Y \to X}$	Transfer entropy from $\mathbf{Y}$ to $\mathbf{X}$
$\mathrm{TE}_{\mathrm{out}}$	Outgoing transfer entropy
$\mathrm{TE}_{\mathrm{in}}$	Incoming transfer entropy
k	Markov order for the previous states of $\mathbf{X}$
l	Markov order for the previous states of $\mathbf{Y}$
d	The length of time-delay

Our ultimate research goal is to obtain a comprehensive understanding of topic-related diffusion across heterogeneous populations. This is enabled by comparing our model-driven and model-free approaches with respect to information topics in different real-world application domains, such as social media and academic publications. Chapter 9 discusses overall insights from the comparisons in depth.

2.3 Nomenclature

Table 2.1 describe symbols which are commonly used in this thesis. In addition, we separate chapter-specific symbols and present them in each chapter. A complete list of symbols throughout the chapters can be referred in the Appendix A.1.

Part I

Real-world Diffusion Phenomena

Chapter 3

News Diffusion in Social Media

We are constantly bombarded by human interest, new scientific findings, sports, and current events as they move quickly through social media. Studying the diffusion of news through social media is particularly important in understanding how stories and information evolve and interact with other pieces of information. This chapter focuses on identifying trending news topics, capturing their spreads in user-contributed content, studying the structural connectivity of the underlying networks from the aspects of document, user, and information networks, and finding interconnectivity between diverse news topics.

3.1 Introduction

News stories quickly spread worldwide, some immediately affects the political, economic, and social lives of millions (e.g., the Blitz in London during WWII), some bears long-term cultural and ideological influence (e.g., The Declaration of Independence), and some plays a significant role in both (e.g., the 1933 recession and subsequent changes in finance practices). Tracking such emergent phenomena is a significant step towards understanding consequential social circumstances. Recent availability of large collections of real-time user-contributed content in social media has facilitated to study such collective behavior at scale.

Tracking an ever-changing list of often unanticipated events is difficult, and most prior work has focused on specific events such as quotes or site-specific interests. For instance, online diffusion tracking has been conducted by finding shared keywords or similar text between documents [11, 66], by tracing hyperlinks [43, 105] or URLs [123], by defining meaningful reaction chains from site-specific actions such as *retweets* or *hashtags* in Twitter [75, 98, 147] and *shares* or *likes* in Facebook [24, 167], by identifying commonly used *meme phrases* [103], or by performing event versus non-event classification on clustered input Twitter stream [33]. In this study, we propose a method for identifying real-world news in social media, and we analyze news diffusion patterns across diverse media types such as mainstream news (News), social networking sites (SNS), and blogs (Blog).

As a working example of real-world diffusion phenomena, we focus on news diffusion in social media by investigating the Spinn3r dataset [4]. This dataset contains

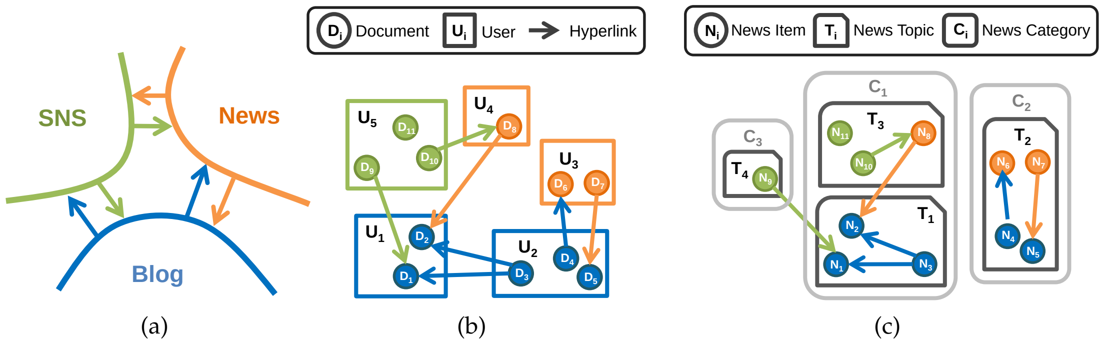


Figure 3.1: Overview of diffusion in social media. **(a)** Complex information pathways across different types of social media such as social networking sites (SNS), mainstream news (News), and blogs (Blog). **(b)** Underlying network structures: two different levels of networks are defined as *document* (D_i) and *authorship* (U_i) networks. **(c)** Information networks: each document is labeled with a *news item* (N_i), continuous news items are updated and aggregated as a common *news topic* (T_i), and each news item also belongs to a *news category* (C_i). Different social media types are color-coded.

386 million web documents, covering a one-month period in early 2011. In this study, we propose a novel method for identifying news-related documents across diverse social media platforms on the Web. This is achieved by referring to the Wikipedia Current Events [9] as a noteworthy news registry and by presenting each news item with names entities covering journalism-inspired features "5W1H", i.e. *Who, What, Where, When, Why and How*. As Figure 3.1 shows, we construct different levels of underlying networks such as document and authorship networks in Figure 3.1(b), and information networks in Figure 3.1(c), by classifying Web documents with the identified real-world news.

For more accurate and explicit network structures, we trace hyperlinks and/or URLs in the main text of each document. From the constructed document networks, we observe that a single document links to different news-related documents, and these linkage patterns account for the majority of the total links. Such cross-topic citations sometimes reveal surprising yet reasonable connections among different real-world news stories which co-evolve over time, such as "Australian Open"-to-"Queensland Floods"-to-"Cricket Game Cancellation". We also study authorship network structures with over 350 thousand nodes. The majority of users (55%) cite to a small fraction of prolific users (4%), a notable departure from the balanced traditional bow-tie model of web content. Such observations, to the best of our knowledge, is seen for the first time.

In the remainder of this chapter, Section 3.2 presents our data processing methods and network statistics, Section 3.3 proposes a method for identifying real-world news and document classification, Section 3.4 analyzes news diffusion patterns from constructed information networks, Section 3.5 observes underlying user networks using the bow-tie model, and finally Section 3.6 concludes this study.

Table 3.1: Key fields of each document record used in the dataset.

Data Field	Main Usage
Publication time	To validate the direction of links from source to destination documents.
Document URL	To obtain the unique identity of documents and also extract user identity from its regular expressions.
Full HTML body	To extract hyperlinks and/or written URLs in the main text of a document.
Written language	To target English documents for avoiding noisy translation.
Publisher type	To target documents of three representative types of social media such as News, SNS, and Blog.

3.2 Preparation and Analysis of Social Media Data

To explore a real-world example of diffusion phenomena, we investigate instances of news diffusion in social media. We first describe our data collection and preprocessing methods, and then examine the fundamental statistics of the data.

3.2.1 Data Collection

We base our work on the Spinn3r dataset which is freely available to the research communities via ICWSM'11 [4]. This dataset consists of over 386 million social media content, such as blog posts, news articles, microblog content, classifieds, and forum posts, during a one-month period in early 2011 (13 January to 14 February). It was originally collected by Spinn3r [8], which is a licensed social media crawler. One document record in the dataset contains information about a title, publication timestamp, written language, document URL, estimated spam probability, and the full HTML body. Key fields of the dataset used for this study are described with their usage in Table 3.1 .

3.2.2 Target Documents Selection

We focus on analyzing documents from mainstream news (News), social networking sites (SNS), and blogs (Blog), which account for 98.37% of the original dataset. These documents are important in that they are not only the most relevant to real-world news stories, but we can also observe dynamic interactions among the representative social media types, i.e. News, SNS, and Blog. We choose to keep documents written in English to avoid error in translation. We then filter out duplicate documents (documents with the same contents and Web address) for their unique identity and disregard the documents with a non-zero spam probability provided. As a result, nearly 6 million documents are left after the filtering.

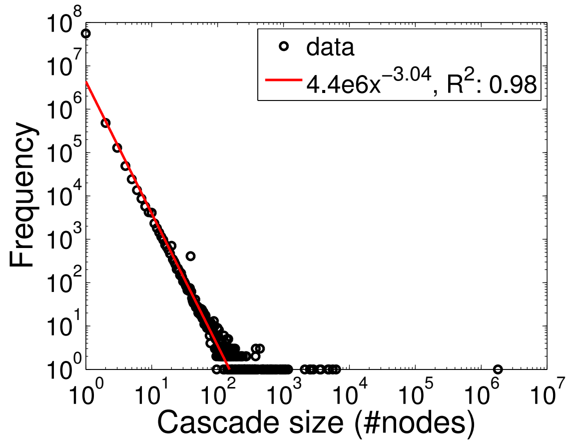


Figure 3.2: Distribution of hyperlink cascade sizes.

Link Extraction: The full HTML bodies of documents contain large spam links and lengthy header and footer information, which can lead to wrong interpretations of document linkage patterns. Thus, our data processing first needs to remove such boilerplates and keep the main document contents. We use the effective *boilerpipe* library [93] to achieve this goal. We then follow out-links (hyperlinks and/or written URLs) in the main text of a document for tracking diffusion. Hyperlinks in the document contents, however, often contain shortened URLs, masking the true identity of link destination. There are over 300 URL shortening services, and this makes it infeasible to query their distinct APIs to recover the original links. Moreover, some links have been shortened more than once, further complicating the recovery. To tackle this issue, we extract the original location from the HTTP header. After extracting out-links, we remove self-links and out-of-scope links that connect to documents outside of the dataset. We also disregard links that point to documents created after the referring documents.

Non-isolated Documents: Figure 3.2 shows the distribution of hyperlink cascade sizes for the selected 60 million English documents, and it shows a heavy tailed distribution. The x-axis indicates the number of weakly connected documents from a single document (isolated due to no hyperlinks or URLs in its main content) to the largest connected documents, and the y-axis shows its frequency. The non-isolated portion has 4,138,283 documents which accounts for 6.9% of the original 60 million documents. Such a small percentage of connected documents tells us that the vast majority of documents have no citations and thus have no linkage to other documents. This is not a trivial fraction when compared with the literature when only 2% of blog posts are not isolated [105]. In fact, our higher percentage results from links between three different types of document sources (News, SNS and Blog) based upon a wide range of content types of the Spinn3r dataset.

Table 3.2: Identified users in social media from 60 million English documents covering a one month period in early 2011. The term "user" indicates an agent or actor who produces documents. A news agency or site is also considered as a super-user (selected news sites are from the largest strongly connected news network).

Media Type	Domain	User Count
News	second-level domains	9,225
SNS	facebook.com	4,560,800
	myspace.com	822,998
	flickr.com	25,613
	twitter.com	6,169
	posterous.com	1,876
Blog	blogspot.com	691,175
	livejournal.com	158,361
	wordpress.com	90,803
	tumblr.com	23,967
	typepad.com	7,603
Total		6,398,590

3.2.3 User Identification

In this study, an individual who has an account in any social media platform is regarded as a *user*, but identifying individuals with more than one account is out of the scope of this study. In addition, each news site is regarded as a super-user. Throughout the thesis, the term "user" indicates an agent or actor of each media type who produces Web documents.

There is no universally valid user information, due to the diverse sources that the dataset draws from. In this regard, we chose five SNS and Blog domains for each media type, as they are not only popular spaces for social networking and blogging, but we can also write regular expressions for extracting user identities from their produced document URLs. This method generates a significantly large set of users, and it is consistent with prior blog user extraction methods [43]. To identify news sites, we regard second-level domains (e.g., cnn.com, nytimes.com) from document URLs as unique identifiers of news sites only when a document's publisher type (in Table 3.1) is mainstream news. We finally extracted 9,225 news sites, which constitute the largest strongly connected network. This strong connection implies that each news site can reach every other news site, which provides news sites with more frequent chances to be connected with each other and further to be exposed to other social media types, such as SNS and Blog.

As shown in Table 3.2, we identified 6.4 million users in total from the selected 60 million English documents. These identified users generated 57% (34 million) of the target documents. In Figure 3.3, red circles describe the distribution of the number of connected documents that are produced by the identified users, while black squares

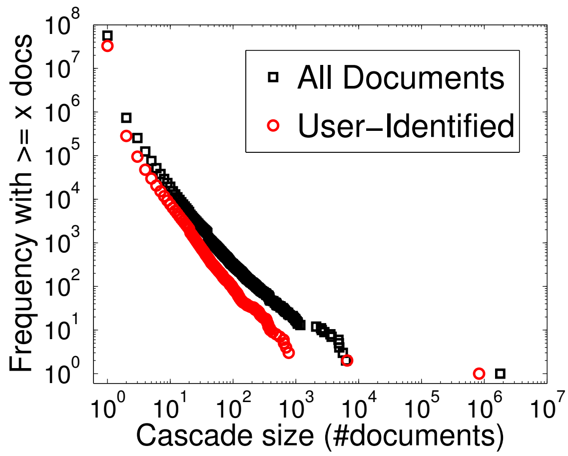


Figure 3.3: Distributions of hyperlink cascade sizes. Black squares indicate the complementary cumulative distribution of hyperlink cascade sizes for all 60 million English documents, while red circles are for documents created by the 6.4 million identified users in Table 3.2.

are for the selected 60 million English documents. The latter is shown in Figure 3.2, but it is redrawn here for comparison. The x-axis indicates the hyperlink cascade (a maximal set of connected documents) size as in Figure 3.2, while the y-axis shows the frequency that the connected document size is equal to or greater than the value of x. As the figure illustrates, the majority of hyperlink cascades are attributed to the documents generated by the identified users. This means that popular social media platforms of News, SNS and Blog likely contribute to a wide diffusion in social media, which also implies that these identified users can provide robust statistics for studying diffusion mechanisms in social media.

3.3 Document Labeling with Real-world News

There are two challenges in extracting information networks from the underlying document networks: (1) to identify trending news items, and (2) to associate each document in our data collection with one of the identified news items that it describes. Our approach starts by learning daily news items from a crowd-sourced online registry. We design document features motivated by journalistic practice to make it possible to compute the similarity between a document and a news item. This is notably different from tracking meme phrases [103] or approximating some events using text clustering [68].

3.3.1 Real-world News Identification

Choosing a relevant data source of noteworthy real-world news requires some caution. Traditional news outlets such as the NYTimes or the BBC are subjected to institutional and geographical biases; a social news reader of, say, Digg is presented

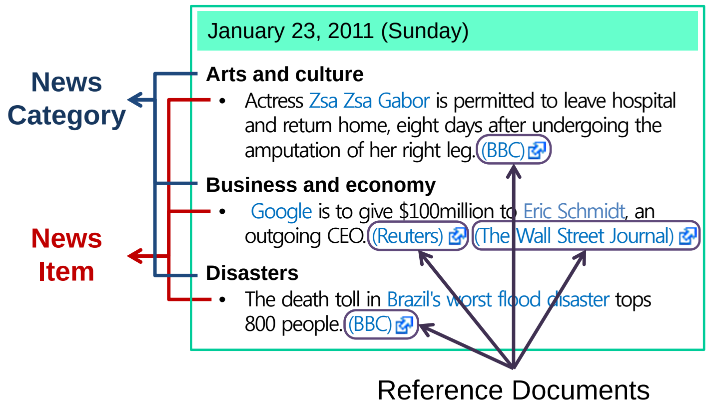
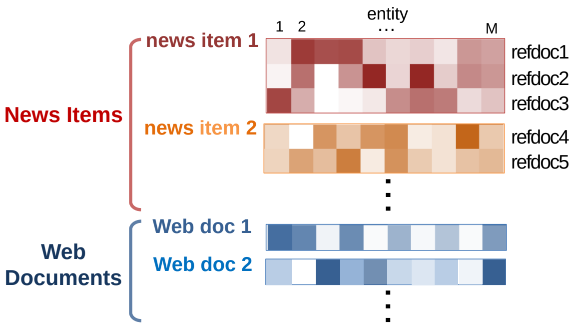


Figure 3.4: An example snapshot of the Wikipedia Current Events [9]. Each bullet point is referred to as one *news item*, which shows the short summary of a happening on that day, along with reference hyperlinks to relevant articles (rounded rectangles). Each title (bold font) is referred to as the corresponding *news category* of a news item.

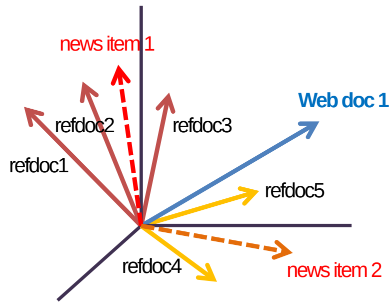
by articles and not by news topics; news aggregation sites, such as Google News, have better coverage but retroactive crawling is not easy due to no information of previous news registries. On the other hand, the Wikipedia Current Events [9] is a chronologically organized event registry of public interest, continuously updated and discussed by volunteers. This seems to be the best news source of public interest across different types of social media, despite potential selection bias of volunteers who self-select to be editors.

Conceptual News Hierarchy: We define a news hierarchy at three increasing levels of generality, as shown in Figure 3.1(c). For instance, a *news item* is the smallest unit of daily new happenings. Some happenings possibly continue for a short or long term, not necessarily day by day, with regional variations (e.g., local, nationwide, or worldwide). Thus, we need to consider grouping such subsequent news items with a common subject, location, person, and/or organization into one *news topic* as people collectively think of them as one event. Also, each news topic can be categorized into one general *news category*. For example, news items such as “Brisbane River banks break in January 2011” and “North Queensland hit by Cyclone Yasi in February 2011” can be grouped into “Queensland Floods” as one news topic which can be classified into the news category “Disaster”.

As Figure 3.4 shows, a news item is mapped into each bullet point which describes the short summary of a daily new happening. Every news item comes along with hyperlinks to reference articles for its detailed description (rounded rectangles in Figure 3.4). These references are all crawled in order to train the model of each news item. Details are discussed later in this section. Titles (bold fonts in Figure 3.4) are also referred to as news categories, and we choose five major categories which are representatives across professional news agencies and have a sufficient number of news items over time. The selected news categories are: Politics, Business & Economy, Disaster, Arts & Culture, and Sports.



(a) News representation with named entity vectors



(b) Document labeling with vector space model

Figure 3.5: Document labeling with identified news items. **(a)** News items and Web documents are from the Wikipedia Current Events [9] and our dataset respectively, and they are all represented with M (= 4,411) dimensional named entity vectors. M entities are from the reference documents (*refdoc*#: rounded rectangles in Figure 3.4) of all selected news items. A vector's element indicates the *TFIDF* of an entity, and it is color-coded. **(b)** The similarity between a Web document and a news item's centroid vectors is calculated in order to label the document with the most similar news item.

Target Real-world News: We parse the Wikipedia Current Event page of January 2011 by considering our dataset period (13/Jan/2011 – 14/Feb/2011). In order to avoid missing diffusion traces, we target news items that occurred between the 6th January (one week before the beginning of our dataset period to account for diffusion of near past news items) and the 30th January, 2011 (two weeks before the end of our dataset period to observe news diffusion for at least two weeks). Finally, we identified 284 news items for this period, among which subsequent news items are manually grouped into one news topic, and accordingly we obtained 161 news topics. Note that the automatic clustering of ongoing news items is not only challenging but also noisy for the diffusion tracking of a news topic. This is because common entities between unrelated news items can mislead to be grouped into one news topic, and it is hard to define one universal rule from diverse factors to determine the relevance of different daily news items. We found that the manual approach is feasible and unambiguous for less than 200 news items in this investigation. For a larger number of news items, automatic clustering and inter-user agreement on its results can be topics for further research.

News Representation with Named Entities: A news item, the smallest unit of daily new happenings, is well defined by the "5W1H", i.e. *Who, What, Where, When, Why and How*, of journalistic practice. Note that among the five Ws, at least three (who, where and when) directly correspond to entities, such as person name, organization, location, date and time indicators. Moreover, the rest of the 5W1H (what, why and how) often contain entities to make statements precise and credible. Therefore, we propose a way of representing a news item with a named entity vector, as shown in Figure 3.5(a), whose elements consist of the TFIDF (term frequency-inverse document frequency) score [111] of each entity extracted from the reference documents of the news item (rounded rectangles in Figure 3.4).

Named entity recognition in documents has been an active research area of natural language processing for over a decade. There are still challenges being actively tackled by the community, and efforts to date have produced high-quality tools. We conducted named entity recognition by using the OpenCalais API [7] which provides up to 116 types of entities (from *Anniversary* to *Voting Results*). We extracted 4,411 unique entities (also, using entity resolution techniques as described in the next section) for 284 news items from the crawled reference pages. For each news item, 4,411-dimensional entity vectors are generated for all reference documents of the corresponding news item, and their centroid vector is compared with Web document vectors for the document classification. Details are discussed in the next subsection.

3.3.2 Document Labeling with Identified News Items

We also represent each Web document with a named entity vector with the same dimensions as the vectors of news items as shown in Figure 3.5(a). We then use the vector-space model for classifying documents into news items by calculating the similarity between document and news class vectors (more precisely, the centroid of news item vectors) as shown in Figure 3.5(b); the most similar centroid vector specifies the most similar news item. Each document can potentially be labeled with zero, one, or more classes of news items. For simplicity, we assign a maximum of one news class per document in this work.

Entity Resolution: An entity name can occur in many different ways among Web documents, resulting in multiple dimensions for the same entity. For example, in our data collection we identify nine name variations for Tunisia's former president Zine El Abidine Ben Ali, including "Zine Al-Abedine Ben", ''Zine Al-Abedine Ben Ali", and "Zine Al-Abdine Ben Ali". To alleviate this problem, we employ approximate string matching techniques to cluster similar entity names. Such techniques are commonly used in entity resolution and data matching to identify similar strings that refer to the same entity.

We investigate four techniques that have been found to be effective for matching names [47]: (1) edit distance, which counts the number of character edits (insertion, deletion, and substitution) to convert one string into another; (2) Jaccard distance based on character bigrams, which counts how many bigrams (sub-strings of lengths two) two strings have in common; (3) Winkler, which is a comparison technique specifically for matching English surnames; and (4) the longest common sub-string approach which recursively extracts the longest sub-strings two strings have in common and then counts the number of characters in these common sub-strings. Each of these comparison functions returns a normalized similarity value between 0.0 (for two totally different strings) and 1.0 (for two strings that are the same).

Evaluation of Document Labeling: For the evaluation of document labeling, we use 653 crawled reference documents for 284 news items as ground truth. We divided them into 538 (80%) documents for training 284 news items and 115 (20%)

Table 3.3: Results of document labeling. Document labeling is conducted on 4 million documents, each of which has at least one connected document, for tracking news diffusion. About 69% of documents contain named entities of selected news items, i.e. (a)+(d), but 21% of them are filtered out based on the threshold value ($\tau = 0.14$).

Doc. Labeling	**Types of Documents**	**Doc. Count**	**Proportion**
Labeled	(a) documents whose similarity scores of news items are over the threshold value (τ)	2,254,049	54.47%
Unlabeled	(b) documents with no content or not supported languages by entity recognition API	341,634	8.26%
	(c) documents containing no news-related named entities	957,957	23.15%
	(d) documents whose similarity scores of news items are under the threshold (τ)	584,643	14.13%
Total		4,138,283	100.00%

documents for testing. The baseline classification accuracy based on cosine similarity between document and news class vectors is 68%. The entity resolution brought the improvement of the accuracy up to 74% using the best-performing Winkler string matching technique. Since not all documents in our data collection are related to the identified news items, we need to choose a similarity threshold value (τ) that filters out documents that correspond to none of the known news items. We empirically determine the threshold value as 0.14 which maximizes the F1 (the harmonic mean of precision and recall) score with a multi-label classification evaluation metric [174]. Document labeling is conducted on 4,138,283 documents which are connected by at least one hyperlink or URL in the main text of a document. This helps us to build more accurate and explicit citation networks of news items and trace news diffusion. As a result, a subset of 2.8 million documents have text containing at least one named entity of a news item, and within this subset, 2.2 million documents are classified into our selected news items with a confidence score above the threshold ($\tau = 0.14$). Details are shown in Table 3.3.

3.4 News Diffusion Patterns

We extracted non-isolated documents and labeled them with the identified news items. Based on these labeled non-isolated documents, we observe the linkage patterns of information networks by aggregating labeled document (N_i) networks into topic (T_i) networks, as shown in Figure 3.1(c).

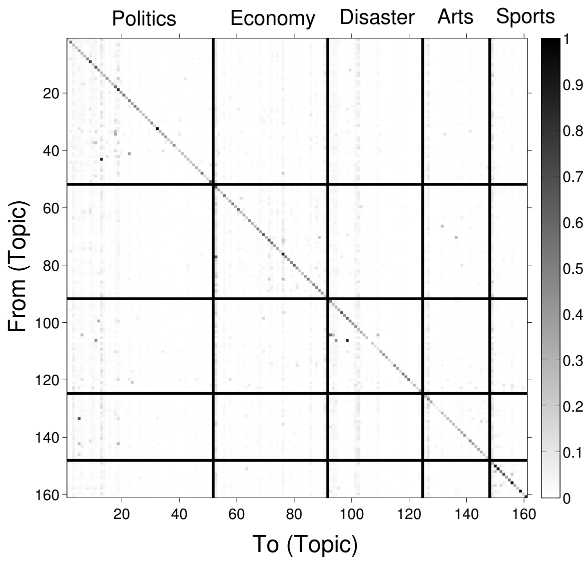


Figure 3.6: Normalized linkage patterns between news topics. Different news categories are separated by black horizontal and vertical lines.

Table 3.4: Document clusters by conducting the normalized cut [155] on the largest cascade of all news categories combined.

	Politics	**Business & Economy**	**Disasters**	**Arts & Culture**	**Sports**
Cluster1	0.0%	0.0%	20.0%	10.0%	70.0%
Cluster2	14.8%	34.6%	23.5%	21.0%	6.2%
Cluster3	54.5%	9.1%	18.2%	18.2%	0.0%
Cluster4	61.5%	7.7%	30.8%	0.0%	0.0%
Cluster5	53.1%	24.5%	12.2%	8.2%	2.0%

News Topic Linkage Patterns: Figure 3.6 shows the normalized linkage patterns between news topics where the value of element (i, j) of the matrix is defined as the number of linkages from news topic T_i to T_j divided by the total number of linkages of T_i. The matrix is rearranged to group news topics based on their news category and expressed in the gray-scale map, where black is the maximum value (1.0) and white is the minimum one (0.0). Note that some dark-gray vertical lines are found in the normalized linkage matrix. These indicate some particular news topics that receive links from most of the news categories such as "US banking crisis", "2011 AFC Asian Cup", "Academy Awards", and "Haiti earthquake anniversary".

From the observation of inter-category linkages in Figure 3.6, we investigate how closely news topics in different categories are connected by applying a clustering algorithm, the normalized cut on the network [155]. Table 3.4 shows the distributions

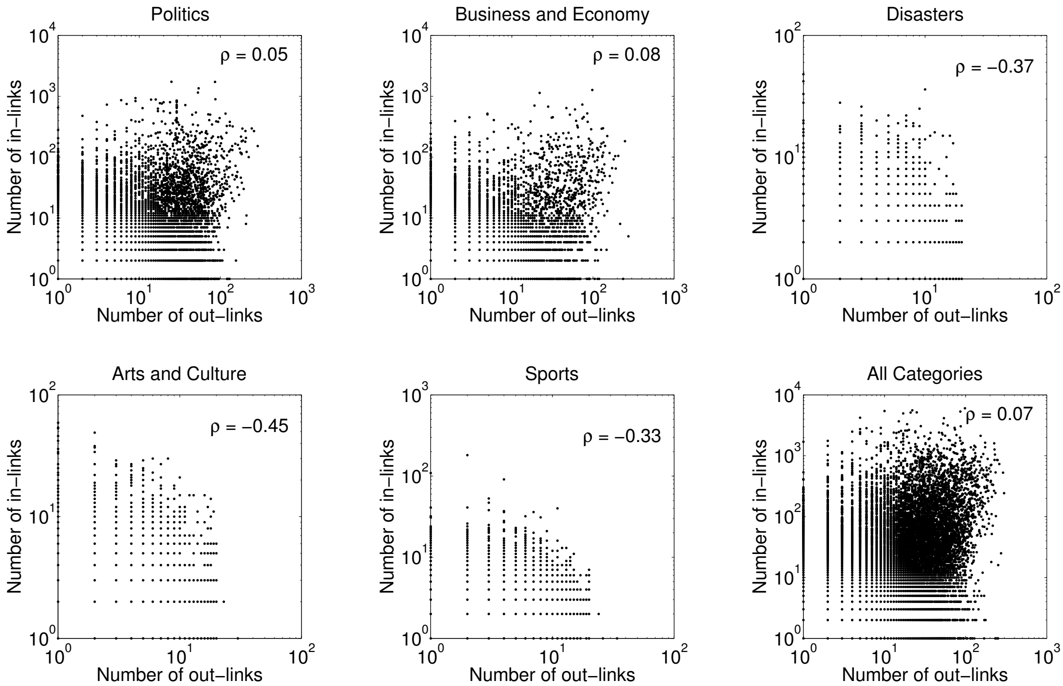
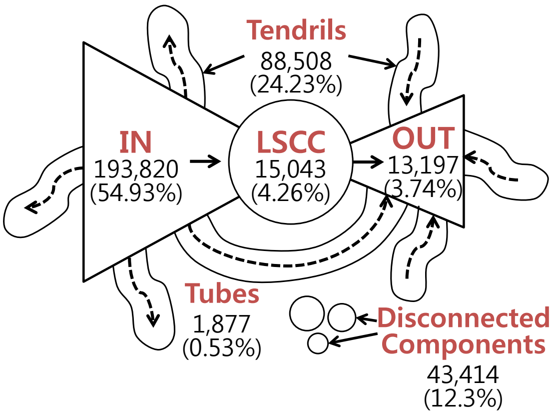


Figure 3.7: Scatter plots of in- and out-degrees of documents for each category and all categories combined. ρ denotes the correlation coefficient.

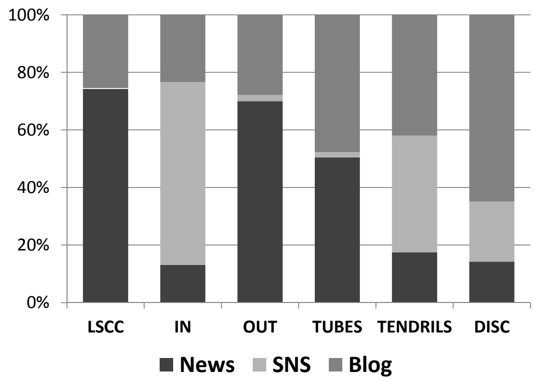
of news categories in each cluster, resulting from the normalized cut. It is difficult to find one-to-one matching between clusters and categories, which reconfirms that news topics are tightly connected across categories. That is, there is no cluster consisting of one news category. The news topics of Cluster1 are about the "2011 Australian Open", ''Queensland Floods", "Cricket Game Cancellation", ''Victory of England in the 2010-2011 Cricket Series in Australia", and "England Hooligans on Trial". In addition, interestingly Cluster1 has news topics mostly related to Australia and England, Cluster3 to Israel and Iran, and Cluster4 to Pakistan and India. It shows that the news topics are interconnected with each other across different news categories, and these connections likely reveal geographical relationships.

Figure 3.7 shows a scatter plot of document in-degrees versus out-degrees for each of the five categories, and across all categories. As the figure shows, there are no strong correlations between in-degree and out-degree sizes, which means that a document which is largely cited does not necessarily cite a lot, and vice versa.

This section discussed linkage patterns among news topics and categories. The findings can serve as parameter estimates for the diffusion rates between news topics or categories, which is one of the topics for future study of event evolution modeling.



(a) Skewed bow-tie



(b) Proportions of users by media types

Figure 3.8: Reachability of user network. **(a)** Macroscopic structure of user network by the bow-tie model. **(b)** Media type distributions of each bow-tie component.

3.5 Underlying User Network

In this section, we examine over 350,000 users who generated non-isolated documents, since they participate in sharing and spreading information. We analyze the user network at the macro level in order to obtain a global view of reachability in the network by using the *bow-tie model* [40], as discussed in Chapter 2, which breaks the network into six components according to the network's structural connectivity.

Skewed Bow-tie Structure: The user network is based on the citation relationships as shown in Figure 3.1(b), so a user network itself reflects document citation structures at an abstract level. Figure 3.8 shows the bow-tie structure of this authorship network. Compared to web structure [40], we can see that the proportion of the central core (LSCC) is 4.3% which is much smaller than 27% in th web structure. On the other hand, the authorship network has a much larger IN component than the OUT component (54.9 vs 3.7%), whereas both are around 21% for the web. Such the user distributions of this surprising small core and the imbalance of IN and OUT are explained by Figure 3.8(b). The small LSCC consists mostly of users from News (74.37%) and Blog (25.45%). The IN accounting for more than half of all users, and the majority (64%) of users in this component are SNS users. The TENDRILS also have a significant SNS presence (41%). Blog users are positioned at all over the components with a minimum 20% of proportions for each component. The majority of SNS users tend to cite to news media, but they are not cited back from it, while Blog users have bidirectional interactions with News. Also, notable is the dominance of Blog users in DISC, suggesting that blogs are popular forums between themselves regarding common interests in non-news items.

Implications from the Underlying Network's Connectivity: Overall, unsurprisingly mainstream news seems to play a central role in spreading news in social media, but surprisingly the connectivity of the user network is unbalanced, i.e. a skewed bow-tie, which is a notable departure from the balanced traditional bow-tie model of web content. In other words, the large proportion of users obtain news stories from out-of-the-network sources rather than from internal contact networks. Thus, external influence is not ignorable. As discussed in Chapter 1, the increasing accessibility to diverse information sources makes social interactions far-reaching across multiple social systems, and thus the underlying social networks become more heterogeneous. Thus, this skewed but far-reaching connectivity of the user network reflects (1) the non-ignorable interactions between different social systems and (2) the heterogeneity of the underlying network. In addition, this structural connectivity suggests an existence of the directionality of influence with different strength in news diffusion across heterogeneous populations.

3.6 Conclusion

We analyzed news diffusion patterns across different types of social media by building different levels of networks such as information and authorship networks which are all based on document networks.

Regarding information networks, news topics are interconnected with each other across different news categories such as Politics, Business & Economy, Disaster, Arts & Culture, and Sports, and across different types of social media such as News, SNS, and Blog. This tells us two important insights to consider in studying diffusion. First, different information categories need to be examined for tracking diffusion. In other words, large diffusion reflects influence on wide areas of our society such as politics, business, economy, and so on. Second, such large diffusion is attributed to heterogeneous populations such as News, SNS, and Blog users not limited to local social networks in a single social platform. In terms of authorship networks, the underlying user network structure exhibits unbalanced but far-reaching connectivity. The majority of users (55%) cite to a small fraction of prolific users (4%) from different social media platforms, reflecting that external influence is not ignorable and the underlying social network is heterogeneous. These all supports the importance of studying diffusion dynamics in terms of the directionality and strength across heterogeneous social systems. Additionally, our analysis proposes approaches for real-world news or event tracking across different types of social media. We expect that this work would shed light on an analysis of diffusion patterns on the Web.

In this chapter, we focused on the structures of diffusion networks. In the next chapter, we investigate knowledge diffusion patterns in academic publications as an another working example of real-word diffusion phenomena. Common patterns will be examined between these two significant examples in the real world. The findings from the empirical analysis in this chapter motivate us to develop our new approaches, model-driven and model-free, in Chapter 5 and 7, and using the two proposed approaches, temporal dynamics of news diffusion are estimated in Chapter 6 and 8, respectively.

Chapter 4

Knowledge Diffusion in Academic Publications

In Chapter 3, we investigated news diffusion patters across different types of social media and observed that the underlying social networks are heterogeneous. In this chapter, we focus on academic publications and analyze knowledge diffusion patterns across different subdomains in computer science. Based on the constructed citation networks, we extract both conference and subdomain networks, and examine their linkage patterns in accordance with popular research keywords in computer science. We then discuss similar diffusion patterns between social media and academic publications.

4.1 Introduction

In the previous chapter, we observed that news spreads across different types of social media such as mainstream news, social networking sites, and blogs. When it comes to academic publications, the situation is not different from user-contributed content on the Web. For instance, we often observe that common research topics have been studied and shared across multiple disciplines, which collectively forms complex information pathways in academic publications.

Scientific knowledge diffusion has been studied focusing on a local research community [190], a research domain [19, 156], or multiple domains [38, 92, 140]. In particular, studying multidisciplinary diffusion is significant to obtain a big picture of diffusion trends in academia. However, it is hard to collect or define such heterogeneous networks, since it is challenging to disambiguate paper profiles (e.g., author name, title, and publication venue) [168] and to categorize publications with well-defined meta-information covering all research domains. In this regard, we focus on computer science citation networks and analyze global diffusion patterns across its multiple subdomains.

Our empirical study is based on the DBLP citation networks, covering recent 30 years, provided by AMiner [2]. In order to understand diffusion patterns across different research areas, meta-information is also collected from Microsoft Academic Search (MAS) [5], such as research categories (domains and their associated sub-

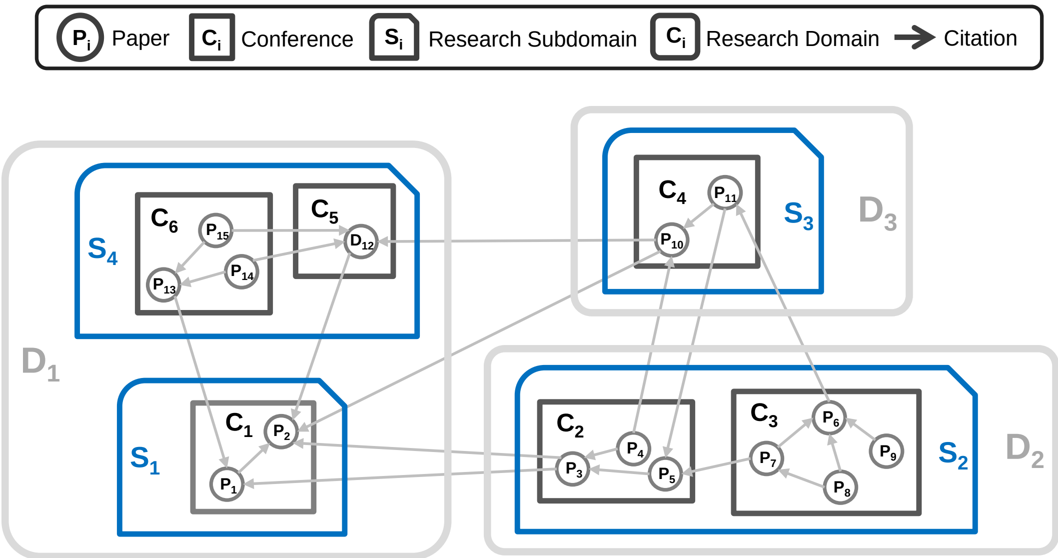


Figure 4.1: Conceptual structure of academic citation networks. Based on paper (P_i) citation networks, conference (C_i), subdomain (S_i), and domain (D_i) networks can be generated according to the attributes of each node in the citation networks.

domains) for each publication venue, and the rankings of conferences or keywords in each category. From the DBLP dataset, we first construct academic citation networks, where each node is assigned attributes by referring to meta-information from the MAS dataset. Based on the conceptual structure of academic citation networks in Figure 4.1, conference (C_i) and subdomain (S_i) networks in the domain (D_i) of computer science are extracted using the labeled node attributes in the constructed citation networks. Note that our dataset is from computer science bibliography, and accordingly only a single domain is available in this study.

We observe that top-ranked conferences lead computer science research, and the conference network structure exhibits a skewed bow-tie as in the user network structure in social media. The majority of conferences (51%) cite a small proportion of other conferences (10%) belonging to different subdomains. That is, the majority of conferences obtain new information from different types of conferences, showing the heterogeneity of the underlying network. Regarding the subdomain network, different subdomains are not only interconnected but also clustered along their common research interests. Some of them receive links from most of other subdomains, such as "Algorithms & Theory", "Databases", "Data Mining", "Distributed & Parallel Computing", and "Operating Systems", which more likely help research topics to be shared across different subdomains in computer science. Finally, we investigate diverse research keywords. They show different diffusion patterns according to their popularity. For instance, popular research keywords (within the top 100 keywords in computer science) much more likely spread across a wider range of subdomains than less popular keywords ranked below the top 100. We also find that the popularity of

Table 4.1: Information used from the DBLP dataset provided by AMiner [2].

Information	Description	Main Usage
Title	Paper title	To identify research keywords of each publication by using the MAS dataset.
Abstract	Publication abstract	
Year	Publication year	To obtain time-series of publications.
Venue	Publication venue (i.e., conference, journal, book, etc.)	To identify a research domain and a subdomain of each publication by using the MAS dataset.
Reference	Paper IDs of references	To construct academic citation networks.

a research topic needs to be considered with its coverage of research areas, not only with its publication size.

The rest of this chapter is structured as follows. Section 4.2 describes our datasets and preprocessing methods, and present major characteristics of academic citation networks. Then, in Section 4.3, we examine the constructed citation networks at two different levels, conference and subdomain networks, and analyze their linkage patterns. Section 4.4 investigates diverse research keywords and examine their different diffusion patterns across subdomains. Section 4.5 discusses common diffusion patterns between social media and academic publications, and finally Section 4.6 concludes this study.

4.2 Preparation and Analysis of Academic Publication Data

We will describe our data collection and data preprocessing methods for identifying and targeting academic publications in computer science. Also, the major characteristics of citation networks are presented.

4.2.1 Data Collection

We collect two datasets; (1) one dataset is for building academic citation networks, and (2) the other is for obtaining meta-information of publications and further providing the attributes of each node in the constructed citation networks.

Computer Science Citation Networks: Our analysis and observations are based on the DBLP (computer science bibliography) [3] citation network dataset[1], provided by AMiner [2] which is a free online web service for search and data mining operations on academic publications. This dataset is the integrated publication data from DBLP, ACM Digital Library, CiteSeer, and SCI (Science Citation Index) by resolving the issue of author name ambiguity [168, 169, 170, 171]. Each record in this dataset consists of information including a paper title, authors, publication year, publication

[1]The version of the dataset used is 6 which was the latest at the time of conducting our experiments [1].

Table 4.2: Information used from the Microsoft Academic Search (MAS) dataset [5].

Type	Information	Description	Main Usage
Meta-data	Research Category	Research domains and the associated subdomains. A publication venue (conference or journal) is classified with one or more research domains and subdomains.	To identify the research category of each publication in the DBLP dataset.
	Conference Ranking	The rankings of conference venues for each research domain or its subdomains.	To compare premier with non-premier conference publications; the top 10 conferences of each subdomain are investigated.
	Keyword Ranking	The rankings of keywords for each research domain or its subdomains.	To analyze topic-related diffusion patterns across different research subdomains; the top 100 keywords in computer science are investigated.
	Keyword Variations	Each keyword has variations in its presentation but with same meaning.	To extract relevant keywords from titles or abstracts of publications in the DBLP dataset.
Profile	Paper	Title, research keywords, year, venue, research category for each publication	To identify keywords of publications in the DBLP dataset by matching profiles provided if no keywords are available from titles or abstracts in the dataset.

venue, abstract, and reference papers within the dataset, each of which, however, may not always be available. The key attributes used and their main usage are described in Table 4.1 in more detail.

However, meta-information, such as research domains, subdomains, and keywords of publications, is not provided in this dataset, which is important to analyze knowledge diffusion patterns across different research areas. In this study, research domains denote academic disciplines or fields of study such as computer science, physics, social science, and so on, while subdomains indicate specific research subareas belonging to one research domain, such as data mining, operating systems, or software engineering under computer science.

Meta-information of Academic Publications: In order to obtain such meta-information, we used the dataset, provided by Microsoft Academic Search (MAS) [5] which is another free online search engine for academic content. This collected dataset contains meta-information (e.g., research domains and subdomains for each publication venue) as well as publication profiles (e.g., paper title, year, keywords, and venue for each publication), but with no citation relationships. For identifying

Table 4.3: Identified twenty-four computer science subdomains in the DBLP dataset by referring to the MAS dataset. "Abbr" denotes the abbreviation for each subdomain name. These abbreviations will be used for subdomain names in the remaining figures for clarity.

Abbr	Subdomain Name	Abbr	Subdomain Name
AT	Algorithms & Theory	MM	Multimedia
SP	Security & Privacy	NC	Networks & Communications
HW	Hardware & Architecture	WWW	World Wide Web
SE	Software Engineering	DP	Distributed & Parallel Computing
AI	Artificial Intelligence	OS	Operating Systems
ML	Machine Learning & Pattern Recognition	DB	Databases
DM	Data Mining	RT	Real-Time & Embedded Systems
IR	Information Retrieval	Sm	Simulation
NLS	Natural Language & Speech	Bio	Bioinformatics & Computational Biology
Gr	Graphics	ScC	Scientific Computing
CV	Computer Vision	Edu	Computer Education
HCI	Human-Computer Interaction	PL	Programming Languages

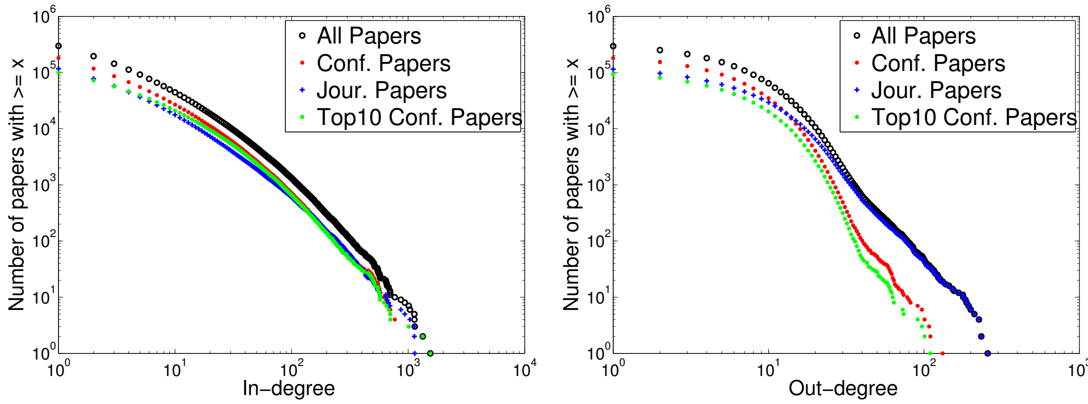
and targeting publications in the DBLP dataset, we also investigate top-ranked conferences and popular keywords in computer science by referring to MAS. Table 4.2 shows referred meta-information and paper profiles, and their main usage in detail.

In summary, from the DBLP dataset we construct citation networks, while node attributes in the networks are obtained by referring to meta-information in the MAS dataset. From these two complementary datasets, we eventually figure out knowledge diffusion across different subdomains in computer science in accordance with trending research keywords.

4.2.2 Fundamental Statistics

We examine main characteristics of academic citation networks based on our two datasets. From the DBLP dataset, we obtained 2,244,018 papers and 2,083,983 citation relationships between papers from 1936 to 2013.

Research domain and subdomain: By referring to the MAS dataset, we finally identified research domains of over 90% (2.1M) of publications (2.2M) in the DBLP dataset. 92% (1.9M) of these identified publications are categorized into the computer



(a) In-degree distributions

(b) Out-degree distributions

Figure 4.2: In- and out-degree distributions of computer science publications in the DBLP dataset, which are separated by the type of venues such as all, conference, journal, and top 10 conferences in each computer science subdomain. In total, 221 premier conferences are identified, belonging to 24 different subdomains in Table 4.3.

science domain, which confirms that DBLP is a computer science bibliography. In terms of a publication venue, about 60% of the identified publications are conference papers, while the rest are journal papers. As the vast majority (92%) of the identified publications are from the research domain of computer science, we accordingly identified 24 different subdomains for these computer science publications by referring to the MAS dataset. Table 4.3 shows the identified subdomains in detail.

Top-ranked Conferences: We identified the top 10 conferences for each subdomain in Table 4.3, by referring to the conference ranking information from MAS. As a result, 221 top-ranked conferences are found in total from the DBLP dataset without duplication between subdomains. It should be noted that some conferences belong to multiple subdomains. For instance, *CIKM*[2] belongs to three subdomains such as "Data Mining", "Information Retrieval", and "Databases", while *ICML*[3] corresponds to two subdomains such as "Artificial Intelligence" and "Machine Learning & Pattern Recognition". For simplicity, we assigned the most highly ranked subdomain to each conference venue such as CIKM to "Data Mining" and ICML to "Machine Learning & Pattern Recognition".

Degree distributions: Figure 4.2 shows the in- and out-degree distributions of publications, which are separated by all, conference, journal, and top-ranked conference papers. As Figure 4.2(b) shows, journal papers include more citations than conference papers due in part to the page limit policies of conference publications. However, as Figure 4.2(a) shows, conference papers are cited more often than journal papers, which can be interpreted that publications in conference venues are more active in computer science than other research domains, and the faster review pro-

[2] ACM Conference on Information and Knowledge Management

[3] International Conference on Machine Learning

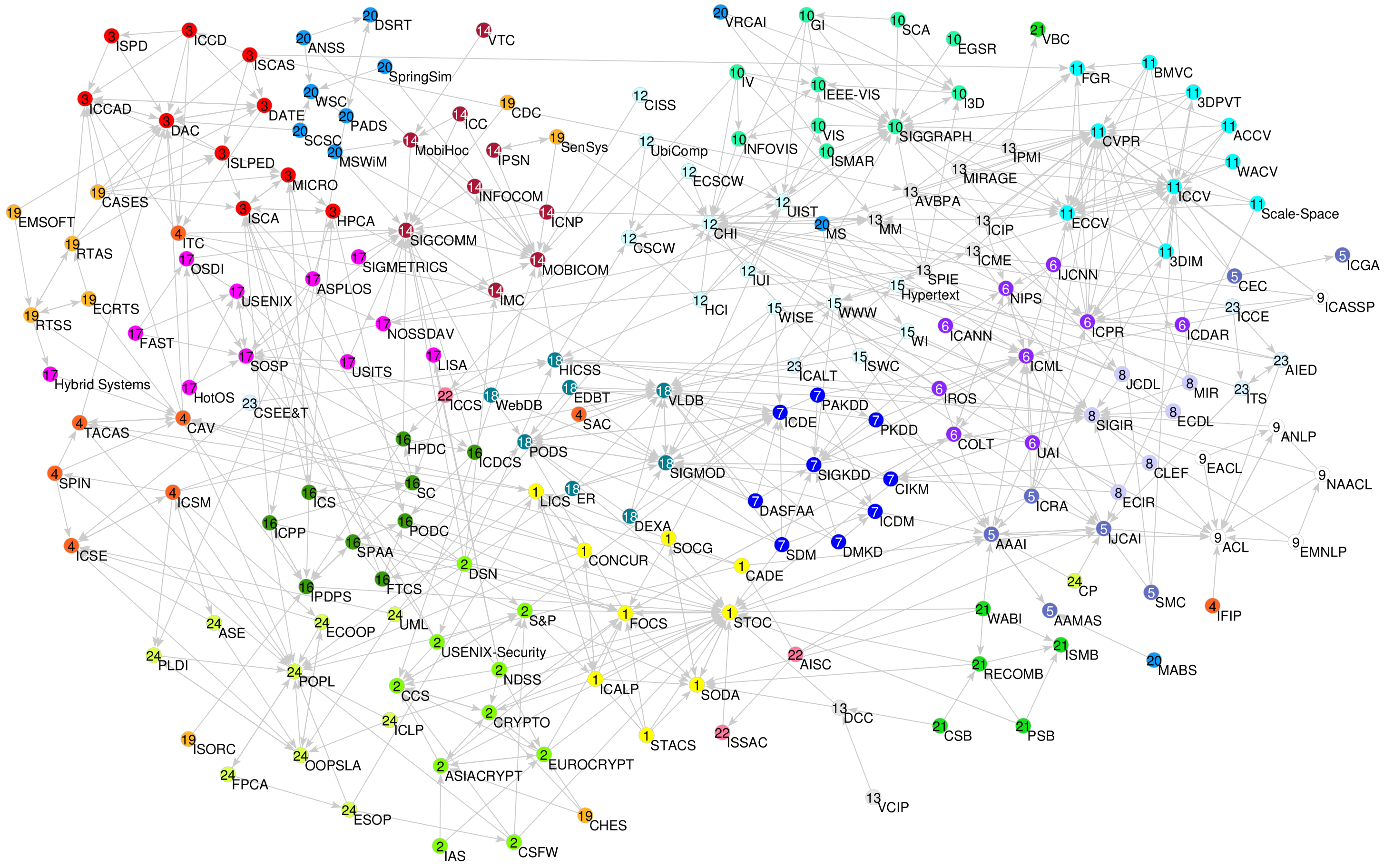
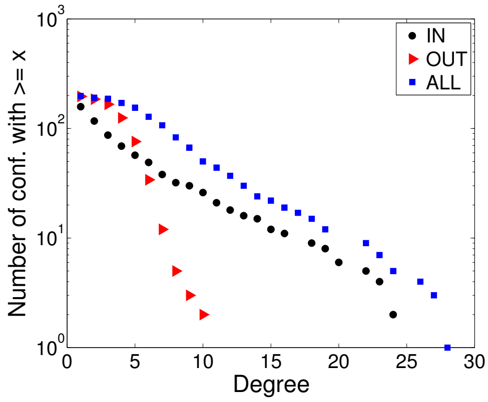
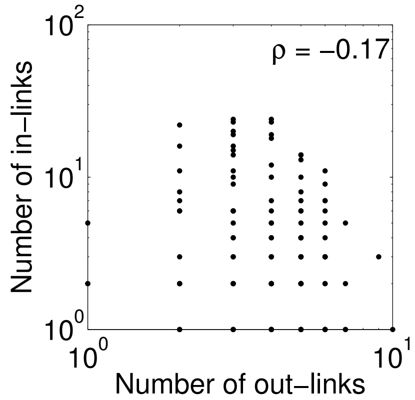


Figure 4.3: Largest connected premier conference network in computer science. This network is extracted from computer science citation networks. Each label beside a node represents an abbreviated conference name. Different subdomains are color-coded and labeled inside the nodes according to the associated subdomains. Numbers inside the nodes are matched with the abbreviated subdomain names in Table 4.3 as follows. 1:AT, 2:SP, 3:HW, 4:SE, 5:AI, 6:ML, 7:DM, 8:IR, 9:NLS, 10:Gr, 11:CV, 12:HCI, 13:MM, 14:NC, 15:WWW, 16:DP, 17:OS, 18:DB, 19:RT, 20:Sm, 21:Bio, 22:ScC, 23:Edu, 24:PL



(a)



(b)

Figure 4.4: **(a)** Degree distributions of the premier conference network in Figure 4.3 which are separated by in-, out-, and all-degrees. **(b)** Scatter plots of in- and out-degrees of the premier conference network (the top 10 conferences of each subdomain are selected). ρ denotes the correlation coefficient.

cesses of conferences lead to their publications to be exposed earlier and thus cited more often compared with journal articles. In addition, conference publications with high citations are from the top 10 conferences in each subdomain. For instance, 90% of popular papers, receiving more than 40 citations, came from the top-ranked conferences. That is, the premier conference publications more likely lead computer science research than others.

4.3 Computer Science Citation Networks

In the previous section, we constructed the citation networks using the DBLP dataset and also identified the nodes in the networks with research domain, subdomain, and top conference publications using the MAS dataset. Based on the identification of the nodes, the citation networks can be presented as conference and subdomain networks as illustrated in Figure 4.1. In this section, we focus on these two networks and analyzed their linkage patterns in detail.

4.3.1 Premier Conference Citation Networks

From the citation networks, we obtained a top-ranked conference network, as shown in Figure 4.3, by identifying the nodes with the top 10 conferences in each subdomain. This figure illustrates the largest connected component among 221 conferences, where each outgoing link of a conference is normalized by its total number of citations. Each outlink implies that more than 10% of publications from a source conference cite target conference publications. These connections help to find significant knowledge diffusion patterns between premier conferences. In this figure, the labels

Table 4.4: The most highly cited conferences from the top-ranked conferences in Figure 4.3. In-degree indicates the number of other conferences giving citations to a target conference. Conferences are listed in a descending order of in-degree.

Conference		Subdomain Name	In-degree
Short Name	**Full Name**		
STOC	ACM Symposium on Theory of Computing	Algorithms & Theory	23
CVPR	IEEE Conference on Computer Vision and Pattern Recognition	Computer Vision	23
VLDB	International Conference on Very large Data Bases	Databases	22
CHI	ACM Conference on Human Factors in Computing Systems	Human-Computer Interaction	22
SIGGRAPH	Special Interest Group on GRAPHics and Interactive Techniques	Graphics	21

beside the nodes in the network indicate the abbreviated conference names, while the labels inside the nodes denote the associated subdomains. From this conference network, it is observed that there are more frequent citations within a conference (self-loop) compared to citations between different conferences. In terms of subdomains, there are also stronger intra-connections within a subdomain than inter-connections between subdomains.

Degree distributions: Figure 4.4 shows the degree distributions of the conference network, separated by in-, out-, and all-degrees, and the scatter plot of in- and out-degrees of the network. As shown in Figure 4.4(a), in- and out-degree distributions exhibit very different tendencies. That is, the top-ranked conferences tend to cite less than 10 other conferences in the network, but they are cited by a larger number of other conferences. In other words, literature reviews are likely conducted by referring to publications from a small number of familiar or closely related conferences, but some new approaches or methodologies can be widely applicable in different areas and are thus cited by a larger number of conferences in diverse subareas. Table 4.4 shows the list of conferences receiving citations from the largest number of other conferences in the network. The listed conferences are cited by about 11% of conferences in the network on average. As the table shows, all these conferences belong to different subdomains, which implies that there is no single dominant subdomain in computer science and that different research subareas influence each other.

Figure 4.4(b) shows the scatter plot of in-degrees versus out-degrees of conferences in the premier conference network. As the figure shows there are no strong correlations between in-degree and out-degree sizes, which means that a conference, which is heavily cited, does not necessarily cite a lot of other conferences, and vice

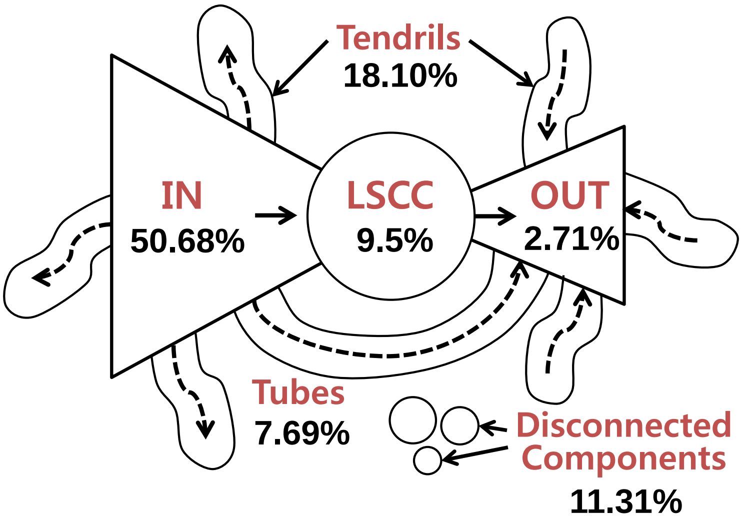


Figure 4.5: Reachability of the premier conference network in Figure 4.3 presented as a skewed bow-tie, obtained by the bow-tie model [40].

versa. This is also observed from news diffusion in social media in Chapter 3.

Table 4.5: Distributions of computer science subdomains for each bow-tie component in Figure 4.5. Each parenthesis after a subdomain indicates its proportion occupied in the corresponding bow-tie component.

Bow-tie	**Subdomain Name (%)**
LSCC	Algorithms & Theory (19%), Data Mining (19%), Artificial Intelligence (14%), Databases (14%), Machine Learning & Pattern Recognition (14%), Others (19%)
IN	Security & Privacy (9%), Graphics (9%), Natural Language & Speech (8%), Software Engineering (6%), Real-Time & Embedded Systems (6%), Hardware & Architecture (4%), Information Retrieval (4%), Multimedia (4%), Simulation (4%), Computer Education (4%), Data Mining (4%), Databases (4%), Algorithms & Theory (4%), Others (26%)
OUT	Graphics (50%), Human-Computer Interaction (50%)
TUBES	Computer Vision (35%), Hardware & Architecture (29%), Others (35%)
TENDRILS	Networks & Communications (23%), Operating Systems (18%), Simulation (13%), Graphics (10%), Others (38%)
DISC	Scientific Computing (28%), Computer Education (20%), Bioinformatics & Computational Biology (16%), Others (36%)

Reachability by the Bow-tie Model: We now view the premier conference network from the aspect of reachability, by using the bow-tie model [40], as we did in Chapter 3. Accordingly, Figure 4.5 shows the bow-tie structure of the conference

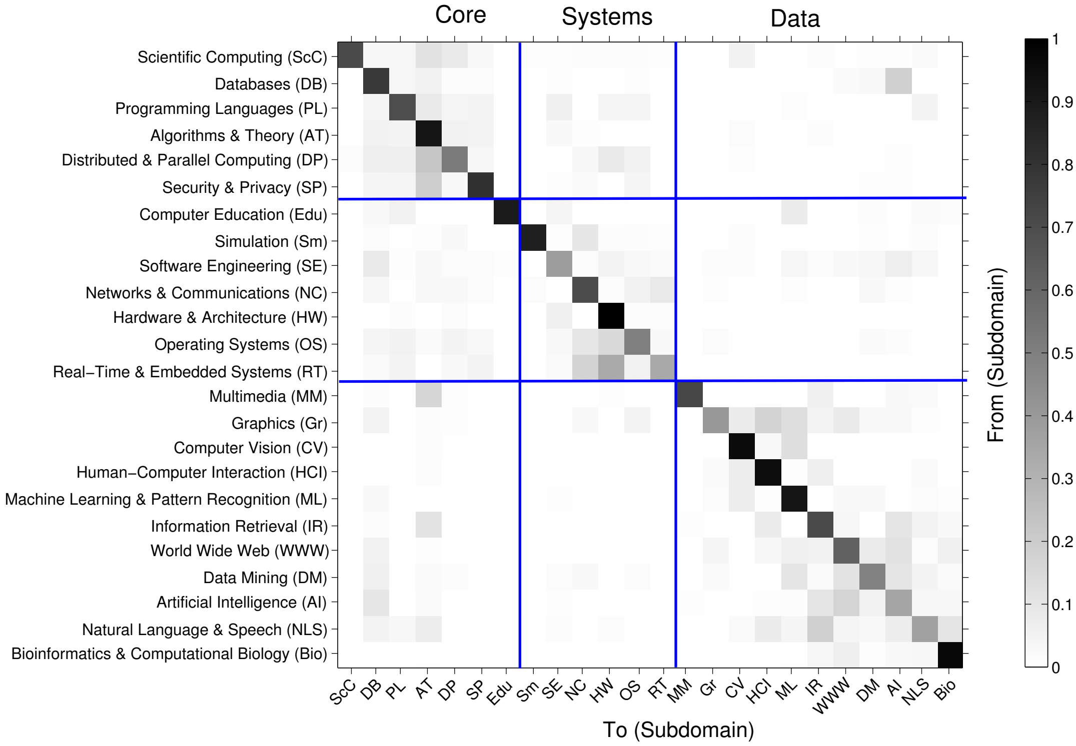


Figure 4.6: Normalized linkage patterns between subdomains in computer science. The x-axis and y-axis represent the abbreviated and full subdomain names respectively, as shown in Table 4.3. Subdomains are grouped by clusters (Core, Systems, and Data) in Table 4.6.

network. As this figure illustrates, the majority (50.68%) of conferences cite a small proportion (9.5%) of other conferences, showing a skewed bow-tie. Regarding this, Table 4.5 shows the distributions of subdomains for each bow-tie component. In this table, major subdomains, occupying the majority of each bow-tie component, are only presented for clarity. The small LSCC component mainly consists of five popular subdomains on theoretical and analytical aspects in computer science, which means that these subdomains can reach every other subdomain in the component, and thus they are more likely exposed to each other. On the other hand, the IN component shows the highest variability of subdomains, compared with other bow-tie components, with no dominant subdomain, while the OUT component exhibits the lowest variability consisting of evenly distributed two subdomains.

This structure is similar to the reachability of the user network in social media in Chapter 3. That is, the large proportion of subdomains is influenced by different subdomains, and they are far-reaching beyond a single domain. This skewed but far-reaching connectivity across different subdomains also implies the interactions between different social systems and the heterogeneity of the underlying network. This unbalanced reachability possibly brings about the directionality of influence

Table 4.6: Clustered 24 subdomains in computer science with the normalized cut [155] on the constructed citation networks from the DBLP dataset.

Label	**Subdomain Name**
Core	Algorithms & Theory, Security & Privacy, Distributed & Parallel Computing, Databases, Scientific Computing, Programming Languages
Data	Artificial Intelligence, Machine Learning & Pattern Recognition, Data Mining, Information Retrieval, Natural Language & Speech, Graphics, Computer Vision, Human-Computer Interaction, Multimedia, World Wide Web, Bioinformatics & Computational Biology
Systems	Hardware & Architecture, Software Engineering, Networks & Communications, Operating Systems, Real-Time & Embedded Systems, Simulation, Computer Education

with different strength in diffusion across heterogeneous social systems, which will be discussed in the next chapter in detail.

4.3.2 Subdomain Linkage Patterns

Based on the identified publications with research categories (domains and subdomains), we constructed computer science citation networks. From the citation networks, the top-ranked conference network was extracted, and its main characteristics were examined in the previous subsection. We now focus on subdomain linkage patterns by extracting subdomain networks from the citation networks.

Figure 4.6 presents the normalized linkage patterns between subdomains, where each cell (i, j) in the matrix is defined as the number of citations from subdomain i to subdomain j, normalized by the total number of citations of subdomain i. The matrix is expressed in a gray-scale map where black denotes the maximum value 1, and white corresponds to the minimum value 0. As the figure shows, in general most of citations tend to be made within same subdomains, while some citations are from multiple subdomains. In particular, "Algorithms & Theory", "Data Mining", "Distributed & Parallel Computing", "Operating Systems", and "Databases" tend to receive more citations from different subdomains than the others. These subdomains more likely lead research topics to be shared and spread across different subdomains in computer science compared with other subdomains.

Based on the subdomain linkage patterns in Figure 4.6, we conducted clustering using the normalized cut [155], which minimizes the inter-connections between partitions and maximizes the intra-connections within partitions, on the subdomain network. In this way, we can examine closely connected subdomains, and they are shown in Table 4.6. As the table shows, the first cluster consists of subdomains which have been at the core of computer science, the subdomains in the second cluster seem to be closely related to approaches on data processing, and the last cluster contains system-related subdomains in computer science. Accordingly, we label them as *Core*, *Data*, and *Systems* respectively for the rest of the thesis, which are revisited in Chapter 6, 8, and 9 in order to examine knowledge diffusion across the three clusters and obtain a macroscopic view of diffusion dynamics in computer science.

4.4 Research Keywords across Subdomains

As discussed earlier in Chapter 2, targeting trending topics is important in order to obtain a global view of diffusion across heterogeneous social systems. In this regard, we investigated the top 100 research keywords in computer science, provided by MAS [5]. From the list, we disregard vague, unspecific, and redundant keywords such as “spectrum”, “source code”, “point of view”, “case study”, and “empirical study”, and we selected 80 keywords as target research topics. Keyword variations are also considered, as described in Table 4.2, for identifying publications containing one of variations of each keyword in their paper titles or abstracts. If no keyword is found from the title or abstract of each paper in the DBLP dataset, then paper profiles (e.g., title, year, and conference venue) are compared between the DBLP and MAS datasets in order to fill in empty keyword information by referring to a matched paper profile in the MAS dataset. As a result, we identified about 13 thousand publications from 221 premier conferences with the selected 80 keywords. We finally chose 20 keywords, each of which is contained in more than 100 publications in our dataset. For a comparison between popular and less popular keywords, we chose another 10 keywords which are ranked below the top 100 in computer science. Each of these keywords also covers more than 100 publications in the dataset as 20 popular keywords for a more fair comparison.

Local vs. Wide Popularity: Table 4.7 shows these 30 keywords in the order of popularity from the aspects of both subdomains and conferences. Each subdomain listed in the table contains more than 10% of publications related to the corresponding keyword, along with the detailed percentage of publications in a parenthesis. Also, each tick box indicates that more than 10% of premier conferences in each cluster (Core, Data, or Systems in Table 4.6) are related to the corresponding keyword. As the table shows, some research keywords are predominantly popular in a specific subdomain such as “System On Chip” in Hardware & Architecture, “Digital Library” in Information Retrieval, and “Machine Translation” in Natural Language & Speech, which correspond to the least popular keywords in Table 4.8. In contrast to these keywords, “Real Time”, “Satisfiability”, and “Large Scale” are popular across diverse research subdomains, with a smaller percentage of publications (less dominant), compared to other keywords. In addition, the large publication size on a specific research topic does not necessarily mean its wide popularity across different research areas. It can be limited to a research community, a conference venue, or a subdomain, showing local popularity. For instance, the keyword “Multi Agent System” is popular in the subdomain of Artificial Intelligence and more specifically in the conference of AAMAS[4].

[4]International Conference on Autonomous Agents and Multiagent Systems

Table 4.7: Selected research keywords by referring to MAS. Keywords are ordered by their popularity (the first 20 keywords are placed within the top 100, and the others are ranked below the top 100). Each subdomain listed accounts for more than 10% of publications on the corresponding keyword (the detailed percentage is shown in a parenthesis). Finally, C, D, and S denote the subdomain clusters, Core, Data, and Systems, respectively in Table 4.6, and each keyword is checked with a tick mark only when more than 10% of premier conferences in each cluster published papers related to that keyword. Keyword rankings ("Rank" in the table) are framed only when C, D, and S are all ticked, implying that the framed keywords are popular across all the clusters. This table continues to Table 4.8 on the next page.

Rank	**Keyword**	**Subdomain Name** ($\geq$ 10% of publications)	**C**	**D**	**S**
$\boxed{1}$	Real Time	Graphics (20%), Real-Time & Embedded Systems (14%)	√	√	√
$\boxed{2}$	Satisfiability	Algorithms & Theory (21%), Hardware & Architecture (18%), Artificial Intelligence (13%), Databases (11%)	√	√	√
$\boxed{3}$	Sensor Network	Networks & Communications (48%), Real-Time & Embedded Systems (18%)	√	√	√
$\boxed{4}$	Large Scale	Simulation (14%), Distributed & Parallel Computing (11%), Data Mining (10%)	√	√	√
5	Wireless Network	Networks & Communications (64%), Simulation (14%)	√		√
$\boxed{6}$	Web Service	World Wide Web (44%), Software Engineering (17%), Databases (10%)	√	√	√
$\boxed{7}$	Genetic Algorithm	Artificial Intelligence (37%), Software Engineering (35%)	√	√	√
8	Support Vector Machine	Data Mining (29%), Machine Learning & Pattern Recog. (26%), Information Retrieval (11%)		√	
$\boxed{9}$	High Performance	Hardware & Architecture (55%), Distributed & Parallel Computing (21%)	√	√	√
$\boxed{10}$	Semantic Web	World Wide Web (59%), Software Engineering (11%), Artificial Intelligence (10%)	√	√	√
$\boxed{11}$	Fault Tolerant	Distributed & Parallel Computing (41%), Hardware & Architecture (16%)	√	√	√
$\boxed{12}$	Distributed System	Distributed & Parallel Computing (36%)	√	√	√
$\boxed{13}$	Approximate Algorithm	Algorithms & Theory (75%)	√	√	√

Table 4.8: Continued from Table 4.7.

Rank	Keyword	Subdomain Name (≥ 10% of publications)	C	D	S
14	Feature Selection	Data Mining (36%), Machine Learning & Pattern Recog. (20%), Artificial Intelligence (10%)	√	√	
$\boxed{15}$	Social Network	Data Mining (26%), Human-Computer Interaction (19%), World Wide Web (18%)	√	√	√
$\boxed{16}$	Search Engine	World Wide Web (25%), Information Retrieval (25%), Data Mining (19%), Databases (10%)	√	√	√
$\boxed{17}$	Learning Algorithm	Machine Learning & Pattern Recog. (35%), Artificial Intelligence (15%), Data Mining (15%)	√	√	√
18	Cluster Algorithm	Data Mining (46%)	√	√	
19	Multi Agent System	Artificial Intelligence (70%), Software Engineering (11%), Simulation (10%)			
$\boxed{20}$	Genetics	Artificial Intelligence (38%), Software Engineering (25%), Hardware & Architecture (10%)	√	√	√
21	Reinforcement Learning	Machine Learning & Pattern Recog. (46%), Artificial Intelligence (42%)		√	
22	Embedded System	Hardware & Architecture (63%), Real-Time & Embedded Systems (23%)	√		√
23	Virtual Environment	Human-Computer Interaction (44%), Graphics (28%)		√	√
24	Virtual Reality	Human-Computer Interaction (45%), Graphics (38%)		√	
25	Recommender System	Human-Computer Interaction (24%), Data Mining (16%), Information Retrieval (14%), World Wide Web (14%)	√	√	
26	Image Retrieval	Information Retrieval (33%), Multimedia (32%)		√	
$\boxed{27}$	Information Flow	Programming Languages (31%), Security & Privacy (31%)	√	√	√
28	System On Chip	Hardware & Architecture (91%)			√
29	Digital Library	Information Retrieval (80%)		√	
30	Machine Translation	Natural Language & Speech (89%)		√	

A framed keyword ID in the table indicates its wide popularity across all the clusters, Core, Data, and Systems. The ratios of the framed keywords for the 20 popular and 10 less popular keywords are 75% and 10%, respectively. This means that popular research keywords much more likely spread across a wider range of research areas than less popular keywords. Thus, the popularity of a research topic needs to be considered with its coverage of research areas, not only with its publication size. However, as we mentioned in the previous chapter, this structural connectivity does not exhibit temporal dynamics of diffusion, which are further examined in the next two thesis parts, by focusing on the widely popular 15 research keywords, i.e. framed keywords in Table 4.7 and 4.8.

4.5 Diffusion in Social Media vs Academic Publications

For analyzing real-world diffusion phenomena, we focused on explicit citation relationships in the two different application domains, social media and academic publications, by extracting hyperlinks in the main content of a web document and bibliography in an academic publication. From these citation networks, we can obtain different levels of linkage patterns, such as document, authorship, and information networks, and they provide different implications on structural diffusion patterns.

There are several empirical findings in common across these two domains. First, we observed that underlying networks exhibit strong intra-connections within a same information category (i.e., news categories in social media and subdomains in academic publications), as well as relatively weak but far-reaching inter-connections between different information categories. Second, there are no correlations between in- and out-degree sizes of web posting or scientific publications, showing that linkage patterns are non-trivial. Finally, the underlying networks in both domains exhibit skewed but far-reaching connectivity from the aspect of the bow-tie model. That is, the majority of individuals in the network obtain new information from a small proportion of other individuals belonging to different social systems, but they are globally inter-connected.

These observations can be interpreted that there are interactions between different social systems, and the underlying networks are heterogeneous. This suggests an existence of dynamic influence with different strength across heterogeneous populations, which motivates to propose our approaches in Chapter 5 and 7. More detailed motivational background will be discussed in the next chapter in detail.

4.6 Conclusion

In this chapter, we have analyzed knowledge diffusion patterns in computer science publications by using the DBLP and MAS datasets. More specifically, the DBLP dataset is used for constructing citation networks and the MAS dataset is used for obtaining meta-information of academic publications in the DBLP datset. From the generated computer science citation networks, we observed that conference papers are cited more often than journal papers and that 90% of popular papers, receiving more than 40 citations, are from the top-ranked conference papers. That is, the premier conference publications lead computer science research.

From the citation networks, we again extracted conference and subdomain networks based on the conceptual structure of academic citation networks. From the extracted premier conference network, we observed that conferences are more likely cited than they cite others, but there is no correlation between in- and out-degrees of the network. This shows that a publication tends to cite previous publications from less than 10 conferences interested, but it is more likely cited by a larger number of conferences. We also found that the conference network exhibits a skewed bow-tie structure, showing the similar pattern to the user network structure in social media in Chapter 3. Such unbalanced but far-reaching connectivity of the underlying networks supports the importance of studying dynamics of influence across heterogeneous populations in the next chapter. Regarding the subdomain linkage patterns, different subdomains are not only interconnected but also clustered along their common research interests. Some of them are cited by a larger number of different subdomains than the others, such as "Algorithms & Theory", "Data Mining", "Distributed & Parallel Computing", "Operating Systems", and "Databases", which can help research topics to be shared and spread across different subareas in computer science. By clustering the subdomain network, we obtained the three clusters which are concerned with basis (Core), data processing (Data), and system architecture and managements (Systems) in computer science. We also investigated 30 research keywords; 20 keywords are placed within the top 100 in computer science, and the others are ranked below the top 100. They exhibit different diffusion patterns; popular keywords are used by a wider range of subdomains covering all the clusters, i.e. Core, Data, and Systems, than less popular keywords. We also found that he popularity of a research topic needs to be considered with its coverage of research areas, not only with its publication size.

In this chapter, we observed structural diffusion patterns in academic publications. However, they do not provide temporal dynamics of diffusion. Rather, unbalanced but far-reaching connectivity across different subdomains, which are also observed in social media in Chapter 3, can support the fact that there possibly exists the directionality of influence with different strength in diffusion across heterogeneous social systems. Based on the empirical analysis in Chapter 3 and 4, temporal dynamics of diffusion will be studied in depth in the next two parts of this thesis.

Part II

Model-Driven Approach

Chapter 5

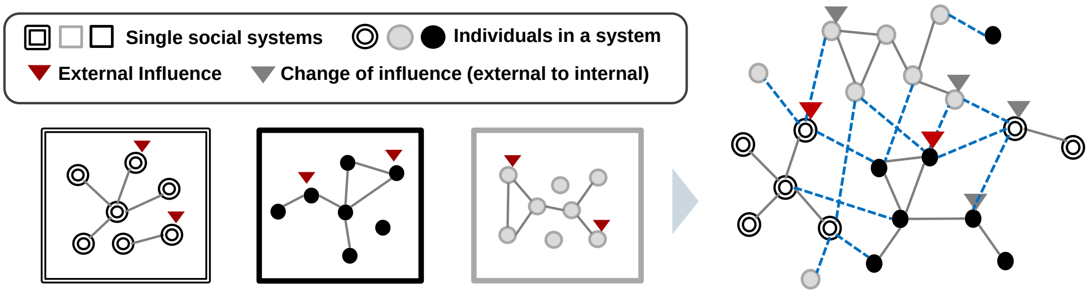
Dynamic Influence Model

In Part I, we analyzed real-world diffusion patterns in two different application domains, social media and academic publications, and obtained a macroscopic picture of structural connectivity of underlying networks. Based on our empirical analyses, in this chapter we first propose a new conceptual framework for diffusion across heterogeneous social networks. Accordingly, we define a macro-level diffusion model, called *Dynamic Influence Model*, in a probabilistic way by incorporating two main features of underlying networks, i.e. heterogeneity and structural connectivity. With this model-driven approach, we will further estimate temporal dynamics of real-world diffusion in the investigated application domains in the next chapter. We expect that the proposed model provides a way of interpreting the dynamics of diffusion in terms of the strength and directionality of influence across target systems and that it applies to a wide class of diffusion scenarios in diverse research areas.

5.1 Introduction

Information spreads beyond a single social system into multiple systems as diverse information sources are increasingly accessible than ever before. From our empirical studies in Part I, we observed that such far-reaching information pathways are formed in social media as well as in academic publications. For instance, in Chapter 3 noteworthy news topics spread across different types of social media such as News, SNS, and Blog, and in Chapter 4 popular research topics draw attention from a wide range of research subdomains in computer science. That is, the underlying networks span across multiple social systems and thus necessarily consist of heterogeneous populations. We call these underlying networks as *heterogeneous social networks* in this thesis, as discussed in Chapter 2. Thus, a single social system alone is not sufficient to understand the mechanisms of global diffusion across heterogeneous social networks, since the effects of interactions between different networks on diffusion is not negligible.

A diffusion process of information is commonly viewed to consist of two distinct phases [31, 67, 95, 103, 107, 123]: (1) the emergence of information by *external influence*, such as mass media (e.g., the triangles in Figure 5.1(a)); and (2) the cascading spread of the information through *internal influence*, such as interpersonal



(a) Separated social systems (b) Dynamic influence framework

Figure 5.1: Conceptual framework for diffusion. **(a)** Traditional diffusion framework: external and internal influences on homogeneous social networks within a single social system. **(b)** Dynamic influence framework: external and internal influences on heterogeneous social networks across multiple social systems.

communications between connected neighbors in a contact network (e.g., the edges in Figure 5.1(a)). Previous studies have mostly focused on diffusion within a single social system alone, as discussed in Section 2.2.2. However, for a better understanding of diffusion mechanisms, it is necessary to consider the effects of interactions between different social networks on diffusion as we observed far-reaching information pathways from our target domains. Such observations suggest more careful considerations on distinguishing influence from different social networks along with its strength and thus defining diffusion processes in a more flexible space. In this regard, this study aims to uncover diffusion mechanisms across heterogeneous social networks in terms of the strength and directionality of influence.

This study requires to resolve the following main challenges. First, external and internal influences need to be redefined in heterogeneous social networks as we extend a diffusion space from a single to multiple social systems. Second, the structure of the underlying networks is hard to collect and define. That is because the underlying networks span multiple social systems, leading to limited data collection, and dynamically changing local interactions make it difficult to define the structures only from a sampled snapshot of the current networks. Third, global diffusion necessarily includes a large collection of different populations, which makes it difficult to estimate and interpret diffusion dynamics among them, and thus it requires us to consider meta-population schemes [29] for obtaining a simple and thus clear overall picture of diffusion. The way of classifying heterogeneous social networks determines the definition of meta-populations, and it provides a different view of diffusion dynamics among meta-populations.

To address these challenges, we first propose a new conceptual framework for diffusion across heterogeneous social networks, called *Dynamic Influence*. Based on this conceptual design, we model macro-level diffusion with a probabilistic approach that incorporates the heterogeneity and structural connectivity of underlying networks into the simple and robust mass-action Bass Model [30, 31]. This proposed model-driven approach will be evaluated and used in Chapter 6 to estimate macro-level diffusion in the investigated application domains from Part I.

This proposed approach can improve the accuracy of diffusion models dealing with a single social system alone, since it does not neglect the effects of heterogeneous interactions on diffusion. In addition, it provides a new way of interpreting macro-level diffusion in terms of the strength and directionality of the influence across heterogeneous social networks. To the best of our knowledge, such a macro-level diffusion model incorporating both heterogeneity and connectivity of underlying networks has not been studied in previous research.

The remainder of this chapter is structured as follows. In the next section, we propose a new conceptual framework for diffusion across heterogeneous social networks with motivational research background. Accordingly, Sections 5.4 proposes our macro-level diffusion model, called Dynamic Influence Model, and finally Section 5.5 concludes this study.

5.2 Conceptual Diffusion Framework

Information diffusion across multiple social systems makes it challenging to discover its underlying mechanisms due to hidden structures of social networks and the diversity of populations in these networks. For a better understanding of global diffusion, in this section we propose a new conceptual framework for diffusion across heterogeneous populations, as shown in Figure 5.1.

Dichotomous View: Figure 5.1(a) illustrates separated social systems, each of which consists of individuals interacting with each other. A set of individuals in a system is called a homogeneous population, and a set of homogeneous populations across multiple social systems are called heterogeneous populations in this thesis, as discussed in Section 2.1.4. From the aspect of a single social system, the world is divided into inside and outside of the system, and thus it does not distinguish the identities of other homogeneous populations beyond the system. Thus, the influences from different populations are all considered as external influence, while the influences from individuals inside a system are regarded as internal influence.

External influence on a social platform such as Twitter has recently been quantified as exogenous out-of-the-network effects [123]. Interestingly, it was shown that 30% of diffusion regarding some trending topics in Twitter is attributed to external influence. In the marketing literature [110], this is ten-times larger than the typical value of external influence (0.03), and it is rather similar to the average value of internal influence (0.38). Such a large proportion of out-of-the-network effects supports the fact that the influence outside of a single social system is not ignorable, since exposure to new information from external sources is highly increasing, not depending on exposure from local contact networks any more. Thus, it is necessary to consider the heterogeneity of populations and the effects of their interactions on diffusion, which helps to obtain a macroscopic view of diffusion beyond a dichotomous view.

Table 5.1: Table of symbols for Chapter 5.

Symbol	Description
$h(t)$	Hazard function of time t
$a(t)$	The number of new adopters at time t
$A(t)$	The number of cumulative adopters until time t
$f(t)$	The proportion of new adopters at time t
$F(t)$	The proportion of cumulative adopters until time t
p	Coefficient of innovation in the Bass Model
q	Coefficient of imitation in the Bass Model
a	A binary random variable for the event of an individual's adoption
$P(a\|t)$	Adoption probability that an average individual adopts at time t
$P(a\|\neg a, t)$	Probability that an average individual, who has not adopted before, adopts at time t
$P_{\text{ext}}(a\|\neg a, i, t)$	Probability that an average individual from a meta-population i, who has not adopted before, adopts at time t by external influence
p_i	Probability of external influence on a meta-population i, implying $P_{\text{ext}}(a\|\neg a, i, t)$ in short
$P_{\text{int}}(a\|\neg a, i, t)$	Probability that an average individual from a meta-population i, who has not adopted before, adopts at time t by internal influence
m	The number of meta-populations
n	Total size of m meta-populations
k	The number of connected neighbors
v_i	The number of neighbors from a meta-population i who have already adopted
$\mathbf{v}$	A vector consisting of v_i for all m different meta-populations as $\mathbf{v} = (v_1, ..., v_m)^{\mathrm{T}}$
c_{ji}	Probability that an individual of a meta-population i adopts when it is influenced by a connected neighbor who is a previous adopter of a meta-population j
α	Power law coefficient
n_i	The size of a meta-population i

Dynamic Influence: In this regard, we define a new conceptual framework, *Dynamic Influence*, for diffusion across heterogeneous social networks. In this diffusion framework, heterogeneous populations directly interact with each other as if they were in the same networks, as shown in Figure 5.1(b). By collapsing the diffusion boundaries of single social systems, homogeneous social networks in each system are not separated but connected into other networks across multiple social systems. Some external influence on original social systems (red triangles in Figure 5.1(a)) is redefined as internal influence (gray triangles in Figure 5.1(b)) between different social networks by their hidden interactions (dashed lines in Figure 5.1(a)), and the remaining external influence (red triangles in Figure 5.1(b)) is from outside of the heterogeneous social networks.

This framework interprets influence between different social networks as direct and simultaneous effects on diffusion. That is, traditional internal influence spans heterogeneous populations, which makes a diffusion framework more flexible and general one.

5.3 Fundamental Framework: Bass Model

In this section, we introduce the basic concept of the Bass Model which is the fundamental framework of our proposed Dynamic Influence Model in the next section.

5.3.1 Background

In the early 1960s, adopting behaviors of a population were defined in social science in terms of the timing of adoption such as innovators, early adopters, early majority, late majority, and laggards [146]. Frank Bass mathematically represented these time-series adopting behaviors as a conditional likelihood of adoption in the Bass Model [30]. It has provided realistic and robust estimation of new product growth patterns, and thus it has been an influential diffusion model used in the marketing literature. As discussed earlier in Chapter 2, the simplicity of the model has enabled intuitive interpretation and has brought a wide range of extensions in diverse research areas [31], such as marketing, computer science, economics and operations research [95, 107, 110, 130, 188].

Its fundamental assumption is that a population is homogeneous and fully connected in the same way as traditional epidemic models [23, 127]. In contrast to this unrealistic assumption, there have been mainly two branches of extensions focusing on either the heterogeneity of populations, or network structures [97, 107, 151].

However, one extension regarding the heterogeneity, such as multinational diffusion of a consumer product [95, 142], disregards the effect of network topologies on diffusion. The other extension on network structures, such as cluster density and reachability [97, 151], and degree distributions [107], has focused only on a homogeneous population. Our study aims at considering the effects of both the heterogeneity and structural connectivity of real-world networks on diffusion by extending the simple and robust mass-action Bass Model.

5.3.2 Bass Model

In the Bass Model [30], the diffusion rate is governed by a hazard function $h(t)$ which is the ratio of the number of new adopters to the number of potential adopters, given a population, at time t as:

$$h(t) = \frac{a(t)}{n - A(t)} = \frac{f(t)}{1 - F(t)} \ , \tag{5.1}$$

where n is the size of a population, and $A(t)$ and $a(t) = \mathrm{d}A(t)/\mathrm{d}t$ denote the number of cumulative adopters and new adopters at time t, respectively. Accordingly, we denote the proportion of the cumulative adopters by $F(t) = A(t)/n$ and the proportion of new adopters by $f(t) = \mathrm{d}F(t)/\mathrm{d}t = a(t)/n$.

As discussed in the previous subsection, the Bass Model assumes that a population is homogeneous and fully connected, and accordingly the hazard function is defined as a simple linear form of the proportion of the cumulative adopters [30, 31] as:

$$\frac{f(t)}{1 - F(t)} = p + qF(t) \ , \tag{5.2}$$

where the parameter p is called the *coefficient of innovation*, since it does not interact with the cumulative adopter proportion $F(t)$, and q is called the *coefficient of imitation*, because it represents the internal influence of previous adopters. Equation (5.2) has a closed form solution as:

$$F(t) = \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}} \ . \tag{5.3}$$

As Equation (5.2) shows, the Bass Model assumes a homogeneous and completely connected population. Our model in the next section incorporates the heterogeneity and structural connectivity of real-world networks into the mass-action Bass Model at a macro level. This helps to improve the accuracy of diffusion models dealing with either single social networks or heterogeneous, but unstructured populations.

5.4 Dynamic Influence Model

In this section, based on the conceptual diffusion framework in Section 5.2, we propose a macro-level diffusion model by extending the Bass Model with two main features (heterogeneity and structural connectivity) of underlying networks.

5.4.1 Problem Statement

We now formally describe the problem and extend the Bass Model in a probabilistic way. In reality, the structure of heterogeneous social networks is hardly obtainable (e.g., limited data collection due to the diversity of information sources) and even dynamic (e.g., changes of interaction patterns in different context of information, or lively changing relationships), which makes it challenging to define the structure from the sampled snapshot of current networks. Regardless, dynamic influences among different social networks are not ignorable. Thus, the goal is to infer the macro-level diffusion processes within and between different populations in a probabilistic way without the need to know detailed network structures, but with a real-world network property.

In more detail, given the number of time-series cumulative adopters $A_i(t)$ for a population i among m populations from time $t = 1$ to T as $\{A_i(t)\}_{t=1}^{T}$, $i = 1, ..., m$, we aim to infer the unobservable influence from a population j to a population i, c_{ji}, which denotes the probability that an individual from the population i adopts when it is exposed to its neighbor who is a previous adopter from the population j (c_{ij} is the other way around). Particularly, we assume that the degree distribution of an individual follows a power law, since a wide range of real-world networks exhibit power-law behaviors in their degree distributions [48, 127].

5.4.2 Model Formulation

For modeling diffusion across heterogeneous social networks, we begin by interpreting the Bass Model from a probabilistic point of view. The proportion of adopters in the Bass Model is in fact its expectation in the mean-field mass-action kinetics of the model, and thus it can be thought of as an adoption probability that an average individual adopts at time t:

$$F(t) = P(adopt \mid t) \ , \tag{5.4}$$

where *adopt* is a binary random variable for the event of an individual's adoption, and it will be abbreviated to "a" in the rest of the chapter for brevity. Similarly, we can view the hazard function as a new adoption probability, $P(a \mid \neg a, t)$:

$$\frac{f(t)}{1 - F(t)} = \frac{\partial_t P(a \mid t)}{1 - P(a \mid t)} = P(a \mid \neg a, t) \ , \tag{5.5}$$

where ∂_t denotes the partial derivative with respect to t and $\neg$ stands for the opposite. Therefore, $P(a \mid \neg a, t)$ indicates the probability that an average individual, who has not adopted before, adopts at time t.

By separating external and internal influences and applying the probability of the union of two independent events, i.e., $P(A \cup B) = P(A) + P(\neg A)P(B)$, we get:

$$\frac{\partial_t P(a|t)}{1 - P(a|t)} = P_{\text{ext}}(a|\neg a, t) + (1 - P_{\text{ext}}(a|\neg a, t))P_{\text{int}}(a|\neg a, t) \ , \tag{5.6}$$

where $P_{\text{ext}}(a|\neg a, t)$ and $P_{\text{int}}(a|\neg a, t)$ denote the new adoption probabilities by external and internal influences, respectively.

Heterogeneity: To deal with the heterogeneity of populations, we introduce a random variable, $i = 1, ..., m$, for different types of m populations. Note that a population can be a meta-population as discussed in Section 5.1, which depends on the definition of a population in a specific application domain.

Thus, a set of new adoption probabilities are constructed for each population $i = 1, ..., m$ as:

$$\frac{\partial_t P(a|i,t)}{1 - P(a|i,t)} = P_{\text{ext}}(a|\neg a, i, t) + (1 - P_{\text{ext}}(a|\neg a, i, t))P_{\text{int}}(a|\neg a, i, t) \ . \tag{5.7}$$

Like the coefficient of innovation in the Bass Model, we consider the new adoption probability of a population i by external influence as:

$$P_{\text{ext}}(a|\neg a, i, t) = p_i \ , \tag{5.8}$$

where $p_i \in [0, 1]$. That is, p_i is the probability that an arbitrary individual in a population i is infected by external influence, and $P_{\text{int}}(a|\neg a, i, t)$ in Equation (5.7) denotes the probability that the individual is infected internally by previous adopters who are neighbors in the heterogeneous networks consisting of m populations.

Structural Connectivity: Now, let us focus on the internal new adoption probability $P_{\text{int}}(a|\neg a, i, t)$ by considering the structural connectivity of underlying networks. Suppose that an individual of type i (belonging to a population i) has k neighbors in which $\mathbf{v} = (v_1, ..., v_m)^{\text{T}}$ neighbors of each individual type have already adopted. Then, from the sum and product rules, the internal new adoption probability is factorized by:

$$\begin{aligned} P_{\text{int}}(a|\neg a, i, t) &= \sum_{k=1}^{n-1} \sum_{\mathbf{v}} P(a, \mathbf{v}, k|\neg a, i, t) \\ &= \sum_{k=1}^{n-1} \sum_{\mathbf{v}} P(a|\mathbf{v}, k, \neg a, i, t) P(\mathbf{v}|k, \neg a, i, t) P(k|\neg a, i, t) \end{aligned} \tag{5.9}$$

where $n = \sum_{i=1}^{m} n_i$, and n_i is the size of a population i.

The distribution of an individual's exposures to previous adopters in its neighbors is modeled as a binomial distribution (refer to Section 2.1.2), which is consistent with prior diffusion models [107, 123]. Thus, each adoption is a Bernoulli trial, and the probability that an individual adopts after $\mathbf{v} = (v_1, ..., v_m)^{\text{T}}$ contacts is:

$$p(a|\mathbf{v}, k, \neg a, i, t) = 1 - \prod_{j=1}^{m} (1 - c_{ji})^{v_j} \ , \tag{5.10}$$

where $c_{ji} \in [0, 1]$ denotes the probability that an individual of type i adopts when it is exposed to a previous adopter of type j in its neighbors. Note that it is the probability that an individual is affected by at least one of its adopting neighbors, i.e. one minus the probability of the complementary event that it is not affected by any of the previous adopters in its neighbors. Comparison between our Bernoulli influence model and the linear influence model of [107] will be discussed later in Section 5.4.3.

From a macro point of view, the probability distribution of having $\mathbf{v}$ adopters in k neighbors is a multinomial distribution as:

$$p(\mathbf{v}|k, \neg a, i, t) = \frac{k!}{v_1! \cdots v_m!(k-v)!} \prod_{i=1}^{m} P(a|i,t)^{v_i} (1-P)^{k-v} \ , \tag{5.11}$$

where $v = \sum_{i=1}^{m} v_i$ and $P = \sum_{i=1}^{m} P(a|i,t)$.

Finally, we assume that the degree distribution of an individual follows a power law, since real-world networks are scale-free networks exhibiting power-law distributions [43, 48, 91, 98, 105, 127] as:

$$p(k|\neg a, i, t) = \frac{1}{\zeta(\alpha)} k^{-\alpha} \ , \tag{5.12}$$

where α is the power law coefficient, and $\zeta(\alpha) = \sum_{k=1}^{n-1} k^{-\alpha}$.

Substituting Equations (5.10) to (5.12) into Equation (5.9) gives the internal new adoption probability as:

$$P_{\text{int}}(a \mid \neg a, i, t) = 1 - \frac{1}{\zeta(\alpha)} \sum_{k=1}^{n-1} \frac{\left(1 - \sum_{j=1}^{m} c_{ji} P(a|j,t)\right)^k}{k^\alpha} \ , \tag{5.13}$$

Note that the neighboring adopters, $\mathbf{v}$, in Equation (5.9) are marginalized out in Equation (5.13) by the multinomial theorem (see Appendix B.1 for details). Therefore, our macro-level diffusion model does not require micro-level information, such as local structures of contact networks.

Again, by substituting Equations (5.8) and (5.13) into Equation (5.7), we obtain the system of partial derivative equations for the Dynamic Influence Model. It is not mathematically tractable, and thus, we need to solve it numerically to get the adoption probabilities $\{P(a \mid i,t)\}_{i=1}^{m}$.

5.4.3 Comparison of Influence Assumptions

Before finishing this section, it is worth comparing our Bernoulli influence model in Equation (5.10) with the linear influence model of [107]. The authors of [107] only considered diffusion in homogeneous networks, and thus, the corresponding linear influence model in heterogeneous networks would be:

$$p(a|\mathbf{v}, k, \neg a, i, t) = \sum_{j=1}^{m} c_{ji} v_j \ , \tag{5.14}$$

where $c_{ji} \geq 0$ is the influence coefficient that an adopter in the neighbors of type j affects an individual of type i. Note that the influence increases linearly with the number of previous adopters v_j in its neighbors.

Technically, it is not a probability distribution, because with fixed $\{c_{ji}\}_{j=1}^{m}$, it is possible that $p(a|\mathbf{v}, k, \neg a, i, t) > 1$, if an individual has many adopters in its neighbors.

Therefore, it is not an appropriate assumption for a probabilistic model. The benefit of the linear influence model is that it helps simplify the internal new adoption probability in Equation (5.9) as a linear form of the adoption probabilities:

$$P_{\text{int}}(a|\neg a, i, t) = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)} \sum_{j=1}^{m} c_{ji} P(a|j, t) \ . \tag{5.15}$$

However, Equation (5.15) is just the first-order Taylor approximation of Equation (5.13), which is explained in Appendix B.2. Thus, the linear influence model is a linear approximation of our Bernoulli influence model when there exist no previous adopters ($\forall i, P(a|i, t) = 0 \Rightarrow x = 1$). Therefore, our Bernoulli influence model is more sophisticated than the linear influence model, and because it is based on a probability distribution, it naturally guarantees the probability constraint, $0 \leq p(a|\mathbf{v}, k, \neg a, i, t) \leq 1$.

The Dynamic Influence Model will be evaluated and applied to our target application domains in Chapter 6.

5.5 Conclusions

In this chapter, we first introduced the new conceptual framework for diffusion across heterogeneous social networks, which extends the traditional diffusion mechanism to a more flexible and general one. Then, we defined a macro-level diffusion model, called the Dynamic Influence Model, by applying this concept to the simple mass-action Bass Model with a probabilistic approach. This is a generalization of the Bass Model into the dynamics of meta-populations in a probabilistic way by incorporating the two essential features, the heterogeneity and structural connectivity of underlying networks.

This generalization enables us (1) to distinguish influences between interconnected heterogeneous social networks from arbitrary external influence on homogeneous social networks, and thus improve the accuracy of a mass-action diffusion model or a model dealing with homogeneous networks, (2) to interpret diffusion dynamics across multiple social systems in terms of the strength and directionality of influence, and finally (3) to estimate macro-level diffusion trends without the needs of detailed network structures, providing benefits when data collection is limited or when a macroscopic picture is needed before targeting a specific social system or region interested for further investigations.

We expect that the proposed model applies to a wider class of diffusion phenomena in diverse areas, including the social sciences, marketing and neuroscience, for interpreting the dynamics of target systems or regions at a macro-level.

In the next chapter, we evaluate our proposed model with synthetic data, conduct experiments on real data from Chapter 3 and 4, and interpret estimated diffusion trends in social medial and academic publications.

Chapter 6

Estimating Real-World Diffusion with a Model-Driven Approach

In this chapter, we estimate real-world diffusion with our Dynamic Influence Model presented in Chapter 5 in terms of the strength and directionality of influence across heterogeneous social networks. We first evaluate our proposed model by using generated synthetic datasets as ground truths and compare the results with the Bass Model as a baseline. In order to show the feasibility of our model, two evaluation metrics are used: model fitting errors and parameter errors. Based on the verification of the model's parameter recovery, we conduct experiments on real data from our target application domains: social media in Chapter 3 and academic publications in Chapter 4. In Part I of this thesis, we analyzed structural diffusion patterns from the constructed citation networks, while in Part II we try to discover topic-related temporal dynamics of diffusion from time-series adopting behaviors in the same application domains. We expect that the way of applying our model to the two example real-world domains helps to interpret diffusion dynamics in other domains at a macro level.

6.1 Introduction

We often encounter consequential statistics brought by collective behavior of a population, such as sales growth of a new consumer product, increasing participants in fund-raising activities, nation-wide mass assemblies for political protests, and so on. From these time-evolving activity sequences, we can obtain growth patterns of adopters over time but without the underlying diffusion mechanisms. In the previous chapter, we extended the mass-action Bass Model into a model of diffusion over heterogeneous and connected social networks. This generalization enables us to infer diffusion processes in a probabilistic way only from cumulative adopters across heterogeneous social networks or meta-populations in more general.

For estimating real-world diffusion with our model-driven approach, we first evaluate the model performance using synthetic data and compare the results with the Bass Model as a baseline. We fit the models by minimizing the sum of squared errors in an iterative way until the error converges. As evaluation metrics, model

Table 6.1: Table of symbols for Chapter 6.

Symbol	Description
T	The total number of time steps
n_i	The size of a meta-population i
$\boldsymbol{\theta}_i$	Model parameters for the i-th equation except for n_i
$\boldsymbol{\Theta}_i$	Model parameters for the i-th equation including n_i, *i.e.* $\boldsymbol{\Theta}_i = \{n_i, \boldsymbol{\theta}_i\}$
$\tilde{e}_k$	Relative parameter error of the k-th parameter
$\hat{\omega}_k$	Estimated value of the k-th parameter
ω_k	Ground truth of the k-th parameter

fitting errors and parameter errors are used as in prior work [95, 107, 123]. Based on the verification of parameter recovery with generated synthetic datasets, we conduct experiments on real data from our target application domains in Part I and interpret diffusion dynamics with the estimated parameter values with real datasets.

From the experiments on social media data from Chapter 3, we find that influences between different media types vary with the context of information, which leads to different diffusion patterns. For instance, News is the most influential in the arts and economy categories, while SNS and Blog are in the politics and culture categories, respectively. Controversial topics, such as political protests in the Middle East and multiculturalism failure, tend to drive concurrent and synchronous diffusion across all social media types, while entertainment topics, such as film releases and celebrities, exhibit stronger internal diffusion within a single social system (homogeneous social networks), i.e. stronger intra-relationships than inter-relationships. Regarding the experiments on academic publication data as presented in Chapter 4, we observed balanced inter-relationships among three meta-populations, Core, Data, and Systems, in the research keywords "Real Time" and "Large Scale", and they exhibit more synchronous and simultaneous cumulative adoption rates. In addition, more balanced intra- and inter-relationships are observed in the most popular keywords, such as "Real Time", "Satisfiability", "Sensor Network", and "Large Scale".

In common, balanced influences among heterogeneous populations more likely lead to synchronous and simultaneous diffusion across them. In other words, the diffusion patterns in a social system can reflect the balance or imbalance of influences among different social systems. Such macro-level observations, to the best of our knowledge, are seen for the first time. We expect that these exemplary applications of our model-driven approach can provide a way of interpreting real-word diffusion from macroscopic perspectives.

The rest of this chapter is organized as follows. Section 6.2 shows the feasibility of the Dynamic Influence Model by conducting experiments on synthetic data. Based on the evaluation of our model, we continue to experiment with the real data from Part I. Accordingly, Section 6.3 and 6.4 estimate and interpret macro-level diffusion

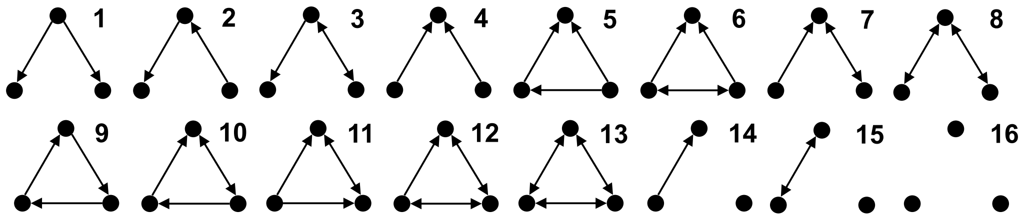


Figure 6.1: Unique structures of influence flows among three meta-populations. Each graph contains self-loops, representing intra-influence within a meta-population, which are omitted for brevity. Empty links between two different nodes indicate very weak connections compared with nonempty links, but they are not ignorable for a better understanding and a more accurate estimation of diffusion across heterogeneous social networks. Thus, there is a directed connection between every pair of nodes in each graph, but with different strengths. Our synthetic datasets are generated based on these sixteen unique structures.

in social media and academic publications, respectively. Section 6.5 discusses our model-driven approach and experimental results, and finally Section 6.6 concludes this study.

6.2 Evaluation of Model Performance

The goal of this section is to recover the hidden diffusion processes from generated synthetic datasets. For testing model performance, model fitting errors and parameter errors are evaluated in the experiments. The former describes how closely our model predicts the cumulative number of adopters for each meta-population, while the latter shows how correctly our model infers the ground truth parameters.

6.2.1 Synthetic Data Generation

As discussed in Chapter 5, the effects of interactions between different social networks on diffusion are not ignorable, and thus we can think of all possible directions of influence flows between meta-populations. When it comes to news diffusion across different types of social media, such as News, SNS and Blog as investigated in Chapter 3, or knowledge diffusion across different clusters of subdomains, such as Core, Data, and Systems in Chapter 4, we can build a 3×3 asymmetric adjacency matrix for presenting influence flows among three meta-populations, where each cell indicates the presence or absence of influence between two meta-populations. Thus, there are 2^9 possible cases of relational structures among three meta-populations in total.

If we also vary the strength of influence, then the number of potential cases becomes intractable. For efficient and meaningful simulation, it is important to generate synthetic datasets reflecting representatives among such numerous possible cases. For avoiding redundant cases, we first consider unique structures of the relations,

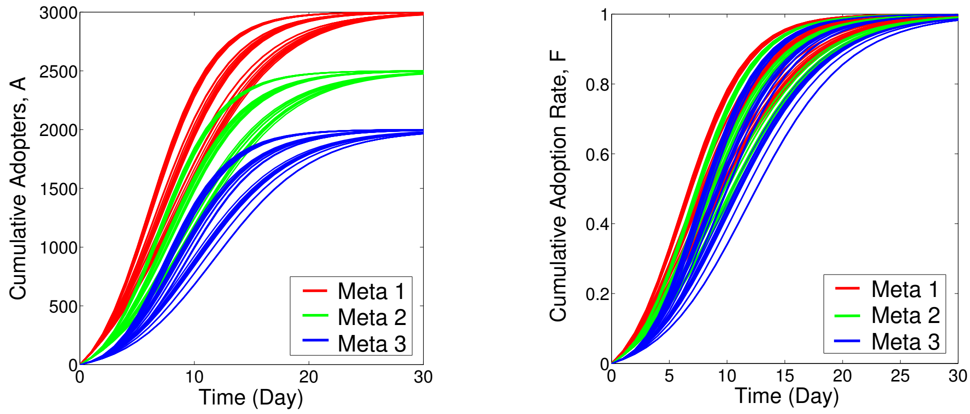
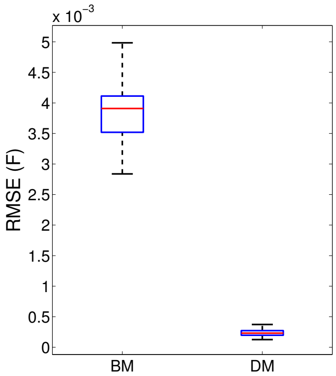


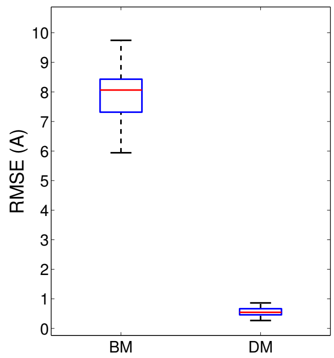
Figure 6.2: Synthetic data generation reflecting dynamic influences among three meta-populations (i.e. Meta 1, Meta 2, and Meta 3). Forty-eight synthetic datasets are generated in total, and the different population sizes reflect real-world situations, such as different number of adopters in News, SNS and Blog. The generated datasets are illustrated with time-series cumulative adopters, A (left) and the proportion of the corresponding cumulative adopters, F (right).

which leaves us with 16 dynamic relations, as shown in Figure 6.1. The dynamic structures include 13 triads (1–13) and additional three disconnected graphs (14–16). In this figure, each graph has three self-loops, indicating intra-relationships within meta-populations, but they are all omitted for brevity. We assume that there always exists influence between two populations, but with different strength. Thus, empty links between different nodes represent very weak influences compared with nonempty links. Note that the 16th graph has a weak connection between every pair of nodes, which avoids the most trivial case, i.e. isolated social systems in Figure 5.1(a) in Chapter 5.

In addition, applying a threshold to the strength of influence can simplify dynamic influence as its presence or absence, which depends on application domains. However, in this study, we do not ignore every weak influence to reflect real-world diffusion exhibiting dynamic interactions between heterogeneous social networks, which heps to estimate more accurate diffusion patterns. Accordingly, three variants of link weights are considered as (non-empty-link-weight, empty-link-weight) $= \{(0.33, 0.01), (0.26, 0.05), (0.18, 0.07)\}$ in order to cover exemplary cases, such as (1) dominant influence within meta-populations, i.e. strong intra- and weak inter-relationships, (2) strong influence of one meta-population on the others, i.e. unbalanced inter-relationships, and (3) balanced influences among three meta-populations, i.e. balanced inter-relationships including balanced intra- and inter-relationships. We finally generated 48 (= 16 $\times$ 3 variants) datasets of cumulative adopters from our diffusion model, each of which consists of $\{(A_1(t), A_2(t), A_3(t))\}_{t=1}^{T}$, as shown in Figure 6.2. The length of time step T is chosen as 30 to reflect the periods of our real datasets (e.g., 30 days and 30 years for the social media and academic publication datasets, respectively), and the subscripts 1, 2, and 3 of $A(t)$ indicate the identities



(a)



(b)

Figure 6.3: Distributions of model fitting errors with 48 synthetic datasets (BM: Bass Model, DM: Dynamic Influence Model). Model fitting errors are calculated with respect to **(a)** cumulative adoption rate F in Equation (6.1) and **(b)** cumulative adopters A by multiplying each population size to corresponding errors in Equation (6.1).

of the three meta-populations. Each link between nodes represents the strength and directionality of influence between two populations. In our model, they are denoted as $c_{ji} \in [0,1]$, which is the probability that an individual from a population i adopts when it is influenced by a neighboring adopter from a population j, as discussed in Equation (5.10) in Chapter 5.

6.2.2 Experiments on Synthetic Data

Let us denote the model parameters by $\mathbf{\Theta}_i = \{n_i, \boldsymbol{\theta}_i\}$, $i = 1, ..., 3$, where n_i represents the size of each meta-population i, and the definitions of $\boldsymbol{\theta}_i$ are different in each diffusion model. For instance, $\boldsymbol{\theta}_i = \{p_i, q_i\}$ in the Bass Model, while $\boldsymbol{\theta}_i = \{p_i, c_{1i}, c_{2i}, c_{3i}\}$ in the Dynamic Influence Model. To fit each model to the generated synthetic datasets, we apply Nonlinear Least Squares (NLS) [80], which minimizes the normalized root mean squared errors (RMSE) as:

$$\text{RMSE} = \sqrt{\frac{1}{3T}\sum_{t=1}^{T}\sum_{i=1}^{m}\left(\frac{A_i(t)}{n_i} - P(a|i,t,\boldsymbol{\theta}_i)\right)^2}\ , \tag{6.1}$$

where $P(a|i,t,\boldsymbol{\theta}_i)$ is the estimated adoption probability of each population at time t, and m is the number of meta-populations. Note that due to the parameter identification problem, where the same results are produced with different settings of parameters, we fix the power law coefficient α in Equation (5.13) in Chapter 5 to be 2.5, as this value in real-word networks is typically in the range $2 < \alpha < 3$ [48, 127].

Figure 6.3 shows the distributions of model fitting errors of two diffusion models, the Bass Model (BM) and our Dynamic Influence Model (DM), with the 48 generated

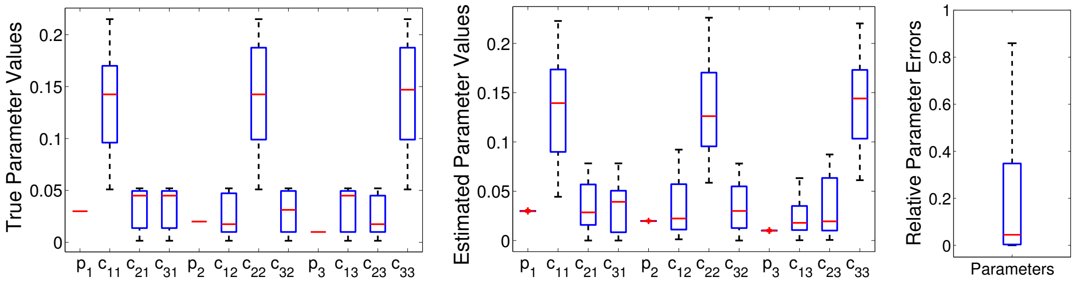


Figure 6.4: Estimation of parameter values of our model with 48 synthetic datasets. (**a**) Distributions of ground truths of parameter values. (**b**) Distributions of estimated parameter values. (**c**) Distributions of relative parameter errors ($\tilde{e}$) in Equation (6.2) for all parameters. (parameters: p_i – external influence of a meta-population i, c_{ji} – internal influence of a meta-population j on a meta-population i, and n_i – size of a meta-population i). Note that n_i is omitted for clarity in (a) and (b) due to the large difference of its magnitude order from others, but it is included in (c).

datasets. As the figure shows, the DM outperforms the BM with more acceptable standard deviation in both RMSEs of F (cumulative adoption rate) and A (cumulative adopters). However, this is not surprising, since the DM has more degrees of freedom, due to having more parameters than the BM.

Thus, we show the parameter recovery of the DM with relative parameter errors. Figure 6.4 shows the distributions of ground truths of parameter values from synthetic datasets in (a), estimated parameter values in (b), and relative parameter errors for all parameters in (c). Due to the different magnitude orders of the true parameter values (e.g., $p_i < c_{ji} \ll n_i$), each parameter error is normalized by a true parameter value to show its relative parameter error ($\tilde{e_k}$) as:

$$\tilde{e_k} = \frac{|\hat{\omega}_k - \omega_k|}{\omega_k} \quad , \tag{6.2}$$

where $\hat{\omega}_k$ and ω_k denote the estimated and true parameter values of the k-th parameter, respectively. Figure 6.4(a) and (b) exhibit very similar distributions of true and estimated parameter values. More precisely, all estimated parameter values and true parameter values differ by less than one order of magnitude, as shown in Figure 6.4(c). These results show the feasibility of our model to reproduce parameters from the datasets, i.e. time-series cumulative adopters.

Based on the evaluation of our proposed model, we estimate real-world diffusion in the rest of this chapter, by taking working examples from news diffusion in social media in Chapter 3 and knowledge diffusion in academic publication in Chapter 4.

Table 6.2: Selected news topics and the corresponding news categories, which have driven the top largest global diffusion in social media. News topics and categories are referred to the Wikipedia Current Events [9].

Category	**News Topics (January 2011)**
Politics	Protests in Tunisia, Egypt, Sudan, and Yemen; Internet shutdown in Egypt; Hosni Mubarak Resignation; Tucson shooting; Julian Assange Wikileaks; Healthcare law, etc.
Business & Economy	US bank crisis; Apple profit record; Borders bankruptcy; New Google CEO; Swiss bank account reveal by Wikileaks; Food crisis; Unemployment; Oil price, etc.
Disasters	Floods in Australia, Sri Lanka, and Brazil; Massive winter storm in US
Technology & Science	10 billion downloads of App Store; Apple iPad2 release; iPads for education; Google versus Bing; Facebook security; Wikipedia 10th Anniversary; Google technology news, etc.
Arts	Academy Movie Awards; Golden Globe Awards; Screen Actors Guild Awards; Film release; Celebrities, etc.
Culture	Multiculturalism failure; Conflicts between Muslims and Christian; Cultural change of female education by Taliban. etc.
Sports	NFL playoffs; BCS Championship; AFC Asian Cup; Australian Open; Ashes series winner; Sky Sport sexism scandal

6.3 Dynamic Influence in Social Media

We analyzed news diffusion patterns in social media in Chapter 3 from the aspects of the structural connectivity across News, SNS, and Blog. In this section, we examine temporal dynamics of diffusion in social media in terms of the strength and directionality of influence. We first describe data preparation, conduct experiments on real data, and discuss experimental results.

6.3.1 Data Preparation

In Chapter 3.2, we described the preparation and analysis of the Spinn3r dataset [4]. Among the 60 million English documents, we selected documents that contain at least one hyperlink or URL in their main text and that are also created by the 6.4 million identified users in Table 3.2. We labeled these documents with 284 identified real-world news items by referring to the Wikipedia Current Events [9]. Eventually, we selected 63 news topics, as shown in Table 6.2, each of which has driven adoptions of at least 150 identified users across different types of social media, i.e. News, SNS, and Blog. These selected news topics led to 3.1 million web documents, and 56% (1.7 million) of them are created by the identified users. The fact that over 50% of the largest diffusion of 63 news topics is led by these identified users reconfirms the validity of our real datasets to study real-world diffusion mechanisms.

Table 6.3: Averages and standard deviations of model fitting errors (RMSE) in Equation (6.1) with real datasets (BM: Bass Model, DM: Dynamic Influence Model).

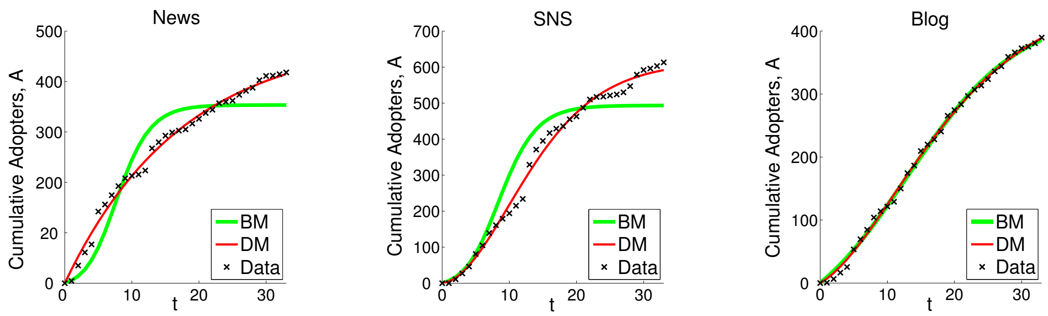
	BM	**DM**
Mean	2.866e-2	2.259e-2
STD	1.902e-2	1.027e-2

We conduct experiments with 63 real datasets, one per topic, each of which consists of daily cumulative adopters for the three media types as three meta-populations in our model during a one month period, i.e. $\{A_i(t)\}_{t=1}^{30}$, $i = n, s, b$, where n, s, and b denote News, SNS, and Blog for each. Note that in Chapter 3 we selectively chose five main categories for obtaining a overall big picture of news diffusion in social media, but here we add and separate the news categories, as shown in Table 6.2, for a better understanding of topic-related diffusion patterns in more detail.

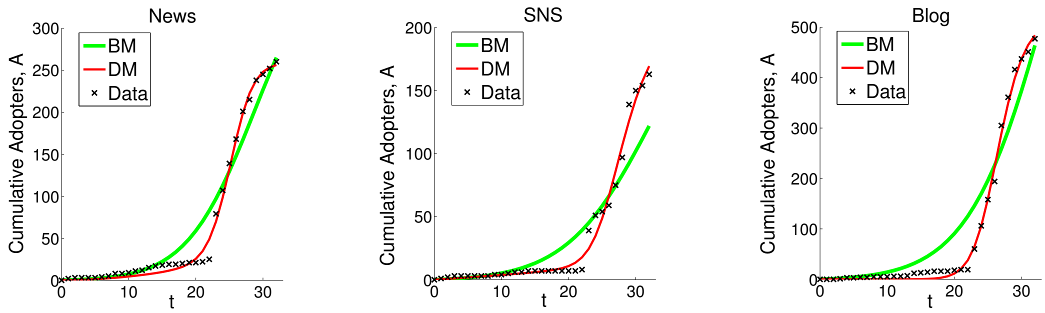
6.3.2 Experiments on Real Data

As we discussed in Chapter 5, our macro-level diffusion model does not require detailed network structures (see Equation (5.13) and Appendix B.1), and what we need to know is only the power-law exponent α in Equation (5.13), based on the assumption of a power-law degree distribution. Most real-world networks in their degree distributions have power-law exponents in the range $2 < \alpha < 3$ [48, 127], as discussed in the previous section. When it comes to social media, the entire Twittersphere, including 41.7 million users, exhibited an exponent of about $\alpha = 2.3$ [98], the blogosphere showed exponents of $2.5 \leq \alpha \leq 2.6$ [43, 105], and authorship networks in our real data also follow a power-law degree distribution with the exponent of $\alpha = 2.3$. Based on the observations from both related works and our study considering possibly missing or noisy connections in our dataset, we set the power-law exponent, $\alpha = 2.5$. With the collected 63 real datasets, we fit the models (DM and BM) and further examine how news spreads across social media by comparing different diffusion patterns between different news categories in Table 6.2.

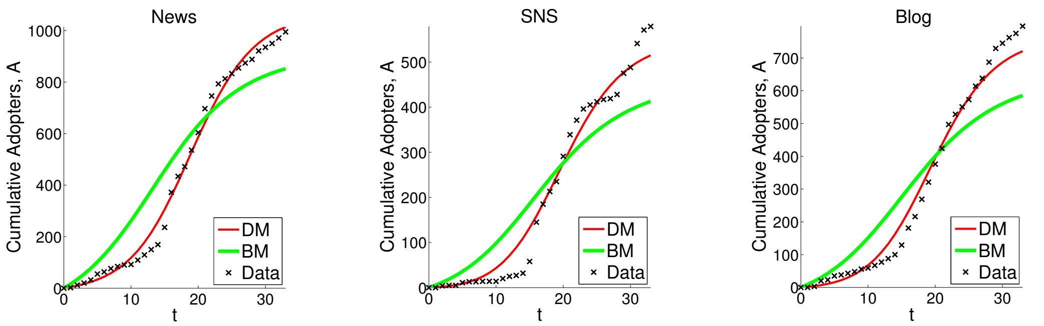
There are no ground truths of parameter values in the real data, so we fit the proposed model (DM) and the baseline model (BM) using nonlinear least squares (NLS), as in the experiments on the synthetic datasets, and evaluate model fitting errors as shown in Table 6.3. Overall, due to noise in the real datasets, the model fitting performance decreased by at least one order of magnitude, compared with those on the synthetic datasets. However, our proposed model still performs better than the BM, with more acceptable standard deviations in all cases. This result can be interpreted as we need to consider influence between heterogeneous populations on diffusion, and thus the proposed model can improve the accuracy of diffusion models dealing with a single social system or a local contact network alone.



(a) Arts: the film "Black Swan"



(b) Culture: multiculturalism failure

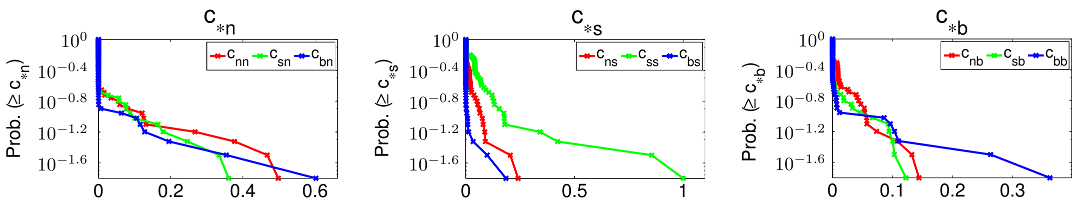


(c) Politics: Yemen protests

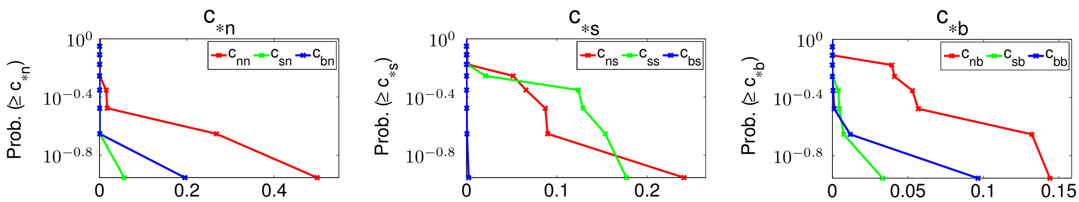
Figure 6.5: Example cases of model fitting results with real datasets from the arts, culture, and politics categories (BM: Bass Model; DM: Dynamic Influence Model). The x-axis indicates time t in days.

6.3.3 Interpretations of Experimental Results

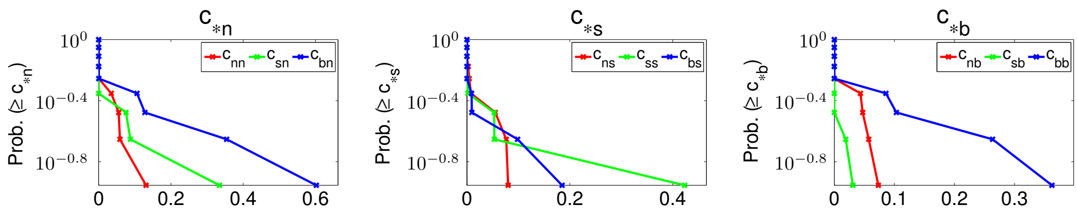
We examine different diffusion patterns by the context of information. Figure 6.5 shows three example cases of model fitting results from the arts, culture, and politics news categories. As Figure 6.5(a) and (b) demonstrate, they show different diffusion patterns. The art and culture categories are often grouped together into the same category in news agencies or the Wikipedia Current Event which our selected news items are drawn from. However, information topics in the same category do not necessarily exhibit similar diffusion patterns.



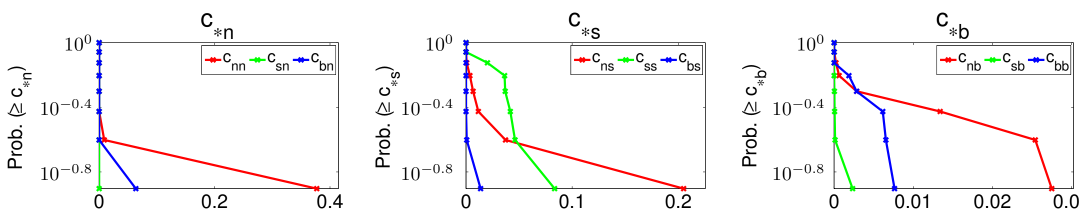
(a) All Categories



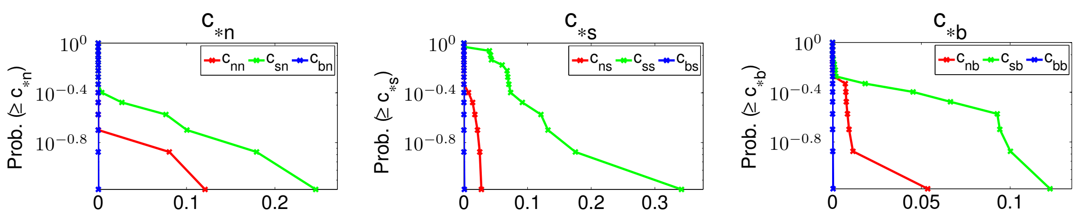
(b) Arts



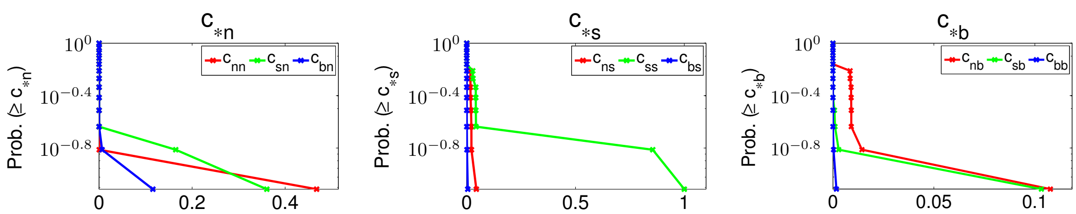
(c) Culture



(d) Business and Economy



(e) Politics



(f) Technology and Science

Figure 6.6: Distributions of estimated parameter values with real datasets, separated by different categories. The x-axis indicates value of parameter c_{ji} (the probability of the influence of media type j on i); the y-axis represents the probability of the parameter value greater than or equal to c_{ji}.

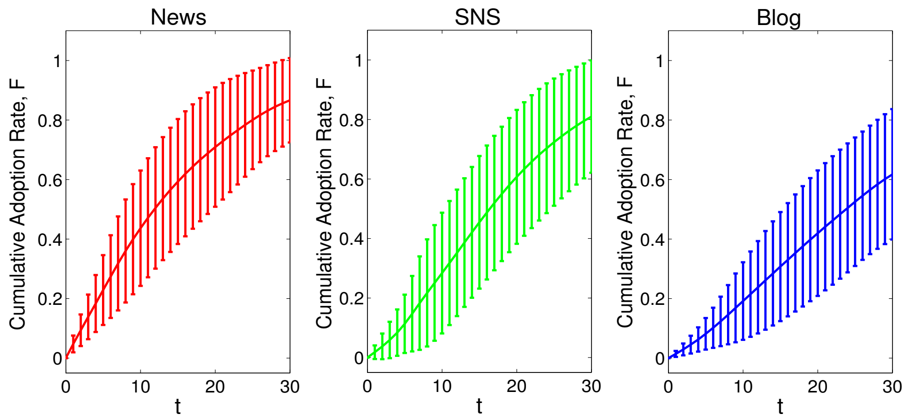


Figure 6.7: Averages and standard deviations of cumulative adoption rates for all 63 pieces of news content by media types.

Synchronous Diffusion across Social Media: In Figure 6.5(a), news about the film "Black Swan" rapidly spreads in news media first, and then it continues to spread to other social media (more rapidly in SNS than Blog). On the other hand, in the case of "multiculturalism" issues in Figure 6.5(b), the growth rate was not rapid from the beginning, but the diffusion begins to grow sharply and simultaneously across all media types after 23 days, when UK Prime Minister, David Cameron, stated the failure of multiculturalism [6]. Similarly, such concurrent behaviors are observed in the diffusion of political movements in the Middle East, such as Tunisia, Egypt, Sudan, and Yemen. As shown in Figure 6.5(c), the "Yemen protests" demonstrate synchronous diffusion patterns after 15 days.

Without direct interactions across social media, such synchronous and simultaneous growth unlikely happens. As the figure shows, the BM cannot follow these concurrent growth patterns without considering the effects of interactions across heterogeneous social networks on diffusion. Thus, influences from different social networks are not ignorable for a better understanding of diffusion processes.

Topic-related Diffusion Dynamics: By categorizing news topics according to Table 6.2, we attempt to distinguish different diffusion patterns in terms of the strength and directionality of influence. Figure 6.6 shows the distributions of estimated parameter values, where c_{ji} in the x-axis indicates the influence of media type j on media type i, and the y-axis represents the probability that the influence of type j on i is equal to or greater than c_{ji}.

Figure 6.6(a) shows overall trends of interactions among three media types by aggregating parameter values of all news categories. In general, News is influenced by all media types in a balanced way, while SNS and Blog, in that order, exhibit stronger internal interactions within the same media types. This can be interpreted that news media need to monitor trending topics in other social media in order to keep their readers updated on popular news topics. As shown in Figure 6.6(a) and 6.6(b), the news topic on the film "Black Swan" in the art category demonstrates the different diffusion pattern from the topic "multiculturalism failure" in the culture category. Such different patterns are governed by different diffusion dynamics. News is the most influential in the diffusion of arts topics, such as Academy Awards, film releases, and celebrities. However, regarding the culture category, Blog tends to show strong influence on News and SNS. Controversial subjects of the culture category, such as multiculturalism failure, religious conflicts, and female education in Afghanistan, likely lead to longer discussions, and thus Blog media can be a more suitable space to deliver personal opinions than SNS allowing a limited text length and News delivering objective and neutral reports. Besides the arts category, News also occupies the influential position in the business and economy category. News can be considered as reliable media to report accurate statistics or impartial facts on economic situations. Interestingly, regarding political topics, SNS exhibits the highest influence on other media types, while the influence of Blog is negligible. Political news generally has great social repercussions, such as the Middle East protests, the Tuscon shooting, and Wikileaks. In this respect, the micro-blogging space can be a better medium to distribute urgent issues rapidly, and their prompt proliferation influences news media to focus on the issues. In the technology and science category, internal buzz in SNS is predominant in contrast to Blog.

Diffusion Rate of Social Media: Figure 6.7 shows the averages and standard deviations of cumulative adoption rates for 63 news topics by media types. In general, News and SNS media spread information rapidly, showing very similar growth rates. That is, SNS users tend to be very responsive to new information in contrast to the persistent behavior of Blog users. Blog shows much slower growth rates than the other types, but it is relatively insensitive to timing of new information and adopts older news topics in a persistent manner.

In summary, News is the most influential in the arts and the business and economy categories, while in the politics and the culture categories, SNS and Blog are influential, respectively. Thus, dynamics of diffusion needs to be explained within the context of information.

6.4 Dynamic Influence in Academic Publications

We analyzed diffusion patterns in academic publications in Chapter 4. As observed in news diffusion in social media, academic citation networks are also interconnected across different subdomains in computer science. While we analyzed the overall structural connectivity of academic publications in Chapter 4, in this section we examine temporal dynamics of topic-related diffusion in terms of the strength and directionality of influence. We first describe the preparation of our dataset, conduct experiments on the dataset, and then discuss experimental results as we did in the previous section.

6.4.1 Data Preparation

In Chapter 4, we constructed computer science citation networks and observed that conference publications are cited more often than journal papers. Among popular conference publications receiving more than 40 citations, 90% of them came from the top-ranked 221 conferences. Based on these observations, we targeted the citation networks consisting of papers from these 221 premier conferences.

Meta-populations: In order to obtain macro-level diffusion patterns across different research areas, we clustered 24 subcomains from the target citation networks into the three clusters: *Core*, *Data*, and *Systems*, as shown in Table 4.6. Accordingly, there are three clusters of conferences based on their associated subdomains. These clusters of conferences are considered as the three meta-populations ($m = 3$) in the Dynamic Influence Model, and accordingly each conference in the i-th cluster corresponds to an individual in the i-th population (refer to Equation (5.13)).

There are several reasons that conferences are considered as individuals of a meta-population in this study. First, adopter distributions of different research topics can be biased when we consider an author as an individual adopter. That is because the number of authors in a paper is highly variable from one to more than ten authors, and thus the adopter count of a specific research topic is possibly overestimated or underestimated. On the other hand, in the previous section an online user is considered as a member of a meta-population, since every document in social media is produced by a single user[1]. Second, author disambiguation from extensive publications can be noisy [47]. As discussed earlier in this chapter, disambiguating paper profiles such as author, affiliation, and publication venues is a challenging issue, which is beyond the scope of this thesis. Finally, regarding each publication as an adopter still can be a problem because each conference publishes a different number of papers, which also cause biased adopter distributions as discussed in the first reason. Besides, a large number of publications share common authors, which makes it more difficult to interpret adopting behaviors.

Research Topic Selection: As discussed in Chapter 2, targeting trending topics is important in order to not only obtain a global view of diffusion across heterogeneous populations but also to distinguish topic-related diffusion patterns. In order to achieve this goal, we investigated the top 100 research keywords in computer science, provided by the Microsoft Academic Search (MAS) [5] in Chapter 4. Among these keywords, we obtained 15 research keywords (framed keywords in Table 4.7 and 4.8), each of which is used in more than 100 publications from the premier conferences and, more importantly, is widely used across all three clusters. In this section, we target these popular keywords as trending research topics, and they are listed in Table 6.4 in the order of popularity. Each keyword listed in the table has occurred in publications from more than 10% of conferences belonging to each cluster.

Dataset Generation: For the selected research keywords in Table 6.4, we generated fifteen time-series datasets (one per keyword) of cumulative adopters $A_i(t)$ for the most recent 30 years as $\{(A_c(t), A_d(t), A_s(t))\}_{t=1}^{T}$, where T denotes the length of time steps ($T = 30$), and c, d, and s represent the three clusters, Core, Data, and Systems, respectively. In each dataset, a conference is counted as an adopter only when its publications are involved in the citation networks consisting of papers related to

[1] There are occasionally co-authored news articles. Even such cases in news media are not an issue in our study because we regard one news site as a super user, as discussed in Chapter 3.

Table 6.4: Selected research keywords among the top 100 keywords in computer science by referring to Microsoft Academic Search (MAS) [5]. Selected keywords are listed in the order of popularity, each of which is widely used across all clusters, i.e. Core, Data, and Systems (refer to Table 4.7 and 4.8 for more details).

Rank	**Research Keyword**	**Rank**	**Research Keyword**
1	Real Time	9	Fault Tolerant
2	Satisfiability	10	Distributed System
3	Sensor Network	11	Approximate Algorithm
4	Large Scale	12	Social Network
5	Web Service	13	Search Engine
6	Genetic Algorithm	14	Learning Algorithm
7	High Performance	15	Genetics
8	Semantic Web		

Table 6.5: Averages and standard deviations of model fitting errors (RMSE) in Equation (6.1) with real datasets (BM: Bass Model, DM: Dynamic Influence Model).

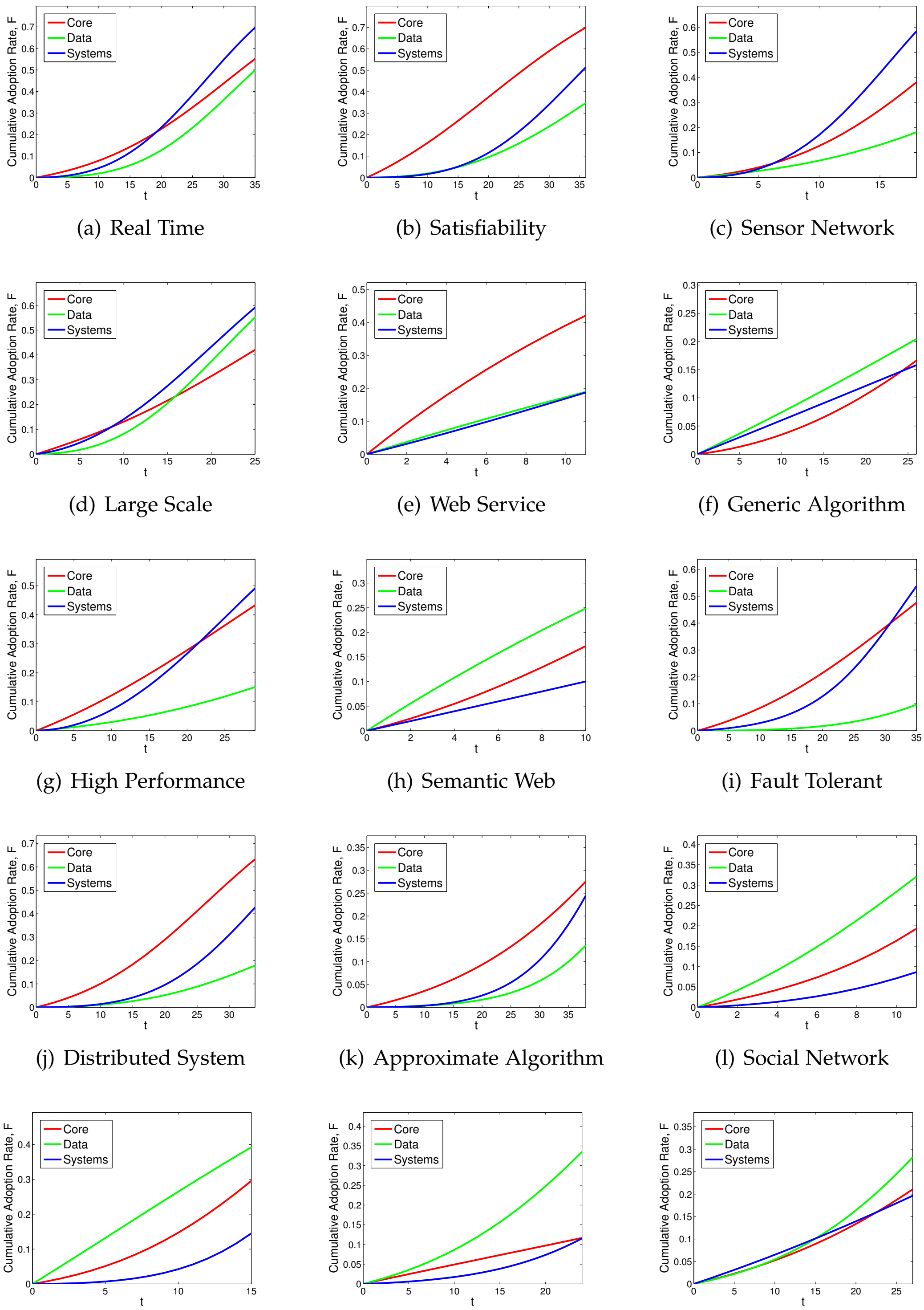
	BM	**DM**
Mean	1.884*e*-2	1.754*e*-2
STD	8.156*e*-3	7.268*e*-3

a corresponding keyword. This helps us to estimate more accurate topic-related diffusion trends. This is the same way as the previous section where we counted online users as adopters only when their web documents constitute document networks related to a selected news topic. Note that, some keywords are relatively young, such as "Sensor Network", "Semantic Web", and "Social Network", and thus the length of T is not always 30. In the next subsection, with these generated time-series datasets we estimate topic-related diffusion patterns in terms of the strength and directionality of influence.

6.4.2 Experiments on Real Data

As mentioned earlier, there are no ground truths of parameter values in the real data, so we fit the models (DM and BM) using Nonlinear Least Squares (NLS) as we did in the previous section. With the generated 15 real datasets, we evaluate model fitting errors as shown in Table 6.5. Overall, our proposed model performs better than the BM, with more acceptable standard deviations as in the case of social media.

Figure 6.8 shows the model fitting results in terms of cumulative adoption rates (F) of the three meta-populations, i.e. Core, Data, and Systems, for each research keyword. As the figure shows, the growth rates of adoptions are varied with different



(a) Real Time (b) Satisfiability (c) Sensor Network

(d) Large Scale (e) Web Service (f) Generic Algorithm

(g) High Performance (h) Semantic Web (i) Fault Tolerant

(j) Distributed System (k) Approximate Algorithm (l) Social Network

(m) Search Engine (n) Learning Algorithm (o) Genetics

Figure 6.8: Time-evolving cumulative adoption rates (F), inferred by the Dynamic Influence Model, for the selected research keywords in Table 6.4. Core, Data, and Systems represent three clusters of conferences according to their associated subdomains in Table 4.6. The x-axis indicates time t in years.

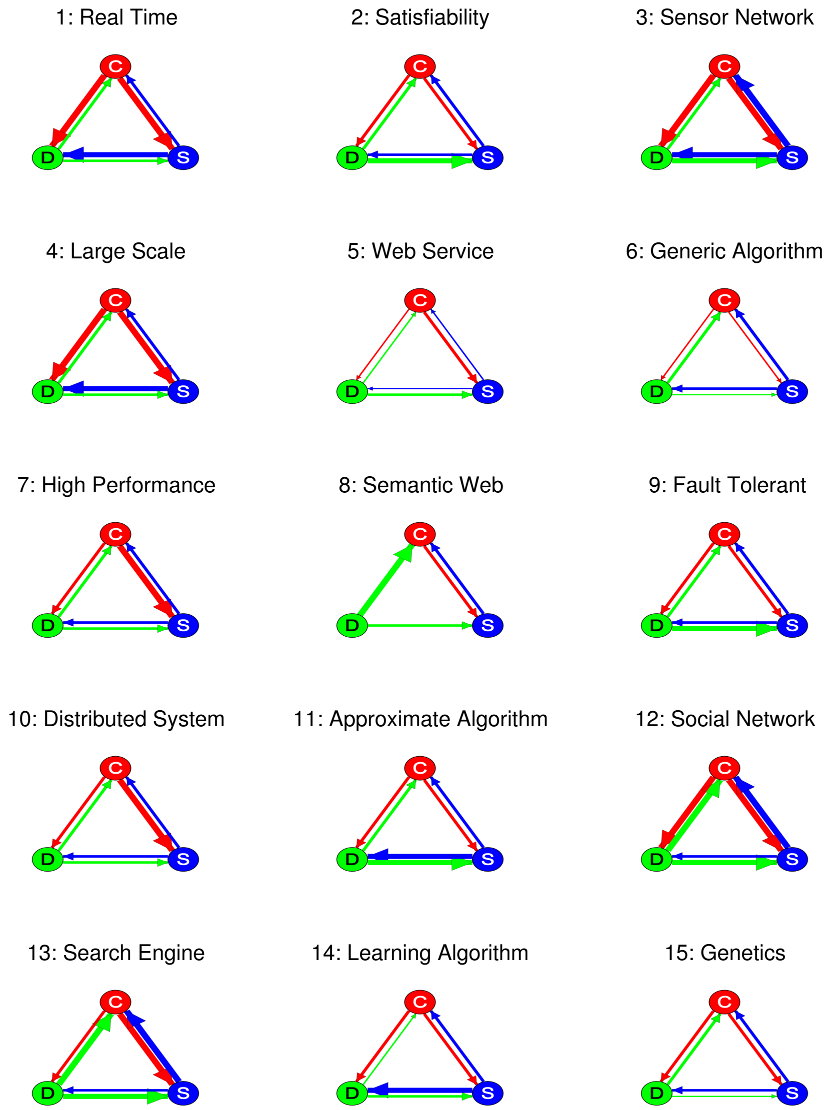
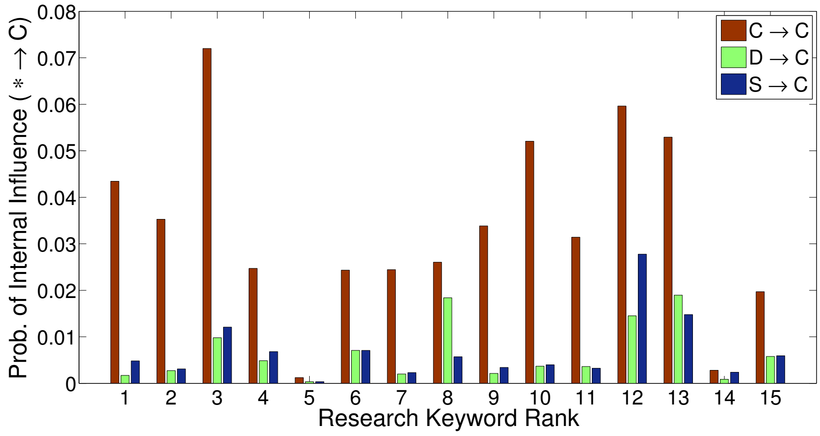
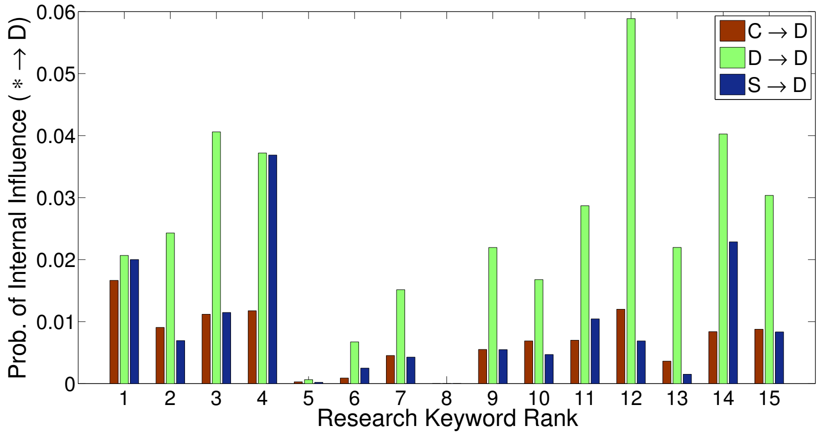


Figure 6.9: Dynamic influence among three clusters (Core, Data, and Systems) for the selected research keywords in Table 6.4. In each graph, the node labels indicate the three clusters (C:Core, D:Data, and S:Systems). Link widths and arrows present the strength and directionality of influence respectively, which is based on the estimated parameter values (c_{ji} in Equation (5.10)) in our Dynamic Influence Model. For clarity, self-loops and links with weights less than 1.0e-3 are omitted.

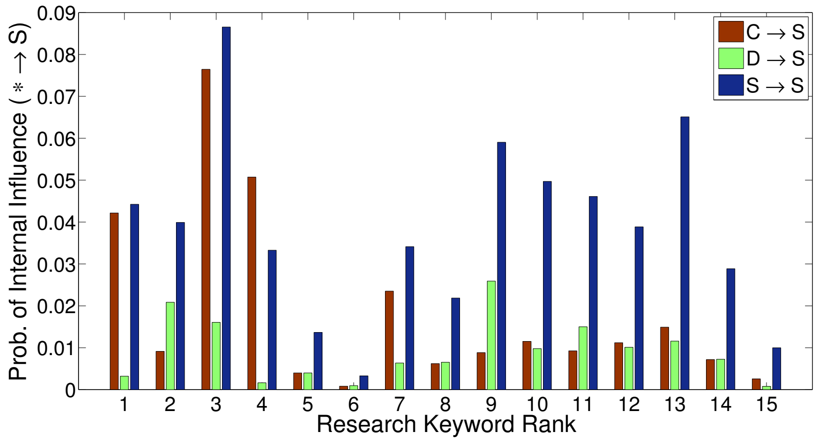
keywords. In general, established keywords, which have drawn attention for more than 30 years, such as "Real Time", Satisfiability", and "Distributed System", show higher cumulative adoption rates than relatively younger keywords such as "Web Service", "Semantic Web", and "Social Network". Interestingly, the keyword "Sensor Network" shows the most rapid growth pattern compared with other younger



(a) Influence of all meta-populations on Core



(b) [Influence of all meta-populations on Data



(c) [Influence of all meta-populations on Systems

Figure 6.10: Estimated internal influence. The x-axis represents the ranks of selected research keywords in Table 6.4, while the y-axis represents the probability of internal influence c_{ji} (influence of a meta-population j on a meta-population i). Graphs are separated by each cluster which is influenced by all meta-populations (C – Core, D – Data, S – Systems): (a) $c_{*\mathbf{c}}$, (b) $c_{*\mathbf{d}}$, and (c) $c_{*\mathbf{s}}$.

keywords. Particularly, the adoption rate in the Systems cluster increased by 40% during the last 10 years. Such abrupt growth pattern of Systems is also shown in other keywords regarding system performances, such as "High Performance", "Fault Tolerant", and "Approximate Algorithm". These cumulative adoption behaviors are governed by diffusion mechanisms, and we try to infer the processes based on the estimated parameter values of our model. Such interpretations are visualized in Figure 6.9, and the detailed quantifications are also presented in Figure 6.10.

6.4.3 Interpretations of Experimental Results

Time-series adopting behaviors of meta-populations can be interpreted with our probabilistic diffusion mechanism from the aspects of the strength and directionality of influence.

Inter-relationships: In general, heterogeneous populations influence each other but with different strengths, as shown in Figure 6.9. This figure provides an overall picture of topic-related diffusion patterns, showing weak or strong influence for all or partial directions across the clusters. Such balance or imbalance of influences may bring about different cumulative adoption rates (F) between different keywords, as shown in Figure 6.8. Even though the selected keywords are most popular in wide research areas, but they are not explained with same dynamics of influence. However, more balanced inter-relationships, observed in the keywords "Real Time" and "Large Scale", more likely lead to similar curves of cumulative adoption rates (i.e. more synchronous and simultaneous growth rates across populations) than less balanced inter-relationships, shown in the keywords "Web Service" and "Semantic Web".

Intra- and inter-relationships: Figure 6.10 presents intra- and inter-relationships (c_{ji} in Equation (5.10)) in more detail from the aspects of all meta-populations' internal influence on each population: Figure 6.10(a) for Core, (b) for Data, and (c) for Systems. As the figure shows, in general intra-relationships (c_{ii}) are stronger than inter-relationships (c_{ji}, $i \neq j$). Core shows relatively stronger intra-relationships than the others. Data is influenced by Core and/or Systems as much as intra-influence for the research keywords "Real Time" and "Large Scale", while Data shows the strongest intra-relationship for the keyword, "Social Network". Systems is strongly influenced by Core in particular for the keywords on system performances, such as "Real Time", "Sensor Network", "Large Scale", and "High Performance". These keywords show similar strength between Core's influence on System (c_{cs}) and intra-relationship of Systems (c_{ss}), while the keyword "Large Scale" exhibits the stronger inter-relationship than the intra-relationship, i.e. $c_{cs} > c_{ss}$. In addition, more balanced intra and inter-relationships are observed in the top popular keywords, such as "Real Time", "Satisfiability", "Sensor Network", and "Large Scale".

In general, balanced inter-relationships more likely lead to a synchronous and simultaneous growth of adoptions, while balanced intra- and inter-relationships tend to be driven by widely popular research keywords.

6.5 Discussion

From the experimental results, we observed that the heterogeneity of social networks has an effect on diffusion, but it also raises issues. First, the collapse of the boundaries of social media platforms brings a concern of identifying borderline users who have more than one account in different social media platforms. However, distinguishing such borderline users is beyond the scope of this study. It is one of the research topics of identifying multi-layered social networks [132]. Second, the structure of diverse social networks are hidden, due to privacy issues, various communication channels (e.g., news feeds, mobile applications, and web search), and frequently changing relations between online users. Even if we obtain user networks in social media, we cannot say that they are real diffusion networks. In this context, our model assumes a power-law degree distribution, which brings another limitation, due to unknown or missing topologies. However, we can obtain macro-level trends of diffusion across populations without the need of detailed network structures. Studying the structural properties of diffusion networks across heterogeneous populations can improve the proposed model further.

Regardless of these limitations, influence between heterogeneous social networks helps to better describe diffusion within homogeneous social networks. This is because external influence on a single social system is not ignorable (e.g., 30% of exposures in Twitter were attributed to external sources [123]), and some of the external sources play an important role as internal influence, as discussed in Chapter 5. Frequent and close interactions between heterogeneous networks are possibly due to web technologies that enable various information sources to be more easily accessible and shared, and thus diverse online social networks to be more interconnected across multiple social media platforms. In this complex environment, this study can provide the benefits of excluding detailed network topologies, but considering realistic circumstances by combining the structural property and heterogeneity of real-wold networks in our proposed model.

6.6 Conclusions

In this chapter, by conducting experiments on both synthetic and real data, we showed a way of interpreting diffusion in terms of the strength and directionality of influence across meta-populations, and we compared different diffusion patterns among a great variety of information topics.

Our model-driven approach helps to better describe diffusion due to the considerations of the effects of interactions between heterogeneous populations. Supportive evidence can be found in the diffusion of news topics regarding political protests and multiculturalism failure, since they tend to drive concurrent and simultaneous diffusion across different types of social media. Such phenomena unlikely happen without direct interactions between different social networks. We also found that there are different diffusion patterns by different news categories. News media are the most influential in the arts and economy categories, while SNS and Blog media are in the politics and culture categories, respectively. Regarding dynamic influence in academic publications, we found that balanced inter-relationships across Core, Data, and Systems more likely lead to a synchronous and simultaneous growth of adoptions, which is observed in the keywords "Real Time" and "Large Scale". In addition, balanced intra and inter-relationships tend to be driven by the most popular keywords, such as "Real Time", "Satisfiability", "Sensor Network", and "Large Scale". Thus, diffusion patterns reflect topic-related dynamic influence, i.e. balance or imbalance of influences, across heterogeneous social networks.

We applied our model-driven approach to the two exemplary application domains of social media and academic publications. We expect that these applications can provide a way of interpreting real-word diffusion in other domains at a macro level. In contrast to the model-driven approach in Part II, in the following we will present a model-free approach independent on any assumptions of network structures in Chapter 7 and also apply this approach to the same applications domains in Chapter 8 for further comparisons between the model-driven and model-free approaches in Part IV.

Part III

Model-Free Approach

Chapter 7

Macro-level Information Transfer

Social interactions on the Web have become more dynamic and far-reaching across multiple social media platforms than ever before. This is because of frequent changes of online contact networks and increasing accessibility to diverse information sources. Accordingly, massive and continuous user-generated content spreads through heterogeneous social networks. In this chapter, we propose a model-free approach for estimating macro-level information transfer across different social systems without any assumptions on such dynamic social interactions.

7.1 Introduction

In Part II of this thesis, we proposed the model-driven approach which estimates dynamic influence across heterogeneous social systems, each of which consists of a large number of individuals interacting with each other. This approach does not neglect the structural connectivity of underlying networks, since it drives a diffusion process in a stochastic way. Thus, this model needs to assume social interactions to obtain underlying stochastic pathways and ultimately quantify non-linear dynamics of complex systems. In order to minimize the limitations of collecting local network structures, our model assumes a universal property of real-world networks, i.e. a power-law degree distribution. Thus, it does not require the knowledge of detailed network topologies, but it still needs to estimate the power-law coefficient which is the scaling exponent for the power-law tail [127].

Information theory, introduced by Shannon [154], provides the benefits of quantifying non-linear dynamics of complex systems such as information of a random variable, a collection of variables, and exchanges between variables. Such quantification does not require any assumptions, which makes information theoretic-measures model-free [50]. Due to this flexibility, *ad hoc* models are not necessarily to be defined whose assumptions are often based on the sampled snapshots of current social networks. These sampled structures have largely relied on site-specific actions (e.g., *mentions* and *retweets* in Twitter, and *likes* and *comments* in Facebook), which makes it hard to generalize these models and to apply them to other application domains.

The goal of this chapter is to propose a model-free approach for estimating macro-level information transfer across heterogeneous social systems without any assump-

Table 7.1: Table of symbols for Chapter 7.

Symbol	Description
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Stochastic processes (an ordered set of random variables)
X, Y, Z	Random variables
T	An ordered set of time t ($t \in T$)
x_t, y_t	The realizations of X and Y at time t respectively
$\mathcal{A}_t$	Acceleration, i.e. the changes of adoption rates at time t
τ	Threshold value to discretize $\mathcal{A}_t$ of a stochastic process
$p(x)$	Probability of the event x
$p(x, y)$	Probability of the joint event x and y
$p(x\|y)$	Probability of the event x given event y
$I(x)$	Information of observing the event x
$H(X)$	Shannon entropy of X
$H(X\|Y)$	Conditional entropy of X given Y
$I(X;Y)$	Mutual information between X and Y
$I(X;Y\|Z)$	Conditional mutual information between X and Y given Z
$I(\mathbf{X};\mathbf{Y})$	Mutual information between $\mathbf{X}$ and $\mathbf{Y}$
$I(\mathbf{X};\mathbf{Y}\|\mathbf{Z})$	Conditional mutual information between $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$
$\mathrm{TE}_{Y\to X}$	Transfer entropy from $\mathbf{Y}$ to $\mathbf{X}$
$\mathrm{Bias}[\mathcal{I}]$	The Panzeri-Treves bias of an information-theoretic quantity $\mathcal{I}$ such as $H(X)$, $H(X\|Y)$, $I(X;Y)$, and $\mathrm{TE}_{Y\to X}$
$D_{KL}(P \parallel Q)$	Kullback-Leibler divergence of two probability distributions P and Q
k	Markov order for the previous states of $\mathbf{X}$
l	Markov order for the previous states of $\mathbf{Y}$
d	The length of time-delay

tions on complex social interactions. This is enabled by (1) designing a social system as a stochastic process which produces a time-series activity sequence, and by (2) considering the transfer entropy with both time-delay and memory (Markov order [117]) effects. This approach helps to obtain a macroscopic view of global diffusion in terms of the strength and directionality of influence, and to understand behavioral characteristics of a social system from the effects of time-delay and memory. We expect that this model-free approach can provide a different way of estimating macro-level diffusion with new perspectives.

The rest of this chapter is structured as follows. In Section 7.2, we first explains fundamental information-theoretic measure as a background. In Section 7.3, we propose a conceptual framework for macro-level information transfer across social systems and defines a model-free approach based on this conceptual design. Finally, Section 7.4 concludes this study.

7.2 Information-theoretic Measures

In this section, we briefly explain fundamental information-theoretic measures, since these are used in quantifying macro-level information transfer in the next section.

7.2.1 Fundamental Measures

In Shannon's information theory [154], the *entropy* is the basic quantity to measure the amount of information or uncertainty. As discussed in Chapter 2, no uncertainty of an event provides no information. Thus, the term "information" in the theory implies the degree of uncertainty of predicting the event x's occurrence, given a probability of the event $p(x)$, produced by an information source as:

$$I(x) = \log_b \frac{1}{p(x)} \quad , \tag{7.1}$$

where b denotes the basis of the logarithm. Note that we express all entropies as bits, so all logarithms, denoted by log, are taken to the basis of two ($b = 2$) for the rest of this thesis. The natural logarithm ($b = e$) will be denoted by ln.

The entropy of a random variable X is defined as the expected amount of uncertainty or information of X as:

$$\begin{aligned} H(X) &= E[I(x)] \\ &= -\sum_x p(x) \log p(x) \quad . \end{aligned} \tag{7.2}$$

In addition, the amount of information of the random variable X's states, given knowledge of another variable Y's states, can be measured using *conditional entropy* as:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y) \quad . \tag{7.3}$$

7.2.2 Bias Correction

These information-theoretic quantities require the probability distributions of target random variables, but the estimation of the probabilities are based on finite samples. In reality, it is known that such a limited sample size brings about a biased estimation of the entropy [173]. In order to resolve this issue, we subtract the Panzeri-Treves bias [134, 135], quantifying the magnitude of the systematic error, from the estimated entropy as in [177, 180].

The Panzeri-Treves bias is defined as:

$$\begin{aligned} \mathrm{Bias}[H(X)] &= -\frac{1}{2N\ln(2)}[\bar{X}-1]\,, \\ \mathrm{Bias}[H(X|Y)] &= -\frac{1}{2N\ln(2)}\sum_{Y}[\bar{X}_Y-1]\ , \end{aligned} \tag{7.4}$$

where $\bar{X}$ and $\bar{X}_Y$ denote the number of observed states of X and the number of observed states of X with nonzero probability of Y respectively, and N is the total number of samples.

This bias correction method is applied to other information-theoretic quantities, and the corresponding Panzeri-Treves bias is described in the rest of this chapter.

7.2.3 Mutual Information

The *mutual information* between the two random variables measures the average reduction in the uncertainty of one variable due to the knowledge of the other variable. The mutual information between the variable X and Y is defined as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= -\sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)p(y)}\ . \end{aligned} \tag{7.5}$$

From Equation (7.5), the Panzeri-Treves bias for mutual information is accordingly calculated with those for entropy and conditional entropy as:

$$\mathrm{Bias}[I(X;Y)] = \mathrm{Bias}[H(X)] - \mathrm{Bias}[H(X|Y)]\ . \tag{7.6}$$

The *conditional mutual information* between X and Y given Z is the average information shared between X and Y except for Z as:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \\ &= \sum_{x,y,z} p(x,y,z)\log\frac{p(x|y,z)}{p(x|z)}\ . \end{aligned} \tag{7.7}$$

The mutual information can be extended to a collection of n random variables, i.e. a stochastic process, as $I(\mathbf{X};\mathbf{Y})$ which exhibits the reduced uncertainty due to the knowledge of the sates of one of the processes. However, it cannot capture the directionality of information between two processes. In this regard, we introduce the transfer entropy in the next section, which enables to obtain the directionality between two random processes.

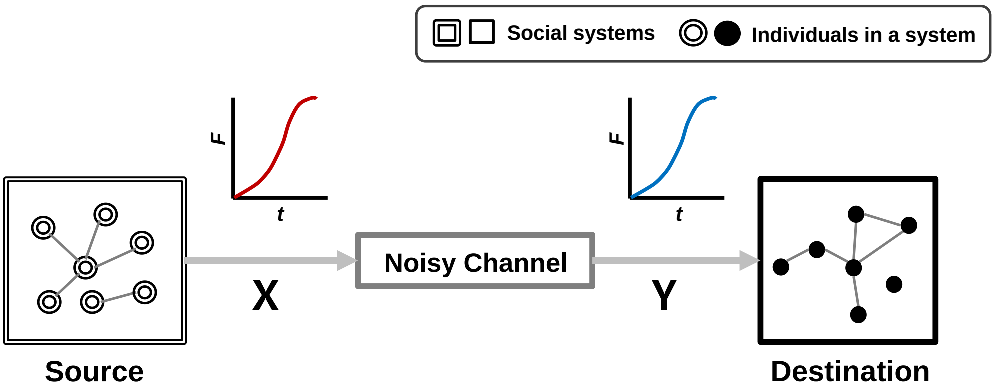


Figure 7.1: Conceptual framework for macro-level information transfer across social systems in the setting of Shannon's information theory. A signal (a time-series activity sequence of a population in a social system) is produced by a source system, an encoded signal ($\mathbf{X}$) is transferred to a destination system through a noisy channel, and then a decoded noisy signal ($\mathbf{Y}$) is received by a destination system. A social system's signal is defined in Section 7.3.1.

7.3 Macro-level Information Transfer across Social Systems

In this section, we define a model-free approach to estimate macro-level information transfer across different social systems by using information-theoretic measures.

7.3.1 Conceptual Framework

We first introduce the concept of a stochastic process and then propose our conceptual framework for macro-level information transfer across social systems.

A Social System as a Stochastic Process: A stochastic process is an ordered set of random variables [52], and thus it can be considered as time-evolving random variables when they are indexed by time. In this thesis, we focus on a discrete-time stochastic process $\mathbf{X}$ as:

$$\mathbf{X} := \{X_t : t \in T\} \ , \tag{7.8}$$

where T is an ordered set of time t, and X_t is a random variable indexed by time t.

Accordingly, a social system can be defined as a stochastic process, since a population in the system produces a time-series activity sequence in a variety of forms such as new adoption rates of a information topic over time. Such time-evolving collective behaviors of a social system influences other social systems. For example, the abrupt growth of adopters in one social system influences the growth patterns in other social systems. We observed synchronous and simultaneous diffusion patterns across different types of social media in Chapter 6, showing that the time-series cumulative adoption rates F in News, SNS, and Blog regarding the controversial news topics (e.g., "Multiculturalism Failure" and "Yemen Protests") are all similar to each other. Such scenarios can also be found in multinational diffusion of a consumer product between neighboring countries [31, 59, 95, 142]. For instance, a technology-leading country (or a big city), as a social system, produces a sales growth pattern of an electronic device over time, and its abrupt or steady increase influences the sales of the same device in neighboring countries (or neighboring cities).

Figure 7.1 illustrates our conceptual framework for macro-level information transfer across social systems from the aspects of Shannon's information theory. As discussed in Chapter 2, the main purpose of information theory is to reproduce original messages generated by a source. Based on this fundamental framework, we consider an information source and destination at a system level.

In the figure, a source social system generates a signal such as time-series growth rates of adopters in the system, and the signal is encoded and transmitted to a destination social system through a noisy channel. The encoded signal $\mathbf{X}$ is noise-corrupted after passing through the communication channel, and its decoded signal $\mathbf{Y}$ is received by the destination system. Thus, identifying a social system' signals is important to catch invisible information transfer between processes. In this regard, we define a signal as follows.

Identification of a System's Signal: Let $a(t)$ be the number of new adopters at time t and $f(t)$ be the proportion of $a(t)$ given a population's size n in a social system. The new adoption rate $f(t)$ of a specific information topic changes over time, and we call these changes of adoption rates as *acceleration* $\mathcal{A}$. We then define three discrete states of a social system X at time t, x_t as:

$$x_t = \begin{cases} -1 \text{ (decrease)}, & \text{if } \mathcal{A}_t \in (-\infty, -\tau] \\ \ \ 0 \text{ (transition)}, & \text{if } \mathcal{A}_t \in (-\tau, \tau) \\ +1 \text{ (increase)}, & \text{if } \mathcal{A}_t \in [\tau, \infty) \end{cases}, \tag{7.9}$$

where $\mathcal{A}_t = f(t) - f(t-1)$, and the value of τ is determined based on real data so that the three states are uniformly distributed.

Based on this conceptual framework, we propose a model-free approach to estimate macro-level information transfer across heterogeneous social systems by using the transfer entropy with both time-delay and memory effects, which will be presented in the following sections.

7.3.2 Macro-level Information Transfer

In the previous subsection, we considered a social system as a stochastic process and defined its signal, transmitted to the other social system, as Equation (7.9). This conceptual framework enables to estimate information transfer at a macro level.

In Chapter 5, we discussed a meta-population scheme, which clusters heterogeneous populations into a small number of relevant groups, for obtaining simple and clear diffusion dynamics across meta-populations. Here, a social system can be extended to a more flexible space that contains a meta-population, consisting of a larger number of individuals interacting with each other. Thus, our definition of a stochastic process can be considered at a meta-population level, and it can be applied in a wide range of application domains according to their appropriate meta-population schemes.

Definition of Transfer Entropy: As discussed in the previous section, the mutual information cannot capture the directionality of information flow between two random processes, since the definition in Equation (7.5) is symmetric. In contrast, the *transfer entropy*, introduced by Schreiber [153], is defined to consider causal relations between two stochastic processes, $\mathbf{X}$ and $\mathbf{Y}$ as:

$$\begin{aligned} \mathrm{TE}_{Y\to X} &= I(X_t; Y^l_{t-1} | X^k_{t-1}) & (7.10)\\ &= H(X_t|X^{t-1:t-k}) - H(X_t|X^{t-1:t-k}, Y^{t-1:t-l}) \\ &= \sum p(x_t, x^{(k)}_{t-1}, y^{(l)}_{t-1}) \log \frac{p(x_t|x^{(k)}_{t-1}, y^{(l)}_{t-1})}{p(x_t|x^{(k)}_{t-1})} \quad , & (7.11) \end{aligned}$$

where t is a time index, and k and l denote the Markov order for the previous states of $\mathbf{X}$ and $\mathbf{Y}$, respectively such that $x^k_t = \{x_{t-1}, ..., x_{t-k}\}$ and $y^l_t = \{y_{t-1}, ..., y_{t-l}\}$. Note that in our study the realizations of the random variable X correspond to the three possible states (decrease, transition, and increase) in Equation (7.9), and accordingly the entropy can be considered as expected acceleration of adoption rates of a random variable X at time t, X_t.

The transfer entropy $\mathrm{TE}_{Y\to X}$ describes the reduction of uncertainty of the state of X_t, given k previous states of the destination process $\mathbf{X}$, by introducing l previous states of the source process $\mathbf{Y}$. In addition, the transfer entropy can be considered as the conditional mutual information as Equation (7.10), which implies the average information shared between the past (Y^l_{t-1}) of the source process $\mathbf{Y}$ and the next state (X_t) of the destination process $\mathbf{X}$ except for $\mathbf{X}$'s past (X^k_{t-1}).

As discussed in Chapter 2, the mutual information $I(X;Y)$ is the *Kullback-Leibler* divergence [109] of the joint distribution $p(x,y)$ from the product distribution $p(x)p(y)$. Accordingly, the transfer entropy is the Kullback-Leibler divergence between two conditional probability distributions $p(x_t|x^{(k)}_{t-1})$ and $p(x_t|x^{(k)}_{t-1}, y^{(l)}_{t-1})$ as:

$$\mathrm{TE}_{Y\to X} = D_{KL}(p(x_t|x^{(k)}_{t-1}) \parallel p(x_t|x^{(k)}_{t-1}, y^{(l)}_{t-1})) \ .$$

In order to estimate system-level information transfer, we also need to alleviate the issue of a biased estimation from a limited sample size. As discussed in the previous section, we subtract the Panzeri-Treves bias [134, 135] from the estimated entropy from samples. Similarly, the Panzeri-Treves bias for the transfer entropy is calculated as:

$$\begin{aligned} \mathrm{Bias}[\mathrm{TE}_{Y\to X}] &= \mathrm{Bias}[I(X_t; Y^l_{t-1}|X^k_{t-1})] \\ &= \mathrm{Bias}[H(X_t|X^k_{t-1})] - \mathrm{Bias}[H(X_t|X^k_{t-1}, Y^l_{t-1})] \ . & (7.12) \end{aligned}$$

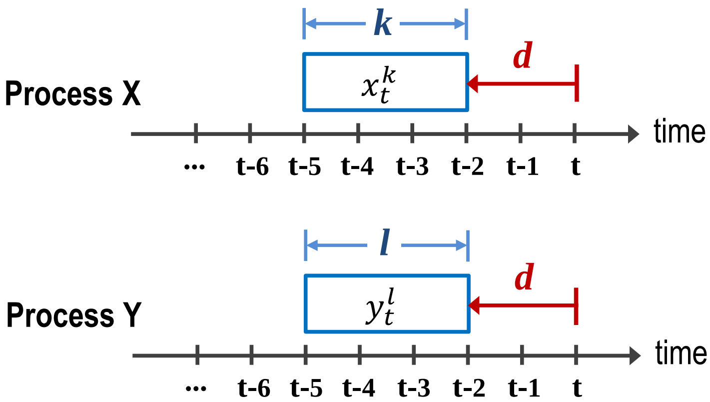


Figure 7.2: Transfer entropy with the effects of time-delay d and memory k, l between discrete-time stochastic processes **X** and **Y** ($\text{TE}_{Y\to X}$ or $\text{TE}_{X\to Y}$); d indicates the length of time shift (red arrows) from time t, while k and l denote the Markov orders (the length of blue frames) of the **X**'s past and **Y**'s past respectively such that $x_t^k = \{x_{t-1}, ..., x_{t-k}\}$ and $y_t^l = \{y_{t-1}, ..., y_{t-l}\}$.

7.3.3 Time-delay and Memory Effects

We consider time-delay and memory effects on diffusion, since heterogeneous populations exhibit different behavioral characteristics when adopting new information or ideas from other populations. For instance, as we estimated news diffusion in social media in Section 6.3, Blog shows more persistent behavior than News and SNS. That is, the growth rates of adoptions in Blog are relatively slower than the others (Blog tend to cite older news compared with News and SNS), while News and SNS exhibit rapid and similar growth patterns. Such common or different behavioral characteristics can reflect the time-delay and/or memory effects on information transfer. Besides, both effects cannot be estimated by our model-driven approach, and thus this approach can provide new perspectives on diffusion.

In this regard, we estimate the transfer entropy by varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k = l \leq 5$) so that we can analyze the effect of the d time-shifted recent l or k states of one social system on the future state of the other social system, as illustrated in Figure 7.2.

Accordingly, Equation (7.11) is modified as:

$$\begin{aligned}\text{TE}_{Y\to X} &= I(X_t; Y^l_{t-\mathbf{d}} \mid X^k_{t-\mathbf{d}}) \\ &= H(X_t|X^{t-\mathbf{d}:t-k-\mathbf{d}+1}) - H(X_t|X^{t-\mathbf{d}:t-k-\mathbf{d}+1}, Y^{t-\mathbf{d}:t-l-\mathbf{d}+1}) \\ &= \sum p(x_t, x^{(k)}_{t-\mathbf{d}}, y^{(l)}_{t-\mathbf{d}}) \log \frac{p(x_t|x^{(k)}_{t-\mathbf{d}}, y^{(l)}_{t-\mathbf{d}})}{p(x_t|x^{(k)}_{t-\mathbf{d}})} \quad , \end{aligned} \tag{7.13}$$

where d denotes the length of time-delay, and it is the only difference between Equation (7.11) and (7.13). In this study, we assume that the Markov orders of the **X**'s past and **Y**'s past are same, i.e. $k = l$, for simplicity.

With this model-free approach, we will estimate macro-level information transfer in the real-world diffusion in the next chapter. Also, the estimated diffusion patterns will be compared with outcomes from our model-driven approach in Part IV.

7.3.4 Applications of Information-theoretic Measures

More recently, information-theoretic measures have been used in predicting user behavioral patterns in social media which is one of our target application domains. For instance, the entropy is used for classifying Twitter user behaviors [60] and for predicting mobility in real [162] or virtual [159] human lives, the mutual information is used for predicting individual or group level future interactions [180], and the transfer entropy is used for detecting relationships between Twitter users independent of the knowledge of follower-followee social structures in Twitter [177, 178].

These studies are all from the aspects of intra-relationships within a single social platform alone. However, as discussed in Chapter 5, underlying networks in diffusion spans multiple social systems, and the effects of interactions between heterogeneous populations on diffusion are not ignorable. Our model-free approach extends a way of estimating micro-level inter-relationships in a single social system into a macro-level information transfer between heterogeneous systems. This approach can be comparable with our model-driven approach in that macro-level transfer can be considered as inter-relationships between two social systems as in the Dynamic Influence Model, i.e. $c_{ij}, i \neq j$ in Equation (5.10).

To the best of our knowledge, such macro-level information transfer has been first applied in both social media and academic publication domains. This can provides a new way of understanding cross-population diffusion without the need of knowledge of social network structures. In Part IV, our model-driven and model-free approaches are compared based on the estimation results in Chapter 6 and 8 with the discussions about the relations between the two approaches, which provides a more comprehensive understanding on global diffusion.

7.4 Conclusion

In this chapter, we introduced a conceptual framework for macro-level information transfer across social systems and accordingly defined a model-free approach using the transfer entropy. This is enabled by considering a stochastic process at a system level and by defining the system's signals at a population level. Identifying such signals is important to catch invisible information flow between two random processes. For this identification, we used the transfer entropy which provides the strength and directionality of influence between social systems.

For a more accurate estimation of information transfer, we do not ignore time-delay and memory effects on diffusion. These effects cannot be obtained by our model-driven approach, and thus our model-free approach can provide complementary new perspectives on diffusion. This model-free approach provides benefits of quantifying nonlinear dynamics with non-parametric estimation. However, the estimations are based on the probability distributions of target random variables from real data, which brings about biased estimations. In this regard, we use the Panzeri-Treves bias correction method for estimating all information-theoretic quantities.

The generality of our framework makes that it can be applied to a wide range of application domains for estimating macro-level information transfer. We expect that the proposed model-free approach can provide a different way of understanding global diffusion with news perspectives in addition to the model-driven approach. In the next chapter, we estimate real-word diffusion with our model-free approach. The estimated diffusion patterns in Chapter 8 are compared with outcomes from our model-driven in terms of the strength and directionality of influence in Part IV.

Chapter 8

Estimating Real-World Diffusion with a Model-Free Approach

In the previous chapter, we proposed a way of estimating macro-level information transfer between social systems with time-delay and memory effects. In this chapter, we estimate real-world diffusion with this model-free approach. For this study, we conduct experiments on real data from our target application domains, social media and academic publications, as we did in Part II. We then analyze the experimental results from macro-level information transfer in terms of the strength and directionality of influence across social systems as in our Dynamic Influence Model for further comparisons in Part IV. In addition, we examine time-delay and memory effects on diffusion which cannot be obtained from our model-driven approach, and finally we interpret the results. Based on the estimation results in Chapter 6 and 8, we will compare our model-driven and model-free approaches in the next chapter in depth.

8.1 Introduction

As discussed in Chapter 5, global diffusion emerges through collective behaviors of a large collection of heterogeneous populations, which makes it difficult to obtain a clear macroscopic picture of diffusion. In this regard, we suggested a meta-population scheme which clusters heterogeneous populations into a small number of relevant groups according to their similarity. The measures of similarity between populations can vary with application domains or research objectives. In Chapter 7, we considered a social system as a stochastic process, since it produces a time-series activity sequence. This concept can be applicable at a meta-population level when its individual members collectively generate time-evolving histories of behaviors in various real-world scenarios, such as daily adoption histories of a news item from News, SNS, and Blog, or yearly publication histories from Core, Data, and Systems. In this chapter, we consider a stochastic process at this meta-population level for estimating macro-level information transfer in our target application domains.

We conduct experiments with our model-free approach using the real datasets from Chapter 6. We also conduct statistical tests for all estimations in order to evaluate the significance of information-theoretic quantities. Estimated diffusion patterns

Table 8.1: Table of symbols for Chapter 8.

Symbol	Description
$\mathbf{X}, \mathbf{Y}$	Stochastic processes
$\mathbf{N}, \mathbf{S}, \mathbf{B}$	Stochastic processes of News, SNS, and Blog in social media respectively
$\mathbf{C}, \mathbf{D}, \mathbf{S}$	Stochastic processes of Core, Data, and Systems in academic publications respectively
k	The length of memory
d	The length of time-delay
$\mathrm{TE}_{Y \to X}$	Transfer entropy from $\mathbf{Y}$ to $\mathbf{X}$
$\mathrm{TE}_{\mathrm{out}}$	Outgoing transfer entropy
$\mathrm{TE}_{\mathrm{in}}$	Incoming transfer entropy

are analyzed in terms of the strength and directionality of influence and are also interpreted from the aspects of behavioral characteristics of meta-populations.

From the experiments on social media data, distinct and common behavioral patterns are observed across News, SNS, and Blog. For instance, Blog is more influenced by longer trajectories of adoption trends of a source system, while News and SNS show relatively weak memory effects. Regarding news topics, the culture, disaster, and sports categories show the strongest memory effects, while the science category shows the weakest. In terms of strength and directionality of influence, the politics category leads to relatively balanced inter-relationships across the systems, which is in contrast to the culture and technology categories.

From the experiments on academic publication data, in general longer histories of publication trends of a source system are more influential to a destination system. In terms of research keywords, relatively time-insensitive topics (e.g., "High Performance" and "Fault Tolerant") show a stronger memory effect than time-sensitive topics (e.g., "Web Service" and "Social Network"). In terms of the strength and directionality of influence, the keywords "Real Time" and "Large Scale" show balanced inter-relationships in contrast to the keywords "Web Service" and "Semantic Web".

In common across the two domains, the effects of time-delay and memory, as well as diversity of information topics, all need to be considered as important factors of diffusion. In other words, meta-populations exhibit different behavioral characteristics which not only lead to different diffusion patterns but are also governed by different information topics. Such macro-level observations in our application domains, to the best of our knowledge, are seen for the first time. We expect that these applications of our model-free approach can provide a different way of estimating macro-level diffusion from our model-driven approach and a way of understanding behavioral characteristics of target systems.

The rest of this chapter is organized as follows. Section 8.2 presents the method of a statistical significance test to evaluate estimations with our model-free approach.

Based on this statistical test, Section 8.3 and 8.4 estimate and interpret macro-level information transfer in social media and academic publications, respectively. Section 8.5 discusses the model-free approach and experimental results, and finally Section 8.6 concludes this study.

8.2 Statistical Significance of Information Transfer

Information-theoretic measures quantify the information of a random variable, a collection of variables, or exchanges between variables based on the probability distributions of the target random variables. However, these distributions are approximated by limited observations from real data, which makes such measurements random. Thus, we need to evaluate the significance of information-theoretic quantities.

For this evaluation, we use statistical significance tests as in prior work [99, 177, 180]. The null hypothesis H_0 is that the future state of a destination random process $\mathbf{X}$ is not predictable by the knowledge of the previous states of a source process $\mathbf{Y}$. We construct a null distribution from surrogate data by shuffling $\mathbf{Y}$'s ordered elements 1,000 times. This enables the surrogate data to follow the same probability distributions with the source process $\mathbf{Y}$, but to diverge from the temporal dependence of destination on the source. We then compare p-value with α with a 95% confidence interval. When we reject H_0, i.e. $p < 0.05$, we consider the estimation is significant.

Based on this significance testing, we analyze the experimental results from our model-free approach in the rest of this chapter.

8.3 Macro-level Information Transfer in Social Media

In this section, we estimate macro-level information transfer in social media by using the model-free approach proposed in the previous chapter. For this study, we first describe data preparation, conduct experiments on real data, and analyze topic-related diffusion trends from diverse aspects.

8.3.1 Data Preparation

In Section 6.3, we described the real datasets used for analyzing diffusion patterns in social media with our Dynamic Influence Model. In the previous experiments, we selected the 63 news topics in Table 6.2, each of which has driven adopters of at least 150 identified users across different types of social media, i.e. News, SNS, and Blog. Corresponding to the selected topics, we generated 63 real datasets, and each dataset consists of daily cumulative adopters of News, SNS, and Blog during a one month period as $\{A_i(t)\}_{t=1}^{30}$, $i = n, s, b$, where n, s, and b denote News, SNS, and Blog, respectively.

In this section, for the experiment with our model-driven approach, we also use these 63 datasets after additional modifications as follows. As we discussed in Section 7.3, we consider a social system as a discrete-time stochastic process which pro-

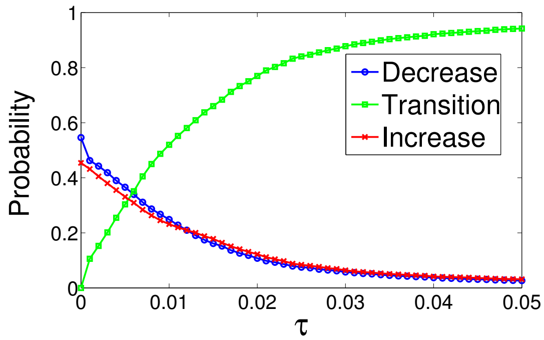


Figure 8.1: Probability of states (decrease, transition, and increase) of stochastic processes, i.e. News, SNS, and Blog, as a function of τ in Equation (7.9).

duces new adoption rates $f(t)$ at time t. Also, as we mentioned in Section 8.1, we consider this stochastic process at a meta-population level for obtaining a more simple and clearer picture of macro-level diffusion. Accordingly, we convert each dataset $\{A_i(t)\}_{t=1}^{30}$ into time-series new adoption rates as $\{f_i(t)\}_{t=1}^{30}$, $i = n, s, b$. We finally discretize the new adoption rates $f(t)$ into three discrete states, i.e. *decrease, transition*, and *increase* according to Equation (7.9). The value of τ in the equation is determined as $\tau = 6.0e\text{-}3$ using all the datasets so that the three states are the most equally likely, as shown in Figure 8.1. With these 63 datasets, we conduct experiments in the following subsections.

8.3.2 Transfer Entropy with Time-delay and Memory Effects

As discussed in the previous chapter, mutual information cannot capture the directionality of influence between two stochastic processes, **X** and **Y**. Transfer entropy overcomes this limitation. We estimate the transfer entropy by varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k = l \leq 5$) so that we can analyze the effect of the d time-shifted most recent l or k states of one social system on the future state of the other social system, as illustrated in Figure 7.2.

We consider the three different online social systems, News, SNS, and Blog, as discrete-time random processes, **N**, **S**, and **B**, respectively. Accordingly, we estimate transfer entropy (TE) of all possible combinations of the processes, i.e. $\text{TE}_{N\to S}$, $\text{TE}_{N\to B}$, $\text{TE}_{S\to N}$, $\text{TE}_{S\to B}$, $\text{TE}_{B\to N}$, and $\text{TE}_{B\to S}$, for the selected 63 news topics, as shown in Figure 8.2. The first plot (ALL) in the figure is obtained by aggregating all the cases to show an overall trend. As discussed in Section 8.2, all estimations are evaluated by statistical significance tests, and thus only significant quantities are considered. Each cell in Figure 8.2 indicates the proportion of the significant transfer entropy, and it is expressed in a gray-scale map with 15 variants (3 variants of time-delay $\times$ 5 variants of memory), where white and black denote the maximum (1) and minimum (0) values, respectively. Also, the framed cells present the highest value in each map, and they are color-coded according to the map positions. As the first matrix (ALL) shows, in general a more recent and longer activity sequence of a process more likely influ-

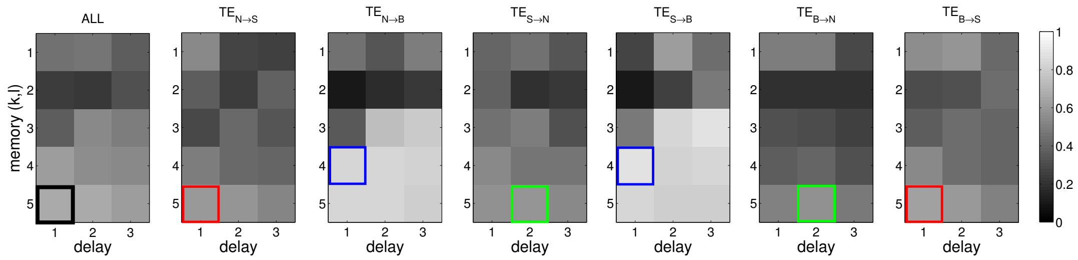


Figure 8.2: Transfer entropy between online social systems (N:News, S:SNS, and B:Blog) by varying the length of time-delay (d) and memory, k and l ($k = l$) in Figure 7.2. Each cell represents the ratio of significant cases out of all 63 news topics in Table 6.2, where its transfer entropy is greater than those of 1,000 shuffled data with 95% confidence intervals, which is gray-color coded (white for the maximum value 1, and black for the minimum value 0). The framed cells present the highest value among 15 variations in each graph, and they are color-coded by the cell positions.

ences other processes; when $d = 1$ and $k = l = 5$, the proportion of the significant transfer entropy is the highest.

$\text{TE}_{N\to B}$ and $\text{TE}_{S\to B}$ exhibit the highest color contrasts along the y-axis compared with the others, which means that the Markov order (memory size) is an important factor when estimating diffusion in the Blog system. That is, the longer trajectories of adoption trends (at least four days long) in the News and SNS systems have a stronger effect on diffusion in the Blog system. Also, diffusion in the News and SNS systems is more likely influenced by longer histories of adoptions in other systems, but the changes are relatively small compared to the Blog system. Thus, in terms of time-delay and memory configuration, setting $d = 1$ and $k = l = 1$ (the latest one-day adoption history of a system) can be used to estimate the transfer entropy when a destination system is SNS or News, i.e. $\text{TE}_{N\to S}$, $\text{TE}_{B\to S}$, $\text{TE}_{S\to N}$, and $\text{TE}_{B\to N}$, while setting $d = 1$ and $k = l = 4$ can be used for estimating the transfer entropy when a desination system is Blog, i.e. $\text{TE}_{N\to B}$ and $\text{TE}_{S\to B}$.

In addition, color-coded cells in Figure 8.2 show common information destinations in the framework of Shannon's information theory. That is, information transfer in social media reflects the properties of destination processes rather than source processes. In other words, behavioral characteristics of a destination system determine the trends of news diffusion in social media. As discussed in Section 2.2.5, node characteristics (e.g., individual interests or receptivity of information) in a social network are one of the main diffusion factors. In this context, a destination system's behavioral characteristics can be considered as node characteristics at a meta-population level and thus as an effect on macro-level diffusion in social media.

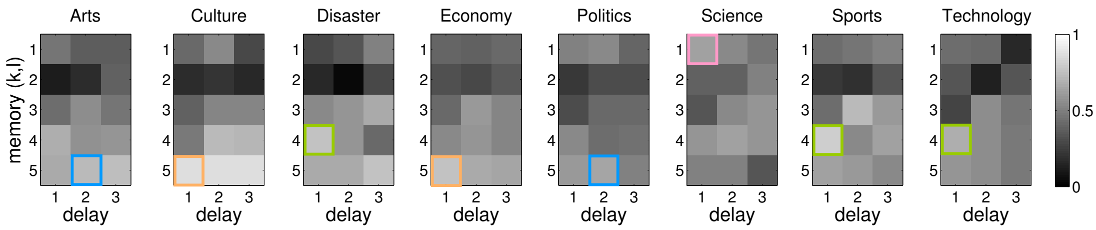


Figure 8.3: Transfer entropy between online social systems (News, SNS, and Blog) for each news category, by varying the length of time-delay (d) and the size of memory, k and l ($k = l$) in Figure 7.2. Each cell indicates the ratio of significant cases out of all possible cases, where its transfer entropy is greater than those of 1,000 shuffled data with 95% confidence intervals. The framed cells present the highest value in each graph, and they are color-coded by the cell positions.

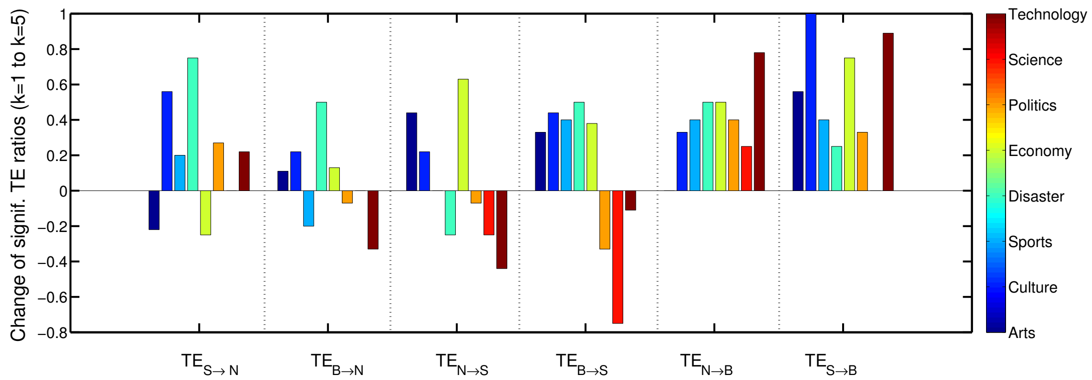


Figure 8.4: Changes of significant transfer entropy (TE) ratios for all 63 news topics when increasing the memory size from $k = l = 1$ to $k = l = 5$ in Equation (7.11) and fixing the time-delay as $d = 1$ in Figure 7.2. The eight news categories are color-coded (N: News, S: SNS, and B: Blog).

8.3.3 Topic-related Diffusion: Time-delay and Memory Effects

We now investigate topic-related diffusion patters for each news category considering the effects of time-delay and memory on the diffusion.

Figure 8.3 illustrates the proportions of the significant transfer entropy for each news category, showing different or common patterns among the eight categories. In this figure, cells with the highest value in each map are color-coded according to their cell positions, and they can be roughly divided into four groups based on their common cell positions of time-delay and memory as: (1) disaster, sports, and technology, (2) arts and politics, (3) culture and economy, and (4) science. As the figure shows, when increasing the size of memory, the changes of the significant TE ratios are larger in the categories of culture, disaster, and sports than the other categories, exhibiting the positive effects of a longer memory size. On the other

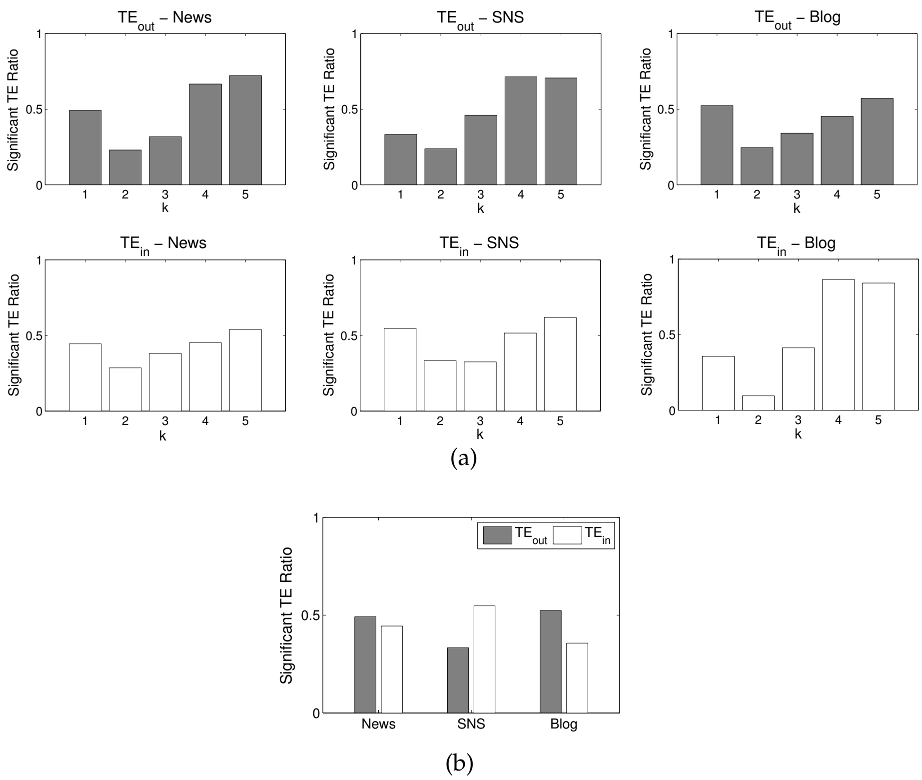


Figure 8.5: **(a)** Outgoing and incoming transfer entropy (TE) for each media type by varying the length of k (=l) in Equation (7.11). The y-axis represents the proportions of significant cases whose transfer entropy is greater than those of 1,000 shuffled data with 95% confidence intervals. **(b)** Comparison of outgoing and incoming transfer entropy for each media type when $k = l = 1$ in Equation (7.11). The y-axis represents the same as (a).

hand, the science category is not too much dependent on the memory size, and its diffusion trend can be estimated with the latest one-day adoption records in other social systems. As shown in Table 6.2, the culture category consists of controversial topics, the disaster category has abrupt but continuous natural disasters, and the sports category contains periodic sports games. These categories draw continuous attention, which convinces the stronger effects of the longer memory size.

To be more specific, Figure 8.4 shows the changes of the significant TE proportions when increasing the memory size k from 1 to 5. As the figure shows, the increase of the memory size generally comes up with positive effects on diffusion. In particular, the Blog system exhibits the positive changes across all categories ($\mathrm{TE}_{N\to B}$ and $\mathrm{TE}_{S\to B}$), and regarding the culture category the ratio increases by more than 80%. On the contrary, the SNS system shows a majority of negative changes ($\mathrm{TE}_{N\to S}$ and $\mathrm{TE}_{B\to S}$), and for the science category the proportion of significant $\mathrm{TE}_{B\to S}$ substantially decreases by 75%. These results are consistent with the previous findings: the strongest effect of a longer memory size in the Blog system ($\mathrm{TE}_{N\to B}$ and $\mathrm{TE}_{S\to B}$) in

(a) Short-term history (1 day) (b) Longer-term history (5 days)

Figure 8.6: Memory effects on the directionality of influence when the memory size k in Equation 7.11 increases from 1 to 5. **(a)** Baseline influence (k=1). **(b)** Changes of influence when k increases from 1 to 5. Arrows represent the strength and directionality of influence based on the findings in Figure 8.5; black solid arrows indicate stronger influence than dashed arrows in (a), and red solid arrows in (b) are for representing the distinct changes in the strength of influence when k increases.

Figure 8.2, and the opposite patterns between the culture (highest proportion when $k = 5$) and science (highest proportion when $k = 1$) categories in Figure 8.3.

8.3.4 Topic-related Diffusion: Strength and Directionality of Influence

Let us examine information transfer in terms of the outgoing and incoming transfer entropy for different media types as well as news categories, by varying the memory size k with the time-delay d set as 1. Figure 8.5(a) shows the estimated outgoing (TE_{out}) and incoming (TE_{in}) transfer entropy for each social system (e.g., $\text{TE}_{N \to S}$ and $\text{TE}_{S \to N}$ denote outgoing and incoming transfer entropy of News to SNS and of News from SNS, respectively). In this figure, the y-axis represents the proportion of the significant transfer entropy when varying the memory size k in the x-axis. As the figure shows, in general, the predictability of diffusion in one social system likely increases as the previous states of the other social system are known as much as possible (longer memory). However, there are differences between the systems. News and SNS more likely influence other social systems (more significant outgoing transfer entropy of News and SNS), and Blog tends to be more influenced by other social systems (more significant incoming transfer entropy of Blog) as the memory size increases.

In particular, when it comes to $k = l = 1$ in Figure 8.5(b), the baseline directionality of influence shows that News is neutral (both outgoing and incoming), SNS is incoming, and Blog is outgoing. This is the simplest memory-less case, where the latest one-day previous states of two stochastic processes are only considered for estimating information transfer between them. Accordingly, these behavioral characteristics are visualized in Figure 8.6(a), and the changes of the baseline influence are represented in Figure 8.6(b).

Figure 8.7 estimates the outgoing (TE_{out}) and incoming (TE_{in}) transfer entropy, separated by different news categories, when $k = l = 1$ and $d = 1$. The x-axis indicates news topics for each category, and they are repeated for each system, which are divided by vertical lines. As the figure shows, for the culture category, News and Blog have higher outgoing transfer entropy than that of SNS, while the incoming

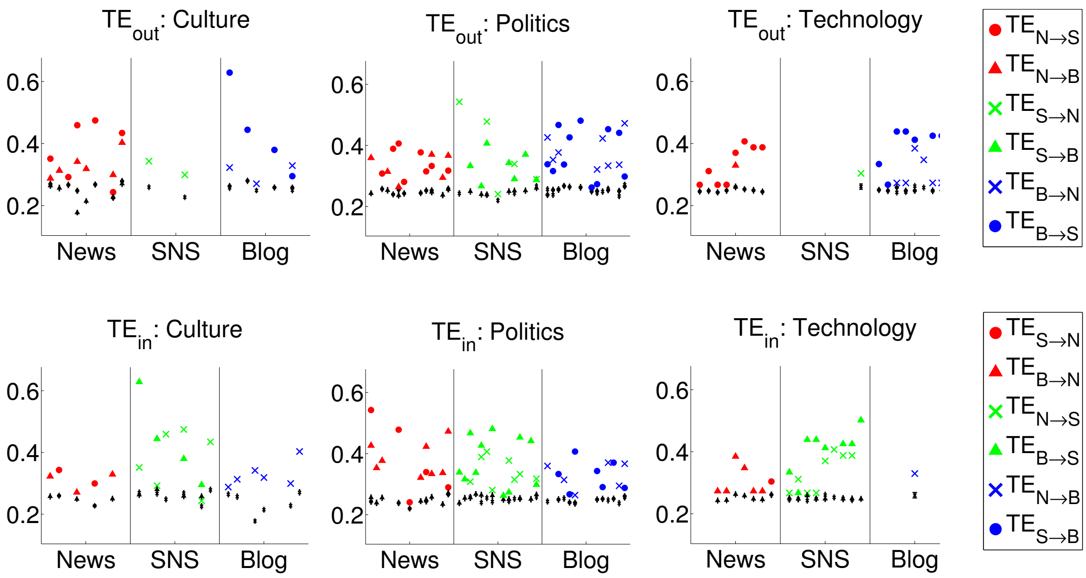


Figure 8.7: Outgoing (TE_{out}) and incoming (TE_{in}) transfer entropy for different news categories when k=l=1 and d=1 in Equation 7.13 (baseline condition). The x-axis indicates the news index for the topics in a corresponding category as shown in Table 6.2, which is repeated for each social system (vertical lines). Thus, there is no special ordering in the news index within each system. Different social systems are color-coded, and the directionality of influence (target systems in TE_{out} and source systems in TE_{in}) is also distinguished by shapes. The y-axis represents the values of TE_{out} and TE_{in} (color-coded markers), and the 95% confidence intervals of 1,000 shuffled data are shown in black. Note that only significant cases are shown for clarity.

transfer entropy of SNS is generally higher than those of the others, i.e. N→S←B. Similarly, regarding the technology category, we can find the directionality of information flow as N→S←B→N. On the other hand, the politics category exhibits the relatively balanced interactions between different media types, compared to the other categories. The estimated diffusion patterns with our model-free approach in this section will be compared with outcomes from our model-driven approach in Chapter 9.

8.4 Macro-level Information Transfer in Academic Publications

In this section, we estimate macro-level information transfer in academic publications with the model-free approach as in the previous section. For the experiments on our second application domain, we first describe data preparation, conduct experiments on real data, and analyze topic-related diffusion trends from diverse aspects.

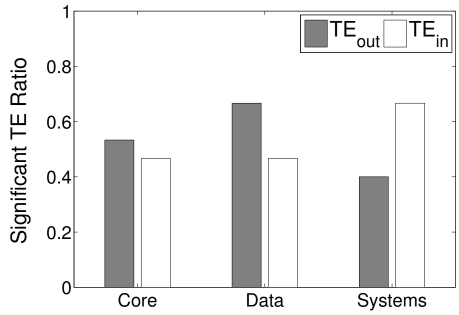


Figure 8.8: The proportions of significant cases of outgoing and incoming transfer entropy (TE) for each cluster when $k = 1$ and $d = 1$.

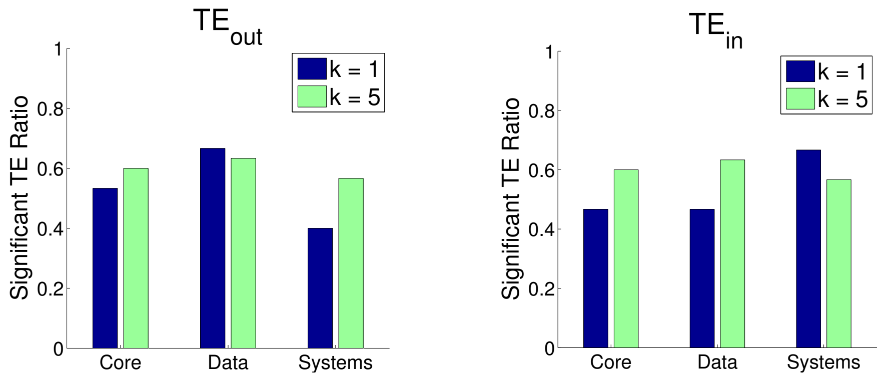
8.4.1 Data Preparation

In Section 6.4, we described the real datasets for estimating diffusion in academic publications with our model-driven approach. We targeted the top-ranked 221 conferences in computer science and classified them into the three clusters, Core, Data, and Systems, each of which consists of closely connected subdomains, as shown in Table 4.6. We consider these clusters as meta-populations for obtaining a macroscopic picture of diffusion as in the case of social media. For understanding topic-related diffusion, we selected 15 research keywords in Table 6.4 among the top 100 popular keywords by referring to Microsoft Academic Search (MAS) [5]. These selected keywords are widely used across all clusters; for each keyword, at least 10% of conferences in a cluster have published papers containing that keyword. Accordingly, we generated 15 datasets, each of which consists of yearly cumulative adopters in a meta-population i during the recent 30years as $\{A_i(t)\}_{t=1}^{30}$, $i = c, d, s$, where c, d, and s denote Core, Data, and Systems, respectively.

In this experiment, we also use these 15 datasets with additional modifications as described in the previous section. We convert each dataset $\{A_i(t)\}$ into time-series new adoption rates as $\{f_i(t)\}_{t=1}^{30}$, $i = c, d, s$. We finally discretize the new adoption rates $f(t)$ into three discrete states according to Equation (7.9). The value of τ in this equation is determined as $\tau = 1.0e\text{-}3$ as in the way we did in social media in Section 8.3.1. With these datasets, we conduct experiments for understanding macro-level information transfer in academic publications in the following subsections.

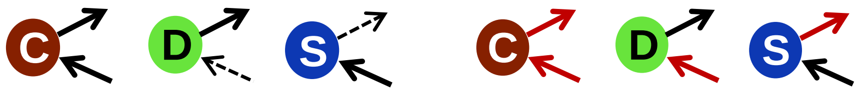
8.4.2 Transfer Entropy with Time-delay and Memory Effects

We estimate the transfer entropy by varying the time-delay d and the memory size k as shown in Figure 7.2 so that we can analyze the effect of the d time-shifted recent k states of a stochastic process on the future state of the other process. For this study, the three clusters, Core, Data, and Systems, in Table 4.6 are considered as discrete-time stochastic processes, **C**, **D**, and **S**, respectively. We estimate transfer entropy of six possible combinations of the processes such as $\text{TE}_{C \to D}$, $\text{TE}_{C \to S}$, $\text{TE}_{D \to C}$, $\text{TE}_{D \to S}$, $\text{TE}_{S \to C}$, and $\text{TE}_{S \to D}$, by varying the time-delay $(1 \leq d \leq 3)$ and memory size



(a) Outgoing transfer entropy (b) Incoming Transfer Entropy

Figure 8.9: The proportions of significant cases of outgoing and incoming transfer entropy (TE) for each cluster by varying the length of memory k from 1 to 5 (d=1).



(a) memory-less ($k = 1$) (b) with memory ($k = 5$)

Figure 8.10: Summary of memory effects on diffusion (d=1); (a) solid arrows exhibit the stronger tendency than dashed arrows, and (b) red arrows indicate distinct changes in the tendency when increasing k from 1 to 5. (a) and (b) correspond to Figure 8.8 and 8.9 respectively.

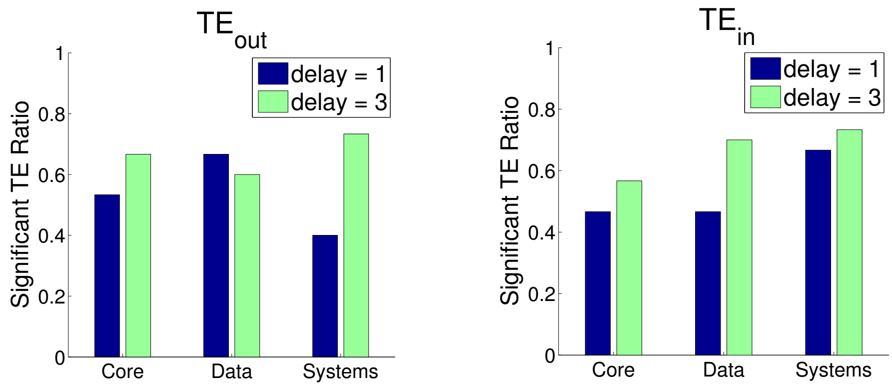
$(1 \leq k \leq 5)$. These six possible cases of transfer entropy are estimated for each selected keyword, but only significant transfer entropy is considered by conducting the statistical significance tests discussed in Section 8.2. Otherwise, the synchronization of diffusion patterns between two stochastic processes is thought of as a coincidence.

Figure 8.8 shows the proportions of the significant transfer entropy for each cluster in terms of the outgoing (TE_{out}, influencing other clusters) and incoming (TE_{in}, being influenced by other clusters) transfer entropy when k=1 and d=1. This figure reflects the fundamental directionality of influence for a cluster, since it is based upon the most recent state of a process (d=1) without information of its activity sequence, i.e. memory-less (k=1). As the figure shows, Core is both outgoing and incoming (the proportions of outgoing and incoming transfer entropy are similar), which implies that a Core's recent state of publications likely influences other clusters and is also influenced by others. In other words, Core keeps a balance between leading and following the other clusters. On the other hand, Data and Systems show the opposite influence directions; Data more likely influences others (leading the other clusters), while Systems is more likely influenced by others (following the other clusters).

However, such behavioral characteristics of each cluster change when we vary the length of memory. As Figure 8.9(a) shows, the proportions of the significant TE_{out}

Table 8.2: Behavioral characteristics of clusters based on the estimated transfer entropy as the time-delay d and the memory k increase.

Cluster	**Baseline Influence** (k=d=1)	**Memory Effect** (as k grows)	**Time-delay Effect** (as d grows)
Core	neutral	more leading & more following	more leading & more following
Data	leading	more following	more following
Systems	following	more leading	more leading & more following



(a) Outgoing transfer entropy (b) Incoming transfer entropy

Figure 8.11: The proportions of significant cases in outgoing and incoming transfer entropy (TE) for each cluster by varying the length of time-delay d from 1 to 3 (k=1).

increases as the memory size k increases from 1 to 5. In general, a longer activity sequence of a cluster tends to more influence other clusters except for Data. This means that the longer we know about the past popularity of a research topic, the better we can tell its popularity in the present. In other words, academic publications are generally based on longer research trends. Particularly, for the Data cluster, its state-of-the-art more likely influences others. In terms of incoming transfer entropy TE_{in} in Figure 8.9(b), Core and Data become more incoming when increasing the memory size k, except for Systems. That is, Data has a tendency to follow longer histories of publications of others, while its short-term history tend to more influence others. In contrast, Systems shows the opposite patterns. All in all, the longer trajectories of publication trends have effect on diffusion in academic publications; systems becomes more outgoing, Data becomes more incoming, and Core becomes both more outgoing and more incoming. Figure 8.10 accordingly summarizes the memory effects on diffusion when $k = 1$ (baseline influence) and $k = 5$, showing the changes of each cluster's behavioral characteristics.

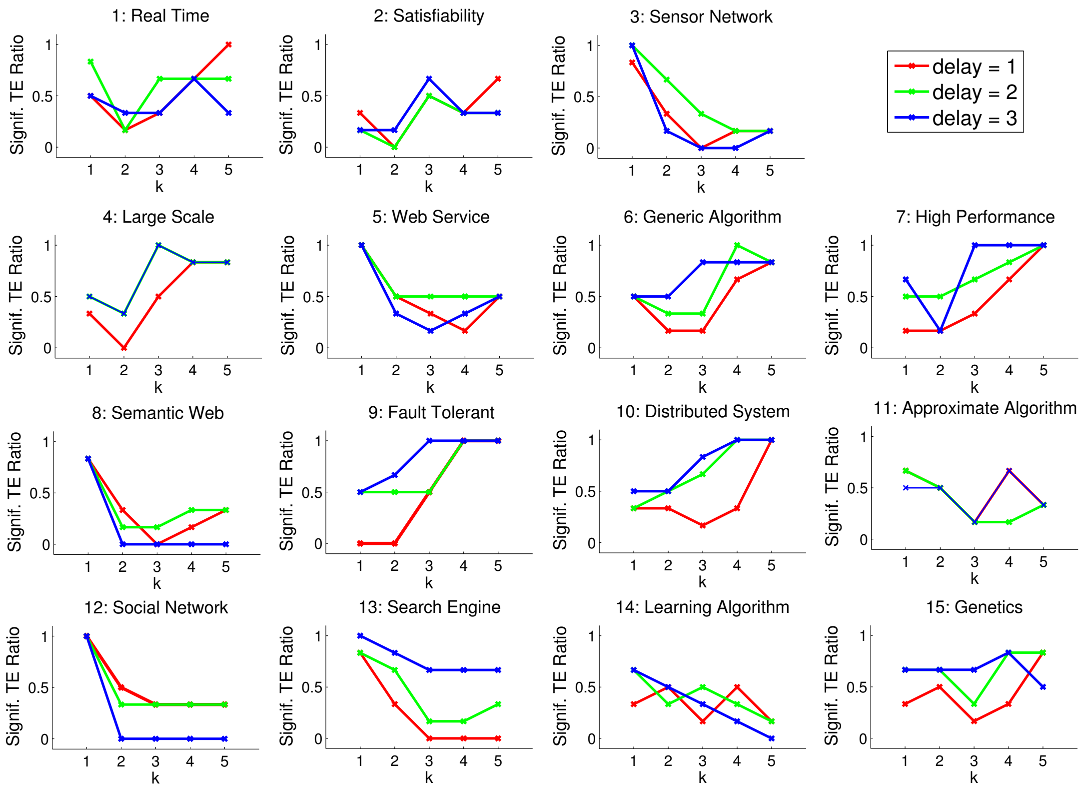


Figure 8.12: The proportions of significant transfer entropy (TE) for the selected research keywords when varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). The x-axis indicates the memory size k, and the y-axis represents the proportions of significant transfer entropy. Time-delay is color-coded in each graph.

We also examine time-delay effect on diffusion by varying the length of the time-delay d from 1 to 3, as shown in Figure 8.11. Regarding the outgoing transfer entropy in Figure 8.11(a), the very recent publication states ($d = 1$) of a cluster are less influential, except for Data. For incoming transfer entropy in Figure 8.11(b), all clusters show more persistent behavior in that they are more influenced by not very recent but earlier publication states of others. Accordingly, Table 8.2 summarizes the effects of both memory and time-delay.

8.4.3 Topic-related Diffusion: Time-delay and Memory Effects

Now, we examine topic-related diffusion patterns for the selected research keywords in Table 6.4 from the aspects of time-delay and memory effects. Figure 8.12 shows the changes of the significant TE proportions for each keyword, by varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). As the figure shows, both effects on diffusion vary with different keywords.

In terms of the memory effect, there are mainly two opposite patterns, positive and negative effects of a longer memory size on information transfer (i.e. the increase and decrease of the significant TE ratios when increasing the memory size k). The positive effect is observed in the keywords “Real Time”, “Satisfiability”, “Large Scale”, “Genetic Algorithm”, “High Performance”, “Fault Tolerant”, “Distributed System”, and “Genetics”. On the other hand, the negative effect is shown in the keywords “Sensor Network”, “Web Service”, “Semantic Web”, “Social Network”, “Search Engine”, “Approximate Algorithm”, and “Learning Algorithm”. Interestingly, the second group of keywords are relatively time-sensitive research topics which need to keep up to date with recent trends of the literature due to the quickly advancing web environment. On the other hand, the first group of keywords are closely related to topics on fundamental computational theories in computer science, whose original concepts are hardly changing, and thus more likely traced and cited than the others.

Also, there are variations in time-delay d across all the topics. In general, publication trends of a cluster start to influence other clusters at least after 2 years ($d >= 2$). Meanwhile, the strongest time-delay effects are observed in the keywords “Fault Tolerant”, “Distributed System”, and “Search Engine” which exhibit the relatively distinct changes in the significant TE proportions when increasing time-delay. These keywords are also influenced by memory effects. Thus, for a better understanding of information transfer in academic publications, we need to consider both time-delay and memory effects as well as diversity of research topics.

8.4.4 Topic-related Diffusion: Strength and Directionality of Influence

Figure 8.13 visualizes macro-level information pathways across the Core, Data, and Systems clusters for each keyword, based on the estimated information transfer. The link weights in each triad represent the averaged significant transfer entropy for all 15 variants of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$).

From this visualization, we can obtain a macroscopic view of information flows in academic publications, rather than the detailed values of link weights. This figure can be compared with Figure 6.9 which is based on the trained parameter values of the Dynamic Influence Model in Chapter 6. As the two figures illustrate, diffusion patterns resulting from the model-driven and model-free approaches exhibit very similar trends in terms of strength and directionality of influence. They do not show exact matches, but overall trends are comparable. For instance, in Figure 8.13, the keywords “Real Time” and “Large Scale” show relatively balanced inter-relationships between clusters, while the keywords “Web Service” and “Semantic Web” exhibit unbalanced inter-relationships. These patterns are also observed in our model-driven approach in Chapter 6.

Such similar patterns suggests that two approaches can help each other to determine its own threshold values of link weights so as to maximize the similarity of outcomes from the approaches and to obtain a more meaningful diffusion patterns. Detailed comparisons will be discussed in the next chapter.

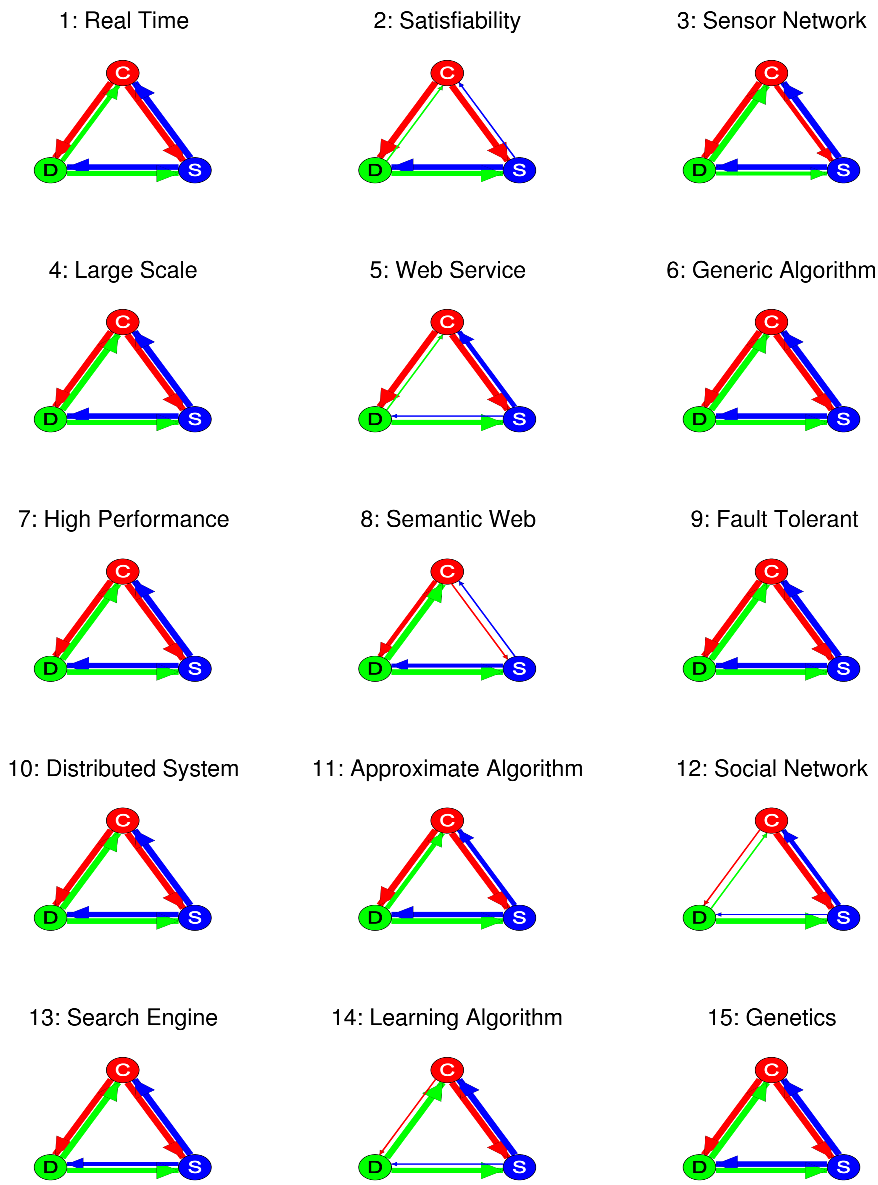


Figure 8.13: Information pathways across clusters, Core, Data, and Systems based on the estimated transfer entropy which are averaged for all time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). Node labels represent the initials of the clusters, and link width and arrows indicate the strength and directionality of influence between clusters. Each cluster and its outgoing links are color-coded.

8.5 Discussion

Information-theoretic measures provides us with benefits of quantifying nonlinear dynamics of complex systems. That is because such quantifications do not require any assumptions on underlying network structures in diffusion, which makes such measures model-free. Based on this flexible setting of information theory, we proposed a conceptual framework where a stochastic process is defined at a system level, and a social system's signal is defined at a population level. This enables us to estimate macro-level information transfer and to obtain a more simple and clearer picture of global diffusion.

By conducting experiments on real data from different application domains, we find that time-delay and memory effects need to be considered as diffusion factors. Both effects reflect the behavioral characteristics of a population in a social system. For instance, the memory effect reveals the degree of persistence of a destination process, which means that a stronger memory effect reflects a tendency to follow a longer activity sequence of a source process: e.g., Blog in social media, and Core and Data in academic publications. In addition, the time-delay effect reflects the timing of influence of a source process on a destination process. For example, in social media the most recent previous states of a process influence others more strongly. In academic publications, it takes more than a year for a source process's activity sequence to influence a destination process.

Heterogeneous populations exhibit different behavioral characteristics which are governed by different information topics. Thus, the diversity of information topics is not ignorable for a better understanding of diffusion. In social media, the politics category shows more balanced inter-relationships across News, SNS, and Blog than the culture and technology categories. In academic publications, the keywords "Real Time" and "Large Scale" show balanced inter-relationships in contrast to the keywords "Web Service" and "Semantic Web". Some keywords on computational theories (e.g., "High Performance", "Fault Tolerant", and "Distributed System") show stronger memory effects than the keywords on relatively time-sensitive topics (e.g., "Web Service", "Social Network", and "Search Engine") which need to keep up with recent trend of the literature due to frequently changing web environment.

Behavioral characteristics of a system lead to different diffusion patterns. Such properties of systems are obtained by quantifying the transfer entropy with time-delay and memory effects which cannot be obtained by our model-driven approach. Comparisons between model-driven and model-free approaches will be discussed in detail in the next chapter.

8.6 Conclusion

In this chapter, we conducted experiments on real data with the model-free approach proposed in Chapter 7 and analyzed macro-level information transfer with time-delay and memory effects. In common between the two application domains, the effects of time-delay and memory, as well as diversity of information topics, govern diffusion patterns across heterogeneous social systems.

From the experiments on social media data, we observed distinct and common behavioral characteristics among News, SNS, and Blog. In general, more recent and longer activity sequences of a process more likely influence other processes. The strongest memory effect on diffusion is observed in the Blog system, while relatively weak memory effects are observed in the News and SNS systems. That is, Blog is more influenced by longer trajectories of adoption trends of News and SNS. In terms of news topics, the culture, disaster, and sports categories show the strongest memory effects, while the science category shows the weakest. In terms of the directionality of influence, News is neutral (both outgoing and incoming), SNS is incoming, and Blog is outgoing when the length of memory k and time-delay d is a minimum ($k = d = 1$). However, as the memory size increases, SNS becomes more outgoing, and Blog is more incoming, showing the persistent behavior of Blog. In terms of the strength and directionality of influence, the politics category leads to more balanced influences across the systems than the culture and technology categories.

Regarding the experiments on academic publications, baseline influence ($k = d = 1$) showed that Core is neutral (both outgoing and incoming), Data is outgoing, and Systems is incoming. When increasing the memory size, Systems becomes more outgoing, Data becomes more incoming, and Core becomes both more outgoing and more incoming. Thus, longer trajectories of publication trends have stronger effects on diffusion in academic publications. In terms of research keywords, relatively time-insensitive topics (e.g., "High Performance" and "Fault Tolerant") show stronger memory effect than time-sensitive topics (e.g., "Web Service" and "Social Network"). Time-sensitive topics have the negative effects of a longer time-delay where the most recent previous state of a process is the most influential. In terms of the strength and directionality of influence, the keywords "Real Time" and "Large Scale" show more balanced inter-relationships than the keywords "Web Service" and "Semantic Web", which is consistent with the outcomes from our model-driven approach.

We applied our model-free approach to our target application domains as in Part II. We expect that the way of applying our approach to these domains can helps to understand macro-level diffusion in different real-word scenarios. In the next chapter, the estimated diffusion patterns from the model-free approach will be compared with results from our model-driven approach.

Part IV

Comparison of Approaches

Chapter 9

Comparisons of the Model-Driven and Model-Free Approaches

In this chapter, we compare our proposed model-driven and model-free approaches with the estimated diffusion patterns in social media and academic publications. For this comparison, we first extract distinct diffusion patterns based on the experimental results from Chapter 6 and 8 so that we can examine how these two approaches uncover the patterns from their own perspectives on diffusion. For each application domain, we analyze and interpret the outcomes from the approaches and try to find common insights between the domains.

9.1 Introduction

In Part I of this thesis, we conducted empirical studies on real-world diffusion in social media and academic publications, and we observed that social interactions are far-reaching across multiple social systems. This makes the discovery of the diffusion mechanisms challenging, since the underlying networks are heterogeneous and thus hard to collect or define the network structures.

In this context, in Chapter 5 we proposed the Dynamic Influence Model which extends the internal influence beyond a homogeneous population into heterogeneous populations by incorporating the heterogeneity and structural connectivity of underlying social networks in a probabilistic way. In Chapter 6, using this proposed approach, we estimated the real-world diffusion patterns in our target application domains and observed topic-related diffusion patterns. This model is based on the assumption of a real-world network property, a power-law degree distribution. Thus, it does not require the detailed network structures but needs to estimate the parameter value of scaling a power-law tail.

In contrast to this model-driven approach, in Chapter 7 we proposed the model-free approach using information-theoretic measures, and thus it is independent of any assumptions on social interactions. This approach enables us to estimate macro-level information transfer across social systems by considering a stochastic process at a system level and defining a system's signal at a population level. In Chapter 8, using this proposed approach, we estimated the transfer entropy with the time-delay and memory effects, which provides different aspects of diffusion from the model-driven approach.

In this chapter, based on the experimental results from Part II and Part III, we compare our model-driven and model-free approaches from the aspects of the strength and directionality of influence across heterogeneous social systems. For this comparison, we first extract distinct diffusion patterns from the estimations so that we can examine how these two approaches uncover the patterns and how the resulting trends are different or common between them.

From this comparison study, we find that in general transitive relations have more opportunities to bring about similar diffusion patterns across social systems, but the strength as well as balance of intra- and inter-relationships are also important to lead to global diffusion. That is, synchronous diffusion needs strong intra-relationships as momentum to make individuals to move to a certain state, i.e. critical mass, which is triggered by strong inter-relationships between different social systems. Dynamic influence with different strength and directionality leads to unlimited diffusion patterns, but from the patterns we can draw an overall picture of momentum formed among heterogeneous populations. We expect that this comparison study can provide a way of understanding macro-level diffusion in a comprehensive way.

The rest of this chapter is structured as follows. Section 9.2 and 9.3 compare the model-driven and model-free approaches based on the estimation results in social media and academic publications, respectively. Section 9.4 discusses the outcomes that the two approaches provide, and finally Section 9.5 concludes this chapter.

9.2 News Diffusion in Social Media

In this section, we compare our model-driven and model-free approaches based on the estimated news diffusion patterns in social media from Section 6.3 and 8.3. The comparisons are conducted by focusing on the distinct diffusion patterns across the News, SNS, and Blog systems so that we can examine how these two approaches distinguish the patterns with their own perspectives on diffusion. We first examine the outcomes from the model-driven approach regarding the external and internal influences of the Dynamic Influence Model. We then analyze the results from the model-free approach with respect to the time-delay and memory effects of macro-level information transfer. Finally, we compare the estimated diffusion patterns in terms of the strength and directionality of influence. The first two and the last investigations correspond to different and common aspects of the two approaches, respectively.

9.2.1 Diffusion Patterns from Model-Driven Approach

In Section 6.3, we conducted experiments on social media data from Chapter 3 with our proposed Dynamic Influence Model (DM) and observed different diffusion patterns between the eight news categories. From the estimated topic-related diffusion in social media, we try to discover distinct patterns so that we can better understand diffusion dynamics in a more principled way. The discovered patterns are also viewed with our model-free approach in the next section.

Distinct diffusion patterns: For extracting distinct diffusion patterns, we clustered the cumulative diffusion rates (F) of 63 news topics in Table 6.2 for each social system (News, SNS, and Blog). As a clustering method, we used the k-means clustering [69] in order to partition the time-series adoption sequences into k similar diffusion patterns. As a result, we obtained five distinct patterns, as shown in Figure 9.1. In this figure, the graphs in the first three columns present the averages and standard deviations of clustered F values for each system, and these averages are separately shown in the graph in the last column for a clear comparison between the systems. Accordingly, Figure 9.2 shows the distributions of the estimated parameter values of the DM, corresponding to each diffusion pattern in Figure 9.1.

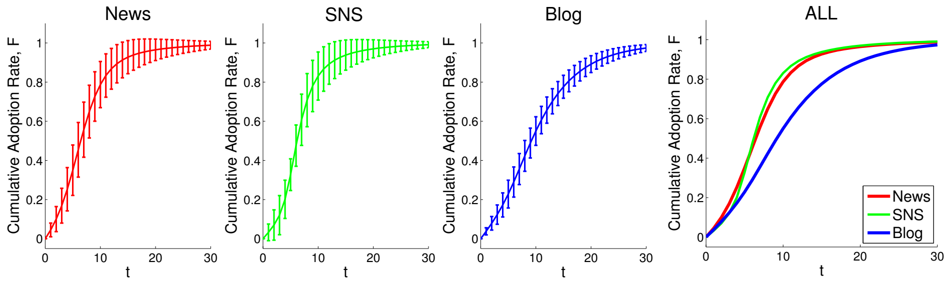
From these two figures, the five diffusion patterns can be characterized as below. Note that, for a clear comparison, we call internal influence c_{ji} of the DM in Equation (5.10) as (1) *intra-influence* or *intra-relationship* when $i = j$, (2) *inter-influence* or *inter-relationship* when $i \neq j$, or (3) *internal influence* for the both cases, for the rest of this chapter.

- **Pattern #1** – *unbalanced intra-influences with a significant difference*: In Figure 9.2(a), News and SNS show very strong intra-relationships (c_{nn} and c_{ss}), and they influence with each other. On the other hand, Blog shows relatively very weak intra-relationship (c_{bb}). As Figure 9.1(a) shows, the growth rate in SNS is even faster than News. Example news items are "Brazil Floods", "Golden Globe Awards", "Mammoth Revive", and "Betelgeuse".

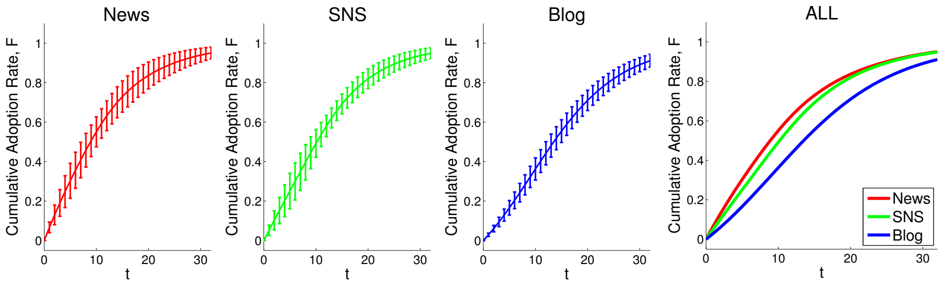
- **Pattern #2** – *stronger external influence than intra-influence*: In Figure 9.2(b), external influence (p_n and p_s) in News and SNS is stronger than their intra-influence (c_{nn} and c_{ss}) except for Blog, but intra-influence (c_{bb}) in Blog is not strong. As Figure 9.1(b) shows, all diffusion curve exhibit concave shapes rather than traditional S-curves. Related news items are mostly about famous people (e.g., "Actress Zsa Zsa Gabor", "Google's outgoing CEO Schmitdt", and "Support for Julian Assange") or even a popular product (e.g., iPad).

- **Pattern #3** – *stronger internal influence than external influence with balanced intra-influences*: In Figure 9.2(c), News, SNS, and Blog all show stronger intra-relationships (c_{nn}, c_{ss}, and c_{bb}) than their own inter-relationships ($c_{ji}, i \neq j$) which are all greater than external influence (p_n, p_s, and p_b). This pattern includes new items on political protests in the Middle East such as Tunisia, Egypt, Sudan, and Yemen. In Figure 9.1(c), this pattern leads to very similar S-curves of diffusion across the systems as we observed the concurrent diffusion pattern in Yemen protests in Section 6.3.

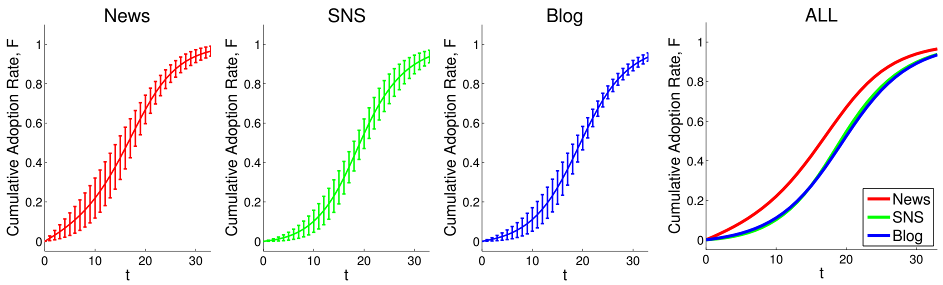
- **Pattern #4** – *unbalanced intra-influences with a relatively small difference*: The large proportion (50%) of the selected news items belong to this pattern. In Figure 9.2(d), News shows stronger external influence (p_n) than internal influence



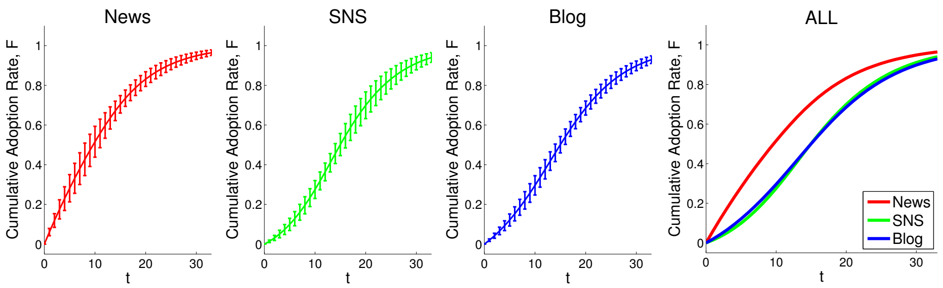
(a) Pattern #1. unbalanced intra-influences with a significant difference



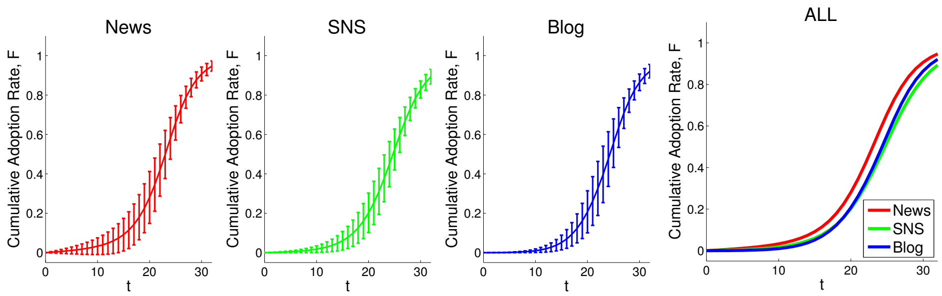
(b) Pattern #2. stronger external influence than intra-influence



(c) Pattern #3. stronger internal than external influence with balanced intra-influences

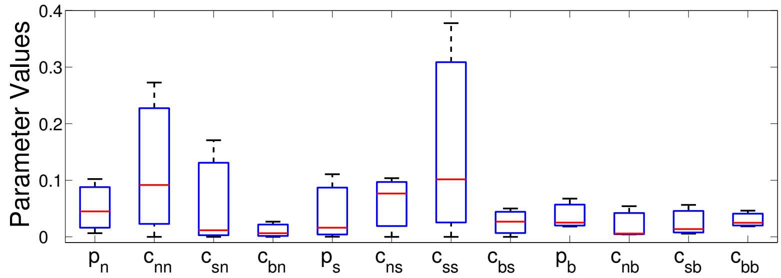


(d) Pattern #4. unbalanced intra-influences with a relatively small difference

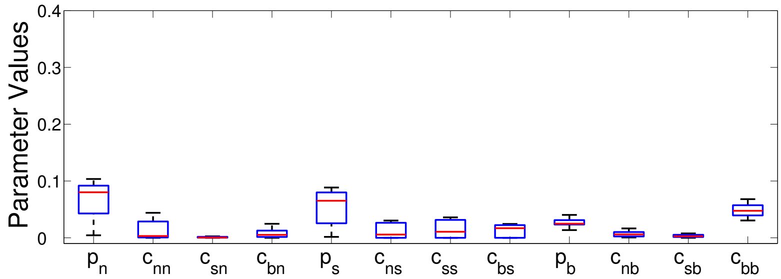


(e) Pattern #5. much stronger internal than external influence with balanced strong intra-influences

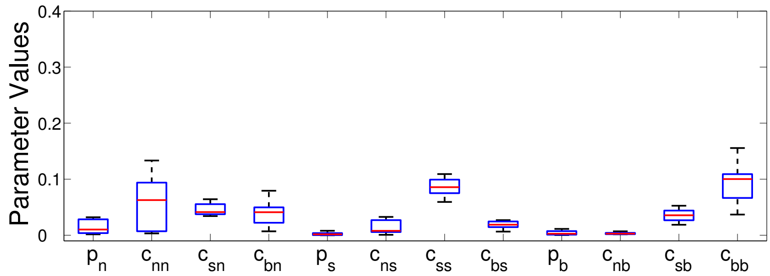
Figure 9.1: Distinct diffusion patterns for each social system by clustering the cumulative adoption rates F of 63 news topics in Table 6.2. In each pattern, the first three graphs show the averages and standard deviations of cumulative adoption rates for each system whose average is separately presented in the fourth graph for clarity.



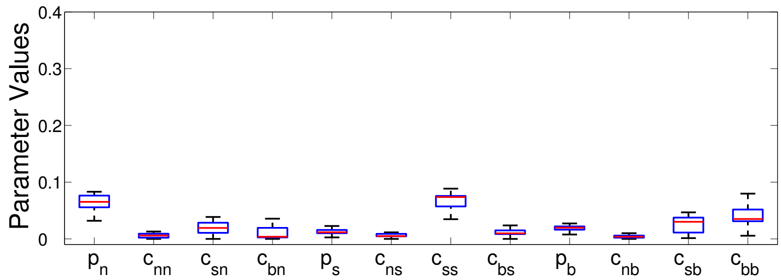
(a) Pattern #1. unbalanced intra-influences with a significant difference



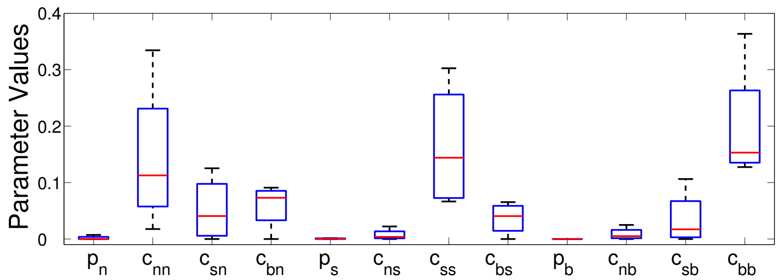
(b) Pattern #2. stronger external influence than intra-influence



(c) Pattern #3. stronger internal than external influence with balanced intra-influences



(d) Pattern #4. unbalanced intra-influences with a relatively small difference



(e) Pattern #5. much stronger internal than external influence with balanced strong intra-influences

Figure 9.2: Distributions of the estimated parameter values of the Dynamic Influence Model for each clustered diffusion pattern in Figure 9.1. (parameters: p_i – external influence on a meta-population i and c_{ji} – internal influence of a meta-population j on a meta-population i). Refer to Equation (5.8) and (5.10) for details.

with very weak intra-relationship (c_{nn}). On the contrary, SNS and News show stronger intra-relationships, but they are relatively weak compared to other patterns. As shown in Figure 9.1(d), SNS and Blog produce relatively flat S-curves compared with Pattern #1, #3, and #5, and News shows the concave-shaped curve as in the Pattern #2.

- **Pattern #5** – *much stronger internal influence than external influence with balanced strong intra-influences*: This pattern shows the most synchronous diffusion curves across the systems, as shown in Figure 9.1(e). In Figure 9.2(e), intra-relationships of News, SNS, and Blog are very strong and balanced, while external influence is ignorable and inter-relationships are relatively balanced compared to other patterns. Related news items are on disputable topics such as "Multiculturalism Failure", and "Muslim-Christian Conflicts". As discussed in Chapter 6, such controversial issues drive a synchronous and simultaneous diffusion across different social systems.

Traditionally, news media have been considered as external influence, but Pattern #1, #3, and #5 show that they interact with each other as if they were in a social network. Pattern #3 and #5 shows similar diffusion patterns across the systems, which are driven by both strong intra- and inter-relationships in a balanced way. The big difference between these two patterns is the strength of intra-influence. That is, stronger intra-relationships more likely reflect influence from other social systems and fuel diffusion within each system, leading to a more synchronous and simultaneous diffusion across the systems.

9.2.2 Diffusion Patterns from Model-Free Approach

In the previous subsection, we discovered the five distinct diffusion patterns in social media across News, SNS, and Blog. In this section, we analyze these patterns with our model-free approach.

We estimated macro-level information transfer with the time-delay (d) and memory (k) effects, as described in Figure 7.2, so that we can analyze the effect of the d time-shifted recent k states of a social system (a stochastic process) on the future state of the other system. Also, only significant transfer entropy is considered by conducting statistical significance tests discussed in Section 8.2. In this way, the synchronization of diffusion patterns between two stochastic processes are not considered as a coincidence.

Time-delay and memory effects: We now examine the discovered diffusion patterns in the previous subsection with our model-free approach, as shown in Figure 9.3. This figure presents the changes of the significant TE proportions for each diffusion pattern when varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). As the first graph in the figure shows, in overall a longer diffusion trend of a social system more influence others, but the time-delay effects are ignorable. In particular, Pattern #2 and #5 exhibit the largest variations when the memory size

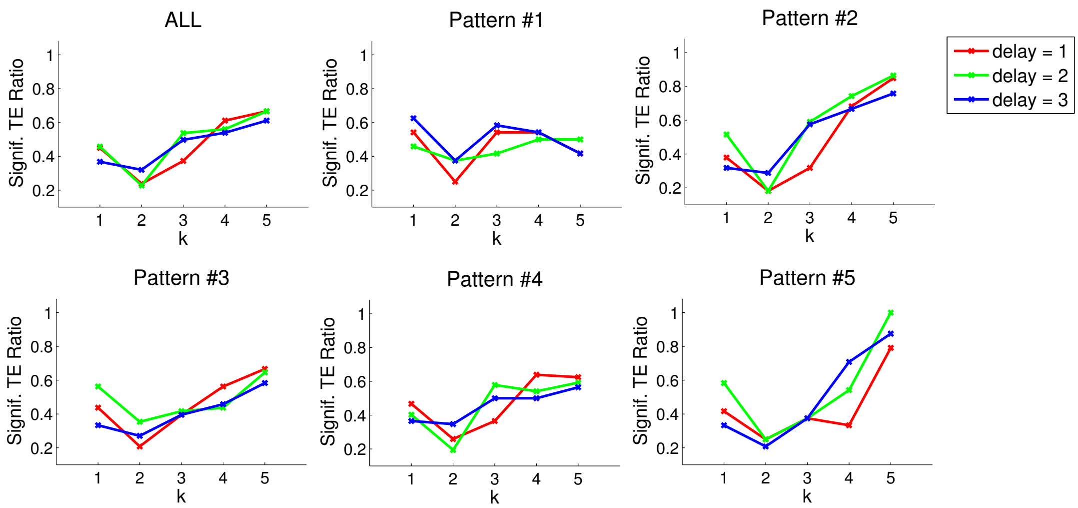


Figure 9.3: The proportions of significant transfer entropy (TE) for the five diffusion patterns in Figure 9.1 by varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). The x-axis indicates the memory size k, and the y-axis represents the proportions of significant transfer entropy. Time-delay is color-coded in each graph.

increases, showing the positive effects of a longer memory size k. This can be interpreted that news topics on famous people or products in Pattern #2 and controversial issues in Pattern #5 tend to draw attentions for a longer term. In contrast to these two patterns, Pattern #1 shows negative effects of a longer memory size. Related news topics are from the science category, which typically does not drive longer debates or discussions compared to the previous patterns. Pattern #3 and #4 also shows the positive effects of a longer memory but not as strong as Pattern #2 and #5.

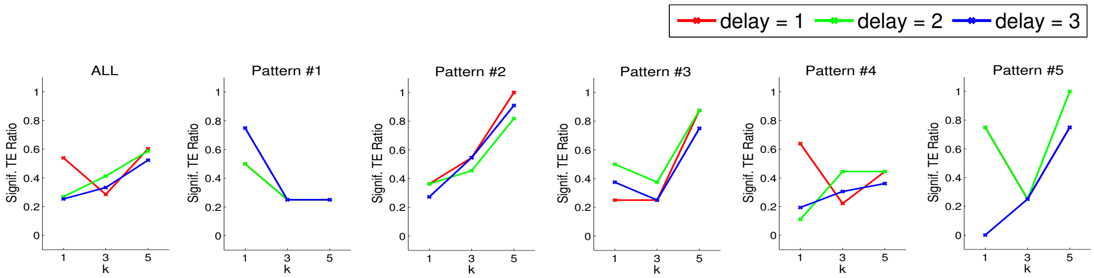
To be more specific, Figure 9.4 shows the detailed time-delay and memory effects for each possible direction of the transfer entropy. This figure provides the detailed behavioral characteristics of each social system in accordance with a source system as well as each diffusion pattern. From this figure, time-delay effects are recognizable compared with overall trends in Figure 9.3.

When increasing the length of time-delay and memory, noticeable changes for each pattern are summarized as below.

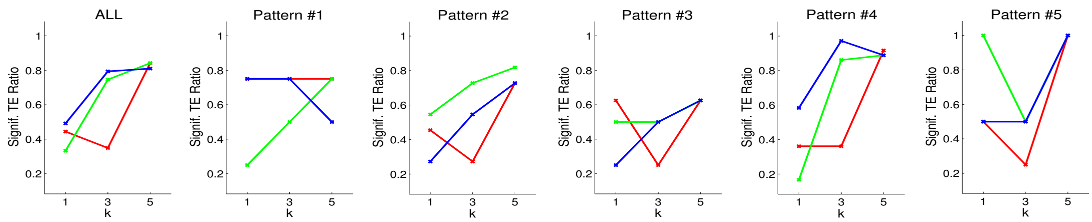
- **Pattern #1** – *unbalanced intra-influences with a significant difference*: SNS shows negative effects with a longer memory size ($\text{TE}_{N \to S}$ and $\text{TE}_{B \to S}$).

- **Pattern #2** – *stronger external influence than intra-influence*: SNS shows positive effects with a longer memory size. Blog shows the highest variations with time-delay effects, and it is most influenced by other systems when $d = 2$ ($\text{TE}_{N \to B}$ and $\text{TE}_{S \to B}$).

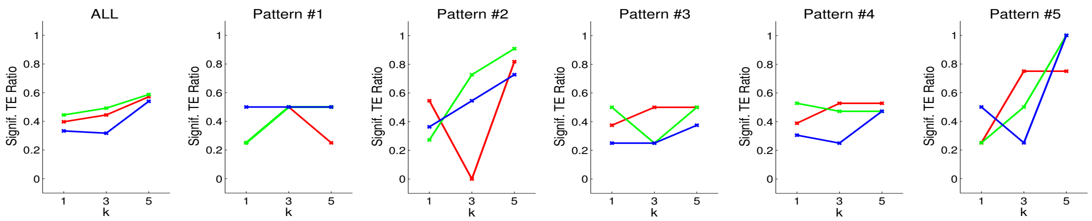
- **Pattern #3** – *stronger internal influence than external influence with balanced intra-influences*: SNS shows positive effects with a longer memory size.



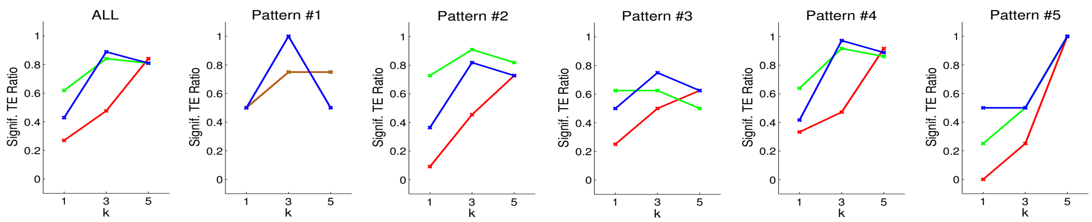
(a) $\mathrm{TE}_{N \to S}$



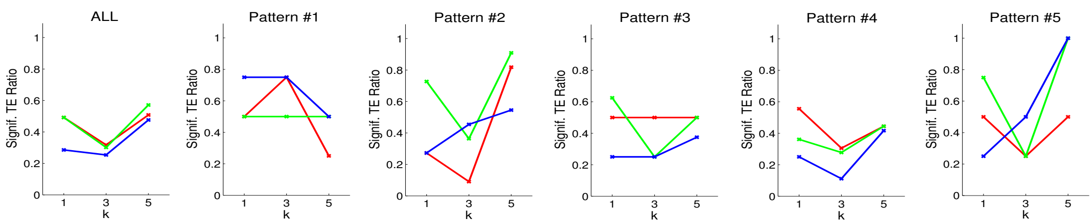
(b) $\mathrm{TE}_{N \to B}$



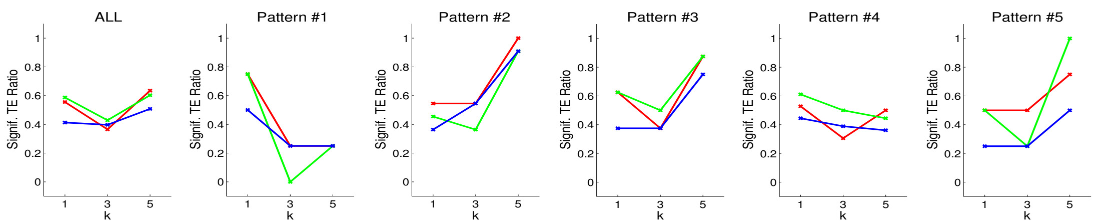
(c) $\mathrm{TE}_{S \to N}$



(d) $\mathrm{TE}_{S \to B}$

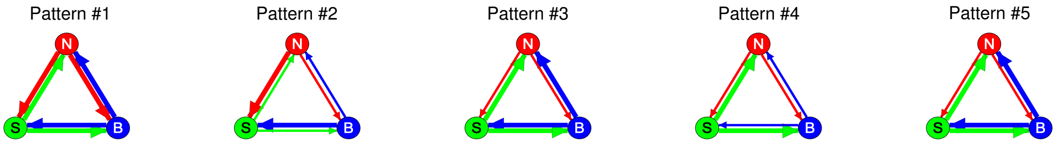


(e) $\mathrm{TE}_{B \to N}$

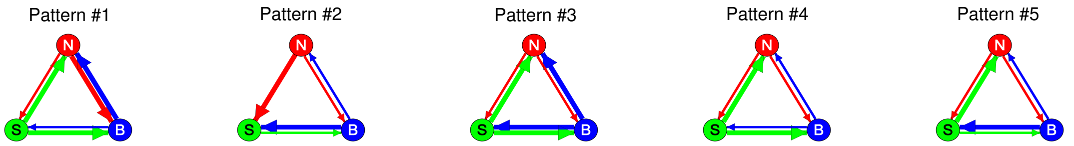


(f) $\mathrm{TE}_{B \to S}$

Figure 9.4: The proportions of significant transfer entropy (TE) for each possible direction and diffusion pattern when varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). Time-delay is color-coded in each graph.



(a) Inter-influence between different systems from the model-driven approach



(b) Significant transfer entropy from the model-free approach

Figure 9.5: Comparison of the diffusion patterns obtained from the model-driven and model-free approaches. Triads visualize the distinct diffusion patterns in Figure 9.1 based on the estimations from the approaches. **(a)** Dynamic influence across News, SNS, and Blog based on the estimated parameter values c_{ji} ($i \neq j$) in Equation (5.10). **(b)** Information transfer across News, SNS, and Blog based on the estimated transfer entropy which are averaged for all time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). In each graph, the node labels indicate the three social systems (N: News, S: SNS, and B: Blog). Link widths and arrows present the strength and directionality of influence, respectively.

- **Pattern #4** – *unbalanced intra-influences with a relatively small difference*: SNS shows negative effects with a longer memory size, while Blog shows the positive effects.
- **Pattern #5** – *much stronger internal influence than external influence with balanced strong intra-influences*: News, SNS, and Blog all show positive effects with a longer memory size.

SNS shows dynamic behavior changing across different diffusion patterns, while News in general exhibits neutral behavior. Blog shows more persistent behavior in that it is more influenced by a older and longer-term activity sequence of the other system compared with SNS and News. This is consistent with the baseline influence of each social system and its changes with an increase of the memory size, as discussed in the previous chapter. That is, News is neural, SNS is incoming, and Blog is outgoing as baseline influence. When increasing the memory size, SNS and Blog become outgoing and incoming for each.

9.2.3 Strength and Directionality of Influence from two Approaches

In Section 9.2.1 and 9.2.2, we examined dynamic influence across News, SNS, and Blog and their behavioral characteristics which can be obtained from the model-driven and model-free approaches respectively. In this section, we examine common outcomes from the approaches, the strength and directionality of influence between different social systems.

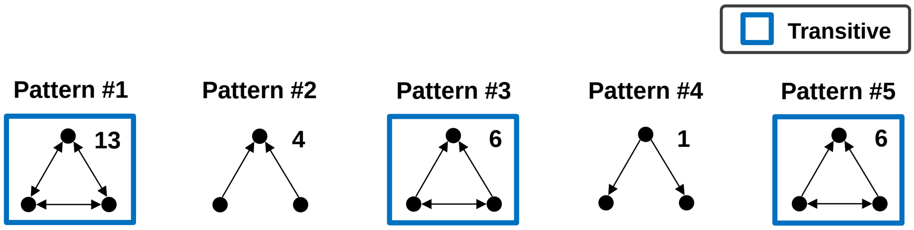


Figure 9.6: Unique structures of relations among social systems. Each triad summarizes the obtained diffusion trends for each distinct pattern in Figure 9.5 by combining only strong connections obtained from the model-driven and model-free approaches. Presented triads are from Figure 6.1 describing 16 unique structures of relations among three meta-populations. Blue frames indicate transitive relations (i.e. if $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$), and the numbers beside triads denote the identities of unique structures in Figure 6.1.

Figure 9.5 shows the macroscopic picture of information pathways for the discovered distinct diffusion patterns discussed in the previous subsections. Figure 9.5(a) visualizes the estimated parameter values c_{ji} $(i \neq j)$ of the Dynamic Influence Model. The parameter c_{ji} indicates the probability that a meta-population j influences a meta-population i in Equation (5.10) and implies inter-relationships between different social systems. Figure 9.5(b) also illustrates macro-level information transfer across social systems based on the estimated significant transfer entropy which tells us that the future state of a destination social system is predictable due to the knowledge of the previous states of a source social system. This can be interpreted as inter-relationships between the systems.

As the figure shows, the estimated diffusion trends from both approaches are in general similar to each other in terms of the strength and directionality of influence. Figure 9.6 summarizes the diffusion trends in Figure 9.5 by combining only strong connections obtained from the approaches. These summarizations indicate the unique structures of relations among three meta-populations in Figure 6.1. Blue frames in the figure present transitive relations (i.e. if $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$). As the figure shows, Pattern #1, #3, and #5 exhibit transitive relations, and particularly Pattern #3 and #5 present the same unique structure. This reflects the similar distributions of intra- and inter-influences between Pattern #3 and #5.

These three patterns showed stronger internal influence than external influence. Even though they all reflect transitive relations, the strength and balance of intra-influences across the systems are also important to distinguish the diffusion patterns. That is, the balanced and strong intra-influences of Pattern #5 in Section 9.2.1 show the most synchronous and simultaneous diffusion across the systems, while the balanced but weak intra-influences of Pattern #3 show less synchronous diffusion than Pattern #5. Unbalanced intra-influences of Pattern #1 show the least synchronous diffusion among these three patterns, even though it shows the most interactive unique structure in Figure 6.1.

Thus, outcomes from both the model-driven and model-free approaches in conjunction can help better understand real-world diffusion in a comprehensive way than either approach alone. In the next section, we also examine knowledge diffusion in academic publications by comparing these two approaches.

9.3 Knowledge Diffusion in Academic Publications

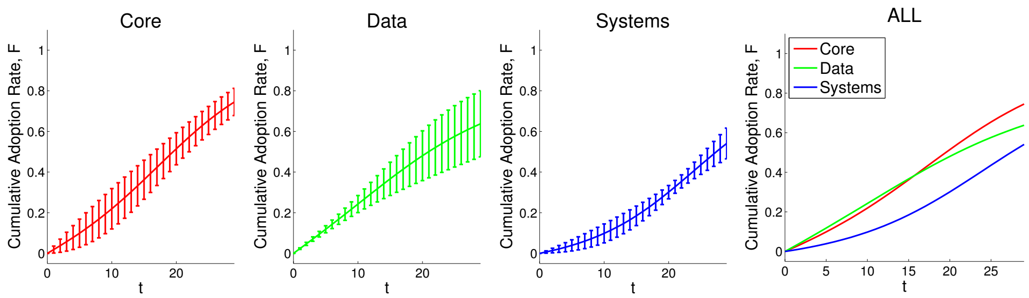
In this section, we compare the proposed approaches by using the estimated knowledge diffusion patterns in academic publications. As in the previous section, the comparisons are based on the distinct diffusion patterns across the Core, Data, and Systems clusters so that we can examine how the two approaches uncover these patterns.

9.3.1 Diffusion Patterns from Model-Driven Approach

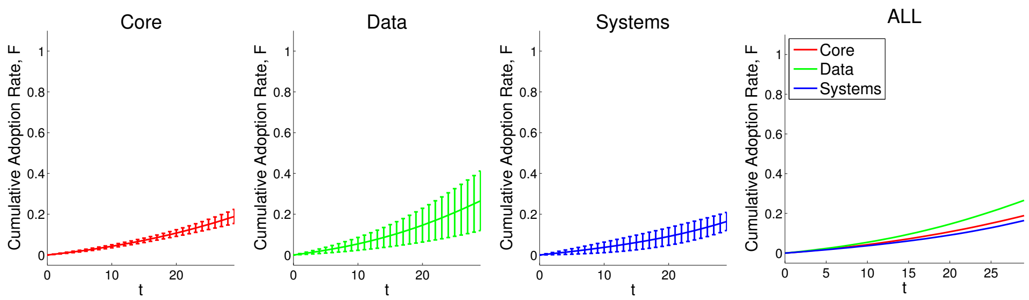
Based on the estimated topic-related diffusion patterns in academic publications from Section 6.4, we try to discover distinct diffusion patterns at a higher level. As in the previous section, by using the k-means clustering method we clustered the cumulative diffusion rates (F) of the 15 research keywords in Table 6.4 for each cluster (Core, Data, and Systems). As a result, we obtained four distinct diffusion patterns, as shown in Figure 9.7. In addition, Figure 9.8 shows the distributions of the estimated parameter values of the Dynamic Influence Model for each clustered diffusion pattern in Figure 9.7. From these figures, the discovered diffusion patterns can be characterized as below.

- **Pattern #1** – *stronger external influence than inter-relationships*: This pattern includes research keywords such as "Social Network", "Search Engine", "Web Service", and "Semantic Web". In Figure 9.8(a), Core and Data show relatively stronger external influence (p_c and p_d) than inter-relationships (c_{dc} and c_{sc} for Core and c_{cd} and c_{sd} for Data), but Core's inter-relationships are stronger compared to the other clusters. Data shows the strongest intra-influence (c_{dd} among the four patterns but relatively weak inter-relationships with the other clusters. Systems shows weaker external influence (p_s) than the others, and rather the strength of external influence is similar to those of inter-relationships (c_{cs} and c_{ds}). Figure 9.7(a) reflects this dynamic influence, showing the slow growth rate of Systems in contrast to Core and Data.

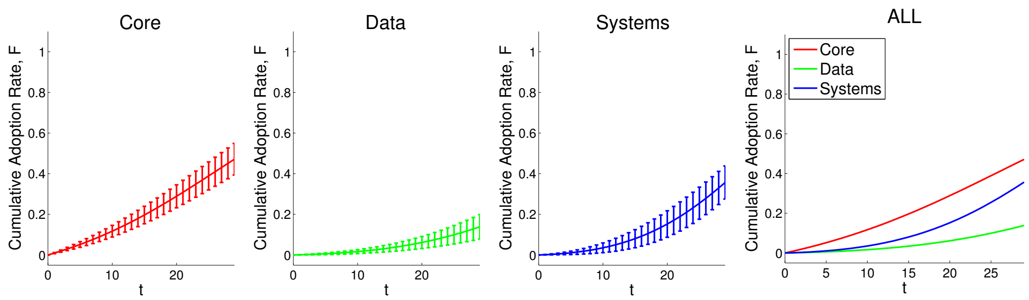
- **Pattern #2** – *balanced but weak internal influences*: Corresponding keywords are "Generic Algorithm", "Genetics", "Learning Algorithm", and "Approximate Algorithm". As shown in Figure 9.8(b), this pattern shows relatively balanced intra-influences across the clusters along with balanced inter-relationships among them, but the strength of internal influence in all clusters is very weak compared to other patterns. As Figure 9.7(b) shows, the growth rates of the clusters are similar but very slow. This can be interpreted that these keywords have drawn attention from small proportions within and between the clusters.



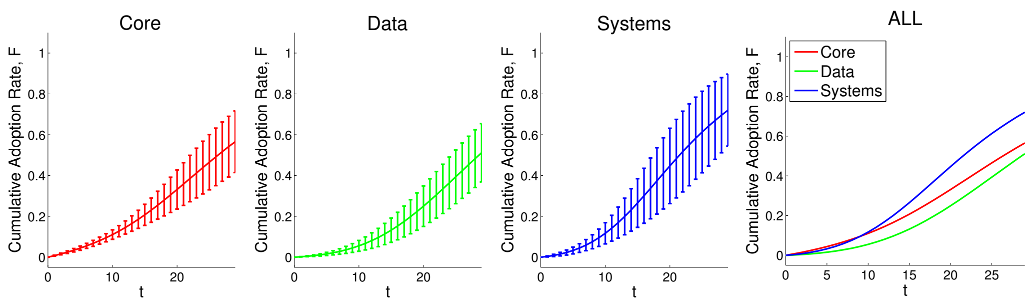
(a) Pattern #1. stronger external influence than inter-relationships



(b) Pattern #2. balanced but weak internal influences

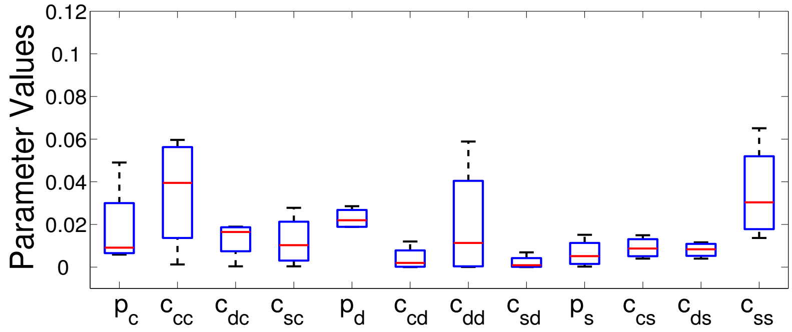


(c) Pattern #3. unbalanced and weak internal influences

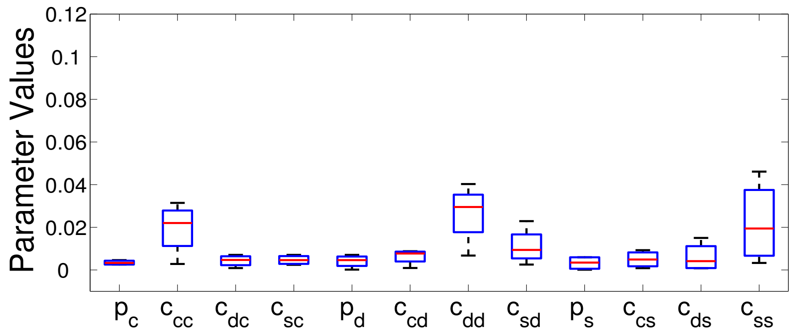


(d) Pattern #4. unbalanced but strong internal influences

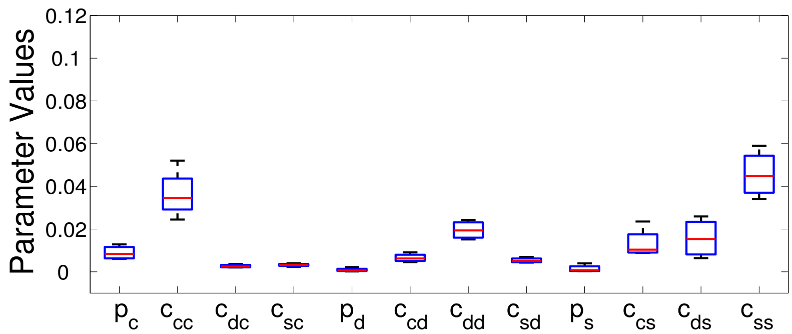
Figure 9.7: Clustered diffusion patterns according to each social system's cumulative adoption rates F of 15 research keywords in Table 6.4. In each pattern, the first three graphs illustrate the averages and standard deviations of cumulative adoption rates for all social systems whose averages are separately presented in the fourth graph for clarity.



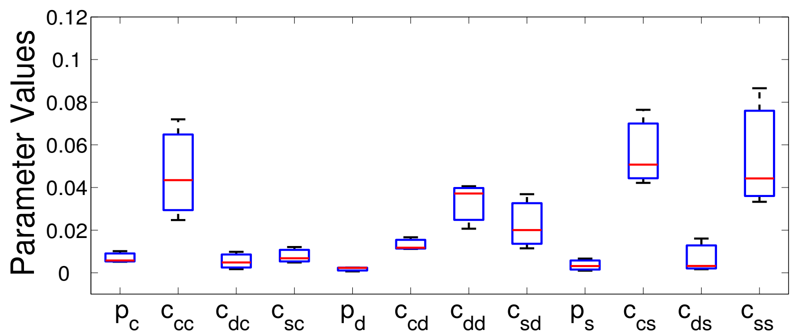
(a) Pattern #1. stronger external influence than inter-relationships



(b) Pattern #2. balanced but weak internal influences



(c) Pattern #3. unbalanced and weak internal influences



(d) Pattern #4. unbalanced but strong internal influences

Figure 9.8: Distributions of the estimated parameter values of the Dynamic Influence Model for each clustered diffusion pattern in Figure 9.7. (parameters: p_i – external influence of a meta-population i and c_{ji} – internal influence of a meta-population j on a meta-population i). Refer to Equation (5.8) and (5.10) for details.

- **Pattern #3** – *unbalanced and weak internal influences*: This pattern contains keywords related to system performance such as "High performance", "Fault Tolerant", "Distributed System", and "Satisfiability". In Figure 9.8(c), Core and System show relatively higher intra-influences than Data, and Systems shows stronger inter-relationships than the others. On the other hand, Data shows low interactions within and between clusters in contrast to Systems. As Figure 9.7(c) shows, the growth pattern of Data is different from the others and its rate is very slow. Systems is influenced by both Core and Data, and accordingly the increases of the growth rates of Data and Core during recent 10 years more accelerate the adoption rate of Systems.

- **Pattern #4** – *unbalanced but strong internal influences*: Relevant keywords are "Real Time", "Large Scale" and "Sensor Network". In Figure 9.8(d), the intra-relationships of Core and Systems are the strongest among the four patterns. In particular, Data and System is influenced by System and Core as much as intra-influence respectively, while the inter-relationships of Core are ignorable compared to the others. Strong intra-influence of Core (c_{cc}) strongly influences System, and System's strong internal influence (c_{ss} and c_{cs}) also highly influences Data. As Figure 9.7(d) shows, this pattern drives the most synchronous diffusion curves among the patterns even though intra-influences are not balanced. That is, inter-relationships are strong enough to fuel such synchronous diffusion.

Compared with the diffusion patterns in social media in the previous section, weak and unbalanced intra- and inter-relationships are observed. However, we can obtain common insights from these different application domains. That is, synchronous diffusion needs strong intra-relationships as momentum to make individuals to move to a certain state, which is triggered by strong inter-relationships between different social systems. Dynamic influence with different strength and directionality leads to form unlimited diffusion patterns, but from the patterns we can draw a rough picture of momentum formed within and between heterogeneous populations.

9.3.2 Diffusion Patterns from Model-Free Approach

We now analyze the discovered diffusion patterns in the previous subsection with our model-free approach. As in Section 9.2.2, we estimate the transfer entropy (TE) by varying the time-delay d and the memory size k as discussed in Figure 7.2 so that we can observe behavioral characteristics of the clusters. Also, only significant transfer entropy is considered by conducting statistical significance tests as discussed in the previous section.

Time-delay and memory effects: Figure 9.9 shows the changes of the significant TE proportions for each diffusion pattern when varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). The first graph in the figure shows overall trends of time-delay and memory effects for all diffusion patterns combined. As the figure shows, the overall trends do not provide much information, since the high variations

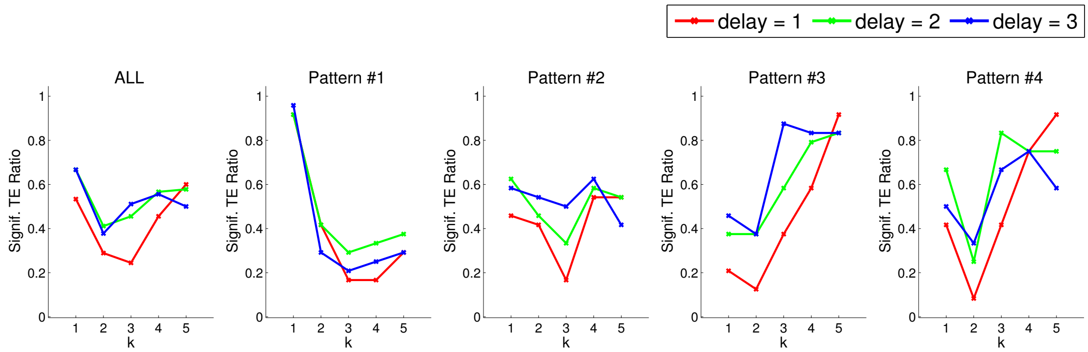


Figure 9.9: The proportions of significant transfer entropy (TE) for four diffusion patterns in Figure 9.7 by varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). The x-axis indicates the memory size k, and the y-axis represents the proportions of significant transfer entropy. Time-delay is color-coded in each graph.

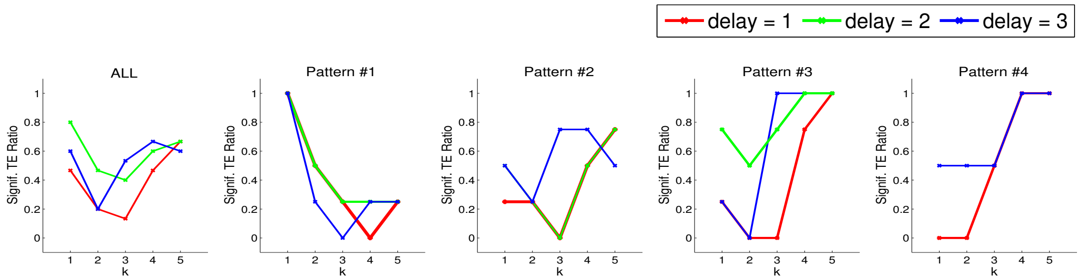
of the effects across the four diffusion patterns cancel out the distinct behavioral characteristics. Note however that in the case of social media in Section 9.2.2, the five diffusion patterns showed common positive effects of a longer memory size, which is reflected in their overall trends.

As Figure 9.9 shows, opposite trends are observed in academic publications such as negative and positive effects of a longer memory size between Pattern #1 and Patterns #3 and #4, while Pattern #2 shows fluctuations. This is consistent with the results from the previous chapter. More details are discussed in this section. To be more specific, Figure 9.10 shows the detailed time-delay and memory effects for each transfer entropy. This figure provides the detailed behavioral characteristics of each cluster regarding the different diffusion patterns.

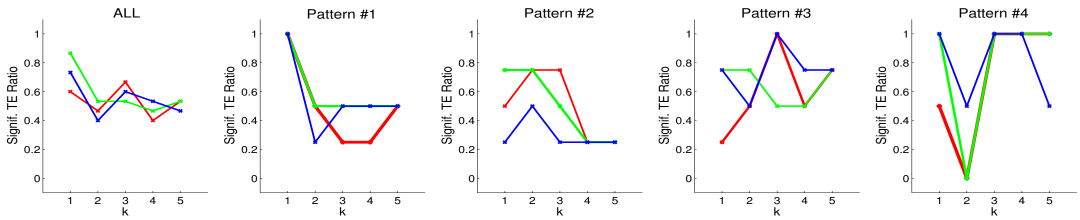
When increasing the length of time-delay and memory, recognizable changes for each pattern are summarized as below.

- **Pattern #1** – *stronger external influence than inter-relationships*: Core, Data, and Systems all show the negative effects of a longer memory size. This pattern includes keywords such as "Social Network", "Search Engine", "Web Service", and "Semantic Web", which all showed the decreases of significant TE ratios when increasing the memory size in Chapter 8. As discussed in Section 8.4, these keywords are related to research topics demanding to be updated with the recent state of the literature compared with theoretical research topics due to frequently changing environments on the Web.

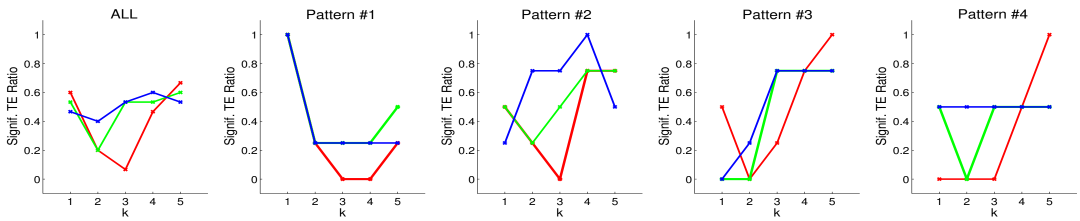
- **Pattern #2** – *balanced but weak internal influences*: Systems shows the negative effects of a longer memory size ($TE_{C \to S}$ and $TE_{D \to S}$). When Data is a source system, the variation of the significant TE ratio increases. This is consistent with the results that in the model-driven approach, Data is influential on Systems regarding this pattern.



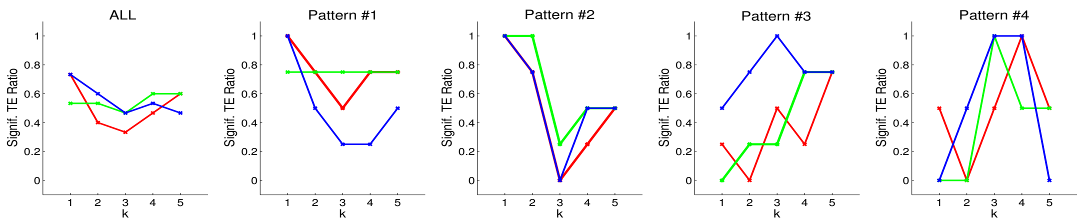
(a) $\mathrm{TE}_{C\to D}$



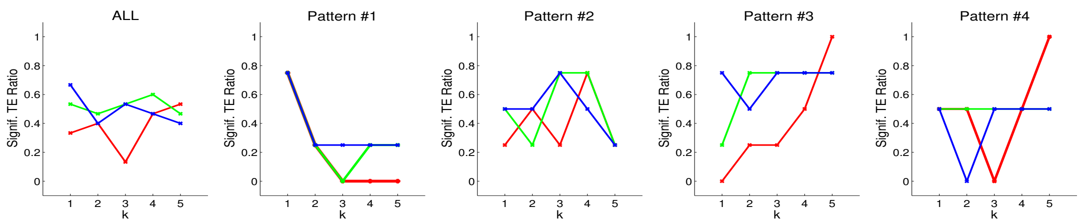
(b) $\mathrm{TE}_{C\to S}$



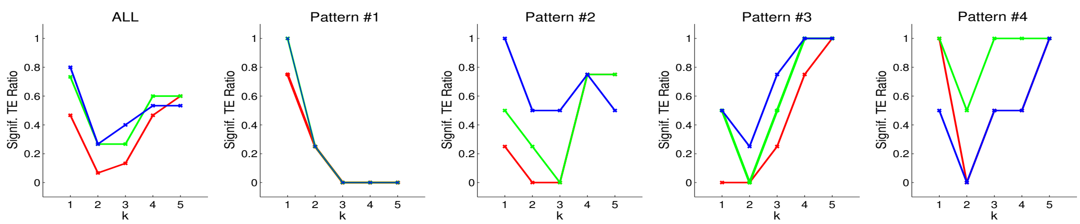
(c) $\mathrm{TE}_{D\to C}$



(d) $\mathrm{TE}_{D\to S}$

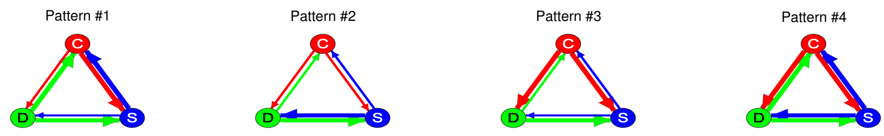


(e) $\mathrm{TE}_{S\to C}$

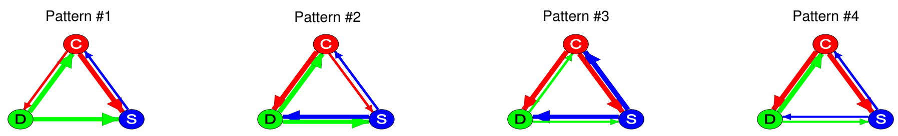


(f) $\mathrm{TE}_{S\to D}$

Figure 9.10: The proportions of significant transfer entropy (TE) for each possible direction and diffusion pattern when varying the length of time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). Time-delay is color-coded in each graph.



(a) Internal influence between different systems from the model-driven approach



(b) Significant transfer entropy from the model-free approach

Figure 9.11: Comparison of the diffusion patterns obtained from the model-driven and model-free approaches. Triads are constructed based on the estimations from each approach and they correspond to the distinct diffusion patterns in Figure 9.1. **(a)** Dynamic influence across Core, Data, and Systems based on the estimated parameter values c_{ji} ($i \neq j$) in Equation (5.10) of the Dynamic Influence Model. **(b)** Information transfer across Core, Data, and Systems based on the estimated transfer entropy which are averaged for all time-delay ($1 \leq d \leq 3$) and memory ($1 \leq k \leq 5$). In each graph, the node labels indicate the three clusters (C: Core, D: Data, and S: Systems). Link widths and arrows present the strength and directionality of influence respectively.

- **Pattern #3** – *unbalanced and weak internal influences*: Core and Data show the positive effects of of a longer memory size. Particularly, Data also shows the positive effects of a longer time-delay as it is more influenced by a older and longer trends of Systems ($\text{TE}_{S \to D}$).

- **Pattern #4** – *unbalanced but strong internal influences*: Data show the positive effects of a longer memory size. In contrast to Pattern #2, System does not show the negative effects of a longer memory size. This reflects the estimation results from the model-driven approach in the previous subsection, where this pattern exhibits that Core is influential on System, but the influence from Data is relatively very weak.

In the previous chapter, we observed two opposite trends of the memory effects. Pattern #1 and Pattern #3 consist of keywords showing the negative and positive effects of a longer memory size for each. In terms of the time-delay effects, Pattern #1 shows the least variations compared with other patterns. In general, at least 2-year before publication trends of a cluster influence other clusters (i.e. $d \geq 2$). Behavioral characteristics of a destination system reflect a source system's influence on that system from the model-driven approach (i.e. c_{ji}, where j and i indicate source and destination systems, respectively). These changes are all closely related to keyword topics, which reconfirms that we need to consider time-delay and memory effects as well as diversity of information topics for a better understanding of information transfer in real-world diffusion.

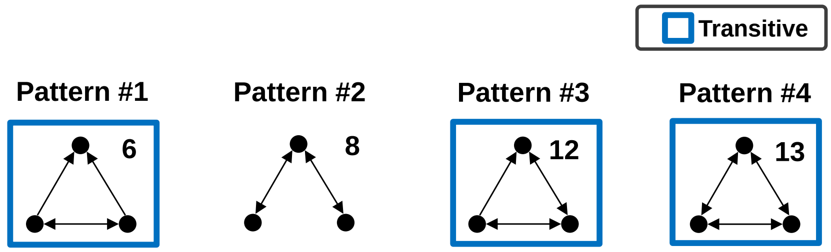


Figure 9.12: Unique structures of influence flows across social systems. Each triad summarizes the obtained diffusion trends for each distinct pattern in Figure 9.11 by combining only strong connections from the model-driven and model-free approaches. Unique structures are referred to Figure 6.1 in Chapter 6. Blue frames indicate transitive relations (i.e. $A \to B$ and $B \to C$, then $A \to C$), and the numbers beside triads denote the identities of unique structures in Figure 6.1.

9.3.3 Strength and Directionality of Influence from two Approaches

In Section 9.3.1 and 9.3.2, we examined dynamic influence and behavioral characteristics of clustered subdomains in computer science with our model-driven and model-free approaches, respectively. In this section, we continue to examine the strength and directionality of influence across Core, Data, and Systems for the discovered distinct diffusion patterns from Section 9.3.1.

Accordingly, Figure 9.11 shows macroscopic pictures of information pathways for each diffusion pattern. Figure 9.11(a) provides the visualizations of the estimated parameter values c_{ji} $(i \neq j)$ of the Dynamic Influence Model as in Section 9.2.3. Figure 9.11(b) also presents macro-level information transfer across the clusters based on the estimated significant transfer entropy. These can all be interpreted as inter-relationships between the systems.

As the figure shows, the estimated diffusion trends from both approaches are in general similar to each other in terms of the strength and directionality of influence. Accordingly, Figure 9.12 summarizes the diffusion trends from Figure 9.11 by combining only strong connections obtained from both approaches. As the figure shows, Pattern #1, #3, and #4 exhibit transitive relations. These three patterns showed stronger intra- and inter-relationships than Pattern #2, while this intransitive Pattern #2 showed the slowest time-series growth rates of adoptions. In other words, widely popular research topics in computer science are more likely led by transitive relations across different subareas. These transitive relations have more opportunities to bring about concurrent diffusion across heterogeneous systems, but the strength as well as balance of intra- and inter-relationships are also important to lead to such global diffusion.

We applied our model-driven and model-free approaches to different real-world application domains, social media and academic publications, which helps to validate the feasibility of the models and further generalize the approaches.

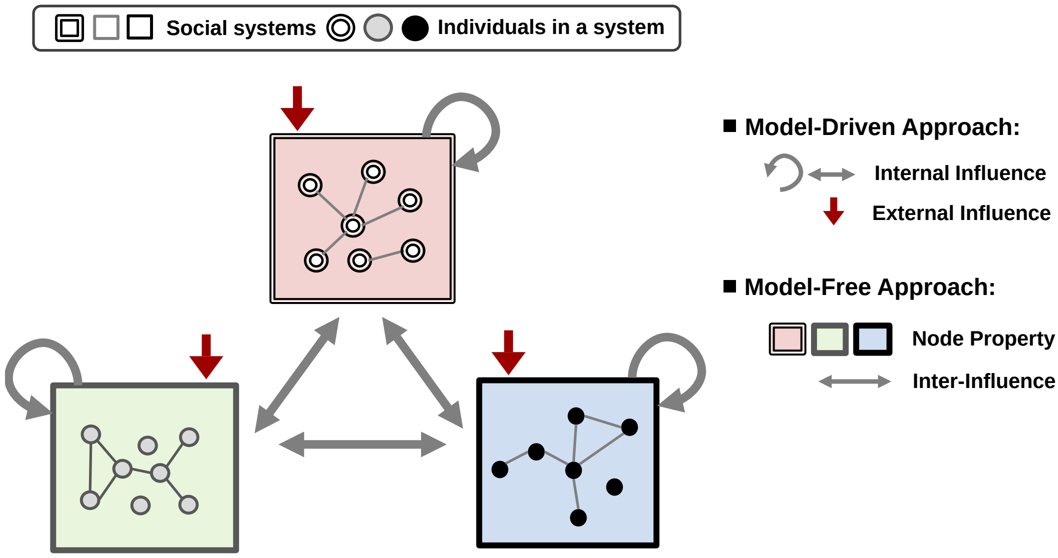


Figure 9.13: Conceptual comparison between the model-driven and model-free approaches. Both approaches in conjunction provide complementary information on diffusion; the model-driven approach provides external and internal influences across heterogeneous social systems, i.e. p_i and c_{ji} of the Dynamic Influence Model, while the model-free approach provides information on each system's behavioral characteristics (node property) and information transfer across the systems, i.e. macro-level information transfer with time-delay and memory effects.

9.4 Discussion

In Chapter 5 and 7, we proposed two model-driven and model-free approaches, respectively, based on the empirical studies on real-word diffusion phenomena in Part I. In this chapter, we compared the two approaches using the estimated diffusion patterns in social media and academic publications in Chapter 6 and 8. We observed that the two approaches provide similar results but with different perspectives, which in conjunction can help better explain diffusion than either approach alone. In this section, we discuss the common and different perspectives of the approaches and their applicability in other research areas.

9.4.1 Model-Driven vs. Model-Free Approaches

Our model-driven approach, called the Dynamic Influence Model, helps better describe diffusion within a homogeneous population in a single social system, since it does not ignore the effects of interactions between heterogeneous social networks on diffusion. As discussed in Chapter 5, some supportive evidence has been observed by a recent study [123] that 30% of exposures in Twitter were attributed to external sources. This percentage is rather close to internal influence as from the marketing literature [110]. In this regard, our approach extends the simple and robust mass action Bass Model into the dynamics of meta-populations in a probabilistic way. This is enabled by incorporating the two essential features, the heterogeneity and structural connectivity of underlying networks in diffusion, and is based on the assumption of a power-law degree distribution which helps to obtain stochastic pathways on heterogeneous networks. Thus, this approach can provide the benefits of excluding detailed network topologies in the complex environment, but it still needs to estimate the parameter value of a degree distribution.

On the other hand, our model-free approach, macro-level information transfer across social systems, as described in Chapter 7 does not require any assumptions on social interactions by using information-theoretic measures. By considering a social system as a stochastic process and defining its signals emitted to other social systems, we estimated macro-level information transfer between meta-populations with both time-delay and memory effects. These non-parametric statistics provide benefits of quantifying non-linear dynamics of complex systems, but they require accurate probability distributions of target random variables. However, it is hard to construct such probability distributions in a realistic situation due to the small sample sizes obtainable from limited observations.

Thus, the two approaches in conjunction can help to obtain more significant diffusion patterns by comparing the estimation results from the approaches and maximizing the similarity between them. Figure 9.13 summarizes information that each approach provides. As examined in Section 9.2.1 and 9.3.1, the model-driven approach enables to reveal the strength and directionality of influence within and between (intra- and inter-relationships respectively) social systems as well as external influence outside these systems. The model-free approach also provides information of inter-relationships between different systems except for intra-relationships and external influence. From the time-delay and memory effects of the transfer entropy in Section 9.2.2 and 9.3.2, we can additionally obtain behavioral characteristics of each system which can be a node property at a macro level, as shown in Figure 9.13.

Thus, the two proposed approaches together provide a complementary macroscopic picture of diffusion by filling the gap which either approach cannot provide alone, and thus they can help better understand real-world diffusion in a comprehensive way. They also suggest alternative options to choose a more appropriate approach according to the experimental conditions of different application domains.

9.4.2 Applications to Other Research Fields

As discussed in the previous subsection, the model-driven approach can provide both intra- and inter-relationships among heterogeneous systems but with assumptions of the underlying social interactions, while the model-free approach reveals inter-relationships between systems along with their behavioral properties but without any assumptions of the social interactions. These two approaches can suggest alternative options in estimating diffusion trends at a macro level in a wide range of application domains.

For instance, from the neuroscience point of view, brain structures can be categorized based on their functionalities [55, 150, 164], which is analogous to heterogeneous social networks categorized into three meta-populations in this study. The nervous system is a complex network of neurons, and their connectivity is hard to collect and define as in the case of dynamic social interactions in social media. Also, neuronal connections may vary with different input signals to the human brain in the way that diffusion patterns in social media are different in accordance with diverse information topics. In this regard, the connectivity between brain functions can be estimated with our approaches in terms of the strength and directionality of influence across target regions in the brain.

In addition, organizational or multinational diffusion has been a common topic across the areas of marketing, social science, and political science [31, 42, 95, 158, 166]. Here, our approaches can help to quantify dynamics of influences among organizations, countries, or continents according to appropriate meta-population schemes.

In this thesis, we estimated knowledge diffusion across computer science subdomains, but these approaches can also be applied to a higher level of research fields. For instance, one research domain can be considered as a social system, such as computer science, social science, and physics, and thus we can estimate diffusion trends of multidisciplinary research topics across different research areas. That is, a flexible meta-population scheme provides more opportunities to apply our approaches to a wider range of application domains.

9.5 Conclusion

In this chapter, we compared our model-driven and model-free approaches based on the estimated diffusion patterns in social media and academic publications from the aspects of common and different features of the approaches. This helped us to obtain a more consistent understanding of cross-population diffusion at a macro level.

From this comparison study, we focused on distinct diffusion patterns in our application domains so that we can examine how the two approaches uncover these patterns. In common, strong and balanced internal influences across the systems more likely drive synchronous diffusion. In general, transitive relations have more opportunities to bring about similar diffusion patterns across heterogeneous systems, but the strength as well as balance of both intra- and inter-relationships are also important to lead to global diffusion. That is, synchronous diffusion needs strong intra-relationships as momentum to make individuals to move to a certain state, which is triggered by strong inter-relationships between different social systems. Dynamic influence with different strength and directionality leads to form unlimited diffusion patterns, but from the patterns we can draw an overall picture of momentum formed within and between heterogeneous populations.

We applied our model-driven and model-free approaches to different real-world application domains, social media and academic publications, which helps to validate the feasibility of the models and further generalize the approaches. The estimated diffusion trends from both approaches are in general similar to each other in terms of the strength and directionality of influence but with different perspectives on diffusion. Thus, outcomes from both the model-driven and model-free approaches in conjunction can help better understand real-world diffusion in a comprehensive way than either approach alone.

We expect that our proposed approaches can provide benefits for understanding dynamics of complex systems whose network structures are hard to collect in reality and provide a way of uncovering cross-populations diffusion in a wide range of application domains.

Chapter 10

Conclusion and Future Directions

A "connection" or "interaction" between individuals can be defined in numerous ways, such as relationships of online communications, spatial vicinity, kinship, friendship, and collaborations. These connections function as communication channels of a new idea or information, which collectively form complex information pathways on heterogeneous (social) networks. Such far-reaching connectivity leads to an unlimited number of unique patterns of global diffusion, and these patterns vary with the context of information.

This thesis aimed to uncover macro-level diffusion mechanisms across heterogeneous social networks. For this study, we conducted empirical analyses on real-world diffusion in two application domains, social media and academic publications, which have drawn attention from diverse research areas. Based on our findings, we proposed the model-driven and model-free approaches and compared estimated topic-related diffusion patterns from the two approaches.

In Part I of this thesis, our empirical analyses focused on the citation relationships in social media and academic publications. We observed that diffusion networks span beyond a single social system into multiple systems, and thus they necessarily contain heterogeneous social networks. The underlying connectivities are not uniform but unbalanced, suggesting an existence of dynamic influence across such systems. These common observations motivated us to propose the model-driven and model-free approaches.

In Part II, we proposed the Dynamic Influence Model which extends the traditional diffusion framework into a more flexible and general one. That is, this model allows direct interactions between different social networks as if they were in the same networks. This is enabled by generalizing the simple and robust mass action Bass Model [30] into a model of dynamics across meta-populations.

In contrast to this model-driven approach, in Part III we proposed a model-free approach by using information-theoretic measures. This approach can estimate macro-level information transfer across social systems by defining a stochastic process at a system level and its signal at a meta-population level, and by considering both time-delay and memory effects.

In Part II and III, we conducted experiments on the real data described in Part I, and we estimated topic-related diffusion patterns using our proposed approaches. In Part IV, we finally compared our two approaches based on the obtained estimation results. We focused on distinct diffusion patterns from social media and academic publications, and we examined how each approach estimates and distinguishes these patterns. In common across the two application domains, transitive relations of social systems have more opportunities to bring about synchronous and concurrent diffusion patterns across such systems. Also, the strength and balance of intra- and inter-relationships play an important role in leading to such simultaneous global diffusion. The estimated diffusion trends from both approaches are in general similar to each other in terms of the strength and directionality of influence but with different perspectives on diffusion. Thus, the outcomes from the model-driven and model-free approaches in conjunction can help to better understand real-world diffusion in a comprehensive way than either approach alone.

We expect that our proposed approaches can provide benefits for understanding the dynamics of complex systems whose network structures are hard to collect and define in reality, and they provide a way of uncovering cross-population diffusion across target systems or regions in a wide range of application domains. For instance, policy makers, urban planners, and service providers can benefit from human mobility patterns when making their timely decisions. Human mobility patterns have been modeled through more unprecedented human activity data than ever before [96, 145, 152, 161], but the coverage of massive activity data often depends on the technology levels of target regions, cities, or countries [18]; e.g., higher mobile phone coverage in industrialized and developed regions than developing regions. Thus, our approaches can provide a way of freeing models from such dependency on diverse data coverage in real world systems.

10.1 Summary of Contributions

In this section, we present the main contributions of this thesis, which are summarized according to the research steps discussed in Chapter 1.

10.1.1 Empirical Analysis

We investigated diffusion in social media and academic publications and provided the major characteristics of cross-population diffusion in the real world. For targeting trending topics, we proposed a way of identifying noteworthy news items from the aspects of journalistic practice "5W1H", i.e. *Who, What, Where, When, Why, and How*. We also showed solid steps of document labeling and profiling, which are important for empirical analyses to be based upon more accurate citation networks. By constructing different levels of citation networks, we showed structural diffusion patterns across heterogeneous social networks in a more principled way. This motivated us to propose the two approaches of estimating temporal dynamics of diffusion in both model-driven and model-free ways.

10.1.2 Approaches

Based on our empirical findings, we proposed the Dynamic Influence Model and macro-level information transfer across social systems as the model-driven and model-free approaches, respectively.

The model-driven approach considers the two essential features, the heterogeneity and structural connectivity of real-world networks, in a probabilistic way, and improves the accuracy of models dealing with a homogeneous population within a single social system alone. This approach is based on the assumption of a universal property of real-world networks, which helps to obtain stochastic pathways on the heterogeneous networks. Thus, it can provide the benefits of excluding detailed network topologies in a complex environment and provide a macroscopic view of cross-population diffusion.

On the other hand, the model-free approach does not require any assumptions on social interactions by using information-theoretic measures. This approach estimates information transfer at a macro level with both time-delay and memory effects, which provides different ways of understanding cross-population diffusion. Time-delay and memory effects are interpreted as behavioral characteristics of social systems, which cannot be obtained from the model-driven approach, and thus can provide different perspectives on diffusion. These non-parametric statistics provide the benefits of quantifying non-linear dynamics across heterogeneous social systems.

A meta-population scheme makes both approaches more applicable to various real-world scenarios, since it extends a social system into a more flexible space containing a meta-population.

10.1.3 Comparisons and Interpretations

We compared the model-driven and model-free approaches based on the estimated diffusion patterns with each approach. The model-driven approach reveals internal influence, intra- and inter-relationships within and between social systems, as well as external influence outside the systems. The model-free approach also provides the strength and directionality of influence between different systems (inter-relationships), except for intra-relationships and external influence, which can be compared with the results from the model-driven approach and help to enhance the model parameters further. That is, the estimated diffusion patterns from both approaches are comparable with respect to inter-relationships, which helps to determine the threshold values of the model parameters in a way that maximizes the similarities of the outcomes.

From the time-delay and memory effects of the model-free approach, we can obtain each system's behavioral characteristics of adopting new idea or information. Behavioral characteristics of a destination system reflect a source system's influence on that system whose influence is obtained from the model-driven approach. As discussed in Chapter 2, node characteristics in a network can be one of the main factors of diffusion, and thus such behavioral characteristics of systems can be considered as node properties in a macroscopic graph of global diffusion.

Thus, both approaches in conjunction provide a macroscopic picture of diffusion by filling the gap which either approach cannot provide alone. They also suggest alternative options to choose a more appropriate approach according to the experimental conditions of different application domains. We expect that our approaches can be applied to discovering mechanisms of emergent phenomena in real world systems as complex systems.

10.1.4 Implications of the Findings from Real World Systems

Based on our proposed approaches, we assessed topic-related influences among heterogeneous meta-populations in social media as well as in academic publications, which lead to global diffusion of news items and knowledge. In both cases, different information topics bring different dynamics, which provides clues of targeting populations appropriate to the context of application domains.

In addition, the future states of social systems are influenced by the previous states or the history of previous states of other social systems. These Markovian or non-Markovian properties of social systems provide clues of targeting data periods appropriate to system properties obtained by the time-delay and memory effects of macro-level information transfer.

10.2 Future Directions

The work presented here has a number of opportunities to be improved and applied in different ways as described in the following future directions.

10.2.1 Inference of Heterogeneous Networks

In this thesis, we assumed a real-world network property, a power-law degree distribution, since extensive research has shown that real-world networks are scale-free networks and exhibit a power-law degree distribution [43, 48, 91, 98, 105, 127]. However, different underlying connectivities of heterogeneous diffusion networks can be found when considering the properties of each homogeneous population and the context of information. Discovering universal properties of heterogeneous networks can provide unlimited opportunities to be applied to diverse research areas.

10.2.2 Trace of Evolving Information

Our study has focused on different trending topics and examined time-series diffusion patterns of each topic in a separate way. However, we cannot ignore that different topics can co-evolve influencing each other. Uncovering the causality of information evolution can help us to predict emerging information topics or events. Different topics showing similar diffusion trends are more likely related to each other as we for example observed such patterns from the political protests in the Middle East. There can be other factors influencing information co-evolution, such as the similarity of contact networks, same location, and timing of an event's occurrence. Thus, developing a mathematical model describing these factors can be one of future directions.

10.2.3 Synchronization of Global Diffusion

In this thesis, we uncovered common dynamics of influence from synchronous diffusion patterns across heterogeneous social systems. The factors of synchronous global diffusion need to be further improved. Based on such improvements, we can define synchronization measures and develop applications maximizing the objective functions of such measures. This can be used in diverse application domains which need to synchronize diffusion across target regions, e.g., standardization across regions showing significant gaps.

10.2.4 Social Roles in Global Diffusion

We have focused on macro-level diffusion and obtained the strength and directionality of influence at a system level. Our approaches can also be applied to micro-level diffusion within a single social system and obtain dynamics of influence at a social community or group level, given an initial setting of a social group or community definition and detection. This application can be improved to further refine a social group definition and detection in an iterative way.

10.2.5 Unified Framework

Uncovering the interplay between diffusion mechanisms and co-evolution of different information topics is challenging. This can be achieved by combining model-driven and model-free approaches into a single framework and by learning model parameters and the properties of target communities, systems, or regions in a more systematic way. Consistent and comprehensive designs of system architectures of such a framework can be one of possible future directions.

10.3 Closing Remarks

The work presented in this thesis helps us to understand diffusion of information across heterogeneous (social) networks by uncovering the underlying mechanisms in both model-driven and model-free ways. We expect that our approaches in conjunction can help to reveal a macroscopic coherent picture of momentum for social movements or knowledge innovations, formed among heterogeneous (meta-) populations in various real-world scenarios.

Appendix A

Nomenclature

A.1 Table of Symbols

Table A.1: Table of all symbols used in this thesis. Symbols used differently across chapters are omitted here (refer to each chapter for its own definition).

Symbol	Description
$h(t)$	Hazard function of time t
$a(t)$	The number of new adopters at time t
$A(t)$	The number of cumulative adopters until time t
$f(t)$	The proportion of new adopters at time t
$F(t)$	The proportion of cumulative adopters until time t
p	Coefficient of innovation in the Bass Model
q	Coefficient of imitation in the Bass Model
a	A binary random variable for the event of an individual's adoption
$P(a\|t)$	Adoption probability that an average individual adopts at time t
$P(a\|\neg a, t)$	Probability that an average individual, who has not adopted before, adopts at time t
$P_{\text{ext}}(a\|\neg a, i, t)$	Probability that an average individual from a population i, who has not adopted before, adopts at time t by external influence
p_i	Probability of external influence on a meta-population i, implying $P_{\text{ext}}(a\|\neg a, i, t)$ in short

Continued on next page

Table A.1 – *Continued from previous page*

Symbol	Description
$P_{\mathrm{int}}(a\|\neg a, i, t)$	Probability that an average individual from a meta-population i, who has not adopted before, adopts at time t by internal influence
m	The number of meta-populations
n	Total size of m meta-populations
k	The number of connected neighbors
v_i	The number of neighbors from a meta-population i who have already adopted
$\mathbf{v}$	A vector consisting of v_i for all m different meta-populations as $\mathbf{v} = (v_1, ..., v_m)^{\mathrm{T}}$
c_{ji}	Probability that an individual of a meta-population i adopts when it is influenced by a connected neighbor who is a previous adopter of a meta-population j
α	Power law coefficient
n_i	The size of a meta-population i
$\boldsymbol{\theta}_i$	Model parameters for the i-th equation except for n_i
$\boldsymbol{\Theta}_i$	Model parameters for the i-th equation including n_i, *i.e.* $\boldsymbol{\Theta}_i = \{n_i, \boldsymbol{\theta}_i\}$
$\tilde{e}_k$	Relative parameter error of the k-th parameter
$\hat{\omega}_k$	Estimated value of the k-th parameter
ω_k	Ground truth of the k-th parameter
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Stochastic processes (an ordered set of random variables)
X, Y, Z	Random variables
x_t, y_t	The realizations of X and Y at time t respectively
$\mathcal{A}_t$	Acceleration, i.e. the changes of adoption rates at time t
$p(x)$	Probability of the event x
$p(x, y)$	Probability of the joint event x and y
$p(x\|y)$	Probability of the event x given event y
$I(x)$	Information of observing the event x

Continued on next page

Table A.1 – *Continued from previous page*

Symbol	**Description**
$H(X)$	Shannon entropy of X
$H(X\|Y)$	Conditional entropy of X given Y
$I(X;Y)$	Mutual information between X and Y
$I(X;Y\|Z)$	Conditional mutual information between X and Y given Z
$I(\mathbf{X};\mathbf{Y})$	Mutual information between **X** and **Y**
$I(\mathbf{X};\mathbf{Y}\|\mathbf{Z})$	Conditional mutual information between **X** and **Y** given **Z**
$\mathrm{TE}_{Y\to X}$	Transfer entropy from **Y** to **X**
$\mathrm{TE}_{\mathrm{out}}$	Outgoing transfer entropy
$\mathrm{TE}_{\mathrm{in}}$	Incoming transfer entropy
$\mathrm{Bias}[\mathcal{I}]$	The Panzeri-Treves bias of an information-theoretic quantity $\mathcal{I}$ such as $H(X)$, $H(X\|Y)$, $I(X;Y)$, and $\mathrm{TE}_{Y\to X}$
$D_{KL}(P \parallel Q)$	Kullback-Leibler divergence of two probability distributions P and Q
k	Markov order for the previous states of **X**
l	Markov order for the previous states of **Y**
d	The length of time-delay
$\mathbf{N}, \mathbf{S}, \mathbf{B}$	Stochastic processes of News, SNS, and Blog in social media respectively
$\mathbf{C}, \mathbf{D}, \mathbf{S}$	Stochastic processes of Core, Data, and Systems in academic publications respectively

Appendix B

Proof of Equations

B.1 Proof of Equation (5.13)

By substituting Equations (5.10) to (5.12) into Equation (5.9):

$$
\begin{aligned}
&P_{\text{int}}(a \mid \neg a, i, t) \\
&= \sum_{k=1}^{n-1} \sum_{\mathbf{v}} P(a|\mathbf{v}, k, \neg a, i, t) P(\mathbf{j}|k, \neg a, i, t) P(k|\neg a, i, t) \\
&= \sum_{k=1}^{n-1} \frac{1}{\zeta(\alpha)} k^{-\alpha} \sum_{\mathbf{v}} \left(1 - \prod_{j=1}^{m} (1 - c_{ji})^{v_j}\right) \frac{k!}{v_1! \cdots v_m! (k-v)!} \prod_{i=1}^{m} P(a|i,t)^{v_i} (1-P)^{k-v} \\
&= 1 - \sum_{k=1}^{n-1} \frac{1}{\zeta(\alpha)} k^{-\alpha} \sum_{\mathbf{v}} \left(\prod_{j=1}^{m} (1 - c_{ji})^{v_j}\right) \frac{k!}{v_1! \cdots v_m! (k-v)!} \prod_{i=1}^{m} P(a|i,t)^{v_i} (1-P)^{k-v} \\
&= 1 - \sum_{k=1}^{n-1} \frac{1}{\zeta(\alpha)} k^{-\alpha} \sum_{\mathbf{v}} \frac{k!}{v_1! \cdots v_m! (k-v)!} \prod_{j=1}^{m} \left((1 - c_{ji}) P(a|j,t)\right)^{v_j} (1-P)^{k-v} \\
&= 1 - \sum_{k=1}^{n-1} \frac{1}{\zeta(\alpha)} k^{-\alpha} \left(\sum_{j=1}^{m} (1 - c_{ji}) P(a|j,t) + (1-P)\right)^{k} \text{ (by the multinomial theorem)} \\
&= 1 - \sum_{k=1}^{n-1} \frac{1}{\zeta(\alpha)} k^{-\alpha} \left(1 - \sum_{j=1}^{m} c_{ji} P(a|j,t)\right)^{k}
\end{aligned}
$$

B.2 Proof of Equation (5.15)

Let the base of the numerator in Equation (5.13) be x:

$$x \triangleq 1 - \sum_{j=1}^{m} c_{ji} P(a|j,t), \ x \in [0,1] \tag{B.1}$$

Then, having the power-law exponent α fixed, the internal new adoption probability can be viewed as a function of x:

$$P_{\text{int}}(a \mid \neg a, i, t) = 1 - \frac{1}{\zeta(\alpha)} \sum_{k=1}^{n-1} \frac{x^k}{k^\alpha} \triangleq f(x) \tag{B.2}$$

Since the derivative of $f(x)$ is:

$$f'(x) = -\frac{1}{\zeta(\alpha)} \sum_{k=1}^{n-1} \frac{x^{k-1}}{k^{\alpha-1}} \tag{B.3}$$

the Taylor expansion of $f(x)$ at $x = 1$ is:

$$\begin{aligned} P_{\text{int}}(a \mid \neg a, i, t) = f(x) \approx f(1) + f'(1)(x-1) &= -\frac{\zeta(\alpha-1)}{\zeta(\alpha)}(x-1) \\ &= \frac{\zeta(\alpha-1)}{\zeta(\alpha)} \sum_{j=1}^{m} c_{ji} P(a|j,t) \end{aligned}$$

which is equivalent with the Taylor expansion of $P_{\text{int}}(a \mid \neg a, i, t)$ at all $P(a|i,t) = 0$.

Bibliography

1. Aminer Citation Network Dataset. http://arnetminer.org/citation.
2. Arnetminer: search and mining of academic social networks. http://arnetminer.org/.
3. The DBLP Computer Science Bibliography. http://dblp.uni-trier.de/db/.
4. International AAAI Conference on Weblogs and Social Media (ICWSM) 2011 Spinn3r Dataset. http://www.icwsm.org/data/.
5. Microsoft Academic Search. http://academic.research.microsoft.com/.
6. Multiculturalism has failed in Britain. http://uk.reuters.com/article/2011/02/05/uk-britain-radicalisation-idUKTRE71401G20110205.
7. OpenCalais Web Service API. http://www.opencalais.com/documentation/calais-web-service-api.
8. Spinn3r.com. http://spinn3r.com/.
9. Wikipedia's Current Event Registry during a one month period in January, 2011. http://en.wikipedia.org/wiki/January_2011.
10. S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E. Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72, 2006.
11. E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 207–214, France, 2005.
12. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 183–194. ACM, 2008.
13. R. Agrawal, M. Potamias, and E. Terzi. Learning the nature of information in social networks. Technical report, Technical Report MSR-TR-2011-59, Microsoft Research, 2011.
14. G. Ahuja, G. Soda, and A. Zaheer. The genesis and dynamics of organizational networks. *Organization Science*, 23(2):434–448, 2012.

15. R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

16. A. Alexander-Bloch, A. Raznahan, E. Bullmore, and J. Giedd. The convergence of maturational change and structural covariance in human cortical networks. *The Journal of Neuroscience*, 33(7):2889–2899, 2013.

17. L. J. Allen, F. Brauer, P. Van den Driessche, and J. Wu. *Mathematical epidemiology*. Springer, 2008.

18. A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):1–20, 2014.

19. Y. An, J. Janssen, and E. E. Milios. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6):664–678, 2004.

20. A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–15, Las Vegas, USA, 2008. ACM.

21. R. Anderson and R. May. *Infectious diseases of humans: dynamics and control*, volume 26. Wiley Online Library, 1992.

22. S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544, 2009.

23. N. Bailey. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 1975.

24. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the International Conference on World Wide Web*, pages 519–528, Lyon, France, 2012. ACM.

25. F. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *The Annals of Applied Probability*, 7(1):46–89, 1997.

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

27. A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.

28. A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

29. A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

30. F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

31. F. M. Bass. Comments on "a new product growth for model consumer durables": The Bass model. *Management Science*, 50(12_supplement):1833–1840, 2004.

32. D. Bassett and E. Bullmore. Small-world brain networks. *The Neuroscientist*, 12(6):512–523, 2006.

33. H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 438–441, Barcelona, Spain, 2011. AAAI.

34. J. Bian, M. Xie, U. Topaloglu, T. Hudson, H. Eswaran, and W. Hogan. Social network analysis of biomedical research collaboration networks in a ctsa institution. *Journal of Biomedical Informatics*, 52:130–140, 2014.

35. P. Blackshaw and M. Nazzaro. Consumer-generated media (CGM) 101: Word-of-mouth in the age of the web-fortified consumer. *New York: Nielsen BuzzMetrics*, 2006.

36. A. S. Blazier and J. A. Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*, 3, 2012.

37. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, 2006.

38. K. W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

39. F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*. Springer, 2011.

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

41. E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

42. R. S. Burt. The network structure of social capital. *Research in Organizational Behavior*, 22:345–423, 2000.

43. M. Cha, J. Pérez, and H. Haddadi. Flash floods and ripples: The spread of media content through the blogosphere. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, USA, 2009. AAAI.

44. D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

45. J. H. Choi, G. A. Barnett, and B.-S. CHON. Comparing world city networks: a network analysis of internet backbone and air transport intercity linkages. *Global Networks*, 6(1):81–99, 2006.

46. N. Christakis and J. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

47. P. Christen. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012.

48. A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

49. M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 89–96, Barcelona, Spain, 2011. AAAI.

50. T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

51. D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the International Conference on Computational Linguistics*, pages 241–249, Beijing, China, 2010. Association for Computational Linguistics.

52. J. L. Doob. *Stochastic processes*, volume 101. New York Wiley, 1953.

53. S. N. Dorogovtsev and J. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, 2013.

54. N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

55. V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian. Scale-free brain functional networks. *Physical Review Letters*, 94(1):018102, 2005.

56. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, 29(4):251–262, 1999.

57. J. Fowler and N. Christakis. The dynamic spread of happiness in a large social network. *British Medical Journal*, 337:a2338, 2008.

58. A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2013.

59. J. Ganesh and V. Kumar. Capturing the cross-national learning effect. *Marketing Science*, pages 328–337, 1996.

60. R. Ghosh, T. Surachawala, and K. Lerman. Entropy-based classification of'retweeting'activity on twitter. *arXiv preprint arXiv:1106.0346*, 2011.

61. S. Goel, D. Watts, and D. Goldstein. The structure of online diffusion networks. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 623–638, Valencia, Spain, 2012. ACM.

62. M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028, Washington D.C., USA, 2010. ACM.

63. M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 23–32, Rome, Italy, 2013. ACM.

64. M. Götz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, USA, 2009. AAAI.

65. A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 241–250, New York, USA, 2010. ACM.

66. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the International Conference on World Wide Web*, pages 491–501, New York, USA, 2004. ACM.

67. R. Guimera, B. Uzzi, J. Spiro, and L. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

68. V. Ha-Thuc, Y. Mejova, C. Harris, and P. Srinivasan. Event intensity tracking in weblog collections. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 24–31, San Jose, USA, 2009. AAAI.

69. J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, pages 100–108, 1979.

70. H. Hethcote. The mathematics of infectious diseases. *SIAM Review*, pages 599–653, 2000.

71. C. A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.

72. S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

73. H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

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

75. K. Y. Kamath, J. Caverlee, Z. Cheng, and D. Z. Sui. Spatial influence vs. community influence: modeling the global spread of social media. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 962–971, Hawaii, USA, 2012. ACM.

76. A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

77. M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

78. M. J. Keeling and K. T. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

79. M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

80. C. T. Kelley. *Iterative methods for optimization*. Society for Industrial and Applied Mathematics, 1999.

81. D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, Washington D.C., USA, 2003. ACM.

82. D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. *Automata, Languages and Programming*, pages 99–99, 2005.

83. A. Khelil, C. Becker, J. Tian, and K. Rothermel. An epidemic model for information diffusion in manets. In *Proceedings of the 5th ACM International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*, pages 54–60. ACM, 2002.

84. M. Kim, D. Newth, and P. Christen. Modeling direct and indirect influence across heterogeneous social networks. In *Proceedings of the 7th ACM SIGKDD Workshop on Social Network Mining and Analysis (SNAKDD)*, page 9, Chicago, USA, 2013. ACM.

85. M. Kim, D. Newth, and P. Christen. Modeling dynamics of diffusion across heterogeneous social networks: News diffusion in social media. *Entropy*, 15(10):4215–4242, 2013.

86. M. Kim, D. Newth, and P. Christen. Modeling dynamics of meta-populations with a probabilistic approach: Global diffusion in social media. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 489–498, San Francisco, USA, 2013. ACM.

87. M. Kim, D. Newth, and P. Christen. Macro-level information transfer across social networks. In *Proceedings of the International conference on World Wide Web Companion*, pages 321–322, Seoul, South Korea, 2014. International World Wide Web Conferences Steering Committee.

88. M. Kim, D. Newth, and P. Christen. Trends of news diffusion in social media based on crowd phenomena. In *Proceedings of the International conference on World Wide Web Companion*, pages 753–758, Seoul, South Korea, 2014. International World Wide Web Conferences Steering Committee.

89. M. Kim, D. Newth, and P. Christen. Uncovering diffusion in academic publications using model-driven and model-free approaches. In *Proceedings of the IEEE International Conference on Big Data and Cloud Computing*, pages 564–571, Sydney, Australia, 2014. IEEE.

90. M. Kim, D. Newth, and P. Christen. Macro-level information transfer in social media: Reflections of crowd phenomena (Accepted). *Neurocomputing*, 2015.

91. M. Kim, L. Xie, and P. Christen. Event diffusion patterns in social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 178–185, Dublin, Ireland, 2012. AAAI.

92. I. Z. Kiss, M. Broom, P. G. Craze, and I. Rafols. Can epidemic models describe the diffusion of topics across disciplines? *Journal of Informetrics*, 4(1):74–82, 2010.

93. C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 441–450, New York, USA, 2010. ACM.

94. R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*, pages 337–357. Springer, 2010.

95. V. Kumar and T. Krishnan. Multinational diffusion models: An alternative framework. *Marketing Science*, pages 318–330, 2002.

96. K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One*, 9(6):e96180, 2014.

97. M. Kuperman and G. Abramson. Small world effect in an epidemiological model. *Physical Review Letters*, 86(13):2909–2912, 2001.

98. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web*, pages 591–600, Raleigh, USA, 2010. ACM.

99. O. Kwon and J.-S. Yang. Information flow between composite stock index and individual stocks. *Physica A: Statistical Mechanics and its Applications*, 387(12):2851–2856, 2008.

100. T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1059–1068, Washington D.C., USA, 2010. ACM.

101. D. H. Lee, I. W. Seo, H. C. Choe, and H. D. Kim. Collaboration network patterns and research performance: the case of korean public research institutions. *Scientometrics*, 91(3):925–942, 2012.

102. J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

103. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, Paris, France, 2009. ACM.

104. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420–429, San Jose, USA, 2007. ACM.

105. J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the SIAM International Conference on Data Mining*, volume 7, pages 551–556, Minneapolis, USA, 2007. SIAM.

106. J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Advances in Knowledge Discovery and Data Mining*, pages 380–389. Springer, 2006.

107. M. Luu, E. Lim, T. Hoang, and F. Chua. Modeling diffusion in social networks using network properties. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 218–225, Dublin, Ireland, 2012. AAAI.

108. E. Maasoumi and J. Racine. Entropy and predictability of stock market returns. *Journal of Econometrics*, 107(1):291–312, 2002.

109. D. J. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

110. V. Mahajan, E. Muller, and F. Bass. Diffusion of new products: Empirical generalizations and managerial uses. *Marketing Science*, 14(3):G79–G88, 1995.

111. C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.

112. R. B. Mars, J. Sallet, U. Schüffelgen, S. Jbabdi, I. Toni, and M. F. Rushworth. Connectivity-based subdivisions of the human right "temporoparietal junction are": evidence for different areas participating in different cortical networks. *Cerebral cortex*, 22(8):1894–1903, 2012.

113. R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(2):275–281, 2002.

114. M. Mathioudakis, N. Koudas, and P. Marbach. Early online identification of attention gathering items in social media. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 301–310, New York, USA, 2010. ACM.

115. J. J. Mcauley and J. Leskovec. Learning to discover social circles in ego networks. In *Proceedings of the Advances in neural information processing systems*, pages 539–547, Lake Tahoe, USA, 2012.

116. A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

117. S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2009.

118. J. W. Mickens and B. D. Noble. Modeling epidemic spreading in mobile environments. In *Proceedings of the ACM workshop on Wireless Security*, pages 77–86, Cologne, Germany, 2005. ACM.

119. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

120. Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6):066130, 2004.

121. Y. Moreno, M. Nekovee, and A. Vespignani. Efficiency and reliability of epidemic data dissemination in complex networks. *Physical Review E*, 69(5):055101, 2004.

122. S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the International Conference on World Wide Web*, pages 913–924, Seoul, Korea, 2014. International World Wide Web Conferences Steering Committee.

123. S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–41, Beijing, China, 2012. ACM.

124. M. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, 2012.

125. M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

126. M. E. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

127. M. E. Newman. *Networks: an introduction*. Oxford University Press, 2010.

128. M. E. Newman, A.-L. Barabasi, and D. J. Watts. *The structure and dynamics of networks*. Princeton University Press, 2006.

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

130. S.-C. Niu. A stochastic formulation of the bass model of new-product diffusion. *Mathematical Problems in Engineering*, 8(3):249–263, 2002.

131. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

132. W. Pan, N. Aharony, and A. Pentland. Composite social network for predicting mobile apps installation. In *Proceedings of the AAAI Conference*, pages 821–827, San Francisco, USA, 2011. AAAI.

133. L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

134. S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, 98(3):1064–1072, 2007.

135. S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

136. H. W. Park. Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, 25(1):49–61, 2003.

137. T. Parsons. *The social system*. Free Press, Glencoe, IL, 1951.

138. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.

139. R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

140. A. L. Porter and I. Rafols. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745, 2009.

141. K. G. Provan, A. Fish, and J. Sydow. Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of Management*, 33(3):479–516, 2007.

142. W. Putsis Jr, S. Balasubramanian, E. Kaplan, and S. Sen. Mixing behavior in cross-country diffusion. *Marketing Science*, pages 354–369, 1997.

143. R. Q. Quiroga and S. Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185, 2009.

144. A. Raj, A. Kuceyeski, and M. Weiner. A network diffusion model of disease progression in dementia. *Neuron*, 73(6):1204–1215, 2012.

145. C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

146. E. Rogers. Diffusion of innovations. 1962.

147. D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics. In *Proceedings of the International Conference on World Wide Web*, Hyderabad, India, 2011. ACM.

148. R. Rowe, G. Creamer, S. Hershkop, and S. J. Stolfo. Automated social hierarchy detection through email network analysis. In *Proceedings of the 1st ACM SIGKDD Workshop on Web Mining and Social Network Analysis*, pages 109–117, San Jose, USA, 2007. ACM.

149. S. Ruan and W. Wang. Dynamical behavior of an epidemic model with a nonlinear incidence rate. *Journal of Differential Equations*, 188(1):135–163, 2003.

150. M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

151. M. Schilling and C. Phelps. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, 53(7):1113–1126, 2007.

152. C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

153. T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

154. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

155. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

156. X. Shi, B. Tseng, and L. Adamic. Information diffusion in computer science citation networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, USA, 2009. AAAI.

157. B. Shulgin, L. Stone, and Z. Agur. Pulse vaccination strategy in the sir epidemic model. *Bulletin of Mathematical Biology*, 60(6):1123–1148, 1998.

158. B. A. Simmons, F. Dobbin, and G. Garrett. Introduction: The international diffusion of liberalism. *International Organization*, 60(04):781–810, 2006.

159. R. Sinatra and M. Szell. Entropy and the predictability of online life. *Entropy*, 16(1):543–556, 2014.

160. D. Smilkov and L. Kocarev. Influence of the network topology on epidemic spreading. *Physical Review E*, 85(1):016114, 2012.

161. S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS One*, 8(12):e81707, 2013.

162. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

163. X. Song, Y. Chi, K. Hino, and B. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *Proceedings of the International Conference on World Wide Web*, pages 191–200. ACM, 2007.

164. O. Sporns, D. Chialvo, M. Kaiser, C. Hilgetag, et al. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, 2004.

165. K. Starbird and L. Palen. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 7–16, Seattle, USA, 2012. ACM.

166. D. Strang and S. A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, pages 265–290, 1998.

167. E. Sun, I. Rosenn, C. Marlow, and T. M. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, USA, 2009. AAAI.

168. J. Tang, A. C. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(6):975–987, 2012.

169. J. Tang, L. Yao, D. Zhang, and J. Zhang. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(1):1–44, 2010.

170. J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic level expertise search over heterogeneous networks. *Machine Learning Journal*, 82(2):211–237, 2011.

171. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, Las Vegas, USA, 2008. ACM.

172. E. C. Titchmarsh. The theory of the riemann zeta-function. 1951.

173. A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, 1995.

174. G. Tsoumakas and I. Katakis. Multi-label classification. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

175. A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

176. A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Physical Review E*, 65(6):066130, 2002.

177. G. Ver Steeg and A. Galstyan. Information transfer in social media. In *Proceedings of the International Conference on World Wide Web*, pages 509–518, Lyon, France, 2012. ACM.

178. G. Ver Steeg and A. Galstyan. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 3–12, Rome, Italy, 2013. ACM.

179. R. Wallace and D. Wallace. Socioeconomic determinants of health: community marginalisation and the diffusion of disease and disorder in the united states. *BMJ*, 314(7090):1341, 1997.

180. C. Wang and B. A. Huberman. How random are online social interactions? *Scientific Reports*, 2 (2012), 2012.

181. S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.

182. D. Watts and S. Strogatz. Collective dynamics of small-world networks. *nature*, 393(6684):440–442, 1998.

183. L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, 2012.

184. S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

185. Z. Xiang and U. Gretzel. Role of social media in online travel information search. *Tourism Management*, 31(2):179–188, 2010.

186. J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the International Conference on Data Mining*, pages 599–608, Sydney, Australia, 2010. IEEE.

187. S.-H. Yook, H. Jeong, and A.-L. Barabási. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Sciences*, 99(21):13382–13386, 2002.

188. H. P. Young. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *The American Economic Review*, 99(5):1899–1924, 2009.

189. X. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance modeling of epidemic routing. In *NETWORKING 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, pages 827–839. Springer, 2006.

190. W. Zhu and J. Guan. A bibliometric study of service innovation research: based on complex network analysis. *Scientometrics*, 94(3):1195–1216, 2013.