ABSTRACT

Title of dissertation: THE DYNAMICS OF MULTI-MODAL NETWORKS

Hossam Sharara, Doctor of Philosophy, 2012

Dissertation directed by: Professor Lise Getoor
Department of Computer Science

The widespread study of networks in diverse domains, including social, technological, and scientific settings, has increased the interest in statistical and machine learning techniques for network analysis. Many of these networks are complex, involving more than one kind of entity, and multiple relationship types, both changing over time. While there have been many network analysis methods proposed for problems such as network evolution, community detection, information diffusion and opinion leader identification, the majority of these methods assume a single entity type, a single edge type and often no temporal dynamics. One of the main shortcomings of these traditional techniques is their inadequacy for capturing higher-order dependencies often present in real, complex networks.

To address these shortcomings, I focus on analysis and inference in dynamic, multi-modal, multi-relational networks, containing multiple entity types (such as people, social groups, organizations, locations, etc.), and different relationship types (such as friendship, membership, affiliation, etc.). An example from social network theory is a network describing users, organizations and in-

terest groups, where users have different types of ties among each other, such as friendship, family ties, etc., as well as affiliation and membership links with organizations and interest groups. By considering the complex structure of these networks rather than limiting the analysis to a single entity or relationship type, I show how we can build richer predictive models that provide better understanding of the network dynamics, and thus result in better quality predictions.

In the first part of my dissertation, I address the problems of network evolution and clustering. For network evolution, I describe methods for modeling the interactions between different modalities, and propose a co-evolution model for social and affiliation networks. I then move to the problem of network clustering, where I propose a novel algorithm for clustering multi-modal, multi-relational data. The second part of my dissertation focuses on the temporal dynamics of interactions in complex networks, from both user-level and network-level perspectives. For the user-centric approach, I analyze the dynamics of user relationships with other entity types, proposing a measure of the ”loyalty” a user shows for a given group or topic, based on her temporal interaction pattern. I then move to macroscopic-level approaches for analyzing the dynamic processes that occur on a network scale. I propose a new differential adaptive diffusion model for incorporating diversity and trust in the process of information diffusion on multi-modal, multi-relational networks. I also discuss the implications of the proposed diffusion model on designing new strategies for viral marketing and influential detection. I validate all the proposed methods on several real-world networks from multiple domains.

THE DYNAMICS OF MULTI-MODAL NETWORKS

by

Hossam Samy Elsaid Ibrahim Sharara

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Lise Getoor, Chair/Advisor
Professor Ashok Agrawala
Professor Carl Kingsford
Professor William Rand
Professor Mark Austin

© Copyright by
Hossam Samy Elsaid Ibrahim Sharara
2012

Foreword

Portions of this dissertation are derived from research and publications co-authored by the candidate and published elsewhere. Chapter 2 is based on the paper *Co-evolution of social and affiliation networks*[122]. The multi-relational clustering work in Chapter 3 is based on the paper *Multi-relational Affinity Propagation*[96]. The loyalty measure proposed in Chapter 4 is based on the journal article *Understanding Actor Loyalty to Event-Based Groups in Affiliation Networks*[99]. The information diffusion and viral marketing work in Chapters 5 and 6 are extensions of the work in *Differential Adaptive Diffusion: Understanding Diversity and Learning whom to Trust in Viral Marketing*[98]. Finally, the active surveying model proposed in Chapter 7 is based on the paper *Active Surveying: A Probabilistic Approach for Identifying Key Opinion Leaders*[97].

Acknowledgments

First and foremost, I would like to sincerely thank my advisor, Prof. Lise Getoor, for her continuous help and support throughout my PhD research. She is the one who introduced me to the field of relational machine learning and network analysis, and opened up a whole new spectrum of ideas and research opportunities in this very interesting domain that build up upon my prior interest in data mining and machine learning. She has been an excellent mentor and tutor who always guided me for structuring and organizing my ideas, setting high standards to achieve the best quality for my work, and enhancing my skills to be a good researcher, bearing my best interest in mind. On the personal side, her moral support, appreciation for my work, and continuous encouragement have been of extreme value for helping me getting through my PhD studies.

I would also like to thank my colleagues in the LINQs group - Galileo Namata, Elena Zheleva, Mustafa Bilgic, Walaa Moustafa, Lilyana Mihalkova, Stanley Kok, Bert Huang, Stephen Bach, Jay Pujara, Ben London, Alex Memory, and Theodoros Rekatsinas, for their useful discussions, and providing a fruitful collaborative work atmosphere. I would like to specially thank Elena, Galileo, Mustafa, and Walaa for their invaluable friendship and continuous support.

I thank all my collaborators, especially Prof. William Rand from Robert H. Smith School of Business, who opened up a whole new perspective on the applications of my research in marketing domains. Prof. Rand has always provided me with helpful insights throughout our collaborative research, and with career

advice and opportunities during my job-hunting process. I also thank Prof. Lisa Singh from Georgetown University for our fruitful collaboration.

I owe my deepest gratitude to my wife, Hend, for taking good care of our little family, and always being there for me. I couldn't have made it through without her patience throughout my prolonged working hours and her continuous support through stressful times. I also thank our little son, Youssef, for always bringing joy and happiness to our life.

I would also like to thank my Egyptian friends here in College Park, especially my social support group who provided me with a warm, loving environment which helped me and my family settle in College Park, endure our homesickness, and and get through various tough times.

Finally, I would like to thank my parents, who have continuously supported my decisions, and encouraged me to pursue the path I desire in life. Without them, this journey would have not been possible.

Table of Contents

List of Figures viii

1 Introduction 1
1.1 Motivation . . . 2
1.1.1 Social Networks . . . 3
1.1.2 Scientific Networks . . . 6
1.2 General Notation and Definitions . . . 9
1.3 Contributions and Organization . . . 10

I Multi-Modal Networks Structure and Organization 13

2 The Co-evolution of Social and Affiliation Networks 14
2.1 Introduction . . . 14
2.2 Related work . . . 17
2.2.1 Evolution of social networks . . . 17
2.2.2 Evolution of affiliation networks . . . 18
2.3 Observations . . . 19
2.3.1 Group size distribution . . . 20
2.3.2 Node degree vs. Average number of group affiliations . . . 20
2.3.3 Distribution of the number of group affiliations . . . 21
2.3.4 Properties of group members . . . 22
2.4 Co-evolution properties and model . . . 24
2.4.1 Events . . . 25
2.4.2 Desired properties . . . 25
2.4.3 Co-evolution model . . . 26
2.5 Experiments . . . 32
2.5.1 Synthetic data . . . 32
2.5.2 Real data . . . 37
2.5.3 Comparison with the naïve model . . . 39
2.6 Conclusion . . . 41

3 Multi-Relational Affinity Propagation 42
3.1 Introduction . . . 42
3.2 Related Work . . . 45
3.3 Method . . . 47
3.3.1 Model Formulation . . . 51
3.3.2 Message Derivation . . . 53
3.4 Experimental Evaluation . . . 59
3.4.1 Synthetic Data . . . 60
3.4.2 Social Media Data . . . 61
3.5 Conclusion . . . 65

II The Temporal Dynamics of Multi-Modal Networks 66

4 Understanding Actor Loyalty to Groups in Affiliation Networks 67
4.1 Introduction . . . 68
4.2 Loyalty Background . . . 70
4.3 Modeling time-varying event-based groups . . . 73
4.4 Loyalty of Individuals to Affiliation Groups . . . 77
4.5 Loyalty Analysis on Individual Data Sets . . . 84
4.5.1 Scientific Publication Network . . . 84
4.5.2 Senate Bill Sponsorship Network . . . 86
4.5.3 Dolphin Social Network . . . 90
4.6 Comparative Loyalty Analysis . . . 93
4.7 Comparison with centrality measures . . . 95
4.8 Conclusion . . . 98

5 Differential Adaptive Diffusion: Understanding Diversity and Trust Dynamics in Complex Networks 99
5.1 Introduction . . . 101
5.2 Background . . . 103
5.3 Case Study: Digg . . . 106
5.3.1 Analysis . . . 107
5.4 Differential Adaptive Diffusion . . . 110
5.5 Influentials . . . 114
5.6 Experimental Evaluation . . . 116
5.6.1 Predicting Adoptions . . . 116
5.6.2 Identifying Influentials . . . 119
5.7 Conclusion . . . 120

6 Adaptive Viral Marketing 122
6.1 Introduction . . . 123
6.2 Background . . . 125
6.3 Conceptual Model . . . 127
6.4 Experiments . . . 130
6.4.1 Fully Observable Mode . . . 131
6.4.2 Learning Preferences Mode . . . 132
6.4.3 Effect of Spammers . . . 133
6.5 Conclusion . . . 135

7 Active Surveying: A Probabilistic Approach for Identifying Key Opinion Leaders 136
7.1 Introduction . . . 136
7.2 Background . . . 139
7.2.1 Opinion Leader Identification . . . 139
7.2.2 Active Learning . . . 140
7.3 Problem Description . . . 142

7.4 Method 147
7.5 Experimental Evaluation 150
7.6 Conclusion 154

8 Conclusion and Future Directions 155
8.1 Summary of Contributions 155
8.1.1 Network Evolution 155
8.1.2 Multi-relational Clustering 156
8.1.3 Bi-modal Interaction Dynamics 157
8.1.4 Information Diffusion 158
8.1.5 Adaptive Viral Marketing 159
8.1.6 Active Surveying 159
8.2 Future Directions 160
8.3 Conclusion 161

Bibliography 162

List of Figures

1.1 Social network example . . . 3
1.2 A dynamic, multi-modal, multi-relational view of the social network example . . . 4
1.3 Scientific network example . . . 7

2.1 Distribution of group sizes on a log-log scale. . . . 21
2.2 Node degree vs. average number of group affiliations . . . 22
2.3 Distribution of the number of group affiliations for nodes with different degrees. . . . 23
2.4 Ratio of the number of singletons to the group size (upper series) and ratio of the maximum degree to the group size (lower series). 24
2.5 Degree distribution in a synthetic network . . . 33
2.6 Densification in a synthetic network . . . 34
2.7 Degree vs. average number of group affiliations on varying the parameter (ρ). . . . 35
2.8 Group size distribution on varying the parameter (τ) . . . 36
2.9 Group size vs. member attributes on varying the parameter (η) (dashed line: % ratio of singletons to group size, solid line: % ratio of maximum degree to group size). . . . 38
2.10 The affiliation properties produced by the naïve model . . . 40

3.1 Sample bimodal network . . . 49
3.2 Multi-relational Affinity Propagation Model . . . 50
3.3 The performance of different clustering approaches for varying the levels of network assortativity. . . . 60
3.4 The effect of varying the cost parameters (θ, ω) on the net similarity and the modularity of the output clustering . . . 62

4.1 center . . . 75
4.2 Single actor dynamic affiliation example . . . 79
4.3 The evolution of loyalty over time for the affiliation network example 82
4.4 The effect of the smoothing factor in calculating group loyalty . . . 83
4.5 The average topic loyalty for the scientific publication network . . 85
4.6 The average topic loyalty grouped by institution type for the scientific publication network. . . . 86
4.7 Average topic loyalty across all topics in the senator bill sponsorship network . . . 87
4.8 Changing loyalty over time for Edward Kennedy in the senate bill sponsorship network . . . 89
4.9 Average topic loyalty of 2008 presidential candidates in the senate bill sponsorship network . . . 90
4.10 Average location loyalty for dolphins . . . 92
4.11 Average location loyalty for different dolphins' age groups . . . 93

4.12 Loyalty Comparison Across Data Sets 94
4.13 Loyalty vs. Centrality for Scientific Publication Network 96
4.14 Author publications in "Information Visualization" topic 97

5.1 Topic distribution of stories in Digg dataset 107
5.2 KL-divergence between the topic distribution of users' submissions and diggs. 109
5.3 KL-divergence between uniform topic distribution and users' submissions 110
5.4 Heat map of the average number of diggs for different values of topic divergence between peers across time. 112
5.5 ROC performance of two comparison models (Bernoulli and Bernoulli-PC) and the proposed model (Adaptive) on the basis of the False Positive Rate (FPR) and True Positive Rate (TPR) for each model. . 118
5.6 Average number of diggs/post for the top 10% influential users in *Digg.com* 119

6.1 Fully observable mode: Varying the conservation parameter α . . . 131
6.2 Learning preferences mode: Varying the conservation parameter α 133
6.3 Varying the percentage of spammers at ($\alpha = 0.5$) 134

7.1 Example *candidates* and *leaders* sets 144
7.2 Feature generation for an example author network 145
7.3 The percentage of respondents (y-axis) needed to identify k% of the opinion leaders (x-axis) at ($\alpha = 2$) 152
7.4 The percentage reduction in required respondents to identify k% of the opinion leaders at ($\alpha = 2$) 153

Chapter 1

Introduction

The unprecedented growth in the availability of network data has recently drawn the attention of researchers from multiple disciplines to network analysis. For instance, with the proliferation of online social networks, researchers are now able to observe and analyze social interactions between individuals on a massive scale. Other examples include biological networks, scientific collaboration networks, and transportation networks. Analyzing these networks enables us to understand the underlying factors that govern the structures and the behavior of the entities involved, and in some cases allow us to predict future interactions.

Much of the existing literature limits the analysis to a static snapshot of the network, focusing on a single type of relationship, or single-mode of interactions, between the target entities. However, networks are dynamic by nature, and often encompass different types of entities and relationships, allowing for complex structures. Thus, limiting the analysis to static, single-mode snapshots of the network interactions results in the loss of a wealth of information that could lead to better understanding and prediction.

In my dissertation, I focus on reasoning about the dynamics of multi-modal, multi-relational networks, analyzing and modeling the different types of interactions that occur within this type of networks, and understanding how these in-

teractions evolve over time. My hypothesis is that incorporating the additional network modalities will enhance the capability of different network models in both interpreting existing phenomena in complex networks and predicting future interactions.

1.1 Motivation

Until recently, much of the research effort in statistics, machine learning and data mining has focused on problems in which data is assumed to be independent and identically distributed (iid). However, as the underlying systems became more complex, especially with the widespread use of the internet, there was a growing need for more advanced methods that can take into account the inherent dependencies between different instances. Hence, statistical relational learning (SRL) [35] methods were developed to leverage these relationships in order to improve the performance of learning and mining methods.

Although leveraging these relationships resulted in significant performance gains over the traditional methods, reasoning about different network modalities in isolation loses a wealth of information present in both the dynamics of different relationships, as well as the mutual effects across different modes.

Next, I discuss examples from two domains to illustrate the utility of analyzing networks at different abstraction levels, taking into account both the dynamic aspects and the different modalities involved at both the entity and the interaction levels.

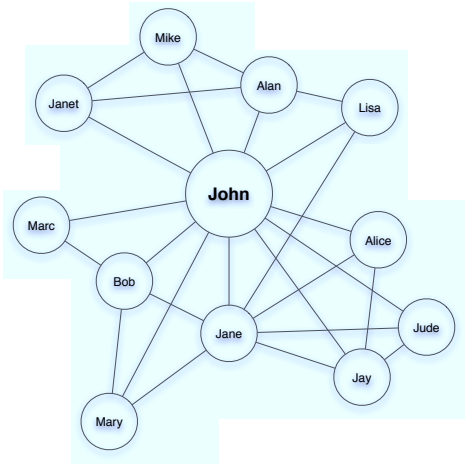


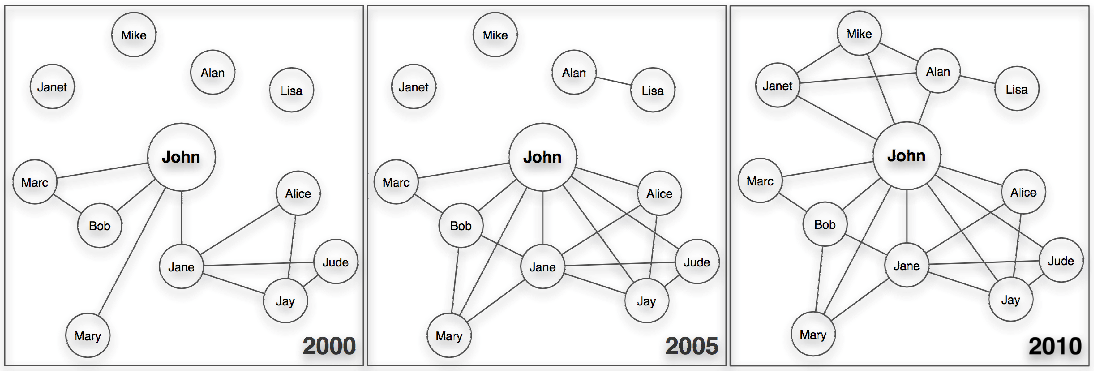
Figure 1.1: Social network example

1.1.1 Social Networks

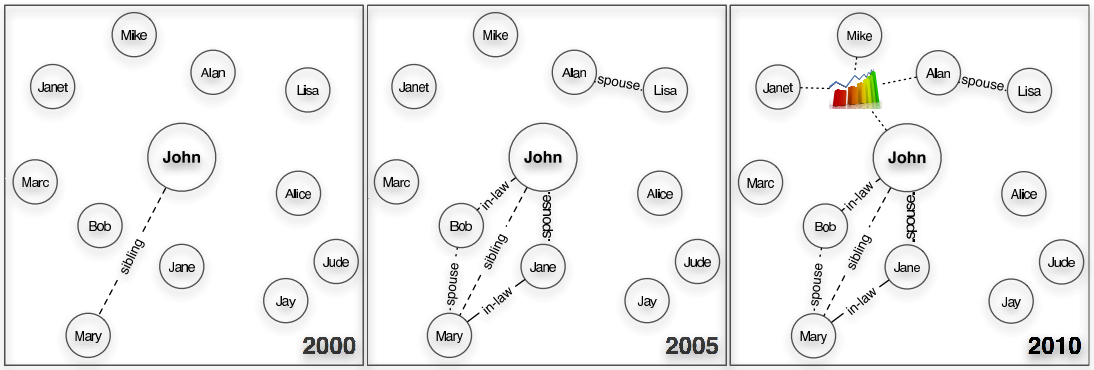
Human networks are complex systems of interactions, encompassing different types of social relationships among individuals, such as friendship, familial, or collegiate relationships. There are numerous factors governing the behavior of individuals in these networks, some of which are dependent on the type of relationships, while others are cross-relational.

Consider a simple example of a friendship network illustrated in Figure 1.1. This social network represents the ego-network of "John," showing all the friendship relationships he has with his peers. As noted in this representation, the network is static and includes only friendship links. The conclusions that can be drawn from this representation are limited to some structural observations, with neither a clear interpretation of when or how these friendships were created, nor whether they have any effect on any other hidden interactions "John" is involved in.

By introducing the temporal dimension, as illustrated in Figure 1.2(a), the



(a) Temporal evolution of the friendship links



(b) Temporal evolution of family ties and organizational affiliation links

Figure 1.2: A dynamic, multi-modal, multi-relational view of the social network example

temporal evolution of this friendship network can be observed. Figure 1.2(a) shows three snapshots of the friendship network at years 2000, 2005, and 2010. By tracking the evolution of "John's" social network, we can see that he started with only four friends in the year 2000, two of which are also friends themselves. Five years later, "John" befriended all the friends of "Jane," while "Jane" became a friend to both "Mary" and "Bob." This observation can be attributed to different factors such as the typical evolution of social networks, where "John" is increasing his social circle, or that "Jane" brought in her social connections to strengthen the friendship with "John," or for any latent factors that are not observed in the

data. In 2010, "John" introduced three additional individuals to his social network that weren't connected to him or any of his friends before that point.

Adding in the temporal aspect helped to understand the steps that the social network has undergone to reach its present structure. This dynamic analysis can then be used to predict which friendship links are probable to occur in the future, as well as gaining insights in studying the diffusion dynamics and the influence between different individuals in the social network. However, this is still insufficient to understand the causal mechanism for the formation of these friendship links, or to study the effects of these relationships on other types of interactions.

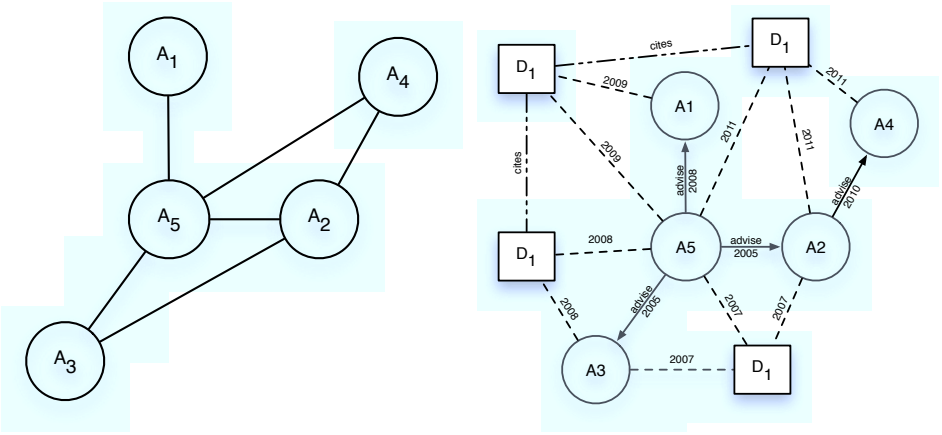
Next, I consider other relationships that exist among the target set of individuals and how they change over time. Figure 1.2(b) shows both the family and organizational affiliation relationships among individuals in "John's" social network. First, note that "Mary" and "John" are siblings, and hence the friendship relationship between them is caused by this family tie. By investigating the evolution of "John's" family network, we discover that "John" and "Jane" got married in 2005, as well as "John's" friend "Bob" and his sister "Mary." In the light of this additional information, we can now hypothesize that the friendship relationships that "John" created with "Jane's" friends were caused by their marriage, which also interprets the new friendship links that occurred between "Jane" and her now sister-in-law, "Mary," and her husband. We might also suggest that the friendship relationship between "John" and "Bob" might have had an effect on having "Bob" and "John's" sister, "Mary," getting married. Finally, the additional

friendship links that "John" introduced with "Jane," "Mike," and "Alan" in 2010, are directly correlated with the *employed-by* relationships they share with the same company, which represent a different node type in the example network. We can then presume that "John's" new affiliation is the reason behind forming these new friendship links.

This simple example illustrates how different types of relationships between different entities have mutual effects on each other, and how these relationships progress over time. The hypotheses and conclusions that could be drawn by incorporating different modalities and the dynamics of the social network, are significantly different from the ones derived from the static, single mode network snapshot. Doing this type of analysis on social networks gives us better insights into the interaction dynamics and the causal mechanism underlying the network evolution.

1.1.2 Scientific Networks

Next, I consider another example from a different domain: scientific networks. These types of networks include scientific and academic collaboration networks, such as citation networks, authorship networks, and scientific collaboration networks. Scientific networks are multi-modal by nature, including different types of entities, such as researchers and publications in authorship networks, and different types of relationships, such as co-authorships and citations between researchers. One of the commonly used methods for modeling scientific



(a) Co-authorship network example

(b) Multi-modal view of authorship, citation and academic advisory networks

Figure 1.3: Scientific network example

networks is projecting the network modalities of interest onto a single mode, then analyzing the resulting single mode network.

Figure 1.3(a) shows a simple co-authorship network between five authors. The edges between authors in the co-authorship network indicate that the corresponding authors have co-authored a publication together. The preliminary analysis of this example network shows that author A_1 is probably working in a different research area than authors A_2, A_3, and A_4, indicated by the fact that there does not exist any collaboration between them. There is little that can be concluded from the information provided by this network.

However, by extending our analysis to the dynamics of different network modalities, shown in Figure 1.3(b), we discover the existence of numerous factors that were not observable in the simpler, single mode version of the network. First, by analyzing the academic advisory network, indicated by the directed solid edges, we note that authors A_1, A_2 and A_3 are actually students advised

by author A_5 at different points in time. Also, the recent advice relationship from author A_2 to author A_4 suggest that A_2 has already graduated and is now holding an academic position, with A_4 being her student. The authorship network, indicated by the dashed line between the authors and the corresponding publications, disambiguates the 3-cliques in Figure 1.3(a), showing that both of them represent the corresponding authors working on a single publication. Finally, the citation relationship between publication D_1, and both publications D_2 and D_3, suggests that our judgement that author A_1 is working in a different research area is probably wrong. The missing collaboration between author A_1 and authors A_2 and A_3 can be attributed to the fact that A_1 has recently joined the group, indicated by the date of the advisory relationship between her and author A_5.

This small example of a co-authorship network in a typical research group illustrates the deficiencies in analyzing a projected, single mode network. Analyzing the full network, taking into consideration both the temporal dynamics as well as the different modalities, is capable of revealing numerous factors that directly impact the accuracy of the predicted model.

The previous examples show the utility of the different network modalities as well as their temporal dynamics in understanding the dependencies that need to be captured in causal, predictive and discriminative network models. In my dissertation, I focus on incorporating this additional information in the latter two types of network models, as well as other tasks related to general network analysis. Next, I discuss the general notations and definitions used to represent multi-modal, multi-relational networks.

1.2 General Notation and Definitions

Network data is often represented as a graph $G(V, E)$, where entities are represented by nodes ($v \in V$), and relationships are represented by corresponding edges ($e \in E$). Both nodes and edges can be associated with a set of features describing the corresponding entity or relationship. For representing time in dynamic networks $G_t(V_t, E_t)$, the graph elements are often associated with a temporal variable representing the creation time of the corresponding entity or relationship.

Multi-modal networks refer to networks comprised of multiple entity types, while multi-relational networks refer to different types of relationships across the underlying entities. For simplicity, I use the term *multi-modal networks* hereafter to refer to multi-modal, multi-relational networks by generalizing the network modalities to include both node and edge types. These networks can be represented by adding a type construct to the graph nodes and edges. Thus, a multi-modal network can be represented a graph $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{V^1, V^2, \ldots, V^n\}$, and $\mathcal{E} = \{E^1, E^2, \ldots, E^m\}$. Each set $V^i \in \mathcal{V}$ represents a set of nodes of type i, and each set $E^j \in \mathcal{E}$ represent a set of edges of type j. This representation enables us to account for the heterogeneity in both the entities and relationships. An example of this type of networks is social and affiliation network, where the network is represented as a multi-modal graph among individuals and organizations. The nodes are represented as $\mathcal{V} = \{V^{person}, V^{organization}\}$, while the social links exist among the "*person*"-type nodes ($E^{social} \subset V^{person} \times V^{person}$), and the affiliation

links exist only across node types ($E^{affiliation} \subset V^{person} \times V^{organization}$).

To account for the network dynamics, I associate a temporal variable with the elements of the complex graph $G_t(\mathcal{V}_t, \mathcal{E}_t)$, denoting the creation time of the associated entity. Thus, a dynamic, multi-modal network can be represented as a set of graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_T\}$, where each graph corresponds to a static, multi-modal network at the corresponding time point. This representation gives us the required flexibility to both analyze the dynamics of each type of relationship separately, as well as investigate the cross-relation effects as dependencies between the corresponding graphs.

1.3 Contributions and Organization

This thesis is organized into two parts that cover different aspects of multi-modal networks. In the first part, I focus on the construction and organization of multi-modal networks. I start by investigating the evolution of multi-modal networks over time in Chapter 2. As an example setting, I analyze the growth patterns and relationships that occur in social and affiliation networks. Based on the observed growth patterns, I then propose a coupled generative model that captures the statistical properties of complex networks. I show that the proposed co-evolution model is able to mimic the evolution of real multi-modal networks, bringing new insights about the role of friendship in joining social groups.

After modeling the evolution process in multi-modal networks, I move in Chapter 3 to investigating network clustering as an initial data exploration ap-

proach for characterizing the organization of complex networks. Though there have been numerous approaches for clustering, they are often limited to data from the same type, and focus on either attribute similarity or structural connectivity. To address these shortcomings, I propose a multi-relational affinity propagation model for clustering complex networks. My proposed approach extends the message-passing affinity propagation clustering framework by encoding additional soft relational constraints to capture the dependencies across different node types. This formulation facilitates the exploration of the middle ground between feature-based similarity clustering, community detection and block modeling in complex networks. I evaluate the performance of the algorithm qualitatively and quantitatively using a variety of evaluation measures, and show that it outperforms a number of other baselines.

In the second part of my dissertation, I focus on the temporal dynamics of the interactions that occur in multi-modal networks, from both user-centric and network-level perspectives. For the user-centric approach, I analyze the dynamics of user relationships with other entity types in Chapter 4, proposing a measure of the ”loyalty” a user shows for a given group or topic based on her temporal interaction pattern. I evaluate the proposed measure on networks from various domains, and show the utility of the loyalty measure for analyzing the dynamic behavior of users and quantifying the degree of their commitment to different time-varying affiliation groups.

I then move to macroscopic-level approaches for dynamic processes that occur on a network scale. As an example of such processes, I analyze the diffusion

of information in complex networks in Chapter 5. I propose a novel information diffusion model which captures diversity across different product campaigns and provides a means for incorporating the trust among users in the network and their effects on the adoption process. I show that the proposed adaptive diffusion model is able to better predict future adoptions compared to other diffusion models. I also propose a method of influential identification based on the proposed differential adaptive diffusion model, and show that it outperforms existing structure-based approaches. I then discuss the implications of the proposed model on viral marketing strategies in Chapter 6.

In cases where user preferences and historical interactions are unavailable, analysts often resort to either primary methods, such as surveys, or secondary methods, such as proxy interactions, for inferring the influence network among users. Primary methods rely on surveys and questionnaires that are directly sent to the people in the target population, while secondary methods rely on the network properties, such as different centrality measures, of alternative "proxy" networks that are assumed to reflect influence, such as collaboration, co-authorship, and citation networks. To augment my study of influential detection, in Chapter 7 I propose an active surveying method for leveraging the different modalities of the secondary networks, to guide the surveying process and minimize the amount of primary data required. I show that active surveying result in significant cost reduction in identifying opinion leaders, without sacrificing the integrity of the process.

Part I

Multi-Modal Networks Structure and Organization

Chapter 2

The Co-evolution of Social and Affiliation Networks

Most studies on network evolution focus on proposing generative models, which capture the statistical properties of real-world networks related only to a single type of link formation. There have been very few studies that analyze the evolution process of multi-modal networks, where different types of nodes and edges exist. In this chapter, I analyze the evolution process of both social and affiliation links in a network comprised of people and social groups. I propose a novel generative model which captures the co-evolution of both social and affiliation networks. I show that coupling the evolution process between different network modes better captures the statistical properties observed in real multi-modal networks than a more simplistic approach that handles each mode separately.

2.1 Introduction

In recent years, there has been a proliferation of online social networks. Many of the networks have millions of users, and allow complex interactions through linking to friends, public messaging, photo commenting, participating in groups of interest, and many others. Studies have been performed to characterize and explain the behavior of users, and most of them concentrate on mod-

eling how users join the network and form links to each other. Little is known about how different types of interaction influence each other. In this chapter, I address the problem of modeling social network generation explaining both link and group formation.

In social networks, users are linked to each other based on a binary relationship such as friendship, co-working relation, business contact, etc. Social networks often contain relationships other than friendship, such as affiliation links, in which users are linked to groups of interest, and groups are linked to their members. In this study, I use three large datasets from online social and affiliation networks, and discover a number of interesting properties. The datasets are from Flickr, LiveJournal and YouTube collected by Mislove et al. [75].

Using the previously studied and newly observed statistical properties of these networks, I propose a generative model for social and affiliation networks. The model explains the complex process of network formation, and captures a number of affiliation network properties which have not been captured by a model before: power-law group size distribution, large number of singletons (group members without friends in the group), power-law relation between the node degree and the average number of group affiliations, and exponential distribution of the number of group affiliations for nodes of a particular degree. My findings are important for understanding the evolution of real-world networks and suggest that the process is more complex than a naïve model in which groups are added to a fully evolved social network. They also show that users join groups for different reasons and having friends in the group is often not nec-

essary. This suggests that information spreads in the network through channels other than friendship links, and this observation has implications on information diffusion and group recommendation models.

In addition, this model can be used for synthetic network generation. This is an important application because real-world network datasets are often proprietary and hard to obtain. Controlling network parameters allows the generation of datasets with different properties which can be used for thorough exploration and evaluation of network analysis algorithms.

My contributions include the following:

- I discover a number of new properties in social and affiliation networks.
- I propose the first generative model for network evolution which captures the properties of both real-world social and affiliation networks.
- I provide a thorough evaluation of the model which shows its flexibility for synthetic data generation.

Notation. I study the interactions of two graphs, the social network graph, G_s, and the affiliation graph, G_a. For the purposes of the study, a social network is a graph $G_s = \{V, E_s\}$ which has one type of node corresponding to the users that participate in it. Nodes can form links which can be directed or undirected; $e_s(v_i, v_j, t)$ denotes the link that v_i and v_j have formed at time t. A directed link is formed whenever one user links to another. An undirected link requires the approval of both parties in order to be formed.

In an affiliation network $G_a = \{V, H, E_a\}$, there are two types of nodes, the social network users V and the groups H that they have formed. I represent the network as a bipartite graph in which undirected links $e_a(v_i, h_j, t)$ are formed between user v_i and group h_j at time t when this user becomes a member of the group. There are a number of reasons why groups are formed. For example, groups can exist because of a common interest, such as philately or book-reading clubs; they can be based on common business relation, such as an employing company; or they can be based on common personal traits, such as geographic location. What is common between the groups that I study is that users have voluntarily chosen to be part of the group, as opposed to clustered together by a group detection algorithm.

2.2 Related work

The evolution of social and affiliation networks exhibits a number of properties previously studied in the literature. I describe some of them in more detail in Section 2.4.2.

2.2.1 Evolution of social networks

The majority of literature on analyzing network properties has focused on friendship networks or actor-actor networks in general. Studying the static snapshots of graphs has led to discovering properties such as the 'small-world' phenomenon [118] and the power-law degree distributions [6, 29]. Time-evolving

graphs have also attracted attention recently, where interesting properties have been discovered, such as shrinking diameters, and densification power law [63].

There have been a number of models proposed to capture these properties. For a survey, one can consult the work by Chakrabarti and Faloutsos [15]. For example, unlike the random graph model, the preferential attachment model proposed by Barabasi et al. [6] captures power-law degree distributions. The forest fire model [63] also captures the power-law degree distribution together with densification and shrinking diameters over time. A more recently proposed, microscopic evolution model [62] is based on properties observed in large, temporal network data, providing insight into the node and edge arrival processes. Another recent model, the butterfly model [73], concentrates on capturing the evolution of connected components in a graph. In this work, I extend the microscopic evolution model by including processes of forming and joining groups of interest.

2.2.2 Evolution of affiliation networks

To the best of my knowledge, there is no model that captures the evolution of affiliation networks in online communities. However, there are studies that describe the relationship between friendship links and group formation properties [4, 75]. They show that the probability of a user joining a group increases with the number of friends already in the group [4], and that higher degree nodes tend to belong to a higher number of groups [75].

Group detection is a related problem (for a survey, see [34]). Its goal is to find new communities based on node features and structural attributes. Unlike group detection work, my work concentrates on unraveling the process governing the formation of existing communities.

2.3 Observations

Though affiliation groups constitute a major part of many social networks, very little work in the literature focuses on analyzing group memberships and evolution. In this section, I analyze different affiliation networks and try to characterize some properties of affiliation groups that are consistent across various datasets. For my analysis, I used three large real-world datasets from LiveJournal, Flickr and YouTube.

LiveJournal is a popular blogging website whose users form a social network through friendship links and form affiliation links to various "communities," which are groups of users having similar interests. The LiveJournal dataset considered contains over 5.2 million users, 72 million links, and over 7.4 million affiliation groups. Flickr is a photo-sharing site based on a social network with friends and family links. Groups in Flickr are also based on users with common interest. The Flickr dataset used in the experiments contains over 1.8 million users, 22 million links, and around $100,000$ groups. The third dataset is from YouTube, which is a popular video-sharing site that includes a social network based on user-defined contacts, and an affiliation network based on the category

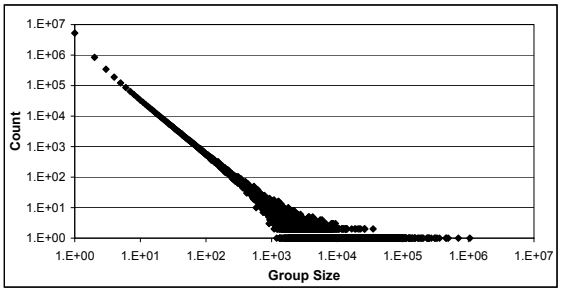
of videos that users post. The YouTube dataset contains over 1.1 million users, 4.9 million links and around $30,000$ groups. The full dataset descriptions can be found in the work of Mislove et al. [75]. Now, I describe the observations that I discovered by analyzing the datasets, and relate them to previously observed properties.

2.3.1 Group size distribution

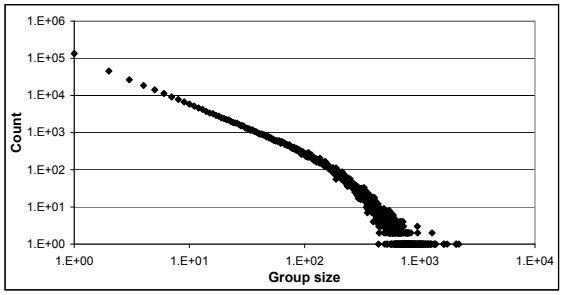
I begin by characterizing the relationship between the size of the affiliation group and its frequency of occurrence. The main observation is that, analogous to the degree distribution, the group size distribution follows a power law with a large number of small groups and a smaller number of large ones. This has also been observed by Mislove et al. [75]. The results are illustrated in Figure 2.1.

2.3.2 Node degree vs. Average number of group affiliations

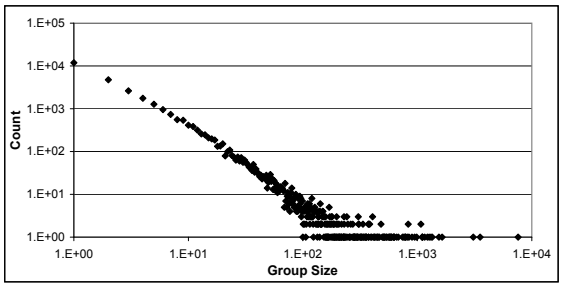
Looking at the relationship between the degree of a node and the number of its group affiliations, I observe that the nodes of lower degree tend to be members of fewer groups than the nodes with higher degree. However, the relation starts declining after a certain point, yielding a lower number of group memberships for very high degree nodes. The relationship is illustrated in Figure 2.2, where the x-axis represents the node degree and the y-axis represents the average number of group affiliations for nodes with that degree. The nodes in the declining part represent a very small portion of the overall number of nodes ($< 1\%$ of the size



(a) LiveJounral



(b) Flickr



(c) Youtube

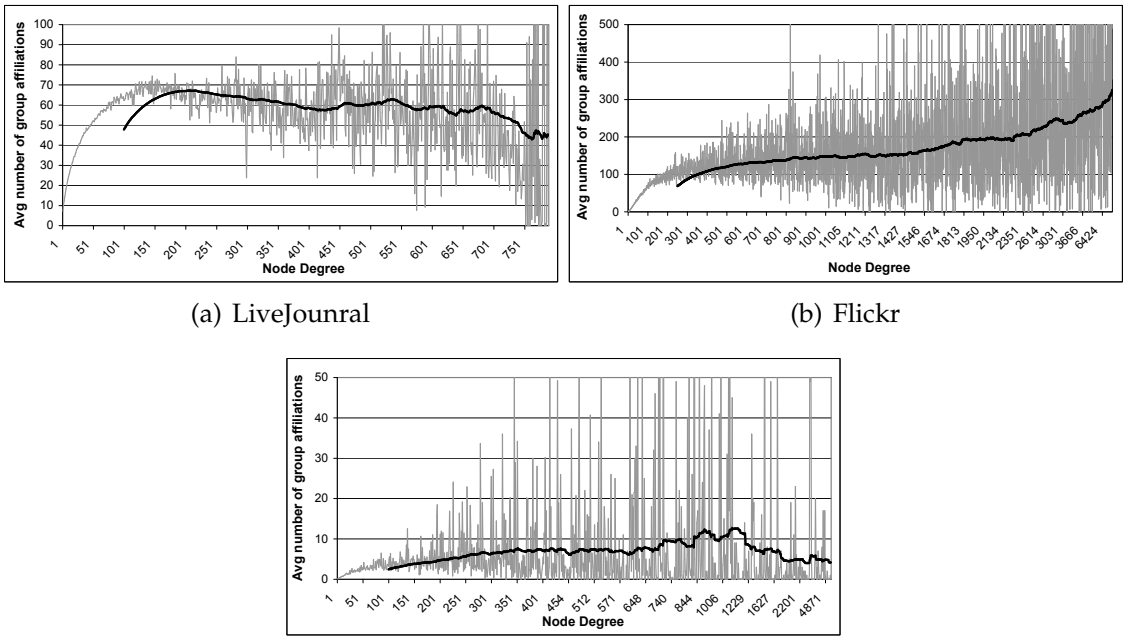
Figure 2.1: Distribution of group sizes on a log-log scale.

of the network in all cases), which is why I fitted only the increasing part of the data points to a function. I evaluated over 55 different distributions including logistic, Dagum and Laplace, using EasyFit [1], a software for distribution fitting. A power-law relation was the best fit according to the Kolmogorov-Smirnov ranking coefficient.

2.3.3 Distribution of the number of group affiliations

The previous observation considered the average number of group affiliations for nodes with different degrees. Here, I look at the actual distribution of the number of group affiliations with respect to the node degree. It turns out that the number of group affiliations for nodes of a certain degree k follows an expo-

[1] At `http://www.mathwave.com`



(a) LiveJounral (b) Flickr

(c) Youtube

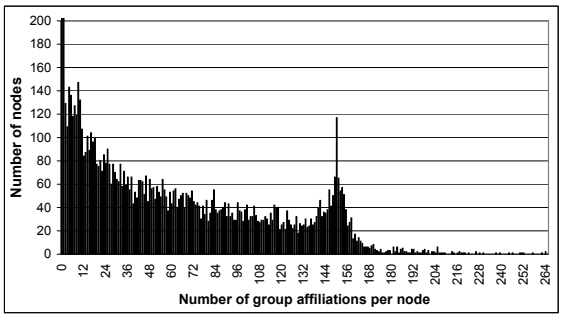
Figure 2.2: Node degree vs. average number of group affiliations

nential distribution. Figure 2.3 reports on $k = 50$ for LiveJournal and Flickr, and on $k = 25$ for YouTube but this was true for other degrees as well.

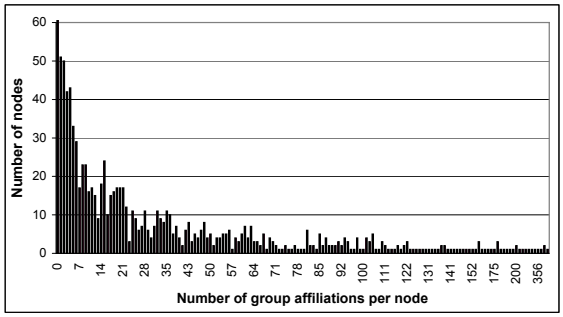
2.3.4 Properties of group members

According to Backstrom et al. [4], nodes are more likely to join groups in which they have more friends. However, it turns out that, in the datasets, there is a large portion of group members without friends in the group ($singletons$), meaning that they did not join the group because of a friend. This is surprising because it shows that users join groups for various reasons with friendship being only one of them.

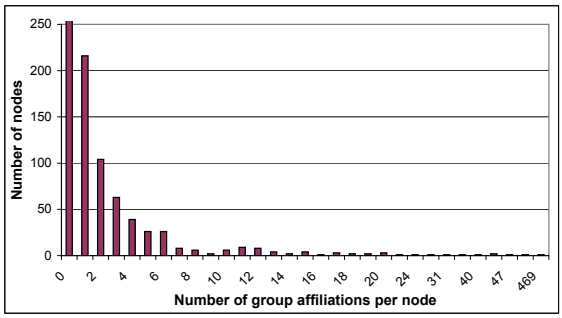
I measure the maximum node degree within groups of various sizes in the



(a) LiveJounral - Degree = 50



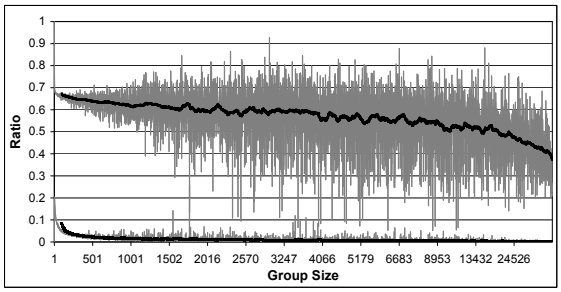
(b) Flickr - Degree = 50



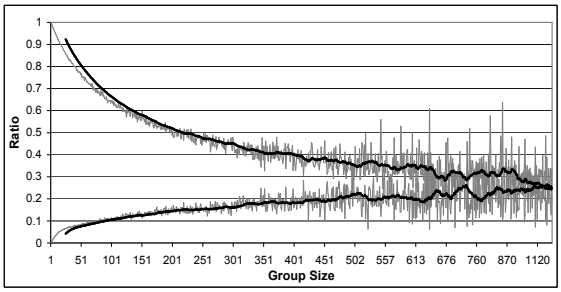
(c) Youtube - Degree = 25

Figure 2.3: Distribution of the number of group affiliations for nodes with different degrees.

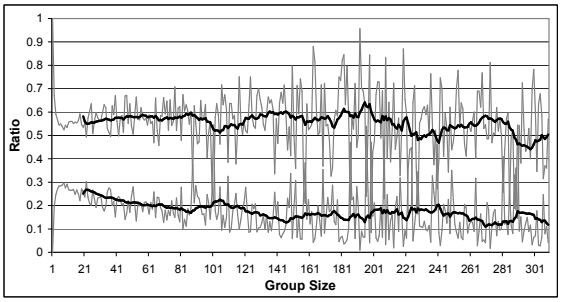
datasets. For all groups of a given size, I measure the average maximum degree per group and the average number of singletons (nodes with no friends within this group) as a percent of the group size. The results show a large number of singletons overall, especially in small groups, indicating that a large percentage of the members of a specific group do not have any friends within this group. This conclusion was confirmed by analyzing the average maximum degree per group. It turned out that the friends of the maximum-degree node within a group do not constitute a large percentage of the group size, even in small groups. The numbers are illustrated in Figure 2.4, where the $upper$ series shows the average ratio of the number of singletons to the group size, and the $lower$ series represents



(a) LiveJounral



(b) Flickr



(c) Youtube

Figure 2.4: Ratio of the number of singletons to the group size (upper series) and ratio of the maximum degree to the group size (lower series).

the average ratio of the maximum degree to the group size. This result shows that the larger the group a user belongs to, the more likely it is for him/her to have a friend in the group. For example, in Flickr, 76% of the members of groups of size 50 are singletons, while for groups of size 500, this number drops to 29%.

2.4 Co-evolution properties and model

A model which describes the evolution of a social network together with the evolution of an affiliation network needs to capture a number of simple events, as well as statistical properties of both networks. Here, I present the events of the co-evolution model and desired properties, some of which have been presented in other work. Then, I present the co-evolution model, which extends the node

arrival and link formation processes of the microscopic evolution model [62] to dynamic social and affiliation networks.

2.4.1 Events

The possible events that the model allows are:

- a node joins the network and links to someone
- a new group is formed with one member
- a node joins an existing group
- a new link is formed between existing users

2.4.2 Desired properties

A co-evolution model needs to capture properties of both social and affiliation networks. Here, I show three types of properties: properties of the social network alone, properties of the affiliation network alone, and properties of both.

Properties of the social network.

- *power law degree distribution* - the node degrees are distributed according to a power law with a heavy tail. This property has been observed in many other studies.
- *network densification* - the density of the network increases with time [63].
- *shrinking diameter* - the effective diameter of the network decreases as more nodes join the network [63].

Properties of the affiliation network.

- *power law group size distribution* - the group sizes are distributed according to a power law with a heavy tail.

Properties involving both the social and affiliation networks.

- *large number of singletons* - many nodes do not have any friends inside the groups they are affiliated with.
- *power-law relation between the node degree and the average number of group affiliations* - see Section 2.3.2.
- *exponential distribution of the number of group affiliations for a particular node degree* - see Section 2.3.3.

2.4.3 Co-evolution model

I now propose a co-evolution model which captures the desired properties. The model is undirected, and it has two different sets of parameters: one is concerned with the evolution of the social network, and the other determines the factors of development of the affiliation network. I also present a naïve model which assumes that the evolution of the affiliation network is independent of the evolution of the social network. Both models utilize the microscopic evolution model [62] for generating the social network because that model is based on observing the temporal properties of large social networks. I present its main components first.

Microscopic evolution model. The main ideas behind the microscopic evolution model are that nodes join the social network following a node arrival function, and each node has a lifetime a, during which it wakes up multiple times and forms links to other nodes. These are the set of parameters needed for the microscopic evolution model: $N(.)$ is the node arrival function, λ is the parameter of the exponential distribution of the lifetime, and α, β are the parameters of the power law with exponential cut-off distribution for the node sleep time gap. Further details of the model can be found in the paper by Leskovec et al. [62]. I utilize these parts:

Node arrival. New nodes $V_{t,new}$ arrive at time t according to a pre-defined arrival process $N(.)$.

Lifetime sampling. At arrival time t, v samples lifetime a from $\lambda.e^{-\lambda.a}$: v becomes inactive after time $t_{end}(v) = t + a$.

First social linking. v picks a friend w with probability proportional to degree(w) and forms edge $e_s(v, w, t)$.

Sleep time sampling. v decides on a discrete sleep time δ by sampling from $\frac{1}{Z}.(\delta^{-\alpha}).e^{-\beta.\text{degree(v)}.\delta}$. If the node is scheduled to wake up before the end of its lifetime ($t + \delta \leq t_{end}(v)$), then it is added to the set of nodes $V_{t+\delta}$ that will wake up at time $t + \delta$.

Social linking. At wake up time t, v creates an edge $e_s(v, w, t)$ by closing a triad two random steps away (i.e., befriends a friend w of a friend).

Naïve model. Before I present the proposed co-evolution model, I first present a naïve model which assumes that the evolutions of the social network

and the affiliation network are two independent processes. As a first step, it creates the social network using the model of Leskovec et al. [62]. Then, it generates and populates groups in such a way that their sizes follow a power-law distribution with an exponent k. Algorithm 2.1 presents the naïve model in detail. I use this model as a baseline.

Algorithm 2.1 Naïve model

Set of nodes $V = \emptyset$
for each time period $t \in T$ **do**
  Set of active nodes at time t, $V_t = \emptyset$
end for
for each time period $t \in T$ **do**
  Node arrival. $V = V \cup V_{t,new}$
  for each new node $v \in V_{t,new}$ **do**
    Lifetime sampling
    First social linking
  end for
  for each node $v \in V_t$ **do**
    Social linking
  end for
  for each node $v \in V_t \cup V_{t,new}$ **do**
    Sleep time sampling
  end for
end for
Set of groups $H = \emptyset$.
for i=1:number of groups **do**
  Group creation. New group h_i is created and its size s is sampled from s^{-k}. $H = H \cup \{h_i\}$.
  for j=1:s **do**
    Group joining. Pick a random node $v \in V$ and form an affiliation link to it $e_a(v, h_i, null)$.
  end for
end for

Co-evolution model. In this model, the affiliation network evolution co-occurs and depends on the social network evolution. When a node wakes up, besides linking to another node, it also decides on a number of groups to join.

Algorithm 2.2 Co-evolution model

Set of nodes $V = \emptyset$
Set of groups $H = \emptyset$
for each time period $t \in T$ **do**
  Set of active nodes at time t, $V_t = \emptyset$
end for
for each time period $t \in T$ **do**
  Node arrival. $V = V \cup V_{t,new}$
  for each new node $v \in V_{t,new}$ **do**
    Lifetime sampling
    First social linking
  end for
  for each node $v \in V_t$ **do**
    Social linking
    Affiliate linking. v determines n_h, the number of groups to join, sampled from an exponential distribution $\lambda' e^{-\lambda' n_h}$ with a mean $\mu' = \frac{1}{\lambda'} = \rho.\text{degree(v)}^{\gamma}$.
    for $i = 1 : n_h$ **do**
      if $rand() < \tau$ **then**
        Group creation. v creates group h, and forms edge $e_a(v, h, t)$. $H = H \cup \{h_i\}$.
      else
        Group joining. v forms edge $e_a(v, h, t)$. Group h is picked through a friend with probability p_v; otherwise, or if no friends' groups are available, it joins a random group with prob. proportional to the size of h.
      end if
    end for
  end for
  for each node $v \in V_t \cup V_{t,new}$ **do**
    Sleep time sampling
  end for
end for

With probability τ, it creates a new group, else, it joins an existing group. There are two mechanisms by which it picks a group to join. In the first one, it joins the group of one of its friends. In the second one, it picks a group at random. Algorithm 2.2 presents the co-evolution model in detail.

Here, I present the parameters of the affiliation network evolution part in

more detail. The first parameter, ρ, represents a tuning parameter that controls the density of the affiliation links in the network. The second parameter, γ, is the exponent of the power law that relates node degree with number of group affiliations. The last parameter to the model, τ, represents the probability by which an actor creates a new group at each time point. All the parameter values range over the interval $[0, 1]$ except ρ which ranges between 0 and the average number of group affiliations per node. I provide some guidelines for picking the right parameter values in the experiments section.

As noted in Section 2.4.2, the relationship between node degree and average number of affiliations is a power-law relation. Even though one can vary the exponent γ of this function, for simplicity, I fixed its value to 0.5, utilizing a square root function to compute this average.

It is also worth noting that other, more sophisticated techniques can be utilized in both social and affiliation aspects of the model that might be able to capture stronger correlation between the evolution of both kinds of networks. One possible modification for the social link creation is considering random steps but with group bias, such that the probability of choosing a node u to close the triad is proportional to the number of groups the two nodes share. Another possible modification is to specify the number of groups a node will join in advance using the estimated power-law function. A disadvantage of such approach is that the approximated degree is hard to compute because it depends on the expected value of a function which changes with the degree. A thorough investigation of the different alternatives is left as future work.

In the group joining step of the algorithm, a node decides to join a group and it has two choices for picking that group. One is through a friend, and the second is by picking a random group with probability proportional to the size of that group. It follows the first choice with some probability p_v, else it resorts to the second one. The intuition behind this is that some nodes in each group are singletons while others have friends in it. The second choice is also based on the observation that the size of the groups follows a power-law distribution; on the principle of "rich get richer," groups with larger size should have a larger probability of getting picked.

There are many options for computing the probability p_v such as making it a constant or dependent on the node degree. One can test which is most appropriate in the presence of temporal data for affiliation networks. Since such data is hard to obtain, I try different possibilities in the model. It turns out that using a constant for p_v yields a relationship between the group size and the singleton ratio that decreases at first but then stabilizes around $1 - p_v$ at higher group sizes. In contrast, what was observed initially is a relationship which decreases with increasing group sizes (see Figure 2.4). When a p_v which is correlated with the degree is used, the resulting relationship becomes closer to the desired one. In particular, I compute:

$$p_v = \begin{cases} \eta * \text{degree(v)} & \text{if } \eta * \text{degree(v)} < 1 \\ 1 & \end{cases} \tag{2.1}$$

though other functions of the degree may be more appropriate. The parameter η represents the friends' influence on the actor's decision to join a group; i.e. the likelihood of an actor joining one of the groups of his/her friends increases by increasing the value of η. The main intuition behind using a degree-correlated probability is the fact that as a node has more friends, the probability that one of its friends belongs to one of the larger size groups increases. Thus, utilizing the friendship bias parameter η actually increases its chances of joining this larger size group of its friend, thus leading to the decreasing relationship noted in the observations.

2.5 Experiments

I present three sets of experiments. The first set shows that the model is able to produce a dataset very similar to one of the real-world datasets, and the second set observes the properties of data generated by the co-evolution model. I also present results for the naïve model which adds groups on top of a social network, showing that this model is not able to produce the real-world affiliation network properties.

2.5.1 Synthetic data

In the first set of experiments, I vary the parameters of the model in order to generate a few synthetic datasets. Then, I check whether each dataset has the properties described in Section 2.4.2.

I have fixed the parameters of the social evolution part throughout this set of experiments, and varied the parameters of the affiliation part of the network. I assume an exponential node arrival function to achieve higher growth rate in the generated network, which is in accordance with what Leskovec et al. [62] showed in some social networks, such as Flickr. However, other arrival functions can also be utilized within the model. The other parameters of the social evolution aspect were fixed as reported by Leskovec et al. for Flickr data: $\lambda = 0.0092$, $\alpha = 0.84$, and $\beta = 0.002$. I also fix the value of the second parameter to the affiliation model, γ, to 0.5.

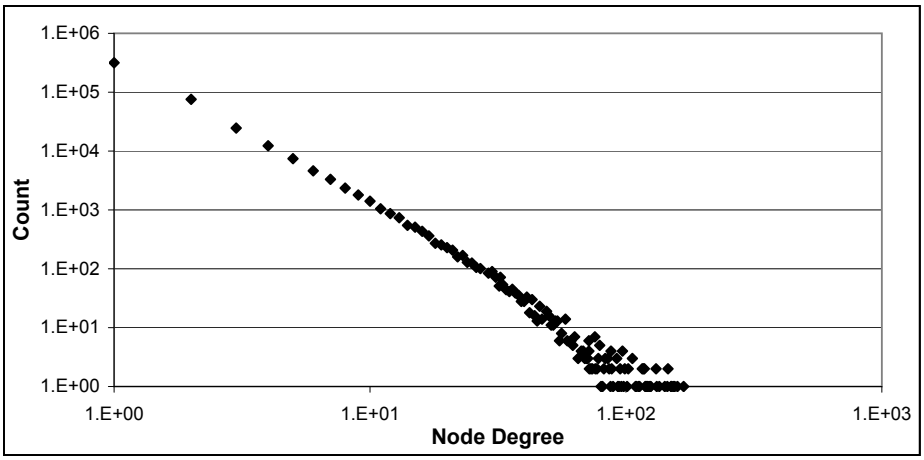


Figure 2.5: Degree distribution in a synthetic network

I first illustrate the results for the social network generated using the specified parameters. The model was run for 400 time steps, resulting in a network with 140,158 actors and 245,043 social links. The degree distribution in the resulting network follows a power-law as Figure 2.5 shows. The network densification property also holds, as illustrated in Figure 2.6 which represents the number of

nodes and number of edges at each time point on a log-log scale.

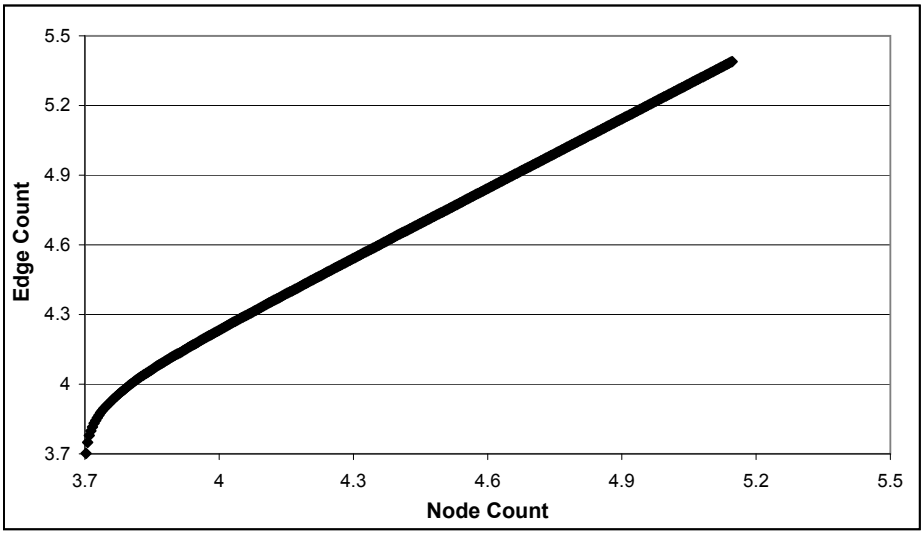
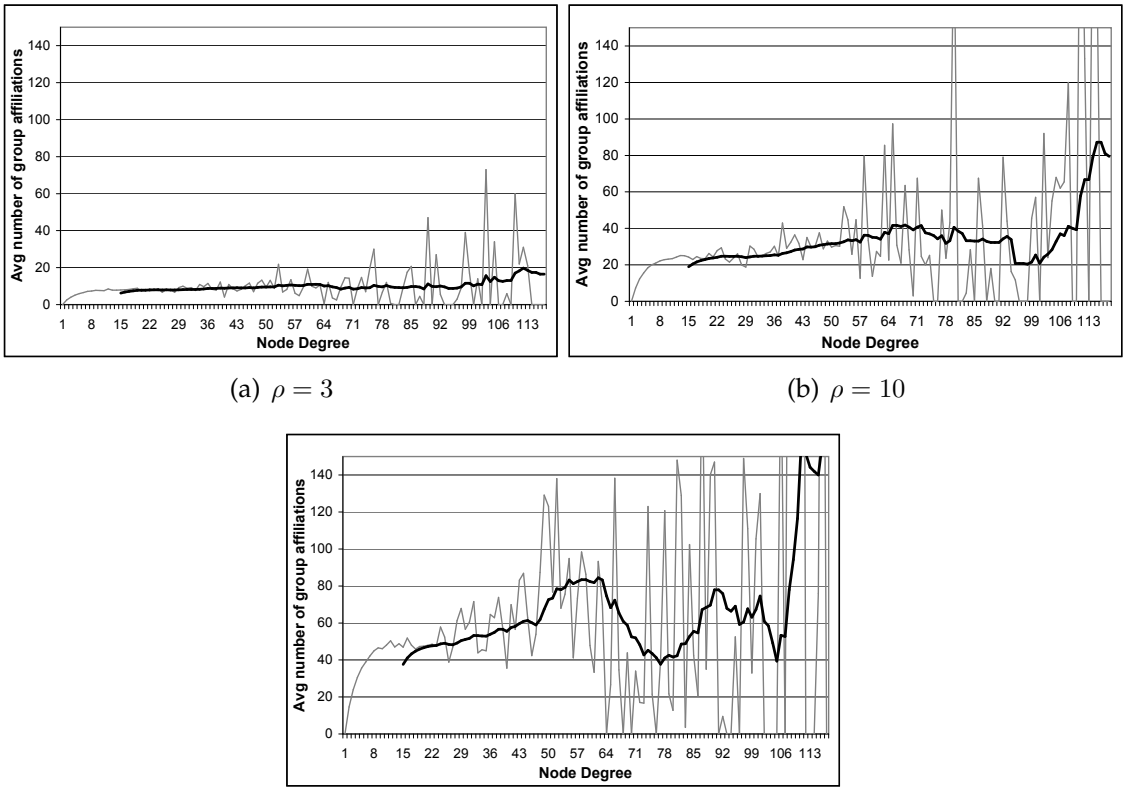


Figure 2.6: Densification in a synthetic network

In order to test the affiliation aspect of the evolution model, I investigated the effect of each parameter in the model on the properties of the resulting affiliation network. I start with the first parameter ρ, which represents a tuning factor of the affiliation links' density. The main properties that are affected by varying the value of ρ are the total number of affiliations and the distribution between the node degree and average number of group affiliations. Figure 2.7 illustrates that the general power distribution persists among different values of ρ, but the main effect is the scale of the distribution; as increasing the value of ρ, more affiliation links are created, and correspondingly increasing the average number of group affiliations per node. Theoretically, the values for this parameter can vary from 0, where no affiliation links are created in the network, to the maximum number of groups, where fully connected affiliation network emerges. Practical values for ρ varies between 0 and 25. The total number of affiliation links for each value of ρ

is reported in Table 2.1.



(a) $\rho = 3$

(b) $\rho = 10$

(c) $\rho = 20$

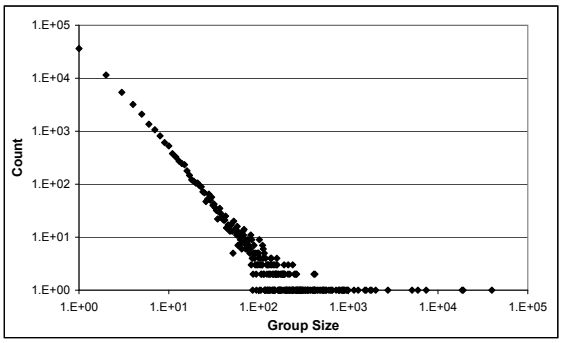
Figure 2.7: Degree vs. average number of group affiliations on varying the parameter (ρ).

ρ	**Affiliation Count**
3	285,536
10	2,411,710
20	4,771,072

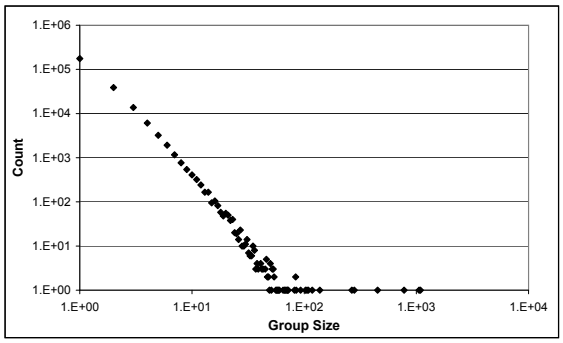
Table 2.1: Number of affiliation links on varying the parameter (ρ)

The next parameter, τ, represents the probability with which a node creates a new group. This parameter directly affects the number of groups in the resulting network, as well as the group size distribution. As illustrated in Figure 2.8, we note that although the power law distribution of the group sizes holds for

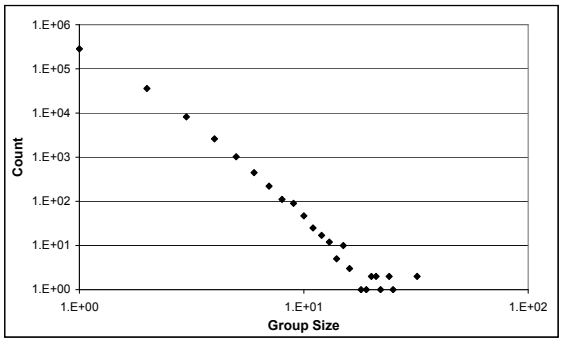
various values of τ (which is one of the desired properties), the maximum group size decreases significantly with increasing the value of τ. This decline in the maximum group size is caused by the fact that for higher values of τ, nodes tend to create new groups more often than joining existing ones, which leads to the existence of a large number of groups with relatively small sizes. This conclusion is also clear in the results illustrated in Table 2.2, where the resulting number of groups in the network and the maximum group size vary significantly with changing the parameter value.



(a) $\tau = 0.1$



(b) $\tau = 0.5$



(c) $\tau = 0.9$

Figure 2.8: Group size distribution on varying the parameter (τ)

Finally, I investigate the parameter η, upon which p_v depends. η represents the extent to which friends influence the decision of a node to join groups. The

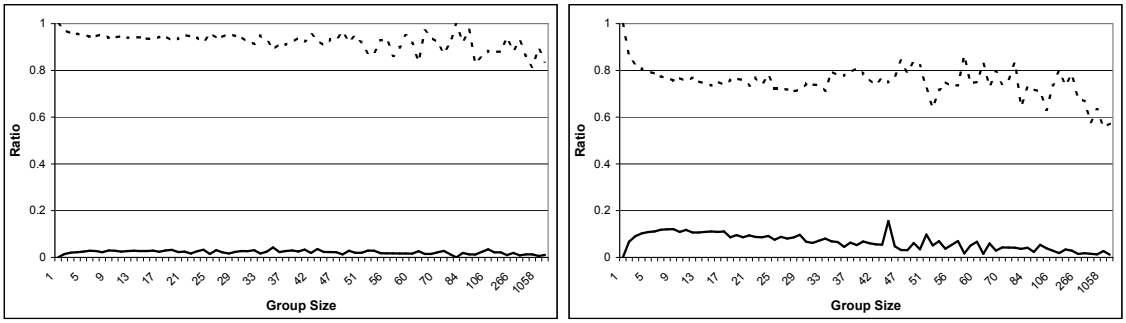
outcome of increasing the value of this parameter is a decreasing number of *singletons* and an increasing relative degree of the nodes within different groups. Figure 2.9 shows that the general distribution captures the desired properties and the observations in real data. The value of η is highly dependent on the social network structure properties, such as the average node degree in the social network and the desired influence of friends on a node's decision. For instance, if the value of $\eta = 0.1$ is used in a setting where the expected value for the average node degree is around 10, then we expect to see high percentage of nodes in the network being affected by their friends.

2.5.2 Real data

In this set of experiments, I look for the model parameters that will produce a network similar to one of the real-world datasets used in the observations of Section 2.3. I searched for parameters that will produce an affiliation network resembling Flickr since the social network evolution parameters for Flickr have already been reported by Leskovec et al. [62]. In order to get an initial seed of the search space for the evolution parameters of the affiliation network, I analyze the affiliation network properties of Flickr as observed in Section 2.3. A summary of the affiliation network statistics of Flickr is given in Table 2.3.

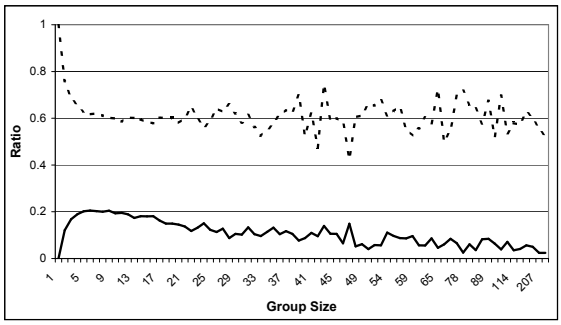
τ	**Groups Count**	**Max Group Size**
0.1	66,887	39,753
0.5	245,143	560
0.9	332,437	32

Table 2.2: Number of groups on varying the parameter (τ)



(a) $\eta = 0.01$

(b) $\eta = 0.05$



(c) $\eta = 0.09$

Figure 2.9: Group size vs. member attributes on varying the parameter (η) (dashed line: % ratio of singletons to group size, solid line: % ratio of maximum degree to group size).

The Flickr dataset is characterized by a relatively small number of groups in comparison to the number of users, where the actual ratio between the group count and the user count is 0.056. As a result, I expect to have a small value of τ close to this ratio. On the other hand, the average number of group affiliations per user in the real dataset is 4.62, and I assign this value to ρ. Finally, as observed in Figure 2.4, the average percentage of singletons in each group is lower than the average for the other datasets, indicating more friendship bias, thus increasing the value of η.

There are other factors to consider when specifying the affiliation network evolution parameters, such as the rate of node arrival and the probabilistic na-

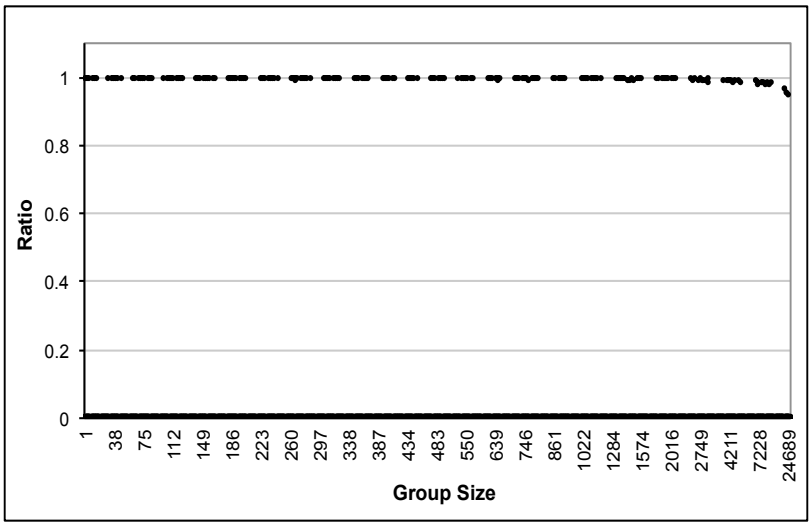
	Real Network (Flickr)	Synthetic Network
Users count	1,846,198	1,707,475
Groups count	103,648	88,749
Affiliations count	8,529,435	7,813,910
Avg groups per user	4.62	4.58

Table 2.3: Statistics of a real network (Flickr) vs. a synthetic one ($\rho = 2.5$, $\gamma = 0.5$, $\eta = 0.1$, $\tau = 0.03$)

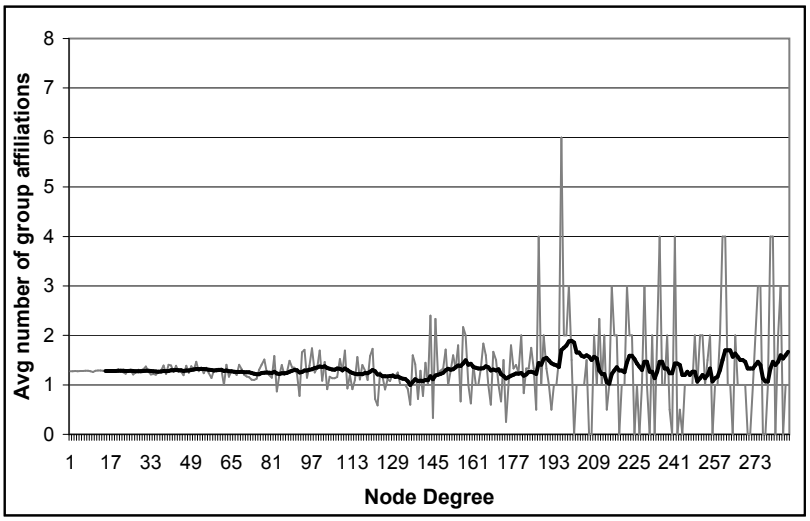
ture of the node's lifetime and sleep time gaps. For example, in Flickr's case, the exponential node arrival rate means that more nodes are created at later times. In this case, the distribution parameters should be a bit lower than desired because many nodes will join towards the end of the evolution process but they will not have sufficient time to create many links and affiliations. By utilizing all these pieces of information to guide the parameter search, I was able to generate a network that has similar attributes to Flickr's, illustrated in table Figure 2.3. I argue that using a similar procedure for parameter selection can result in generating synthetic networks that have many of the properties of a real one.

2.5.3 Comparison with the naïve model

In this set of experiments, I was interested to learn whether the desired network properties can be produced by utilizing the naïve evolution model. The model can clearly capture the social network properties since the process of creating it is the same as in the co-evolution model. In terms of the affiliation network properties, I used the naïve model to produce a social network similar to Flickr, as described in the previous experiment. Then I created the desired number of groups and picked the size of each one from a power-law distribution with the



(a) Average number of singletons (dashed line) and average maximum degree (solid line)



(b) Degree vs. avg number of affiliation groups

Figure 2.10: The affiliation properties produced by the naïve model

parameters observed in Flickr. Each group was populated by picking random users from the social network. As a result, the naïve model is able to capture the group size distribution. However, Figure 2.10(a) shows that it is not able to capture the average number of singletons and the average maximum degree as a percent of the group size. By picking random members, almost all members in

each group end up being singletons (except for groups with very large sizes), and the average maximum degree is close to 0. Figure 2.10(b) shows that the model is also not able to capture the relation between degree and average number of group affiliations for nodes with lower degrees. The naïve model generates a relation between them which is closer to linear than a power law.

2.6 Conclusion

I presented a generative model for creating social and affiliation networks. The model captures important statistical properties of these networks, and provides new insights into the evolution of networks with both social and affiliation links. It shows that groups can be formed for various reasons and friendship links are not the only propagators of influence. I believe that this observation not only affects the design of network evolution models but it may have broader implications on other mechanism designs, such as group recommendation, information diffusion and viral marketing strategies.

Chapter 3

Multi-Relational Affinity Propagation

After analyzing the evolution process of multi-modal networks, I now move to characterizing the structural patterns that exist in these networks. I focus on cluster analysis as one of the common initial data exploration approaches. In this chapter, I propose a multi-relational clustering approach for identifying the patterns and grouping structures that occur in complex networks. My proposed approach extends the affinity propagation clustering framework for encoding different types of relational constraints to capture the dependencies between different node types, and across various relationship types. This formulation allows for combining information from both the features of different entities as well as their inter-relational structure to explore different clusterings of the network. The output clustering can then be used for further analysis of different network-related tasks.

3.1 Introduction

Cluster analysis is one of the foundational components of unsupervised learning and exploratory data analysis. It has long attracted the attention of researchers from multiple disciplines. The classical clustering approaches focus on feature similarity for finding latent groupings in the data. However, with the

emergence of data that is naturally described in more complex ways, particularly in the form of heterogeneous graphs or networks, these classical approaches are no longer sufficient.

In order to address the challenges in this structured data, a number of graph clustering and community detection approaches have been proposed [28, 37, 82]. The methods find groups of nodes that are tightly connected to each other, and loosely connected to nodes in different clusters. Similar ideas for graph partitioning are also used for feature-based clustering, by constructing a network among the data points based on their attribute similarity rather than intrinsic structure (e.g., [101]). In addition, block modeling approaches have also been proposed for grouping nodes that link to similar collections of other nodes [54, 120, 1].

However, many real-world problems include rich, structured relationships that include multiple dependencies among different entity types. Clustering heterogeneous (multiple node types), multi-relational (multiple edge types) networks poses a number of challenges that the proposed algorithms should be able to address. First, the proposed algorithm should be able to account for the structural dependencies between nodes of the same type, as well as the information contained in their descriptive features. Second, the algorithm should be able able to model the relationships across different entity types and incorporate them in the clustering process.

To motivate the multi-relational clustering problem, consider the task of customer segmentation for marketing purposes. Using only the customer demographics as features, the best achievable segmentation is a one based on age,

gender, etc. Although this demographic profiling might help determine suitable products or appropriate marketing design strategy, it does not provide insight into the social structure, which may be important for predicting product adoption or collections of customers to target. By incorporating the social network structure, a relational clustering algorithm can produce segments that are based on both the demographics and the connectivity of the users in their corresponding social communities. This is likely to help in determining the projected adoption and gives some insight into the social influence. In addition, by considering information about additional relationship types, such as affiliations and memberships between people and other organizations or entities in the network, we may be able to develop a more nuanced picture. For example, a multi-relational clustering algorithm can account for customers' affiliations to different industrial segments and their organizational roles. This may lead to a better quality segmentation, that is more helpful for influence estimation or targeted advertising.

To address these challenges, a number of multi-relational clustering approaches have been introduced in the literature [109, 78, 10, 67, 5]. While each of these methods has their advantages and disadvantages, the majority of them either make certain distributional assumptions about the underlying data or require certain characteristics in the feature set. In addition, many of these approaches rely on expensive inference methods, such as Gibbs sampling or other MCMC approaches. In this work, I present a novel, general clustering approach that utilizes both feature similarity and relational dependencies across multiple relationship and entity types to produce a clustering that balances between the

homogeneity of the data points and their relational structure. The main advantages of my approach are that it is simple, elegant, scalable, and does not make any distributional assumptions about the underlying data. My work extends the affinity propagation (AP) clustering algorithm [31] to complex networks domains, by leveraging the relational dependencies in the underlying network data through the introduction of structural constraints in the AP model. These constraints bias the optimization problem to favor clusterings which conserve both the homogeneity of the data points as well as their connectedness.

The proposed multi-relational affinity propagation framework uses signals from the links among both similar and different node types to augment the information gained through features similarity, while allowing the user to control the extent of this effect. This facilitates the exploration of clusterings that account for both feature and structural similarities. I show the advantages of my framework over previous approaches through a set of experiments on a sample network from the social news website, *Digg.com*.

3.2 Related Work

Early work in relational clustering was first done in the ILP community, in which objects of each type are clustered based on the objects of other types linked with them (e.g., [57]). In addition to the logical-based approaches, there has been also a body of literature on probabilistic approaches. Taskar et al.[109] proposed a relational clustering algorithm based on probabilistic relational models that used

both feature and link information in uncovering the latent group structure. However, one of the drawbacks of the algorithm is the acyclicity constraint which is hard to satisfy in general network data. Neville et al.[78] proposed a hybrid approach for graph partitioning that relies on both link and feature information.

Although a number of clustering methods have been proposed to combine both feature and structural information, most of them have focused on clustering a single node type, with the link structure serving as an additional factor in determining similarities or enforcing constraints on the clustering problem. Recently, the problem of clustering general heterogeneous data in multi-relational settings has recently started to attract the attention of more researchers, especially with the increased complexity of the existing data and the associated analysis tasks. An early example is the framework proposed by Zeng et al.[121] for clustering heterogeneous web objects, through an iterative reinforcement clustering process. A different approach was proposed by Xu et al.[120] by introducing an infinite dimensional latent variable for each entity in the network, as part of a Dirichlet process mixture model. As the inference in this approach mainly relies on the Chinese Restaurant Process, the method's performance might not scale favorably for large networks. More recently, a probabilistic framework approach was proposed by Long et al.[68] for clustering different types of entities, taking into consideration the multiple types of relationships among them. However, one of the limitations of this work is that it assumes the underlying statistical distribution of the data belongs to the exponential family.

Bekkerman et al.[10] proposed a framework that simultaneously clusters

variables of different types based on their pairwise interactions. More recently, Banerjee et al.[5] proposed a multi-way clustering approach for relational data, that relies on simultaneously clustering multiple entity types represented as a multi-modal tensor. One of the limitations of this approach is that it is only applicable to Bregman loss functions. However, formulating the problem as tensor clustering is an active area of research that has been recently attracting the attention of multiple researchers (e.g., [50, 107]). Other related work includes the framework proposed by Plangprasopchok et al.[84], which extends affinity propagation to account for structural constraints in inferring consistent taxonomies from shallow personal hierarchies on the web. In addition to the previous approaches, a recent logic-based approach was proposed by Kok et al.[58] for discovering new concepts in ILP settings, using a second-order Markov logic framework. The proposed model forms multiple relational clusterings, while iteratively refining them based on the underlying data.

3.3 Method

I represent the underlying multi-modal network structure as a complex graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{T})$, where $\mathcal{T}$ denotes the different node types in the network, $\mathcal{V} = \{V^t : t \in \mathcal{T}\}$ represents the set of node sets of each type in $\mathcal{G}$, and $\mathcal{E} = \{E^{t_1 \to t_2} : t_1, t_2 \in \mathcal{T}\}$ represents the set of edge sets in $\mathcal{G}$. I distinguish between two types of edges in $\mathcal{E}$: *homogeneous* and *heterogeneous* edges. Homogeneous edges are edges among the same node type (e.g., friendship links among people

in a social network), and take the form $E^{t \to t}$. To simplify the notation, homogeneous edges between nodes of type t are represented as E^t. Heterogeneous edges link entities of different types (e.g., affiliation links between people and organizations), are denoted as $E^{t_1 \to t_2}$ where $t_1 \neq t_2$.

Feature-based clustering approaches focus on clustering data points using similarity measures defined over their features. One simple framework that has been recently proposed is affinity propagation (AP) [31]. AP is an exemplar-based clustering that relies on a message passing algorithm. Given the similarities among the underlying data points, it finds a clustering by identifying a set of exemplars, and finds an optimal assignment of the rest of the data points to these exemplars.

One of the appealing aspects of affinity propagation is its formulation as a max-sum algorithm on a binary factor graph model [39]. This formulation facilitates the incorporation of new constraints via functional nodes in the underlying factor graph. The similarity values among all pairs of data points, along with the 1-of-N constraint which enforces that each node is assigned to a single exemplar, and the exemplar consistency constraint that asserts that exemplar nodes should only choose themselves as exemplars, constitute the core of the affinity propagation algorithm that I use as a base for my approach.

In addition to the feature similarities among the nodes, the edges in in $\mathcal{G}$ also encode a set of relational dependencies across the corresponding entities the nodes represent. These dependencies should be made use of in the proposed multi-relational clustering algorithm. The proposed algorithm takes these de-

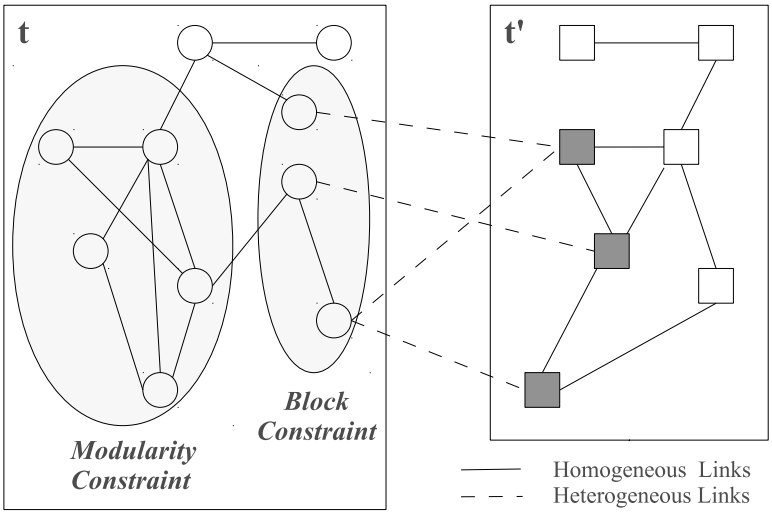


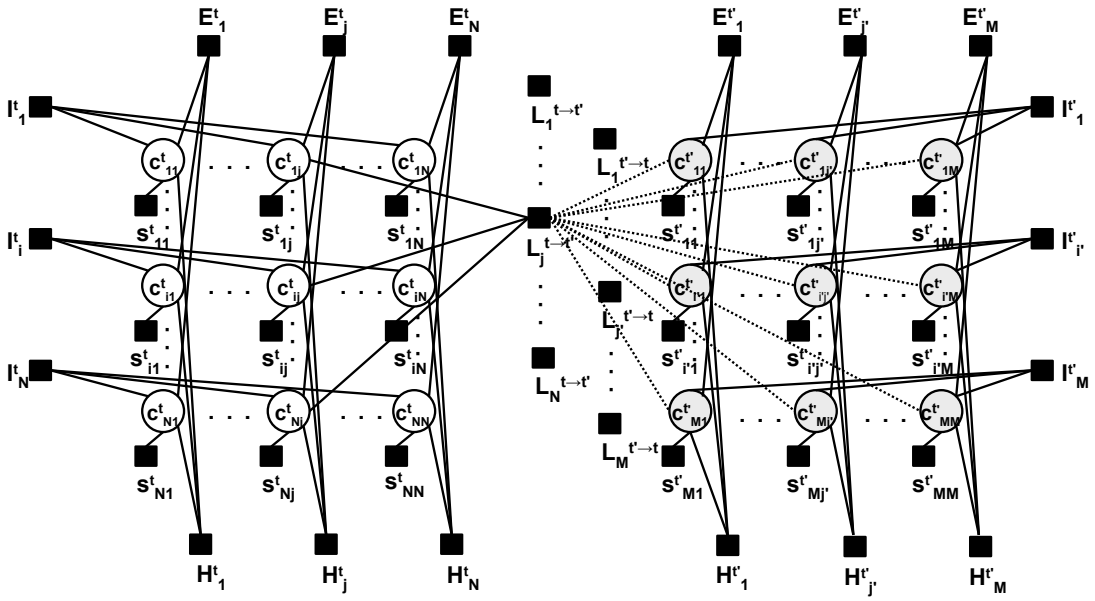
Figure 3.1: Sample bimodal network

pendencies into consideration along with the feature-based similarity during the clustering process. Thus, I require the clustering algorithm to satisfy these additional conditions:

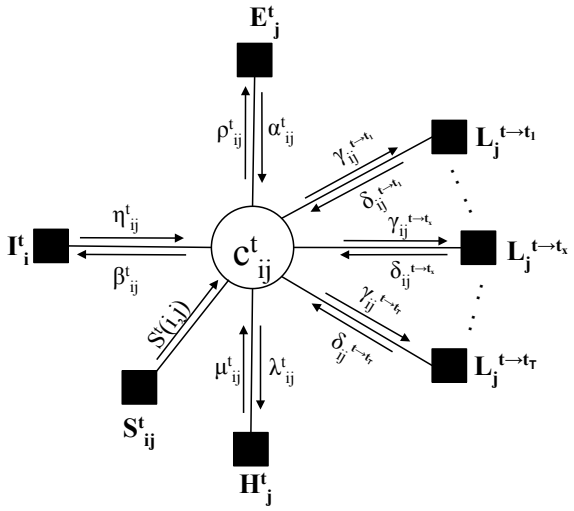
1. Nodes that are connected by homogeneous links should be in the same cluster (*modularity constraint*).

2. Nodes that are connected by heterogeneous links to nodes of a different type residing in the same cluster should be clustered together (*block constraint*).

The first condition, referred to as a *modularity* constraint, favors clusterings that ensure a high degree of connectivity among the members of the same cluster. This is a common assumption made in a variety of community detection algorithms. The second condition ensures that nodes of one type that are connected to the same cluster of nodes of a different type, are also clustered together. This is a common assumption made in a variety of block-modeling algorithms, and I re-



(a) Binary variable diagram with the structural dependencies factors. To simplify the diagram, only one instance of the factor L is included



(b) Factor graph representation

Figure 3.2: Multi-relational Affinity Propagation Model

fer to this constraint as a *block* constraint. My goal is to encode these conditions in a flexible clustering framework that allows users to vary the importance of each.

3.3.1 Model Formulation

Starting from the binary factor graph model introduced by Givoni and Frey [39], I augment it with the additional information needed in the multi-relational setting. Each possible assignment of node i of type t to an exemplar j is modeled as a binary variable c_{ij}^{t}, such that ($c_{ij}^{t} = 1$) iff node i is assigned to the cluster represented by exemplar j. To simplify the discussion, I consider the bi-modal network illustrated in Figure 3.1. The example network $\mathcal{G}$ contains only two different node types $\left(\mathcal{V} = \left\{V^{t}, V^{t'}\right\}; |V^{t}| = N, |V^{t'}| = M\right)$, one homogeneous link type among each node type, and one heterogeneous link type across them $\left(\mathcal{E} = \left\{E^{t}, E^{t'}, E^{t \to t'}\right\}\right)$. Note that the same analysis can be easily extended to settings where there are more than two node and edge types.

Given the above network, the possible assignments of nodes from both types to their corresponding exemplars can be described by the two sets: $\left\{c_{ij}^{t}\right\}; i, j \in \{1, 2, \ldots, N\}$, and $\left\{c_{i'j'}^{t'}\right\}; i', j' \in \{1, 2, \ldots, M\}$. Accordingly, I extend all the constraints defined in the original AP model to the multi-modal settings by replicating the factor nodes for each node type in the model as shown in Figure 3.2.

Next, I introduce two additional factors for each type to enforce my proposed modularity and block constraints. I formulate the structural constraints as soft constraints, parametrized by different costs for violating them. As opposed to the formulation of the 1-of-N and exemplar consistency constraints in the original AP model as hard constraints, the proposed soft structural constraints allow the user to control the level of impact of the relational dependencies on the clus-

tering output and at the same time increases the search space by permitting the model to violate some of the constraints to reach a better solution in the optimization process.

The modularity constraint is represented by the factor H^t, which is defined over each node type t as follows:

$$H_j^t(c_{1j}^t, ..., c_{Nj}^t) = \begin{cases} -\sum_{i \in V^t} \theta_i^t & \forall e^t(i,k) \in E^t : c_{ij}^t = 0, c_{kj}^t = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The factor H_j^t is defined over nodes of the same type, and it penalizes clusterings that assign different exemplars for nodes that are directly linked by an edge. For each potential exemplar j for node k, if there is an edge that links i to k, where j is not the current exemplar for node i, a cost θ_i^t is included in the objective function to reduce the likelihood of the corresponding clustering. This cost value can either be constant or variable depending on some structural properties of the terminal nodes (such as clustering coefficient, degree, etc.).

For the second type of constraint capturing the clustering across edges among different node types, I introduce the block constraint factor $L^{t \to t'}$, defined for each pair of node types t and t' as follows:

$$L_j^{t\to t'}(c_{1j}^t,...,c_{Nj}^t) = \begin{cases} -\sum\limits_{i\in V^t} \omega_i^{t\to t'} & \forall e(i,i'), e(k,k') \in E^{t\to t'}, c_{i'j'}^{t'} = c_{k'j'}^{t'} = 1 \\ & : c_{ij}^t = 0, c_{kj}^t = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

The factor $L^{t\to t'}$ is defined over heterogeneous edges, connecting nodes of type t to all the potential exemplars of the nodes of type t' that they are linked to. It penalizes clusterings that assign different exemplars for nodes of type t which are linked to nodes of type t' residing in the same cluster by introducing a cost $\omega^{t\to t'}$ for such configurations. This guides the clustering process to favor clusterings that account for the structural dependencies across node types.

The global objective function of my proposed multi-relational affinity propagation on a network $\mathcal{G}$ can then be expressed as follows:

$$\begin{aligned} S(c_{11}^1,\ldots,c_{N^\mathcal{T}N^\mathcal{T}}^\mathcal{T}) = &\sum_{\substack{t\in\mathcal{T}\\ i,j\in V^t}} S_{ij}^t(c_{ij}^t) + \sum_{\substack{t\in\mathcal{T}\\ i\in V^t}} I_i^t(c_{i1}^t,\ldots,c_{iN^t}^t) + \sum_{\substack{t\in\mathcal{T}\\ j\in V^t}} E_j^t(c_{1j}^t,\ldots,c_{N^tj}^t) \\ &+ \sum_{\substack{t\in\mathcal{T}\\ j\in V^t}} H_j^t(c_{1j}^t,\ldots,c_{N^tj}^t) + \sum_{\substack{t,t'\in\mathcal{T}\\ j\in V^t}} L_j^{t\to t'}(c_{1j}^t,\ldots,c_{N^tj}^t) \end{aligned} \tag{3.3}$$

3.3.2 Message Derivation

Following the derivation of the original AP model, I use the max-sum algorithm to optimize the objective function in Equation 3.3, by deriving the scalar message updates in the factor graph model shown at Figure 3.2(b). The max-sum message update rules from a variable node to a factor node can be simply defined

as follows [12]:

$$\beta_{ij}^t = \alpha_{ij}^t + \mu_{ij}^t + \sum_{\substack{t' \in \mathcal{T} \\ t' \neq t}} \delta_{ij}^{t \to t'} + S_{ij} \tag{3.4}$$

$$\rho_{ij}^t = \eta_{ij}^t + \mu_{ij}^t + \sum_{\substack{t' \in \mathcal{T} \\ t' \neq t}} \delta_{ij}^{t \to t'} + S_{ij}^t \tag{3.5}$$

$$\lambda_{ij}^t = \alpha_{ij}^t + \eta_{ij}^t + \sum_{\substack{t' \in \mathcal{T} \\ t' \neq t}} \delta_{ij}^{t \to t'} + S_{ij}^t \tag{3.6}$$

$$\gamma_{ij}^{t \to t'} = \alpha_{ij}^t + \mu_{ij}^t + \eta_{ij}^t + \sum_{\substack{t'' \in \mathcal{T} \\ t'' \neq t, t'}} \delta_{ij}^{t \to t''} + S_{ij}^t \tag{3.7}$$

where the value S_{ij} corresponds to the feature-similarity between node i and its potential exemplar j, and the message pairs $(\alpha_{ij}^t, \rho_{ij}^t)$ and $(\eta_{ij}^t, \beta_{ij}^t)$ are the ones defined for factors E_j^t and I_i^t in the original affinity propagation model for encoding the 1-of-N and the exemplar consistency constraints. The message pair $(\mu_{ij}^t, \lambda_{ij}^t)$ is the one associated with the introduced modularity constraint factor H_j^t, and the messages $(\gamma_{ij}^{t \to t'}, \delta_{ij}^{t \to t'})$ are the ones associated with the block constraint factor $L_{ij}^{t \to t'}$.

I now move to the derivation of the message updates from the introduced factors to the corresponding variable nodes. I start with the modularity constraint factor H_j^t defined over nodes of similar types. To simplify the notation, I remove the type qualifiers, as the message derivation for the factor H_j is independent of the node type. To derive the message μ associated with factor H_j, we have to consider the two possible settings for each variable node c_{ij}. First, when $c_{ij} = 1$,

we get:

$$\mu_{ij}(1) = \max_{c_{kj}, k \neq i} (H_j(c_{1j}, ..., c_{ij} = 1, ..., c_{iN}) + \sum_{l \neq i} \lambda_{lj}(c_{lj})) \tag{3.8}$$

For the cases where a node l is not connected to i, the value of the function H_j reduces to zero. However, for the set of neighboring nodes $D(i) = \{k : \exists e(i,k) \in E\}$ that are homogeneously linked with i, there are two different cases: first, if k is in the same cluster j as i (i.e., $c_{kj} = 1$), then the function H_j reduces to zero. However, in the second case where ($c_{kj} = 0$), the function H_j evaluates to the corresponding cost $-\theta_k$. By taking both cases into consideration, Equation 3.8 can then be re-written as follows

$$\mu_{ij}(1) = \sum_{k \in D(i)} \max(\lambda_{kj}(1), \lambda_{kj}(0) - \theta_k) + \sum_{l \notin D(i)} \max_{c_{lj}} \lambda_{lj}(c_{lj}) \tag{3.9}$$

Next, I consider the case when $c_{ij} = 0$:

$$\mu_{ij}(0) = \max_{c_{kj}, k \neq i} (H_j(c_{1j}, ..., c_{ij} = 0, ..., c_{iN}) + \sum_{l \neq i} \lambda_{lj}(c_{lj}))$$

Similarly, the assignment of the nodes that are not directly linked to i is unconstrained. However, for the nodes $k \in D(i)$, H_j evaluates to zero for the nodes that are not assigned to exemplar j ($c_{kj} = 0$), and to the cost value $-\theta_i$ of the node i for the ones associated with the value $c_{kj} = 1$. Therefore, the previous equation reduces to

$$\mu_{ij}(0) = \sum_{k \in D(i)} \max(\lambda_{kj}(0), \lambda_{kj}(1) - \theta_i) + \sum_{l \notin D(i)} \max_{c_{lj}} \lambda_{lj}(c_{lj}) \tag{3.10}$$

By taking the difference between Equations 3.9 and 3.10, we get

$$\begin{aligned}
\mu_{ij} &= \sum_{k \in D(i)} \max(\min(\lambda_{kj}, \theta_i), \min(-\theta_k, \theta_i - \theta_k - \lambda_k j)) \\
&= \sum_{k \in D(i)} \max(\min(\lambda_{kj}, \theta_i), \min(\lambda_{kj}, \theta_i) - \theta_k - \lambda_{kj}) \\
&= \sum_{k \in D(i)} (\min(\lambda_{kj}, \theta_i) + \max(0, -\lambda_{kj} - \theta_k)) \\
&= \sum_{k \in D(i)} (\min(\lambda_{kj}, \theta_i) - \min(0, \lambda_{kj} + \theta_k)) \qquad (3.11)
\end{aligned}$$

It is worth noting that in the final message value for μ_{ij}, if the costs θ_i and θ_k are replaced with infinite value, turning the modularity constraint H_J into a hard constraint, the message value reduces to ($\mu_{ij} = \sum_{k \in D(i)} \lambda_{kj}$), which corresponds to the summation of all the incoming messages to i from its similar-type neighbors. However, if the costs are replaced by zero instead, the value of μ_{ij} reduces to zero, effectively removing the effect of the corresponding constraint.

For deriving the update messages for the second factor type $L_j^{t \to t'}$, I first generalize the definition of the typed neighbor set $D^{t'}(i) = \left\{ i' \in V^{t'} : \exists e(i, i') \in E^{t \to t'} \right\}$ as the set of neighboring nodes of type t' that are directly linked to a given node i of type t. I start by considering the case where $c_{ij}^t = 1$:

$$\delta_{ij}^{t \to t'}(1) = \max_{c_{kj}, k \neq i} \left(L_j^{t \to t'}(c_{1j}, \ldots, c_{ij} = 1, \ldots, c_{Nj}) + \sum_{l \neq i} \gamma_{lj}^{t \to t'}(c_{lj}) + \sum_{i', j' \in V^{t'}} \gamma_{l'j'}^{t \to t'}(c'_{l'j'}) \right) \qquad (3.12)$$

To evaluate the function $L_j^{t\to t'}$ in the previous equation, we need to consider all the potential settings of the other variables c_{kj}^t of type t. For each node $c_{kj}^t = 0$, the function evaluates to the cost value $-\omega_k^{t\to t'}$ for all the nodes k' of type t' that are in the neighbor set $D^{t'}(k)$, and are associated with with a value of $c'_{k'j'} = 1$. The function evaluates to zero in all other cases. To simplify the notation, considering a network with two different entity types, I refer to variables of type t with c_{ij} and the ones of the opposite type t' with c'_{ij}. During the derivation, I also remove the type qualification from the γ, δ, and ω values as I am focusing on deriving the messages of type t, depending on one alternate type t' at a time. So, all the values mentioned in the equations afterwards are presumably qualified with $t \to t'$ type dependency. Thus, the previous equation can be written as follows:

$$\delta_{ij}(1) = \sum_{j' \in V^{t'}} \Bigg[\sum_{l' \notin \{\cup_x D'(x)\}} \max_{c_{l'j'}} \gamma_{l'j'}(c'_{l'j'}) + \sum_{i' \in D'(i)} \max \Bigg[\gamma_{i'j'}(0) + \sum_{k \neq i} \sum_{k' \in D'(k)} \max(\gamma_{kj}(c_{kj}) + \gamma_{k'j'}(c'_{k'j'})), \gamma_{i'j'}(1) + \sum_{k \neq i} \max \Big[\gamma_{kj}(1) + \sum_{k' \in D'(k)} \max_{c'_{k'j'}} \gamma_{k'j'}(c'_{k'j'}), \gamma_{kj}(0) + \sum_{k' \in D'(k)} \max(\gamma_{k'j'}(0), \gamma_{k'j'}(1) - \omega_k) \Big] \Bigg] \Bigg] \tag{3.13}$$

The previous equation consists of two main parts: First, all nodes of type t' that are not connected to any node of type t have unconstrained assignment to the any exemplar j'. The second part is a summation over the neighbor nodes of type t'

that are connected to the current node i; the function $L_j^{t\to t'}$ evaluates to zero for the nodes in $D'(i)$ that are not assigned to the current exemplar j', and thus all other similar and opposite type nodes are now unconstrained. However, for the nodes $i' \in D'(i)$ that are associated with the value ($c'_{i'j'} = 1$), we need to consider all the other nodes k that are of the same type t as node i; if ($c_{kj} = 1$) then the assignment of the nodes in $D'(k)$ is unconstrained, while if ($c_{kj} = 0$) the function $Lt \to t'_j$ reduces to the cost value ω_k for all the nodes in $D'(k)$ that are assigned to exemplar j'.

Similarly, for $c_{ij} = 0$, the same derivation applies, and the value for $\delta_{ij}(0)$ can be represented as follows:

$$\delta_{ij}(0) = \sum_{j' \in V^{t'}} \Bigg[\sum_{l' \notin \{\cup_x D'(x)\}} \max_{c_{l'j'}} \gamma_{l'j'}(c'_{l'j'}) + \sum_{i' \in D'(i)} \max \Bigg[\gamma_{i'j'}(0) + \sum_{k \neq i} \sum_{k' \in D'(k)} \max(\gamma_{kj}(c_{kj}) + \gamma_{k'j'}(c'_{k'j'})), \gamma_{i'j'}(1) + \sum_{k \neq i} \max \Big[\gamma_{kj}(0) + \sum_{k' \in D'(k)} \max_{c'_{k'j'}} \gamma_{k'j'}(c'_{k'j'}), \gamma_{kj}(1) + \sum_{k' \in D'(k)} \max(\gamma_{k'j'}(0), \gamma_{k'j'}(1) - \omega_i) \Big] \Bigg] \Bigg] \tag{3.14}$$

By taking the difference between Equations 3.13 and 3.14, we get

$$\delta_{ij} = \sum_{j' \in V^{t'}} \sum_{i' \in D'(i)} \Bigg(\min\Big(\gamma_{i'j'} + \sum_{k \neq i} \min(0, \gamma_{kj}), \sum_{k \neq i} \min(A_{i,k}, \gamma_{kj})\Big) - \min\Big(\gamma_{i'j'} + \sum_{k \neq i} \min(0, \gamma_{kj}), \sum_{k \neq i} \min(0, \gamma_{kj} - B_k)\Big)\Bigg) \tag{3.15}$$

where the variables $A_{i,k}$ and B_k are defined as follows

$$A_{i,k} = \sum_{k' \in D'(k)} \big(\min(\omega_i, \gamma_{k'j'}) - \min(0, \gamma_{k'j'}) \big)$$

$$B_k = \sum_{k' \in D'(k)} \big(\max(0, \gamma_{k'j'} - \omega_k) - \max(0, \gamma_{k'j'}) \big)$$

3.4 Experimental Evaluation

To evaluate the proposed multi-relational affinity propagation approach, I show its performance on a number of cluster quality measures using both synthetic and real-world data. I compare the proposed algorithm to a variety of baselines, including the MMRC relational clustering algorithm proposed in Long et al. [68], as well as the original affinity propagation model and the modularity maximization algorithm[82]. The experimental results show that while the propsoed approach doesn't achieve the best performance on any single measure, unlike other baselines, it provides good performance across both the feature-based and the structure-based measures. I also show that the clustering that the algorithm generates is often closer to the real grouping of the data, when the ground truth is available, than the clusterings generated by the other baselines.

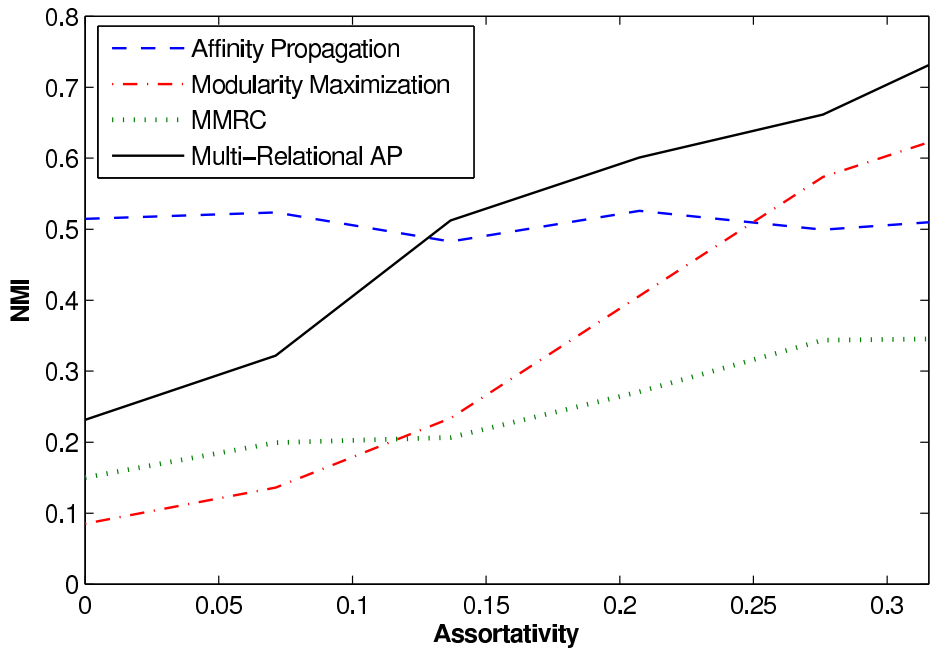


Figure 3.3: The performance of different clustering approaches for varying the levels of network assortativity.

3.4.1 Synthetic Data

One of the main factors that affects the performance of the proposed algorithm is the level of assortativity[80] in the network, which is a measure of similarity among linked nodes. Networks with high assortativity levels tend to have connected communities of nodes that have similar characteristics, as well as highly correlated features of the connected nodes across different types. The proposed multi-relational AP algorithm can leverage the feature-similarity among the nodes, and, using the structural constraints, can capture the community connectivity; using both, it can boost overall performance.

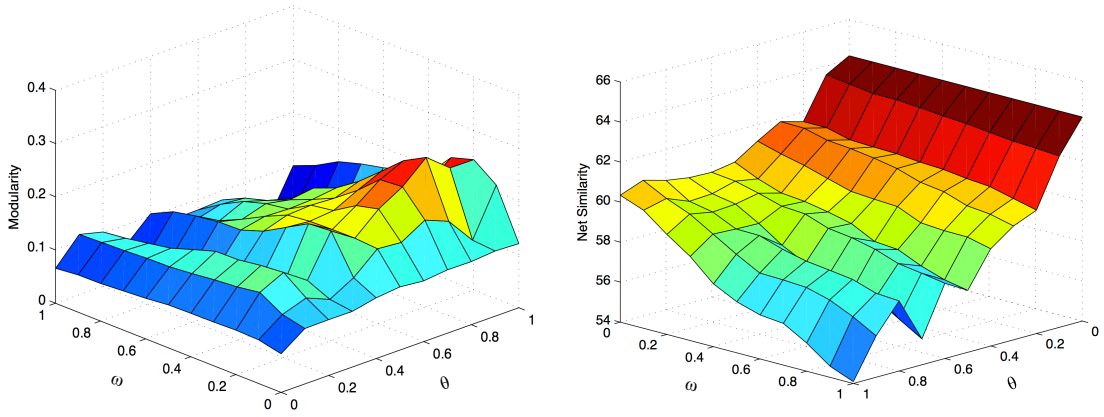
In order to test these effects, I generated a sample network of 250 users with 1209 homogeneous friendship links, and 100 social groups with 692 heterogeneous affiliation links using the co-evolution model proposed by Zheleva et

al. [122]. Next, I used the labeling heuristic proposed by Rattigan et al. [86] to generate a set of labels for both node types. By varying the percentage of seed nodes for each label, different node labelings with different assortativity coefficients can be generated. Finally, I generated a set of features for each node based on the assigned label using a Naive Bayes model.

Figure 3.3 shows the performance of the clustering algorithms using normalized mutual information (NMI) [105] between the output clustering and the assigned labels for the user nodes. As shown in the figure, the proposed multi-relational AP approach outperforms the other baselines for moderate and high levels of assortative mixing. Similar trends were also obtained on the inferred labels for the social groups.

3.4.2 Social Media Data

To evaluate the proposed multi-relational AP algorithm on social media data, I use a dataset from *Digg.com*, a popular social news website, where users can post stories on different topics, and then vote on them in a process referred to as "*digging*" to determine the story's ranking on the front page. Digg users form a social network by "*following*" other users, which in turn results in the target user posts showing on their homepages. I constructed a sample from the Digg network which includes 104,478 "*following*" links among 3750 users and their 438,379 "*digging*" links to 3305 stories. I use the "**title**" and "**description**" of the stories to construct a normalized tf-idf word vector for each post, which is then used



(a) Modularity (b) Net Similarity

Figure 3.4: The effect of varying the cost parameters (θ, ω) on the net similarity and the modularity of the output clustering

to measure cosine similarities between different stories. Similarly for the users, I used the "**about**" field that the user provides upon registration to compute the cosine similarities between users over the corresponding tf-idf vectors.

The AP algorithm relies on setting a preference value for each node in the network that reflects the likelihood of this point being an exemplar, which then affects the number of clusters in the output. In the experiments, I follow the approach that was proposed in the original AP model where there is no prior bias for certain nodes to be exemplars, and thus I set the preference value to the median of the corresponding input similarities. For evaluation, I use both the modularity of the resulting community structure and the total similarity, referred to as net similarity, of the exemplars to their assigned nodes in the output clustering to show the impact of different cost settings.

First, I show the performance of the multi-relational affinity propagation algorithm over a range of cost values for both the homogeneous and heterogeneous

structural constraints. Figure 3.4(a) shows the performance of the proposed algorithm for both the modularity and the net similarity measures of the user nodes clustering in the Digg dataset. Note that for lower values of ω, the modularity of the output clustering increases with increasing the value of θ, which corresponds to having higher costs on violating the homogeneous communities constraint. However, by increasing the value of the heterogeneous link cost ω for a given value of θ, the modularity of the output clustering increases initially, and then starts to decreases on higher values of ω. This can be attributed to the fact that increasing the cost of violating the block constraint initially provides additional evidence for the clustering of the alternate node type, but after a given point it starts fragmenting the clustering output, resulting in an increased number of clusters which decreases the overall modularity of the output clustering. On the other hand, Figure 3.4(b) shows the trade-off in terms of the decrease in the net similarity of the clustering output with increasing the cost values. However, it can be noted from the figure that the average decrease in the net similarity across different cost settings is much lower than the increase gained in terms of the modularity of the output clustering. Due to the lack of edges among story nodes, it is infeasible to show the effect of changing the value of θ on the clustering of stories, or to compute the modularity of the output clustering. However, by varying the cost of the block constraint, similar trends in the net similarity are obtained.

To compare the proposed approach with other baselines, I use both Davies-Bouldin [18] and Dunn [26] indices for internal clustering validation, as well as normalized mutual information (NMI) [105] for external validation with the

	Users			Stories		
	Modularity	**DB Index**	**Dunn Index**	**NMI**	**DB Index**	**Dunn Index**
MMRC	0.005	2.23	0.63	0.106	2.09	0.81
Modularity Maximization	**0.458**	2.4	0.57	N/A	N/A	N/A
Affinity Propagation	0.072	**1.504**	0.67	0.209	1.86	0.86
Multi-relational AP	0.13 (0.28)	1.52 (1.54)	**0.76 (0.78)**	**0.287 (0.34)**	**1.852 (1.859)**	**0.868 (0.87)**

Table 3.1: Comparison of MMRC, Modularity Maximization, AP, and multi-relational AP on different clustering evaluation measures. For multi-relational AP, the reported values are the average over all settings of the cost parameters, while the ones in parentheses are obtained from the optimal parameter settings, identified through an exhaustive search. The entries in bold face correspond to the best performance for the corresponding measure.

ground truth when available. The results in Table 3.1 show that the proposed multi-relational affinity propagation approach results have a superior performance compared to the MMRC relational clustering algorithm on all evaluation measures. By analyzing the clustering of the "users" type, I find that while the modularity maximization algorithm achieves the best modularity score for its output, it performs poorly on similarity-based measures. On the contrary, the original affinity propagation model shows better performance on the similarity based measures than the modularity score. However, the proposed multi-relational AP algorithm shows good performance on both structure-based and similarity-based evaluation measures, illustrating the balance that it is capable of achieving between both paradigms.

Finally, I move to the evaluation of the clustering for the "stories" node type. Due to the fact that there are no links among stories in Digg, I am unable to compute the modularity measure. However, the stories on Digg are manually assigned to a specific topic when posted. Therefore, the story topic can be used as an evaluation measure for clustering this node type, which enables the

computation of the normalized mutual information (NMI) quality measure of the output clustering. As can be seen in Table 3.1, the proposed algorithm achieves the best performance on all evaluation measures, including the NMI measure with the ground truth. This shows the value of the signal from block constraint and the favorable effect of coupling the clustering process across different node types. Another important advantage of the multi-relational affinity propagation algorithm over the MMRC algorithm is computational efficiency, as it is orders of magnitude faster than the MMRC algorithm.

3.5 Conclusion

In this work, I presented a novel, multi-relational clustering framework for identifying latent groupings in complex network domains. The proposed approach provides a simple and elegant way of extending the affinity propagation framework to multi-relational network domains by incorporating different soft constraints for capturing the structural dependencies among different types of nodes in the network. I showed how my proposed multi-relational affinity propagation algorithm could be used to output different clustering with varying degrees of dependence on both the feature similarity of the nodes as well as their relational structure. I conducted a set of experiments on both a synthetic and a real dataset from from an online social news website, and showed that the proposed approach outperformed previous approaches for multi-relational clustering.

Part II

The Temporal Dynamics of Multi-Modal Networks

Chapter 4

Understanding Actor Loyalty to Groups in Affiliation Networks

After characterizing the evolution and clustering aspects of the multi-modal networks, I now proceed to characterizing the temporal dynamics of the interactions occurring in these complex networks, from both user-level and network-level perspectives. I start by analyzing the individual user behavior with respect to other entities in the network, and how this behavior changes over time.

In this chapter, I introduce a method for analyzing the temporal dynamics of affiliation networks as an example of a 2-mode network. I define event-based affiliation groups as those describing temporally related subsets of actors and propose an approach for exploring changing memberships in these affiliation groups over time. To model the dynamic behavior in these networks, I introduce a measure that captures an actor's *loyalty* to an affiliation group as the degree of 'commitment' an actor shows to the group over time. I evaluate the proposed loyalty measure using three real world affiliation networks: a publication network, a senate bill co-sponsorship network and a dolphin network. The results show the utility of the measure for analyzing the dynamic behavior of actors and quantifying the stability of their interactions with different time-varying affiliation groups.

4.1 Introduction

Across many fields, researchers are interested in understanding an individual's commitment to a group [52], the social structure of groups [27], and the changing dynamics of group structure [100]. In marketing, researchers investigate customer behavior, comparing the purchasing behavior of different customer groups in an attempt to determine customer satisfaction and brand loyalty [83]. In sociology, researchers investigate commitment [52], community cohesion [91] and structural embeddedness of social groups [76]. In computer science, researchers have also modeled time-varying links to improve automatic discovery of relational communities or groups [13, 46]. While some statistical models have been developed for longitudinal analysis of social networks (see Snijders [102] for an overview), work remains to better understand the variation in actor commitment or loyalty to groups over time.

Social psychologists have investigated the role played by feelings of loyalty to groups. Druckman explains that "loyalty to a group strengthens one's identity and sense of belonging" [25]. In this chapter, I focus on an operational definition of loyalty to affiliation groups in an attempt to adequately measure this ubiquitous idea. Consistent with sociology literature [76], I believe that high loyalty may be an indicator of group cohesion.

More specifically, I investigate actor loyalty to groups in two-mode affiliation networks. A two-mode affiliation network contains two different types of nodes, one for actors and one for events. Edges between actor nodes and event

nodes are used to indicate relationships between actors and events in which the actors participate [115]. Affiliation networks capture a wide variety of domains, including communication data among people (email, cell phone calls, etc.); organizational data describing peoples' roles on teams or in companies; and epidemiological networks describing people and the specific disease strain with which they are infected. In time-varying affiliation networks, an actor's participation in a particular event is associated with a specific time, representing when this participation occurred. Annotating affiliation networks with temporal information allows us to capture changing actor behavior over time. In this chapter, I focus on this changing behavior as it relates to groups.

Consider an author/publication network describing authors, with the publications represented as events in which the co-authors are participants. If the publications are annotated with topic areas, then I can create groups of actors who publish in the same topic area at the same time. Furthermore, I can see how loyal an author is to specific topic areas over time by examining their changing publication topics. One common scenario is that an author starts publishing in a specific area, then after some time she begins publishing in additional areas, and eventually she might end up switching areas completely. Another common scenario is that an author starts publishing in an area, and, rather than adding additional areas, remains steadfast, and continues publishing regularly in the same area over a long period of time. In this chapter, I introduce a measure that captures this dynamic behavior of actors in time-varying affiliation networks by introducing the concept of **affiliation group loyalty** and define an actor's loyalty

to an affiliation group as the degree of 'commitment' an actor shows to the group over time.

4.2 Loyalty Background

Within literature across different disciplines, terms like *loyalty*, *commitment*, and *cohesion* have been given a number of different theoretical and operational definitions. my goal is not to provide an exhaustive literature review on these subjects, but rather to give a context for the remainder of the discussion on loyalty in affiliation networks.

Sociologists first formalized commitment as a way to link extraneous interests with a consistent line of activity [9]. While other definitions and theories concerning commitment exist in sociology [52] and social psychology [56], a definition proposed by [74] is as follows:

> *Commitment is a force that binds an individual to a course of action that is of relevance to a particular target.*

Loyalty extends the concept of commitment. For example, [83] defines customer loyalty in terms of brand commitment (the strength of the relationship between customers and a particular brand), and gives the following multi-faceted definition:

> *Loyalty is a deeply held commitment to re-buy or re-patronize a preferred product/service consistently in the future, thereby causing repetitive same-*

brand or same brand-set purchasing, despite situational influences and marketing efforts having the potential to cause switching behavior.

A well known operational definition of loyalty in business literature defines brand loyalty as the percent of purchases devoted to one's most often purchased brand [17]. Newman et al. [79] define loyal customers as those repurchasing a brand considering only that brand, while Tellis [110] views loyalty as repeat purchasing frequency or the relative volume of same brand purchasing. Jacoby et al. [49] state that frequent purchasing of a product is not synonymous with brand loyalty and that the notion of commitment is essential for distinguishing between brand loyalty and frequent purchasing of a product.

While business research tends to focus on the economic component, based on purchasing behavior, social psychologists have investigated ways that people relate to groups. One dimension of this is the role played by feelings of loyalty to groups. Druckman [25] explains:

The feelings of attachment that comprise loyalty are not whimsical, but are generally basic to the individual's definitions of themselves. Loyalty to a group strengthens one's identity and sense of belonging.

As will be discussed later, I define a group in terms of related events. I focus on an operational dimension of loyalty to affiliation groups in an attempt to adequately measure a ubiquitous idea. I demonstrate effects of my operational definition as it compares to frequency based brand loyalty in the business literature. Consistent with sociology literature [76], I believe that a group containing

actors with high loyalty may be an indicator of a highly cohesive group.

Because of the size and complexity of social networks, computer scientists, physicists and other scientists have also begun investigating different aspects of social networks. The community detection literature uses measures of cohesion and clustering to find subsets of actors that are densely connected to each other, but less densely connected to others. The majority of research conducted on community detection focuses on static networks and constrains the problem by letting an actor belong to only a single community [13, 30, 38, 48, 64, 81].

Recently, researchers have begun to analyze the dynamics of communities over time [3, 4, 11, 32, 102, 106, 108]. Much of this research focuses on two questions: what are the communities that exist in a particular data set, and how do they change or evolve over time. In contrast, the approach that I propose is a more micro-level analysis that focuses on the dynamics of specific actors or individuals in the network. While I focus on creating groups using affiliation event attributes (as will be described in the next section), my analysis of actors can be conducted using the output from any grouping, clustering, or community detection algorithm. Once the social groups are established, my goal is to understand the dynamics of actors and their social relationships in the context of these predefined social groups.

One approach which also proposes methods for identifying important actors in dynamic networks is the work of Habiba et al. [45]. They identify nodes in a single mode network that are likely to be good spread blockers, individuals that can block the spread of a dynamic process through the population. To

accomplish this, they introduce dynamic measures for density, diameter, degree, betweenness, closeness and clustering coefficient.

The graph summarization method proposed in Sharan et al [95] also uses a measure similar to ours to build a classifier for predicting evolving domains. While both of my measures attempt to quantify temporal aspects of the network, there are differences between their work and ours. First, the graph summarization method is used to create an aggregation of network snapshots over time by weighting the edges according to the point in time in which they occur. In contrast, my loyalty measure is used to quantify an actor's participation pattern in different affiliation groups. Second, though the authors mention that their proposed weighting kernels are able to model both temporal recurrence and temporal locality, which represents the aspects of consistency and recency I am addressing in the loyalty measure, it is unclear how the weighting kernels used account for temporal recurrence. In contrast, the recursive formulation of the proposed loyalty measure encodes this aspect directly.

4.3 Modeling time-varying event-based groups

There are many ways to define a group. Groups can be formed using community detection algorithms, clustering algorithms, etc. Because I am interested in understanding groups based on affiliation networks, I describe an approach that defines groups based on a participation relationship between actors and events. Formally, an affiliation network can be represented as a bi-partite graph

$\mathcal{G}(\mathcal{A}, \mathcal{E}, \mathcal{P})$ containing a set of actor nodes $\mathcal{A}$, a set of event nodes $\mathcal{E}$, and a set of participation edges $\mathcal{P}$ that connect actors in $\mathcal{A}$ to events in $\mathcal{E}$:

$$\mathcal{A} = \{a_1, a_2, a_3, \ldots, a_n\}$$

$$\mathcal{E} = \{e_1, e_2, e_3, \ldots, e_m\}, \text{and}$$

$$\mathcal{P} = \{(a_i, e_j) | a_i \in \mathcal{A}, e_j \in \mathcal{E}\}.$$

I denote participation of actor a_i in event e_j as $p_{i,j}$. For clarity, I use a running example of an author publication network in which the actors are authors, the events are publications, and the participation relation is paper authorship. Figure 4.1 shows an example network with three author nodes, $\mathcal{A} = \{a_1, a_2, a_3\}$, fifteen publication nodes, $\mathcal{E} = \{e_1, e_2, \ldots e_{15}\}$, and twenty paper authorship edges. As an example, participations involving actor a_1 are the following:
$\mathcal{P}_{a_1} = \{p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4}, p_{1,5}, p_{1,7}, p_{1,8}, p_{1,9}, p_{1,10}, p_{1,11}, p_{1,13}, p_{1,14}, p_{1,15}\}$.

Each actor node and event node can have attributes associated with them. For example, each author in Figure 4.1 may have a name and an age. For author a_1, we may have the following attribute values $a_1.name$ = 'Peter Pan' and $a_1.age$ = 50. Each publication event may have a title attribute, e.g. $e_1.title$ = 'Static networks as non-evolving dynamic networks' and a topic attribute, $e_1.topic$ = 'social networks'. In Figure 4.1, I use shapes to indicate topic. Since e_1 is a circle, all the events that are circular have the same value for topic, e.g. 'social networks'. For ease of exposition, I map each shape to the following topics: circle - $topic_1$, square

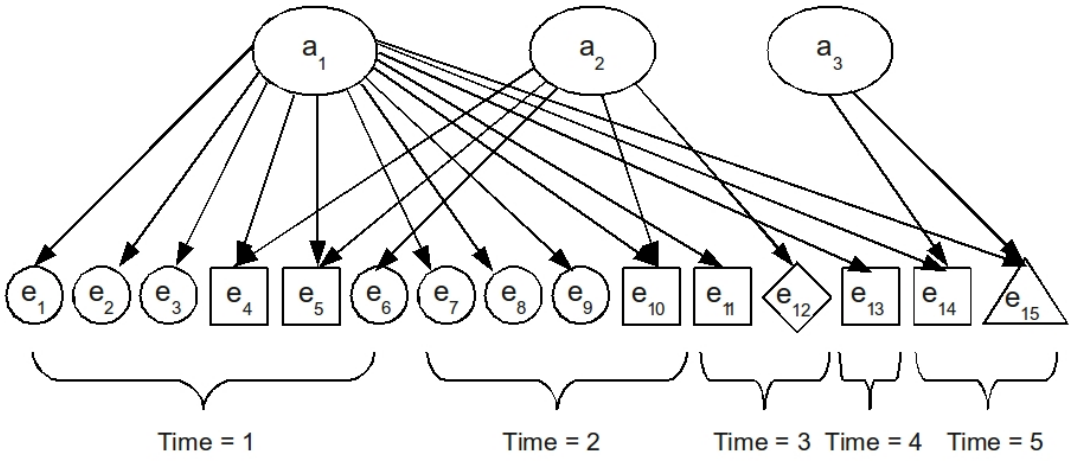


Figure 4.1: An affiliation network example with 3 actors, 15 events and 20 relationships across 5 time points.

- $topic_2$, triangle - $topic_3$, diamond - $topic_4$.

Because the affiliation networks are temporal, a discrete $time$ attribute is associated with each event e_j, and is denoted as $e_j.time$. For affiliation networks, this time is the same as the time of the participation relationship. In the example, the time attribute is the date of publication. I have labeled the time point associated with each event in Figure 4.1.

While the publication event serves as a grouping of a subset of actors, this event only occurs at one particular time. Because my goal is to understand the dynamics of affiliation networks over time, I am interested in analyzing actor participation in groupings of similar events across time. I propose grouping events based on values of an event attribute. In other words, a social group is defined based on a *shared event attribute value*. The choice of a specific method for grouping actors depends on the semantics of the underlying analysis task. Using

shared event attributes is particularly meaningful for affiliation networks since it incorporates the semantics of events into the data model. For other types of social networks, particularly uni-mode networks, it is reasonable to use other methods for defining social groups.

Each event feature or attribute F has an associated domain $Domain = \{g_1, \ldots, g_p\}$, where p is the number of distinct values of F. I denote a particular value g_l of an event e_j for event attribute F as $e_j.F = g_l$. Based on this, I define an affiliation group to be a subset of actors having the same group value g_l at time t for an event e_j: $G(g_l, t) = \{a_i | a_i \in \mathcal{A}, (a_i, e_j) \in \mathcal{P},$ where $e_j.F = g_l \; and \; e_j.time = t\}$. In the example, suppose the grouping attribute is *topic*. Referring back to Figure 4.1, $G(topic_1, 1) = \{a_1, a_2\}$ is the set of actors in topic group $topic_1$ at time 1.

I pause to mention a few advantages of my grouping formulation. First, actors can belong to multiple affiliation groups at a particular time. In other words, membership in different groups can be *overlapping*. In the example, author a_1 participates in five events at time 1. Also, actors are not required to be part of an event (or group) at every time t. This is also illustrated in the example. Author a_1 participates in an event at every time step. Authors a_2 and a_3 do not. In my experience, these assumptions better capture the dynamics of real world affiliation networks.

4.4 Loyalty of Individuals to Affiliation Groups

In order to better understand the loyalty of an actor to groups based on event affiliation, the participation of the actor in different groups over time should be quantified. Based on the example in Figure 4.1, Figure 4.2 shows actor a_1's membership in topic groups, $topic_1$, $topic_2$, and $topic_3$ across five time steps. The rectangles represent different topic groups. An edge from the author to a topic means that an author has published on the linked topic. The count on the edge represents the number of publications an actor a_i has published on this topic during a particular time period. For example, the network snapshot of the first time period shows author a_1 having three publications with $topic_1$ and two publications with $topic_2$. As time continues, author a_1 stops publishing on $topic_1$, continues publishing on $topic_2$ at each time step, and begins publishing on $topic_3$ in the last time step. Intuitively, by considering the loyalty of the author at time step 5, it is preferable to see a higher loyalty score for $topic_2$ since the author has published in this topic since time step 1. At time step 2, a topic shift occurs from $topic_1$ to $topic_2$. My goal is to create a measure that is sensitive to both continual group membership and changing group membership over time.

I begin by considering two simple loyalty measures: frequent participation and recent participation, illustrating how poorly they capture the nuanced nature of loyalty in dynamic networks. Loyalty based on frequent participation, which I refer to as *frequency-based loyalty* considers an actor loyal if she appears in a group frequently. Let $n(a_i, g_l)$ represent the number of participations of actor a_i

in group g_l and $n(a_i, *)$ represent the number of participations of actor a_i in all groups. Then the frequency-based loyalty of actor a_i is defined as the number of participations in a particular group g_l divided by the number of participations across all groups:

$$Loy_{FP}(a_i, g_l) = \frac{n(a_i, g_l)}{n(a_i, *)}$$

Using the running example, author a_1 publishes in $topic_1$ six times, $topic_2$ six times, and $topic_3$ one time. Therefore, $Loy_{FP}(a_1, topic_1) = Loy_{FP}(a_1, topic_2) = 6/13$ and $Loy_{FP}(a_1, topic_3) = 1/13$. $topic_1$ and $topic_2$ are considered equally important even though the author has not published in $topic_1$ since time step 2. Thus, considering frequency alone ignores the temporal component of the group affiliation and results in assigning higher loyalty values to groups that the actor was once active in, but may not be active in any longer. Frequency-based loyalty can be viewed as a static measure of commitment.

Focusing on the temporal aspect of the data, a *recency-based loyalty* measure considers an actor loyal if she has participated recently in a specific group. Let $n(a_i, g_l, t)$ represent the number of participations of actor a_i in group g_l at time step t. The recency-based loyalty of actor a_i is defined as the number of participations in a particular group g_l at the last time step t_f divided by the number of participations across all groups at time t_f:

$$Loy_{RP}(a_i, g_l) = \frac{n(a_i, g_l, t_f)}{n(a_i, *, t_f)}$$

In the example, when $t_f = 5$, $Loy_{RP}(a_1, topic_2) = Loy_{RP}(a_1, topic_3) = 1/2$ and $Loy_{RP}(a_1, topic_1) = 0$. Author a_1 is equally loyal to $topic_2$ and $topic_3$ even though $topic_3$ only appears in the current time step. If the last two time steps are considered (using a recent window as opposed to a recent time point), then a_1 is most loyal to $topic_2$. While this is accurate, the strong early participation of actor a_1 to $topic_1$ is not captured at all since $Loy_{RP}(a_1, topic_1) = 0$. Using recent participation leads to assigning an actor high loyalty values for groups that the actor participates in during current time steps, but it disregards earlier participation.

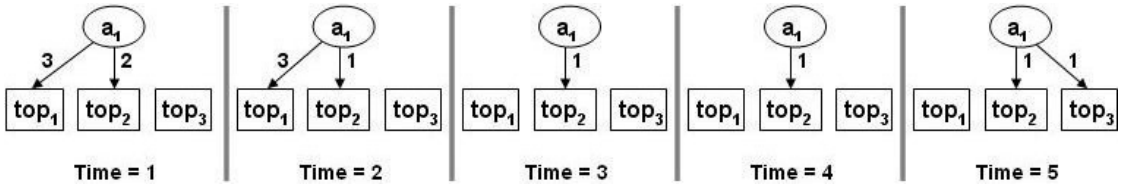


Figure 4.2: Single actor dynamic affiliation example

This simple example shows that a temporal measure of affiliation group loyalty should incorporate participation frequency for giving higher preference to actors with a large number of participations in the affiliation group, consistency for putting more bias toward actors with regular group participations across time over those with more sparse, isolated participation, and recency for favoring actors with current participations. In order to capture all of these aspects, I incorporate frequency and recency based loyalty into a more comprehensive measure of loyalty.

Let T_{tot} represent the total number of time points over which the dynamic affiliation is defined. The loyalty of an actor to a group that she has not partic-

ipated in yet is equal to zero. In order to keep track of consistent participation over time, we need to keep track of the actor's loyalty in the time step that precedes the current one. Thus, I define t_{prev} as the previous time point (relative to the current time point t) that actor a_i participated in group g_l. Let $n(a_i, g_l, \Delta t)$ be the number of participations of actor a_i in group g_l from the starting time point t_0 until the current time point t, and let $n(a_i, *, \Delta t)$ be the number of participations of actor a_i to all groups from t_0 until time t. I define the loyalty of an actor a_i to a group g_l on his first participation in the group as

$$Loy(a_i, g_l, t_0) = \frac{n(a_i, g_l, t_0)}{n(a_i, *, t_0)}$$

where $t = t_0$ and the loyalty on any consecutive participation is given by

$$Loy(a_i, g_l, t) = \frac{n(a_i, g_l, \Delta t)}{n(a_i*, \Delta t)} \times Loy(a_i, g_l, t_{prev})^{\alpha \frac{t - t_{prev}}{T_{tot}}}$$

where α represents a smoothing parameter that will be described shortly.

By examining the different components of the loyalty measure, we note that the first term, $\frac{n(a_i, g_l, \Delta t)}{n(a_i, *, \Delta t)}$, accounts for the frequency of participation of an actor into a specific group. The second term includes the component $Loy(a_i, g_l, t_{prev})$ which takes into consideration the most recent recorded loyalty for an actor in a specific group, g_l, and is used to favor recent participation in that group. Finally, to favor continuous actor participation, the second term includes an exponent term for the recent loyalty. This decreases the effect that the loyalty in the previous time

step has on the calculated loyalty in the current time step based on how long in the past this previous participation occurred. The more recent and continuous the participation, the larger the effect of this component on the overall loyalty of the actor to the group.

The smoothing parameter α is introduced to control the overall effect of time. The value of α can be varied from 0 to T_{tot}. A value of 0 means $Loy = Loy_{FP}$, focusing on the frequent participation component of the measure. A value of T_{tot} means that the recent participation component of the measure is dominant. For exploratory analysis, setting the value of α to 1 represents a good initial point to start off, where the loyalty accounts for both the frequency and the recency factors.

For consistency with the group membership notation, where the actor membership values in various groups sum up to 1, the values of loyalty of a specific actor to various groups that she participated in over the considered time period are normalized. As a result, the final loyalty value of actor a_i to group g_l at the final point in time t_f can be defined as follows

$$Loyalty(a_i, g_l, t_f) = \frac{Loy(a_i, g_l, t_f)}{\sum_j Loy(a_i, g_j, t_f)}$$

where the summation parameter j ranges over all the groups that actor a_i participated in during the entire time period.

Returning to the earlier example, we see that the proposed measure results in the desired effect. Setting the value of ($\alpha = 1$), the results for actor a_1 loyalty to

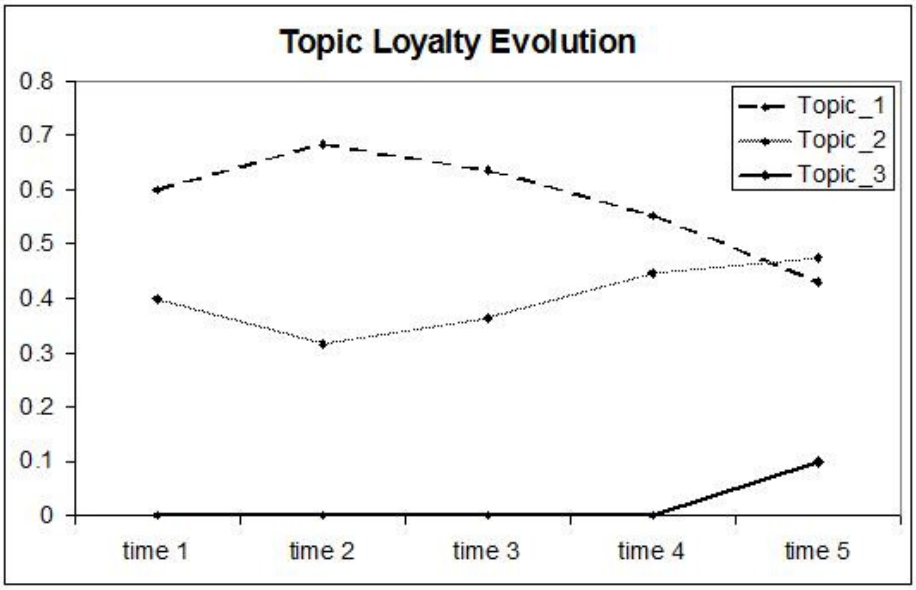


Figure 4.3: The evolution of loyalty over time for the affiliation network example

different topics are as follows:

$$Loyalty(a_1, topic_1, t_5) = 0.429$$

$$Loyalty(a_1, topic_2, t_5) = 0.474$$

$$Loyalty(a_1, topic_3, t_5) = 0.097$$

The evolution of the author's loyalty for each topic at each time step with $\alpha = 1$ is illustrated in Figure 4.3. $topic_1$ begins with the highest loyalty at time 1. Its loyalty increases at time 2 and then begins to decline. After time 4, author a_1's loyalty to topic $topic_2$ overtakes that of $topic_1$ because of the effect of recency.

To further illustrate the effect of the smoothing factor, Figure 4.4 shows the different values for the loyalty of the author to all the topics at the final time step

by varying the value of α. When $\alpha = 0$, the loyalty values are the same as if only the normalized frequency, Loy_{FP}, is considered. As the value of α increases, the effect of recency starts to dominate the frequency. At the maximum value of ($\alpha = 5$), the highest loyalty score is assigned for $topic_3$, which corresponds to the most recent group.

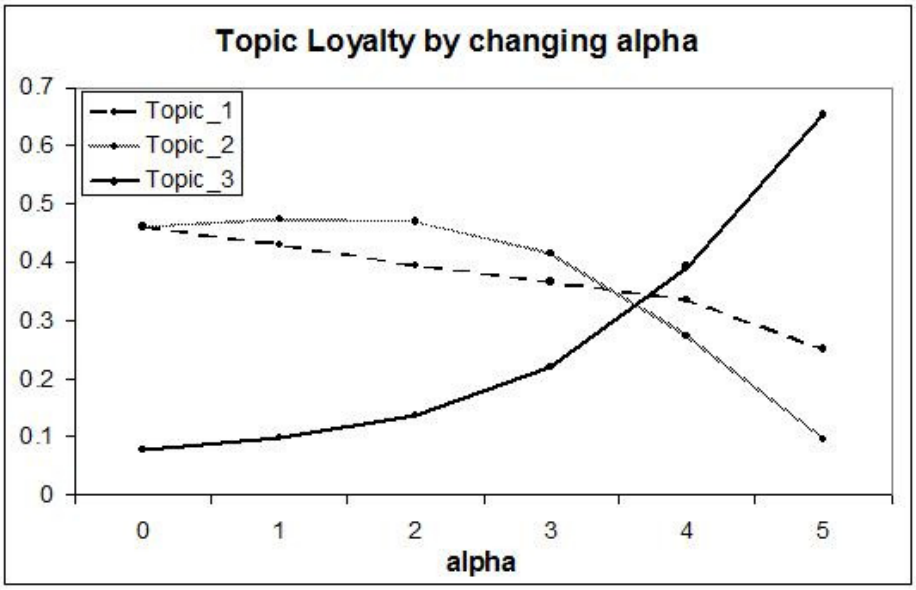


Figure 4.4: The effect of the smoothing factor in calculating group loyalty

The interpretation here is that as the value of α increases, the measure favors new topics occurring at the last time point. Therefore, $topic_3$, which has just occurred at the last time point, dominates all other topics because its loyalty is not decreased by previous occurrences. The loyalty values of $topic_1$ and $topic_2$ are overwhelmed by the large exponential factor resulting from the large α.

4.5 Loyalty Analysis on Individual Data Sets

I analyze my proposed loyalty measure on three data sets - a scientific publication network, a senate bill sponsorship network and a dolphin social network. In order to consider frequency, consistency, and recency, I set $\alpha = 1$.

4.5.1 Scientific Publication Network

The scientific publication network is based on publications in the ACM Computer-Human Interaction (ACMCHI) conference from 1982 until 2004. Similar to my running example, this data set describes an author/publication affiliation network. The data set was extracted from the ACM Digital Library and contains 4,073 publications and 6,358 authors. There are 12,727 participation relationships (edges) between authors and publications. In this data set, I filtered 5230 authors having only one publication over the entire period of time since no 'dynamic' group loyalty exists for these actors. Also, by removing them, I avoid biasing the average loyalty statistics calculated for the data set. The remaining 1,128 authors had 4,688 relationships with publication events.

There are a number of features that the publication events can be grouped on; for this analysis, like the simple running example, publications were grouped by their topic. There are 15 different topics, and the loyalty of authors to different topics was measured. The results of applying the proposed loyalty measure on the ACMCHI data set are shown in Figure 4.5. This box plot highlights the average loyalty, outliers and the amount of spread for actor loyalty to different topics.

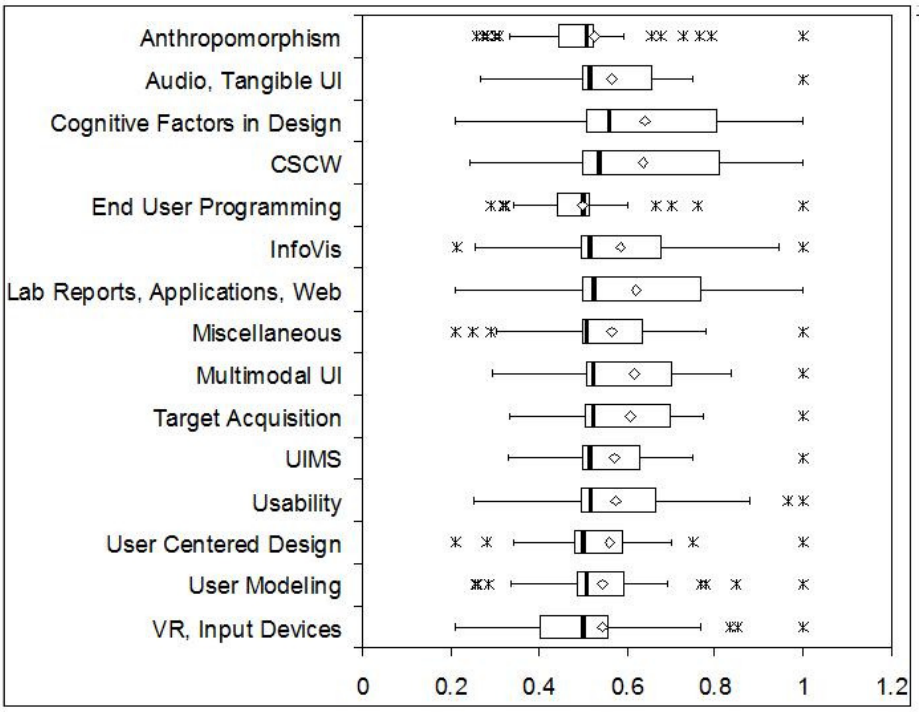


Figure 4.5: The average topic loyalty for the scientific publication network

The topic loyalty of authors range from 0.2 to 1, while the average topic loyalty ranges from 0.5 to 0.65 for all 15 topics. While there are a number of interesting observations to be made, I highlight two of them. First, the average topic loyalty is fairly uniform across the topics. This is an indication of the continued importance of these topics at the ACMCHI conference. Second, the average loyalty of authors to topic groups is very high across all the topics. This is an indication that, in general, authors in this data set consistently published in a particular research area as opposed to oscillating among multiple areas.

To better understand the distribution of author loyalty as it relates to an author's employer type, Figure 4.6 shows the average loyalty of authors categorized by employer type (corporate institutions, universities, research laboratories, and

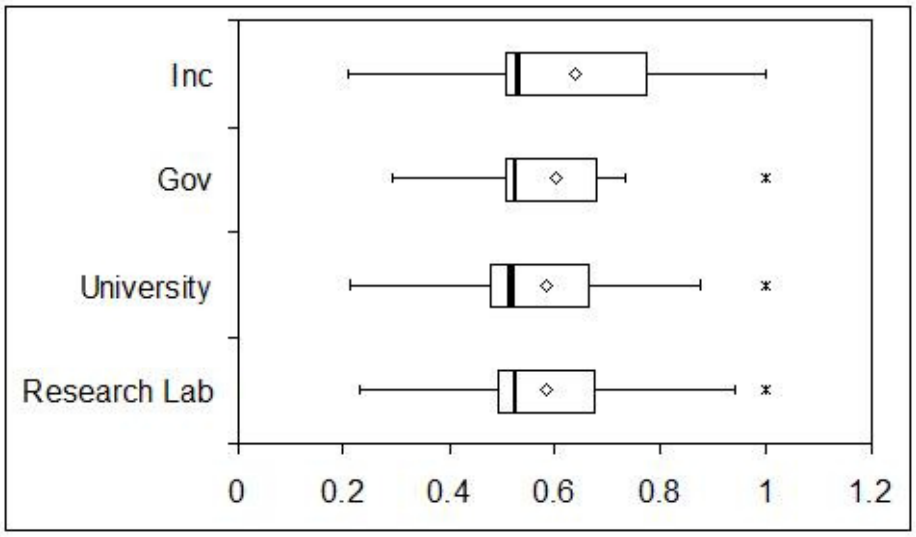


Figure 4.6: The average topic loyalty grouped by institution type for the scientific publication network.

government). One interesting result is that authors from corporate institutions, i.e. Inc., have a statistically significant higher average loyalty to their topic areas than the authors from academic institutions (like universities and research laboratories). One possible explanation for this is that authors from corporate institutions are more likely to publish in an area that coincide with corporate product or research goals, while authors from academia have more flexibility in terms of research agenda.

4.5.2 Senate Bill Sponsorship Network

The senate bill sponsorship network is based on data collected about United States senators and the bills they sponsor ([43]). The data contains senators' demographic information and the bills each senator sponsored or co-sponsored from 1993 through February 2008. Each bill has a date and topics associated with

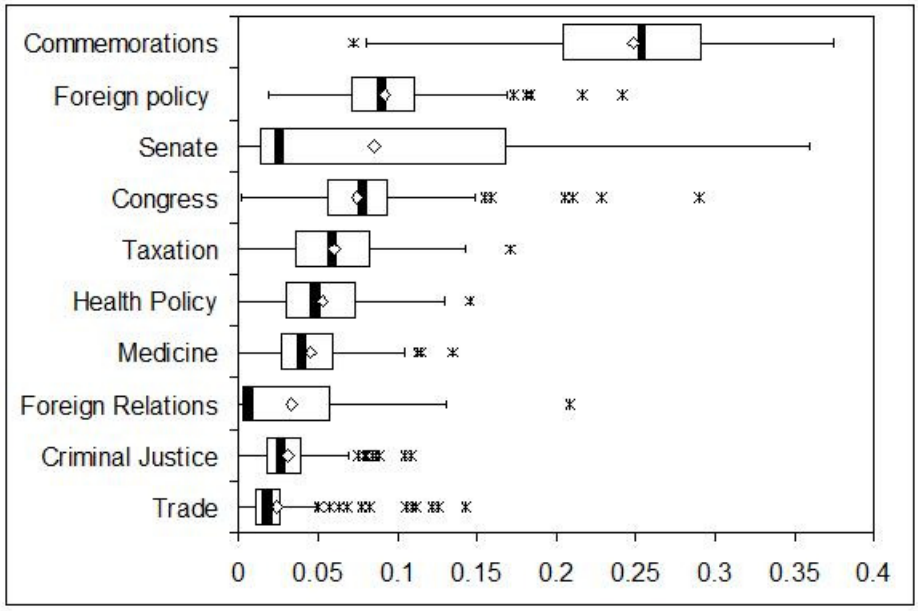


Figure 4.7: Average topic loyalty across all topics in the senator bill sponsorship network

it. I group the bills using their high-level topic, and then measure the loyalty of senators to different topics. After removing the senators that do not sponsor a bill or sponsor only a single bill and removing bills that do not have a topic, my analysis uses 181 senators, 28,372 bills, and 188,040 participation relationships spanning 100 high level topics.

When considering only the topics that each senator is most loyal to, the three bill topics that have the highest average loyalty values are Commemorations, Senate, and Congress. This average loyalty ranges from 0.22 to 0.27. By investigating the data set, I found that these three topics constitutes 56,035 (approximately 30%) of the total number of sponsorship/co-sponsorship relationships. This finding seems consistent since bills with these topics occur frequently, regularly, and have a large number of senators sponsoring them. Figure 4.7 shows

the 10 bill topics with the highest average loyalty across all the topic groups each senator sponsors a bill in. When looking across all topics for each senator, foreign policy has the second highest average loyalty value. Because the United States has been at war in recent years, this result is not surprising. The average loyalty of senators to bill topics is generally low. This is because of the large number of bills sponsored by senators across a large number of topics. Many may find comfort in this result since senators supporting bills across topics can be interpreted to mean that they are servicing a wider constituency.

To better understand the changes in loyalty over time, I investigate the changing dynamics of a particular senator's loyalty over time. I selected the senator that sponsored the largest number of bills - Senator Edward Kennedy, a democrat from Massachusetts. As illustrated in Figure 4.8, I calculated his group loyalty at 5 different time points. Although he sponsors bills across 130 topics, the graph shows nine topics with the highest means and standard deviations for loyalty values across the entire time period. During each time period, he consistently sponsors or co-sponsors roughly 10% of the Senate bills. The figure illustrates that Senator Kennedy starts out with a distribution of loyalty that favors a small number of bill topics. He does not sponsor bills across all the topics listed. Over time his loyalty to some of the topics decreases and increases to others as highlighted in the figure. It is also interesting to note that the variance of his loyalty across the topics decreases over time.

Finally, I briefly consider the 2008 presidential election. Examining results in the spring of 2008, in the time period preceding the 2008 fall presidential elec-

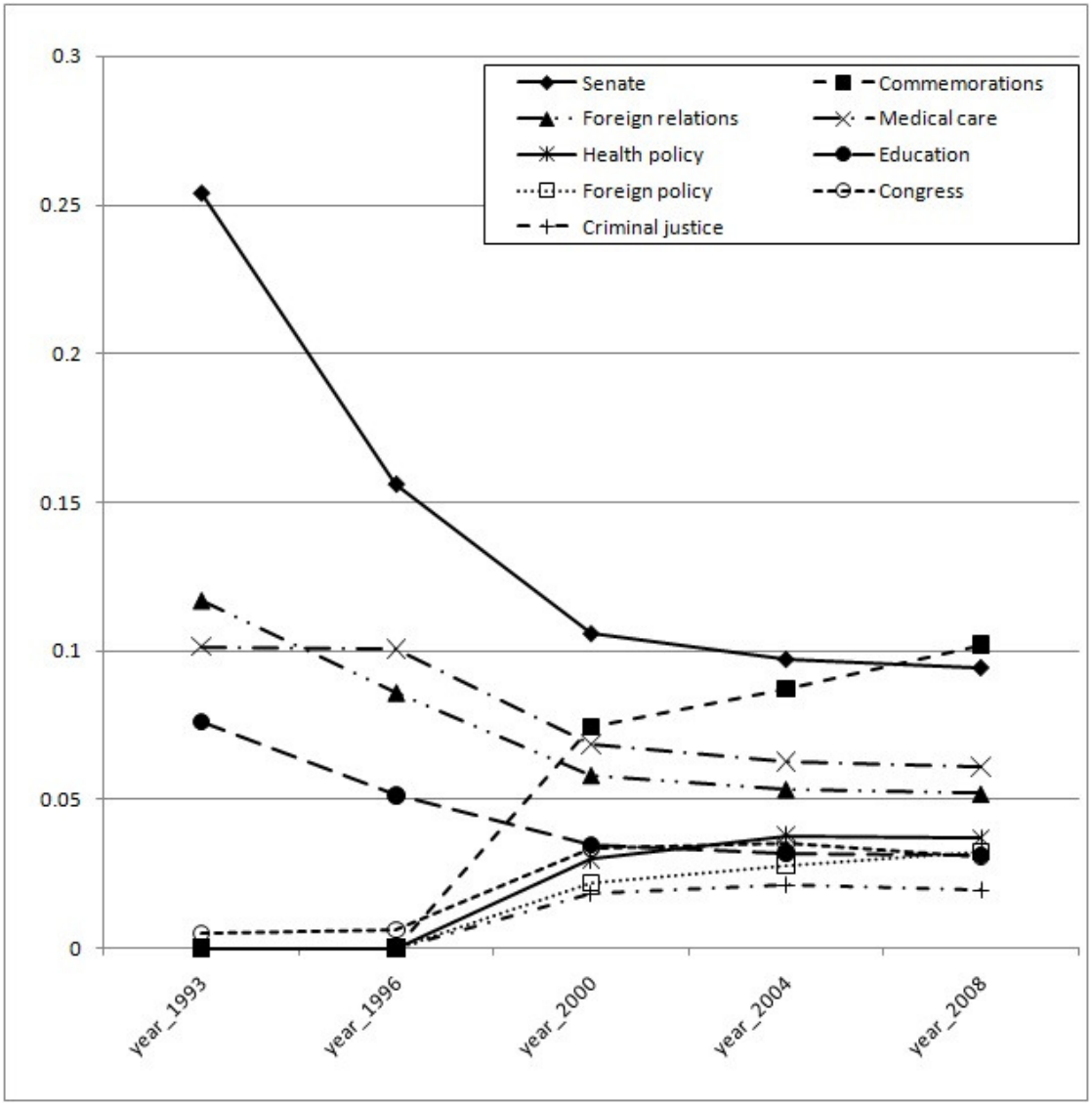


Figure 4.8: Changing loyalty over time for Edward Kennedy in the senate bill sponsorship network

tion, I compared the loyalty of the presidential candidates, John McCain, Barack Obama, and Hillary Clinton across a subset of bill sponsorship topics. The results are shown in Figure 4.9. These bill sponsorship loyalty values are consistent with priorities emphasized on the campaign trail. All the candidates have strong positions on foreign policy. Senator McCain made it a centerpiece of his campaign.

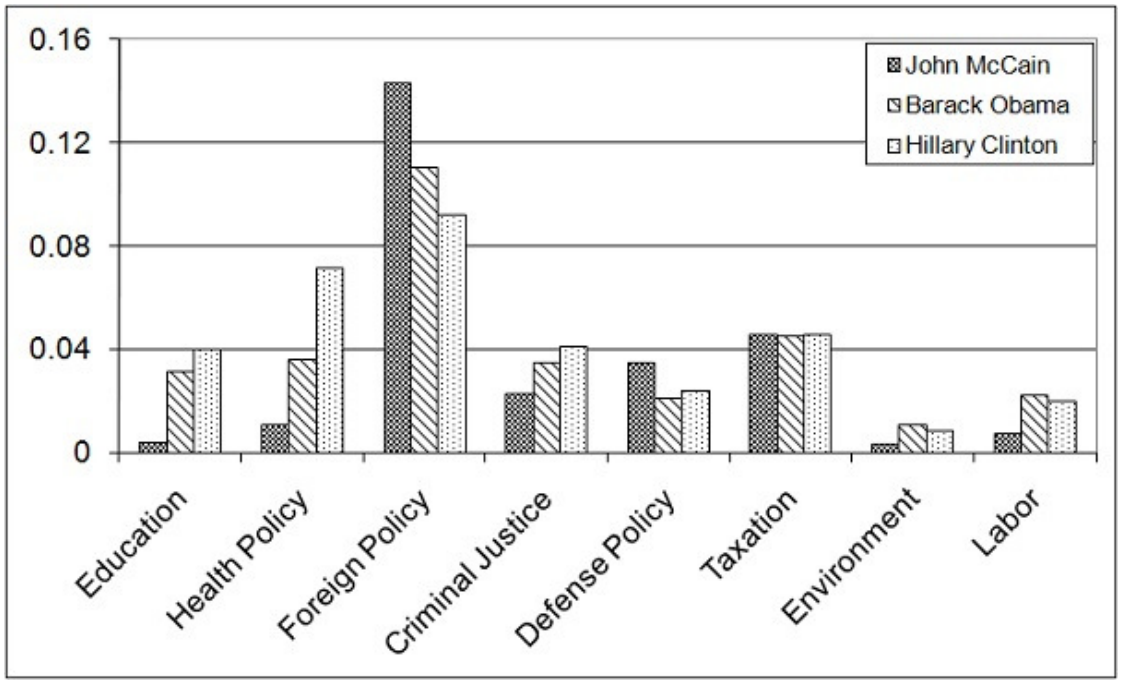


Figure 4.9: Average topic loyalty of 2008 presidential candidates in the senate bill sponsorship network

Senator Clinton had highlighted her commitment to health care. Both Senators Obama and Clinton also spent a lot of time discussing education. Interestingly, Senator McCain's loyalty to sponsoring education bills is very low.

4.5.3 Dolphin Social Network

I also consider an affiliation network based on a data set describing a long-term study of a wild bottlenose dolphin (Tursiops sp.) population in Shark Bay Australia [70]. It is the most comprehensive dolphin data set in research today with over 20 years of behavioral, reproductive, demographic and ecological data on wild bottlenose dolphins. For this analysis, I focus on observational surveys, collected by researchers on the Shark Bay Dolphin Research Project (SBDRP).

Data gathered includes location, animal behaviors, associates, habitat, photographic information, and physical data (e.g., scars, condition, speckles). These surveys are brief, typically lasting 5 to 10 minutes. They are used to present a "snapshot" of associations and behaviors among dolphins.

The affiliation network is defined by using dolphins as actors and surveys as events. Dolphins observed in a survey constitutes the participation relationship. I group survey observations together by the location the observation takes place. There are six different general regions in this data set. Similar to the other analysis, I remove dolphins with few sightings (less than 5) as well as the surveys with no location. After doing this, the analysis includes 560 dolphins, 10,731 surveys, and 36,404 relationships between dolphins and surveys for the loyalty analysis.

Figure 4.10 show the average loyalty of dolphins to different locations based on the observational surveys. First, the average loyalties of dolphins across all locations ranges from 0.45 to 0.9. Some locations appeared to invite higher loyalty than others, e.g. *East* and *Red Cliff Bay*. One explanation for this is the varying habitat structure. For example, *East*, which has the highest loyalty, is mostly deep channels bisected by shallow sea grass banks. Many dolphins spend a large amount of time foraging. The extensive habitat heterogeneity might limit the region to dolphins with certain foraging specializations (channel foragers or sea grass bed foragers) [72]. For example, a subset of the dolphins in this population use sponges as foraging tools, and will forage almost exclusively in the *East* channels [92]. *Peron* is at the tip of the peninsula and is a very open area where the western and eastern gulf meet. This open habitat (to the Indian Ocean) may

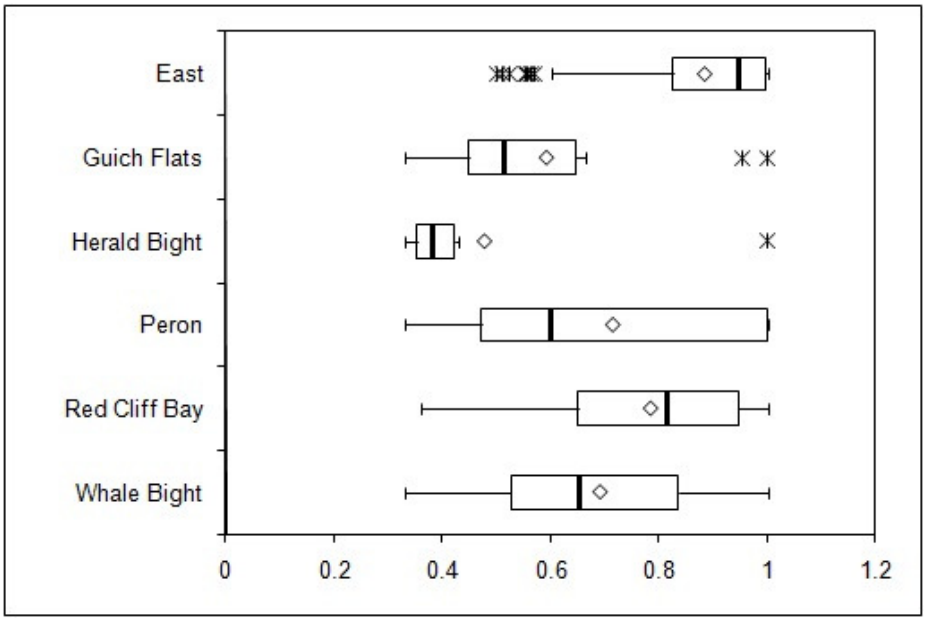


Figure 4.10: Average location loyalty for dolphins

allow for great mobility and less loyalty when compared to other areas.

Previous work by project biologists indicates calves are most tied to the locations of their mothers and maternal foraging type [71]. After weaning, juveniles might range further and develop bonds with others separate from the mother. Figure 4.11 looks at the distribution of location loyalty among different age groups: calves (0-4 years), juveniles (5-11 years), young adults (12-24 years), and old adults(25+ years). The results indicate that loyalty decreases with age, but still remains very high. This may occur because older dolphins travel more during the course of their life and they explore more places, while calves tend to have higher loyalty to a small number of locations (which happen to be the ones their mothers are also in). Location loyalty is a nice indication of long-term residency in the population and allows researchers to track individuals over long

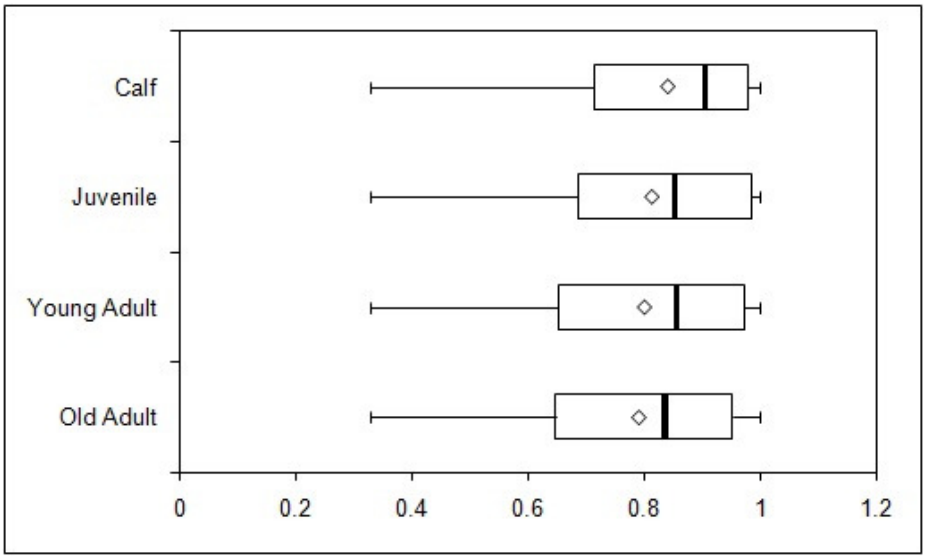


Figure 4.11: Average location loyalty for different dolphins' age groups

periods of time.

4.6 Comparative Loyalty Analysis

I now compare the average actor loyalty across these different communities. As the loyalty metric values can vary from zero to one, I divided the range of loyalty into three bins; low loyalty (scores from 0 to 0.25), moderate loyalty (scores from 0.25 to 0.75), and high loyalty (scores from 0.75 to 1).

The results in Figure 4.12 shows the percentage of actors with loyalty scores falling in each of the three bins for the scientific publication network, the dolphin survey network, and the political bill sponsorship network, respectively. This figure highlights the different distribution of actor loyalty in the different data sets.

The results for the ACMCHI publication network show that 79.2% of the

authors have moderate loyalty and most of the rest (20.4%) have high loyalty to the topic of their publications. In the political data set, 53.3% of the senators have moderate loyalty to the topic of the bill they sponsor, and the rest fall in the low loyalty category. For the dolphin affiliation network, we can observe that most of the dolphins (61.5%) have high loyalty to their locations. This large variation in the distribution of actor loyalty across data sets reinforces the utility of a measure that captures changing loyalty of actors to affiliation groups.

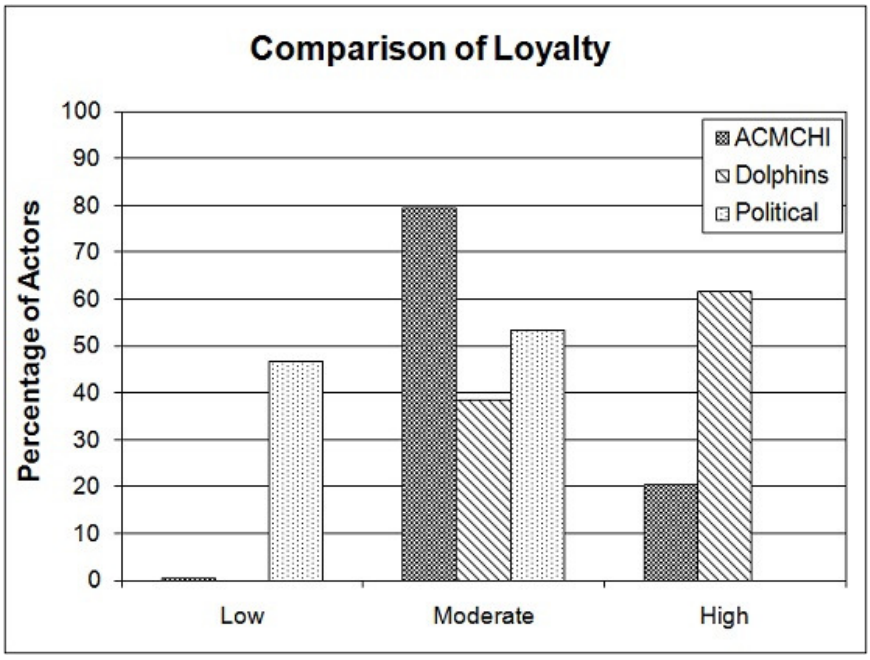


Figure 4.12: Loyalty Comparison Across Data Sets

These classification results are consistent with the interpretations of community loyalty presented in the previous section. The figure highlights the varying distribution of actor loyalty to groups in each affiliation network. As a final analysis, for each affiliation network I consider the average number of events each actor participates in. This allows us to compare the loyalty of these affilia-

tion networks to the density of the connections in the network. The averages are as follows:

1. Average number of Publications per Author = 3.61
2. Average number of Bills per Senator = 159.94
3. Average number of Observations per dolphin = 19.16

The network with the highest density is the senate bill sponsorship network, followed by the dolphin social network. The author publication network is much more sparse than the other two networks. Interestingly enough, the loyalty categories are not completely consistent with these frequency averages, thereby affirming that frequency alone may not be sufficient to capture loyalty.

4.7 Comparison with centrality measures

It is natural to want to understand how loyalty compares to existing centrality measures. Does it capture the same information, or does it provide additional insight? I begin by comparing actor loyalty to the most common centrality measures. The first centrality measure used is betweenness centrality, calculated by computing all pairs shortest paths in the network and computing the number of shortest paths that the target node occurs on. The second centrality measure used is the closeness centrality, defined as the average of shortest paths from the target nodes to all other nodes reachable from it. Lastly, eigenvector centrality measures importance of a node based on the importance of neighboring nodes. For more

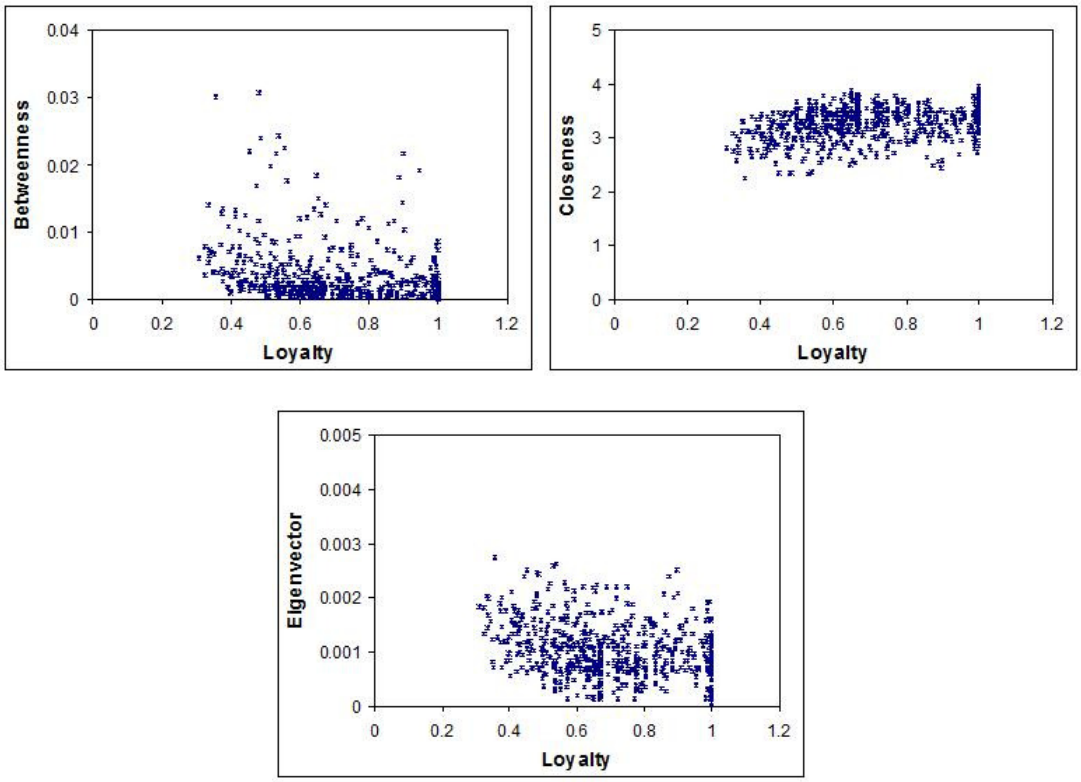


Figure 4.13: Loyalty vs. Centrality for Scientific Publication Network

detail about these measures, refer to [115].

For each affiliation group I consider the set of actors participating in the group and generate the underlying co-membership subnetwork. I do not generate a single clique of actors as the subnetwork of the group, but instead create a static subnetwork using the aggregation of projections of all the participation edges at each time point. We then compute the various centrality measures on each of the generated subnetworks corresponding to each of the affiliation groups and compare the centrality measures to the loyalty scores of the actors in each group. As can be noted in Figure 4.13, the scatter plot between loyalty and various centrality measures on the publication data set shows authors having all combination of values for both measures, with no visible trend in the results.

Note that actors may appear multiple times in this figure because they participate in multiple affiliation groups. The same results holds for the other two data sets.

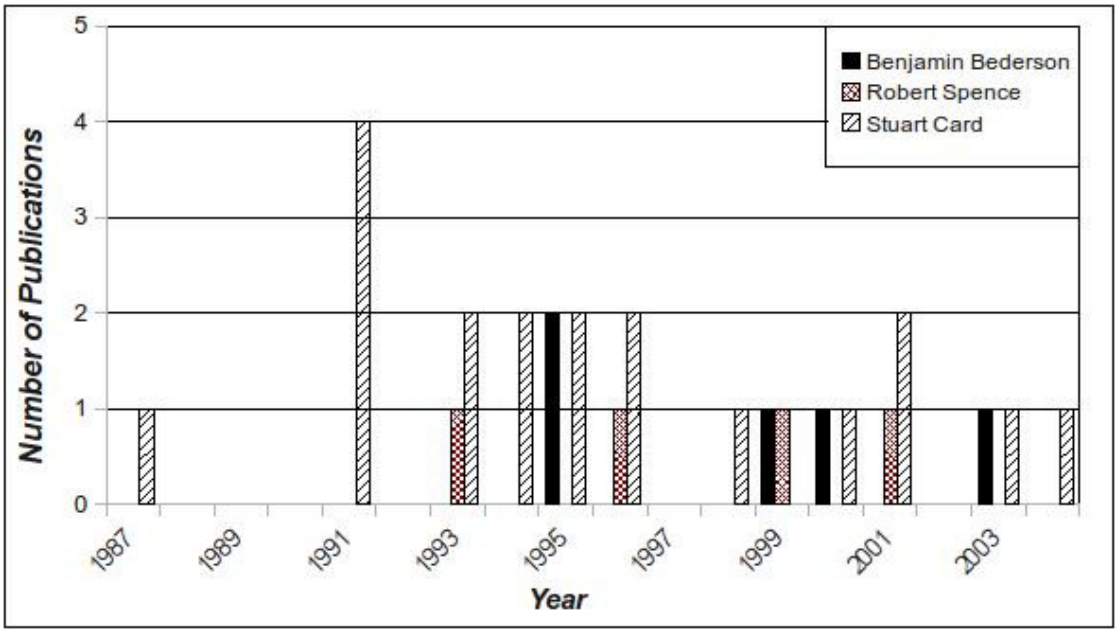


Figure 4.14: Author publications in "Information Visualization" topic

To further investigate loyalty to a particular topic, I take a more detailed look at the 'Information Visualization' topic as a sample affiliation group. For this group, Benjamin Bederson is ranked as the author with highest betweenness and eigenvector centrality. However, by examining Figure 4.14, I notice that his publication pattern is neither consistent across time nor numerous. This is also true for Robert Spence who was ranked first according to the closeness centrality. On the other hand, the time-consistent, recent and numerous publications of the most loyal author, namely Stuart Card, as shown in the same figure, illustrates exactly what the proposed loyalty measure captures that the other centrality measures missed.

4.8 Conclusion

I proposed a new measure for capturing loyalty in time-varying affiliation networks. I begin by defining affiliation groups, which describe temporally related subsets of actors. This is accomplished by grouping events over time based on attribute values. To model the dynamic behavior of affiliations to groups, I consider the concept of loyalty and introduce a measure that captures an actor's loyalty to an affiliation group as the degree of 'commitment' an actor shows to the group over time. I compare the proposed measure to both frequency-based loyalty and recency-based loyalty and find my measure to be more flexible since it incorporates components for frequency, consistency, and recency. I then demonstrate its utility on three real world affiliation networks: a publication network, a senate bill co-sponsorship network, and a dolphin network. It is interesting to note that the distribution of actor loyalty varies substantially across data sets, thereby reinforcing the utility of a measure that captures changing loyalty of actors to affiliation-based groups.

Chapter 5

Differential Adaptive Diffusion: Understanding Diversity and Trust Dynamics in Complex Networks

After analyzing the user-level interactions in affiliation networks, I now move to a macroscopic analysis of network dynamics, investigating the process of information diffusion in complex network settings. Information diffusion techniques focus on modeling the spread of innovations, diseases, products, etc. on the existing social network among users. A number of probabilistic models have been proposed in the literature to model the diffusion process, and estimate the levels of adoption/infection in the network given a set of initial seeds. Although these models capture the dynamics of diffusion within a given scenario, they have a number of shortcomings. The first drawback lies in the fact that these models are independent of the nature of information that is spreading over the network. Treating the information as an orthogonal dimension to the diffusion process loses a wealth of information in the existing relationships between the users and the ideas that are spreading. For example, in viral marketing domains, traditional techniques neglect the existing user-product preference networks. My first hypothesis is that leveraging these additional dependencies allows us to better model the diffusion process.

The second drawback of existing information diffusion models, is the exist-

ing implicit assumption that the underlying social network structure is a static proxy for the influence among users. This assumption fails to model the dynamics of the relationships among users, such as their trust in each others' recommendations, which is crucial to the diffusion process. Returning to the viral marketing example, if there exists a spammer in the network who continuously make irrelevant product recommendations for her peers, traditional diffusion models would not be able to capture the decrease in the influence along the corresponding links. Thus, my second hypothesis focuses on the premise that an adaptive diffusion model that captures the change in confidence values among users as a result of their prior interactions would better model the true underlying dynamics of information diffusion.

To evaluate my hypotheses, I propose a differential adaptive diffusion model for complex networks. I focus on the diffusion of different stories / news posts in a real-world network extracted from the Digg social news website. By analyzing the diffusion in the Digg network, I provide insights into the effects of network dynamics and topic preferences on the adoption of stories of different topics. The experiments show that the proposed model outperforms earlier non-adaptive diffusion models in predicting future adoptions. I also discuss the implications of the proposed adaptive diffusion model on identifying influentials in social networks.

5.1 Introduction

How information diffuses through social networks is a question that has attracted the attention of scholars from a wide variety of research disciplines. A richer understanding of the mechanism governing the spread of new ideas or trends in social media has implications for marketing, sociology, journalism, computer science and many other research areas. Models of network diffusion have been used to study phenomena as widespread as product recommendation systems [65], viral marketing [20, 61], disease transmission [19], herding behavior in financial markets [24], and even the contagion properties of obesity [16]. This is in part because the widespread growth and use of online social networks has created a new opportunity to observe diffusion processes on a very large scale and across different types of interactions from email to microblogging to photo-sharing.

Most of the existing information diffusion literature builds upon the premise that the social network can be used as a static proxy for influence, where the diffusion process is mainly dependent upon the structure of the network. However, in real settings, the influence within social networks is not static. As consumers continue to listen to their friends and family, they learn that some of their social connections have recommendations that are more appropriate for them and that other members of their social network simply do not have the same interests as they do. This is in part because different individuals are interested in different topics.

As an example, for someone who is not interested in sports, if a friend is constantly talking to her about new sports developments, sending her emails and links to promotions for sporting events, then this friend is essentially acting as a spammer and the focal individual will eventually decrease their trust in her recommendations. However, if another friend makes a recommendation and the focal individual adopts the product that they recommend, then the trust of the focal individual in that friend's recommendation will increase. As a result of these processes, the social network of confidence changes over time as a result of the diffusion and adoption process. Although the dynamics of social trust has attracted the attention of multiple researchers [40], most information diffusion models do not fully address the social trust aspect, nor the heterogeneity of preferences that individuals have for different topics.

In this chapter, I present an adaptive model that addresses these shortcomings by allowing individuals to have different preferences for different types of information, while adapting their confidence in other individuals' recommendations on the basis of their historical interactions. I show the novelty in my model over previous ones which assume the confidence that a user has in her peers remains constant over time and that the preference for adoption is not dependent on the type of information being diffused. By incorporating network-level dynamics into a standard diffusion model and allowing for heterogeneous preferences, the proposed model provides a better prediction of expected users' adoption. Finally, I discuss one application of the proposed model for the problem of identifying influentials and seed users for maximizing the adoption process.

5.2 Background

One of the first and most influential diffusion models was proposed by Bass [8]. This model of information diffusion predicts the number of people who will adopt an innovation over time, and though it does not explicitly account for the social network, it does assume that the rate of adoption is dependent on other members of the population, specifically the current proportion who have already adopted. The diffusion equation used by this model describes the cumulative proportion of adopters in the population at any time as a function of the intrinsic adoption rate and a measure of social contagion. The model describes an S-shaped curve, where adoption is slow at first, then takes off exponentially and flattens at the end. The Bass model has been shown to effectively model word-of-mouth product diffusion at the aggregate level [69], but does not explicitly model the decision of an individual consumer.

Though the Bass model can easily be generalized to address individual-level decisions [104], most diffusion models that capture the process of adoption of an idea or a product at an individual level use different mechanisms and can generally be divided into two groups: *threshold models* and *cascade models*. Threshold models are based on the work performed by Granovetter [44] and Schelling [93] in the late 70's. Basically, each individual, v, in the network has a personal adoption threshold $\theta_v \in [0, 1]$, typically drawn from some probability distribution. A given individual v in the network adopts a new idea or product if the sum of the connection weights of its neighboring peers that have already adopted it

$N(v)$ is greater than her personal threshold:

$$\sum_{u \in N(v)} w(u, v) \geq \theta_v.$$

Although the above model represents a *linear threshold model*, it can be easily generalized further with replacing the summation with an arbitrary function on the set of active neighbors of individual v. Dodds and Watts [19] have also shown that a more general model than this can be used to describe both the Bass model and the threshold model.

Cascade models [41] were originally inspired by research on interacting particle systems. In these type of models, whenever a peer u of an individual v adopts a given idea, then individual v also adopts with probability $p_{u,v}$. In other words, each individual has a single, probabilistic chance to activate each one of her currently inactive peers, after becoming active herself. A very common example is the *independent cascade model*, in which the probability that an individual is activated by a newly active peer is independent of the set of peers who have attempted to activate her in the past. Kempe et al. [55] proposed a broader framework that simultaneously generalizes the linear threshold and independent cascade models, having equivalent formulations in both cases.

Regardless of the adoption model, one of the key aspects that affects information diffusion is the interaction structure. For instance, a model for product adoption in small-world networks was proposed by Centola et al. [14], where an individual's probability of adopting a product is dependent on having more than

one neighbor who has previously adopted the product. Wu et al. [119] modeled opinion formation on different network topologies and found that if highly connected nodes were seeded with a particular opinion, this would proportionally affect the long term distribution of opinions in the network. The work of Holme et al. [47] focuses on coupling the evolution of both the social network and opinion formation, where both aspects adapt to each other during the evolution process.

Once a diffusion model and a network topology are specified, the next question is which set of individuals should be targeted to maximize the spread of information throughout the network. The problem of influence maximization was formalized by Domingos et al. [21], who noticed that ordinary data mining techniques that reason about consumer behavior in independent settings do not utilize network information. They proposed a probabilistic model of user-interaction to study influence propagation in networks, and then explored how to identify a group of individuals, who if they adopted a product, would maximize the speed and amount of adoption throughout the network. Even before Domingos et al. formalized this problem, one hypothesis as to how to maximize diffusion centered around the concept of *influentials*, who are individuals that have a disproportionate effect, compared to average individuals, on the amount and rate of information diffusion. In many information diffusion models, it has been shown that the most influential individuals in a network are the most central, where centrality is measured in a variety of different ways, including the most highly connected nodes, i.e. degree centrality [115, 2]. Other solutions have also been proposed, for instance, Stonedahl et al. [104] show that not only is de-

gree centrality important in maximizing diffusion, but in real social networks it is important to consider the clustering of a node's neighbors since tight clustering slows the diffusion process.

5.3 Case Study: Digg

Many popular online social network platforms allow for individuals to recommend items of interest and exchange knowledge. One such example is *Digg.com*, which is a popular social news website, where users can share and vote on different stories, referred to as "*digging*", to elevate the ranking of the story on the website. Digg's users form a social network by "following" other users in the network, which enables automatic tracking of their future diggs and submissions. Each news story on Digg belongs to one of ten topics; *Business*, *Entertainment*, *Gaming*, *Lifestyle*, *Offbeat*, *Politics*, *Science*, *Sports*, *Technology*, and *World News*. I constructed a sample from the Digg network which included both the diggs and follows for 11,942 users and the stories they submitted over a 6 months period (Jul - Dec 2010). The sample include 1.3 million follows relationships among the users, with over 1.9 million diggs, on 48,554 news stories.

The network alone is not enough to describe the diffusion process in a network, it is also important to understand the mechanism by which a user provides recommendations to their peers. These mechanisms differ by platform and marketing strategy. For example, some mechanisms are based on broadcast techniques, where all the peers of a given user are informed when she adopts a prod-

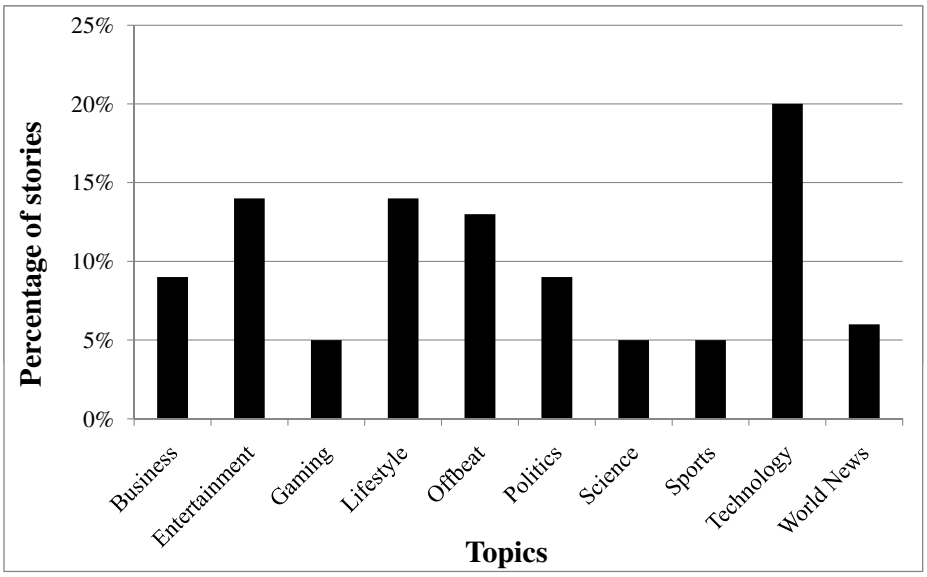


Figure 5.1: Topic distribution of stories in Digg dataset

uct. In other settings, the user has to explicitly select peers to send her product recommendations to after adoption. *Digg.com* uses a broadcast mechanism, where connected users are able to see all the activities of their peers as soon as it is performed.

5.3.1 Analysis

I begin by analyzing the topic distribution of the news stories in the collected data. As shown in Figure 5.1, though there are differences, all ten topics are represented at comparable levels in the dataset, without a single topic dominating the others. *Technology*, *Entertainment*, and *Lifestyle* are among the topics with higher frequency, while *Gaming*, *Science*, and *Sports* are the ones with lowest number of submissions.

I use the topic distribution of individual user submissions (the actual stories

/ links they submitted), as opposed to their diggs, as an influence-independent source for determining a user's topic preferences. Given this topic distribution, I then measure the correlation between the users' topic preferences and their actual adoptions, i.e., their diggs. Figure 5.2 shows the Kullback-Leibler divergence between the topic distribution of the users' submissions versus their diggs. For most users, there is very little divergence between their adoption behavior and their inferred preferences according to their submissions. However, in approximately 10% of the users, there is a quite significant difference between the topic distribution of the stories they digg and the ones they submit. One possible explanation is that while most people adopt only stories of interest to them, there are a smaller percentage of "imitators" who are easily influenced by their peers and do not weight their own preferences as highly. Similar results were obtained using normalized mutual information (NMI) between the topic distribution of users' preferences and adoptions, with imitators appearing to be even more prominent (~16% of the users).

In order to characterize users' topic preferences, I measure the KL-divergence between the topic distribution of each user's submissions and a uniform distribution of topics. Lower values indicates that the user's submission pattern is closer to uniform, while higher values indicate that the user is more interested in certain topics but not in others. From Figure 5.3, we can distinguish three different groups of users in the network: *Focused* users (~53% of the users) who are characterized by having highly skewed preferences towards one or two topics, *Biased* users (~32% of the users) who have less skewed preferences towards a larger set

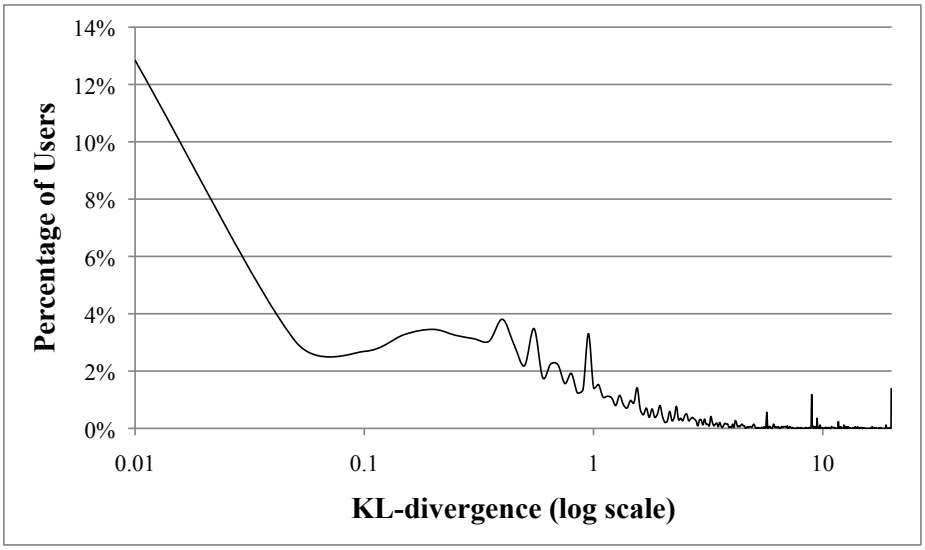


Figure 5.2: KL-divergence between the topic distribution of users' submissions and diggs.

of topics, and *Balanced* users (~15% of the users) who have almost-uniform topic preferences in their submissions.

Finally, I analyze the dynamics of change in the nature of the social relationships between users, and how it affects peer influence over time. I hypothesize that as time passes, peers with similar preferences in topics start gaining confidence in each other's recommendations, yielding higher levels of adoptions, while on the other hand, peers whose preferences are farther apart from each other become less confident in each other's recommendations, resulting in lower adoption levels. To test my hypothesis, I measured, at different time points, the average number of diggs on the same story by different peers for different values of KL-divergence between their topic preferences. Figure 5.4 shows that peers with lower KL-divergence in their topic preferences increase their number of shared diggs over time, while the ones with higher levels of divergence have

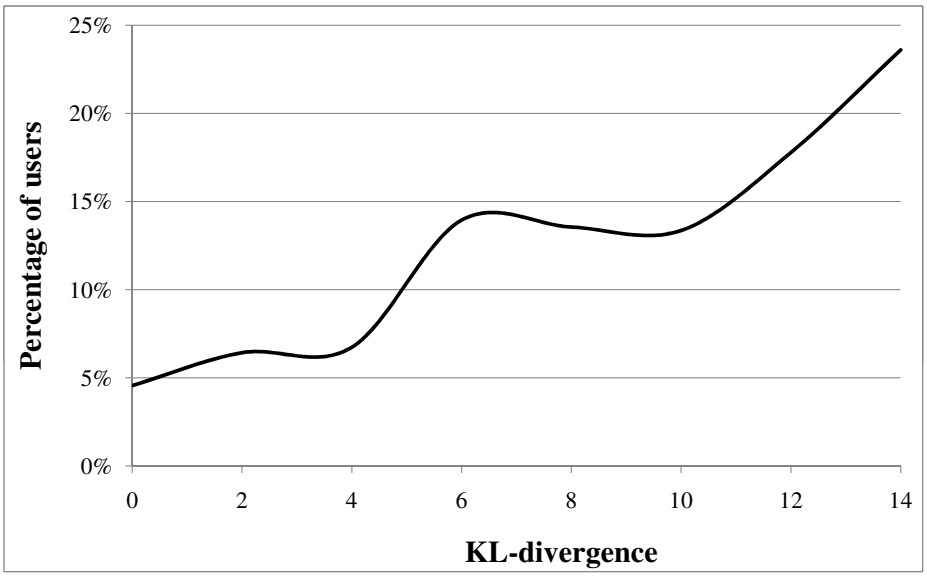


Figure 5.3: KL-divergence between uniform topic distribution and users' submissions

a decreasing pattern of adoptions over time.

5.4 Differential Adaptive Diffusion

The input social network can be viewed as a directed weighted graph $G(V, E)$, where V represents the network users, and E represents the social relationships among them. Each edge $e(u, v) \in E$ is associated with a confidence value $w_i(u, v) \in [0, 1]$ representing the confidence user v has in the recommendations of her peer u during campaign i. This confidence value $w_i(u, v)$ is updated only once per campaign, and in general this update could take place either immediately after a recommendation or at the end of a campaign. In the model results presented here, the confidence weights are only updated at the end of a campaign. Given a preference function $\mathcal{F}(v, c) : V \times C \rightarrow [0, 1]$ that quantifies user preferences for different product categories $c \in C$ for a given user v, the probability of node v

adopting a product of category $c \in C$ within campaign i as a result of node u adopting it can be defined as:

$$p(u, v) \triangleq w_i(u, v) \times \mathcal{F}(v, c)$$

To start a new campaign for a certain product x_c of category c, a marketing incentive is provided to a chosen set of seed nodes in the network to initiate the diffusion. As the diffusion process unfolds, the set of nodes who adopt the product at each time step, t, referred to as the "*active*" nodes, influence their peers through recommendations. These recommendations cause their neighbors to consider whether or not to adopt the product. The adoption function can take any form including any of the functions described in the background section, but throughout the following discussion I will assume an independent cascade process. Thus each active node u in time step t has a single chance of activating a peer v that has not already adopted the product where it succeeds with probability $p(u, v)$, which will result in v adopting the product. Once node u attempts to activate an inactive node v, it can never attempt to activate node v, in any future time step, i.e., node u will return to an inactive but adopted state after this time step. Given the set of active neighbors $N_t(v)$ of a given inactive node v at time t, the posterior probability of v adopting the product at time $t + 1$ can be defined as $p_{t+1}(v, x_c|N_t(v)) = 1 - \prod_{u \in N_t(v)}(1 - p(u, v))$. When a node adopts the product, it becomes active and starts activating its currently inactive neighbors at future time points. The diffusion process continues until no further adoptions occur for

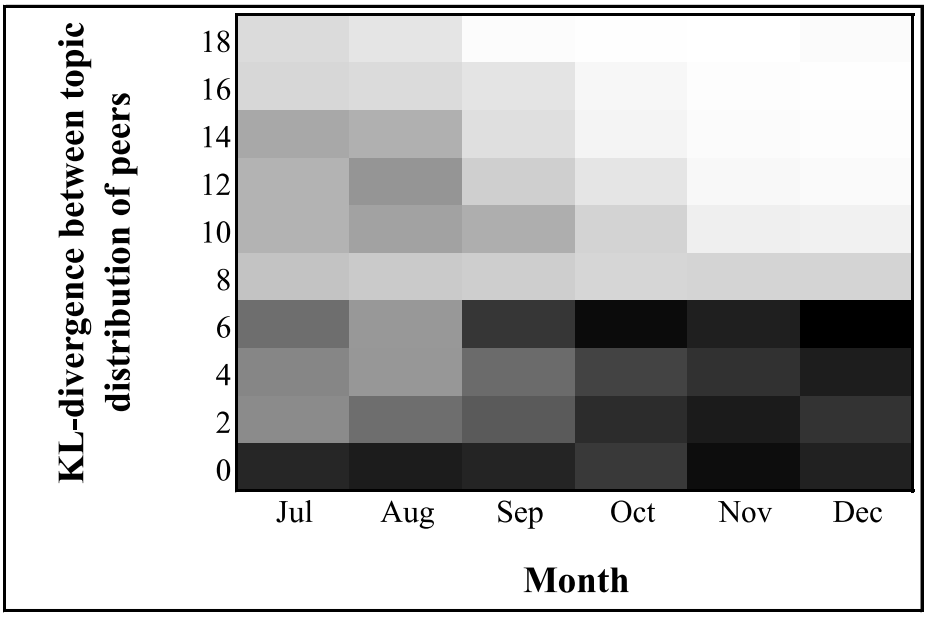


Figure 5.4: Heat map of the average number of diggs for different values of topic divergence between peers across time.

the current product.

At the end of each campaign, the confidence values among peers are updated according to the outcome of the product recommendation across the corresponding edge. I denote by $t_i^*(v)$ the time step within campaign i at which a node v adopts the product. If a given node u ends up not adopting the product by the end of campaign i, $t_i^*(u)$ is set to ∞. Using a kernel function K, the change in confidence values at the end of campaign i for product x_c can be calculated as $\Delta W_{i+1} = K(W_i; \theta)$, where $\theta \in [0, 1]$ is a kernel parameter specifying the rate of

change. For instance, a linear kernel can be defined as:

$$K_L(W_i;\theta) = \begin{cases} \theta \times \frac{1-w_i(u,v)}{t_i^*(v)-t_i^*(u)+1}, & t_i^*(u) < \infty \wedge t_i^*(v) < \infty \\ \theta \times \frac{-w_i(u,v)}{t_i^{max}(v)-t_i^*(u)+1}, & t_i^*(u) < \infty \wedge t_i^*(v) = \infty \end{cases}$$

where $t_i^{max}(v) = max_{t_i^*}\{t_i^*(u) : (u,v) \in E \wedge t_i^*(u) < \infty\}$ represents the time of the last adoption by any of v's peers.

This linear kernel assigns credit to each peer u of a node v proportional to the elapsed time between that peer's recommendation and node v adopting the product. The intuition is that the node u, that last recommended the product, has the highest impact for influencing node v to adopt the product, and thus should be assigned higher confidence in her future recommendations to v. If node v ends up not adopting the product by the end of the campaign, each peer u who recommended the product to node v is penalized relative to the time of the last recommendation. In this case, the last person to recommend the product, even though v still has not adopted it and will not adopt it, gets the maximum penalty for their recommendation.

Different types of kernels can be used to control the dynamics of the confidence levels in the network. For instance, this kernel could be exchanged with a kernel where only the last node to provide a recommendation is penalized or rewarded, as opposed to all nodes, or one where all nodes are punished or re-

warded equally. Regardless, as a new campaign is initiated for a different product, the new, updated confidence values are used to compute the influence probabilities, thus enabling the model to capture the dynamics of the diffusion process across different product types.

5.5 Influentials

One of the tightly related problems to information diffusion is identifying the set of users that should be initially targeted to maximize the spread of information throughout the network. This problem was formalized by Domingos et al.[21] who noted that ordinary data mining techniques that reason about each consumer behavior independently, lead to suboptimal marketing decisions resulting from not accounting for the influence effects among users in the network. They suggested that incorporating the users' network effect into the marketing decision leads to better decisions and, consecutively, higher profit. Thus, instead of deciding whether or not to market to a customer based solely on the expected profit that would be gained from her making a purchase, marketing companies should instead take into account the effect that this consumer would have within her social network. This gives rise to the notion of "*influential*" users within the network; these are users who are capable of spreading the information throughout the network at a higher rate than other members.

One of the standard approaches for identifying influentials in networks is using degree centrality, where high-degree nodes are considered the most influ-

ential as they can reach out to many other nodes in the network [2, 116]. However, most of the centrality-based approaches ignore the dynamics of the user interactions that occur as a result of the diffusion processes themselves, as well as the heterogeneity in user preferences. In order to account for these shortcomings, I propose a confidence-based approach for identifying influentials based on the differential adaptive diffusion model.

The proposed confidence-based approach relies on using the confidence values $w_i(u, v)$ on the edges in the social network, at the start of the target campaign i, to construct a confidence-weighted influence score $s_i(v)$ for each user as follows:

$$s_i(v) = \sum_{u \in N(v)} w_i(u, v)$$

At the beginning of each campaign, the confidence weights are updated according to the utilized kernel function K, and the new scores for the users in the network are calculated. Then, the network is filtered to keep only the set of users with the highest preference to the current product category c. Finally, the remaining set of users are sorted based on their current scores, and the top $k\%$ are chosen as a seed influential set for the product.

By using both the user-preference network for filtering, and the adaptive scores s_i for sorting, the set of influentials chosen at the beginning of each campaign is able to capture both the diversity in user preferences as well as the trust dynamics in the network.

5.6 Experimental Evaluation

To test the proposed model, I used the first four months of interactions, i.e., diggs and submissions, on the Digg network as training data to learn the confidence values between different users, and used the last two months for evaluation. I use the action of "digging" a story as a proxy for product adoption, and the topic distribution of users' submissions to estimate their preferences. Starting from a uniform assignment of confidence values across all peers, I track the propagation of user diggs and update the corresponding confidence values according to the proposed model. The learned values along with the user preferences can then be used to predict adoptions for new stories, and to identify the influentials for future campaigns.

5.6.1 Predicting Adoptions

To evaluate the accuracy of the proposed diffusion model in predicting future adoptions, I compare my approach with two proposed approaches in [42] for learning the influence probabilities from training data. In the first approach (Bernoulli), they consider each recommendation a separate Bernoulli trial, and then estimate the confidence between two users as the maximum likelihood estimate (MLE) of the ratio of successful recommendations over the total number within a given contagion time. In the second proposed approach (Bernoulli-PC), the authors use the same Bernoulli representation but in this approach they give partial credit for each product adoption based to the set of peers who recom-

mended the product within a given time frame. Although both approaches have comparable performance, Goval et al. show that introducing the notion of "contagion time" as a factor in estimating the influence probability outperforms static methods and yields more accurate results.

The above method utilizes a threshold adoption rule as opposed to the cascade rule that is utilized in the model (Adaptive). We can convert between these two models; as shown by Kempe et al. [55], the independent cascade model is equivalent to a threshold model where the adoption threshold is set to the posterior probability of adoption; i.e. for a given user v, if we set $\theta_v = 1 - \prod_{u \in N(v)}(1 - p(u, v))$, the threshold model is equivalent to the independent cascade model. I use this conversion to facilitate in-depth evaluation of my model. I compare the different models by means of ROC curves, which are more appropriate than precision-recall curves in this setting [85]. The ROC curve shows the relative trade-offs between the true positives (correctly identified adoptions) and the false positives (unrealized predicted adoptions) as the discrimination threshold is varied. Each point in the ROC curve corresponds to one possible value of activation threshold for the users.

Figure 5.5 illustrates the performance of all three models using ROC curves where the x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR). The proposed model (Adaptive) outperforms both baselines (Bernoulli and Bernoulli-PC), yielding higher true positive rates at low values of false positives. I also experimented with using a predictor that ignores the peer-influence altogether and relies only on the stories that were promoted to the "top sto-

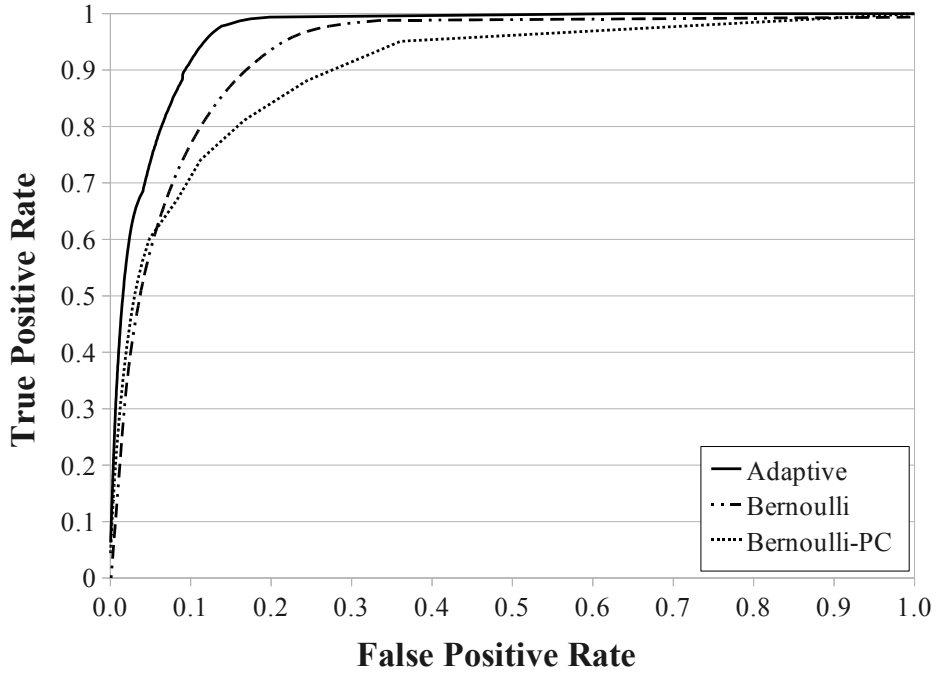


Figure 5.5: ROC performance of two comparison models (Bernoulli and Bernoulli-PC) and the proposed model (Adaptive) on the basis of the False Positive Rate (FPR) and True Positive Rate (TPR) for each model.

ries" section in *Digg.com*. This popularity-based predictor yielded an accuracy of 45.7%, which is lower than random prediction This indicates that individuals' connections and interactions with their content preferences are more important factors than the overall popularity. Similar results were also confirmed by [60]

These results show that by modeling the dynamics of the diffusion process at a finer-grained level, taking into account the heterogeneity of users and the dynamics of the social network, it is possible to create a model which outperforms a more naïve model. This in turn leads to a better understanding of the whole diffusion process.

5.6.2 Identifying Influentials

Using the same experimental setup, I compute the confidence-weighted scores s for all the users in the dataset over the training period. Then, I use the computed scores along with the learned preferences for each user to identify the set of confidence-based influentials at the start of each campaign in the over the last two months in the dataset. I compare the proposed method to a random baseline and a degree-centrality approach for identifying influentials. The evaluation is based on the cascade size, measured by the average number of diggs / post, of the stories posted by the chosen set of influentials in each of the compared methods during the test period.

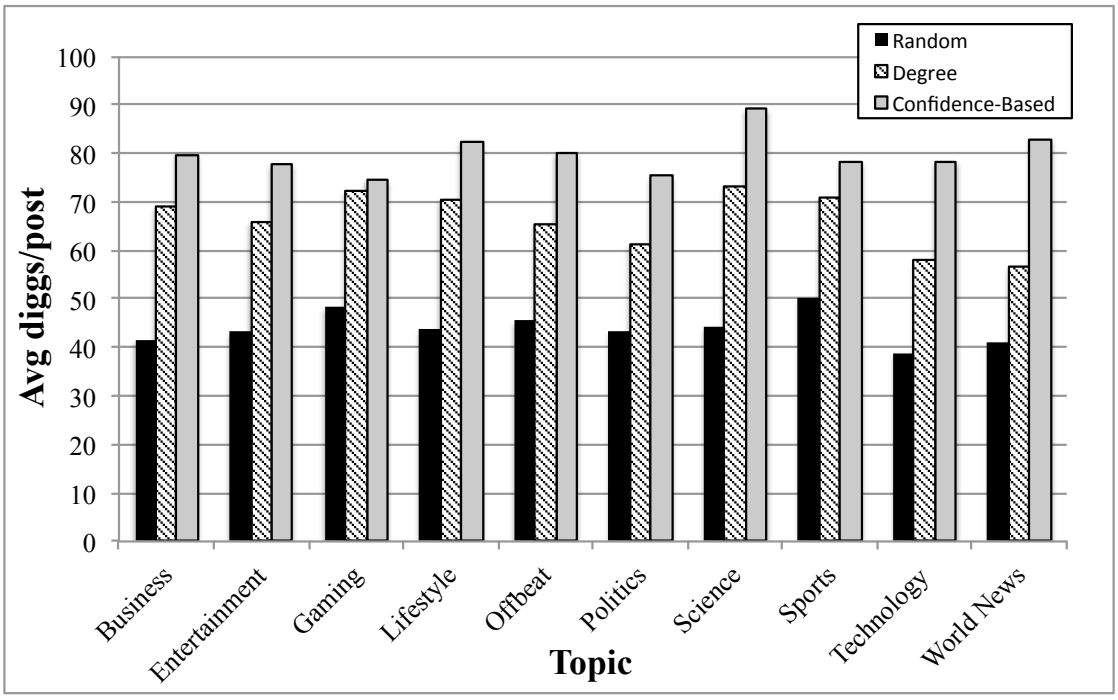


Figure 5.6: Average number of diggs/post for the top 10% influential users in *Digg.com*

By choosing the top 10% influential users according to the compared meth-

ods, 5.6 shows that the proposed confidence-based approach consistently outperform both the random and the degree-centrality baselines across all topics. This confirms our findings that the reputation of the users and their past behavior plays an important role in their future influence and impact over the network. The suggested confidence-approach method also provides the brand managers with a mean for identifying sets of users in the network for Ad-targeting to maximize the spread of different product types.

5.7 Conclusion

In this work, I provided new insights into the effect of network-level dynamics and individual heterogeneity on the diffusion process in real-world networks. Utilizing a sample of users' interactions on the *Digg.com* social news website, I analyzed the effect of peers' confidence in each other's recommendations on the adoption of different news posts over time. I presented an adaptive diffusion model that is able to capture the observed properties, and showed that it outperforms earlier non-adaptive models in predicting future adoptions. I also proposed a confidence-based approach for identifying influentials based on the proposed differential adaptive diffusion model. I showed that the proposed method for identifying influentials outperforms the classical, structure-based approaches across different topics / product categories.

I believe one of the important future steps is studying the implications of the proposed adaptive diffusion model on existing viral marketing mechanisms. The

proposed model suggests that the current incentive structure of most of the existing viral marketing techniques doesn't account for the trust dynamics among users, and might lead to decreased efficacy of these strategies in the long run. Analyzing the performance of the existing viral marketing strategies in the light of the proposed diffusion model will provide new insights about potential methods for designing more efficient incentive structures.

Chapter 6

Adaptive Viral Marketing

Viral Marketing has proved to be one of the most successful marketing strategies that allowed companies to effectively reach a large segment of the potential customers that are resistant to more traditional marketing strategies. The basic idea behind viral marketing is relying on the concept of information diffusion over the existing social network among customers to advertise for different products. One of the main implications of the differential adaptive diffusion model proposed in the last chapter is a better understanding of the effects of existing viral marketing strategies on the underlying social networks in the long term. The model suggests that user recommendations are most effective when recommended to the right subset of friends. If a user is very selective and makes each recommendation to only a few friends, then the chances of success are slim due to limited network exposure. On the other hand, recommending a product to everyone may have limited returns as well, due to the effect of irrelevant recommendations on the confidence levels in the underlying social network.

In this chapter, I focus on analyzing the implications of the differential adaptive diffusion model, discussed in the previous chapter model, on existing viral marketing strategies. I illustrate the effect of classical viral marketing techniques on the trust dynamics among users in the social network, and then propose a

new viral marketing strategy for maintaining the trust levels among users over time. I show the utility of the proposed adaptive viral marketing method over the multiple campaigns of different product, compared to the existing strategies.

6.1 Introduction

Viral marketing builds upon the ideas from network-based diffusion processes. The main goal of viral marketing is to exploit existing social networks among customers by encouraging those customers to share product information with their friends. This goal is based on the premise that consumers' purchasing decisions are heavily influenced by recommendations and referrals from their family, friends, and colleagues; an assumption that has been supported by some of the earliest studies of diffusion [89]. Recently, viral marketing has become more appealing to marketers as consumers have started to show an increasing resistance to traditional forms of advertising such as TV or newspaper ads.

One of the major early success stories of viral marketing was the introduction of "Hotmail." When this web-based email service started in 1996, each message sent by a user included a promotional message with the URL of the service. As a result, "Hotmail" gained its first twelve million subscribers in just eighteen months, on an advertising budget of only $50,000 [51]. Similarly, cell phone companies are another industry where providers take advantage of social network-based diffusion by offering highly discounted rates for customers talking to other customers within the same network. Thus, if a customer's social circle (family,

friends, colleagues) is using a certain provider, there's an added incentive for her to use the same provider.

In order to motivate users to spread product recommendations throughout the network, most viral marketing strategies includes some kind of an incentive or a "*reward*" for sending the product recommendations through the users' personal connections. A major drawback that arises from this mechanism is the emergence of star-like patterns [61] where a set of users recommend the product to all their peers in an effort to increase their expected reward. As a result, due to the heterogeneity of the user-product preferences in the network, a percentage of the users in the network end up receiving recommendations for products that they might not be interested in, just to maximize the possible benefit for the recommender. Moreover, as I discussed in chapter 5, this kind of behavior leads to a change in the trust dynamics among the users in the network, which in turn affect the information diffusion process.

The majority of literature on viral marketing assumes that the way information spreads throughout the network is static, and conveyed solely through the existence of links [36] or some other structural properties of the network [33]. This is evident in the methods for choosing the influentials for initial targeting, which are often determined using some structural property (degree, betweenness centrality, etc.). This strategy for identifying influentials is based on the premise that a large number of connections in the social network directly correlates with a larger impact and more potential for spreading information over the network [2, 116].

In this chapter, I investigate the implications of the differential adaptive diffusion model on the design of the viral marketing strategies for different product types. Based on these implications, I propose an adaptive design for a new viral marketing strategy which is capable of sustaining the influence level between peers in the network. My hypothesis is that utilizing more selective methods which take into account sustaining the trust among users will result in better cumulative adoption of various products over time. In order to test my hypothesis, I compare the proposed adaptive viral marketing strategy with classic strategies that focus only on maximizing the individual product's adoption rate. I show that the proposed adaptive viral marketing strategy is able to incorporate: (1) multiple different product campaigns, (2) the diversity in user preferences among different product categories, and (3) changing confidence in peers' recommendations over time. These factors allows the model to sustain the trust values among users in the network, thus achieving better adoption rates of different products over time.

6.2 Background

Recent work by Leksovec et al. [61] focused on tracking the actual diffusion of recommendations through email, in order to quantify the importance of various factors introduced in the literature. They used a product recommendation dataset from an online retailer who employed a viral marketing strategy based on rewarding the referring customer who makes a successful recommendation for a

product with 10% credit, and the referred customer who accepts the referral with 10% discount on their purchase. Leksovec et al. utilized the recommendation data for different products to model its suitability for viral marketing in terms of both the properties of the network and the product itself.

The first observation the authors made was that the nature of the product highly affected the recommendation pattern, which can be attributed to users having different preferences for different product types. For instance, users tend to buy more DVDs and are more likely to recommend them to their friends, while they seem to be more conservative with books. The authors proposed different potential reasons for this behavior, among which is the fact that books needs further time investment for the user to read it and actually recommend it to a friend, in opposition to the nature of a DVD which can be viewed in a shorter period of time. Other factors include assumptions about the consumer behavior, as people in general are more informed about certain products (like DVDs in that case) through other means of advertisements, which in turn gives the user more confidence in making the recommendation.

The second observation the authors made was that the probability of a user making a recommendation at all, given that she has already adopted the product, declines after an initial increase as one gets deeper into the cascade. However, if this deeply nested individual chooses to make recommendations, she tends to recommend the product to a larger number of peers on average.

The authors also provided a thorough analysis on how the effectiveness of recommendations changes as one received more recommendations from the

same person. The experiments showed that recommendations start to lose effect after more than two or three are passed between two individuals. As the number of exchanged recommendations increases, the probability of buying starts to decrease to about half of the original value and then levels off. From an aggregate perspective, they carried out a set of experiments to measure how the average number of purchases changes with the number of outgoing recommendations, and showed that the result varies with different product types.

Similar findings were also discovered in [98] by analyzing another dataset from *Digg.com* social news website. However, in the work by Sharara et al., the authors provided a formal adaptive model for information diffusion that takes into consideration both the change in trust dynamics between users over time, and the diversity in user-product preferences.

6.3 Conceptual Model

Given a reward value of r units, the classical incentive structure is to grant the recommender a full reward (r units) for each successful recommendation that results in a purchase or an adoption. However, this incentive structure encourages users to spread the product recommendation for all their peers, without accounting for their preferences. This behavior can lead to decreased values of confidence in peer-recommendation throughout the network, which negatively affects the diffusion process as discussed in 5.

In order to avoid the side effects of this incentive structure, we need to de-

sign a new strategy for viral marketing that aligns the immediate utility of the users with the end goal of sustaining the confidence in peer-recommendation over time. The proposed "adaptive reward" mechanism achieves this tradeoff by restructuring the incentives to account for both successful and unsuccessful recommendations. Specifically, when a user makes a successful recommendation to one of her peers, she gets rewarded $(\alpha \times r)$ units, whereas if the recommendation is unsuccessful, the user gets penalized $((1-\alpha) \times r)$. The parameter α acts as a "*conservation parameter*", varying from 0 to 1, with 0 representing fully conservative behavior and 1 representing fully non-conservative behavior.

According to the classic viral marketing mechanism, where users only receive rewards for successful recommendations that result in product adoptions and no penalties for the unsuccessful ones, there is no reason for a user to be selective in the choice of whom to recommend the product to. This behavior encourages the users to send the recommendations to all their peers, as the expected reward can only increase by expanding the domain of users receiving the recommendation. This corresponds to setting the conservation parameter α to one in the proposed adaptive rewards mechanism.

However, by varying the value of α, the penalty for unsuccessful recommendations starts to affect the net reward that the users acquire, as for values of $(\alpha < 1)$, it is no longer the case that additional recommendation can only increase the net reward. To illustrate the effect of α, consider the following example: Suppose a user v adopts a product of a given category, and decides to recommend the product to her neighbors $N(v)$. Assume that only $M(v) \subseteq N(v)$ of her peers

have high preference for this product category, and thus adopt the product as a result of v's recommendation. Therefore, the net reward that v acquires can be expressed as:

$$R_{net}(v) = \max(0, r \times (\alpha|M(v)| - (1-\alpha)|N(v)/M(v)|))$$

Therefore, if a user chooses to follow a nonconservative strategy, the expected reward decreases by a penalty relative to the number of unsuccessful recommendations she makes. Tuning the conservation parameter α varies the tradeoff between the reward and the penalty that the user incurs, and thus allows us to test different mechanisms and analyze their effect on both the rate of product adoption as well as the overall confidence levels among users.

Despite the fact that the main benefits of the proposed adaptive strategy appears on the network level through reducing the spamming behavior within the social network, it also carries an advantage for users by maximizing their rewards over time. While the users have different preferences for different product categories, their judgment in the confidence of their peers is evaluated on an aggregate level. So, if a user chooses to engage in spamming behavior, this will lead to increased resistance by her peers to any future recommendation they receive from her, regardless of their preference for the product category, thus decreasing her future rewards significantly. As a result, by using the proposed method, users must face the penalty of spamming behavior explicitly, and as a result they will be more likely to follow a strategy which will maintain their peers' confidences

in them on the long run, and therefore increase their long term reward.

6.4 Experiments

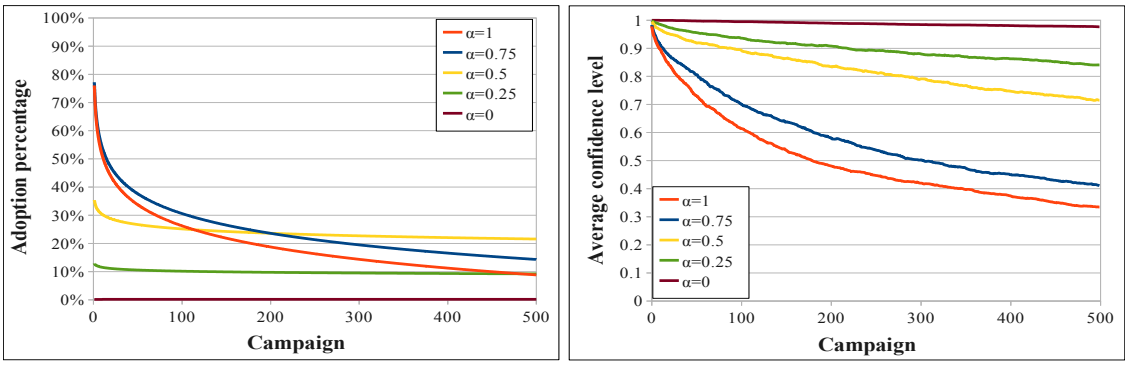
To evaluate the proposed viral marketing strategy, I use an agent-based model to simulate the behavior of customers in real settings. First, I create a synthetic social network using the preferential attachment [7] model. Then, for a given number of product categories, the user-preference network is generated by assigning a set of preference values for each agent in the network. Following the differential adaptive diffusion model[98], each link in the social network is assigned a weight representing the confidence that the target agent has in the source agent's recommendations. Given both the preference values and the peer-confidence, the influence probability of agent u on agent v for a product of category c at campaign i can be fully specified according to the differential adaptive diffusion model as:

$$p(u, v) = w_i(u, v) \times \mathcal{F}(v, c)$$

where $\mathcal{F}(v, c)$ represent the preference agent v has for product category c.

For the purpose of the experiments, the confidence values across all agents are initialized to unity, and are updated using the linear kernel function defined in [98]. The objective of each agent in the model is set to maximize its cumulative reward according to the incentive structure in effect. As the agents' beliefs about the preferences of its peers play an important role in the recommendation process, I conduct two sets of experiments. In the first set, referred to as "*fully observable*"

mode, the agents are allowed to directly observe the product preferences of their peers, and base their recommendation decisions accordingly. The second set of experiments, *"learning preferences"*, is a more realistic setting which allows the agents to learn the preferences of their peers according to the output of prior recommendations. For each set of experiments, I simulate the diffusion of 500 products campaigns using 5 different categories, with an initial target set of 10% of the users.



(a) Adoption Rate (b) Confidence Level

Figure 6.1: Fully observable mode: Varying the conservation parameter α

6.4.1 Fully Observable Mode

Figure 6.1 shows that by decreasing the value of α, encouraging the users to be more conservative in their decisions, the rate of decline in the average confidence level between peers decreases. However, as a side effect of being more conservative, the spread of the product information over the network decreases as well, leading to lower adoption rates. For higher values of the conservation parameter α, the reward for successful recommendations is higher than the penalty, which encourages the users to send out more recommendations for their peers.

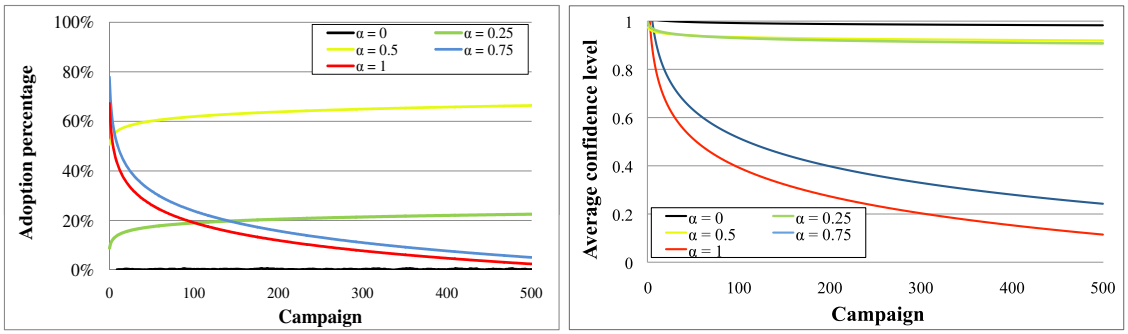
Despite the fact that this behavior leads to an initial increase in the product adoption rate, we notice that the adoption rate declines substantially in later campaigns due to the rapid decrease in confidence levels between peers.

However, we notice that utilizing intermediate values for α (e.g. $\alpha = 0.5$, corresponding to equal chances of reward and penalty) consistently maintains high adoption rates and high overall confidence even over a large number of marketing campaigns. The robustness of this result was tested by varying the number of product categories and the size of the initial seeding set. The same conclusion holds across all of these changes in the parameters of the systems.

6.4.2 Learning Preferences Mode

In real settings, users do not necessarily know the preferences of their peers in advance, but rather learn them through the peers' responses to different recommendations. To account for this more realistic situation, I give agents the ability to learn the preferences of their peers instead of directly observing them. At each time step, if the agent decides to recommend a product to one of its peers, it stores the output of this recommendation (whether or not it resulted an adoption). Then, when deciding to make a new recommendation for a similar product category in the future, the agent uses the stored outcomes to estimate that peer's preference toward different product categories.

The basic hypothesis for this experimental mode is the increase of adoption rate over the fully observable mode, due to the fact that the agents inference of



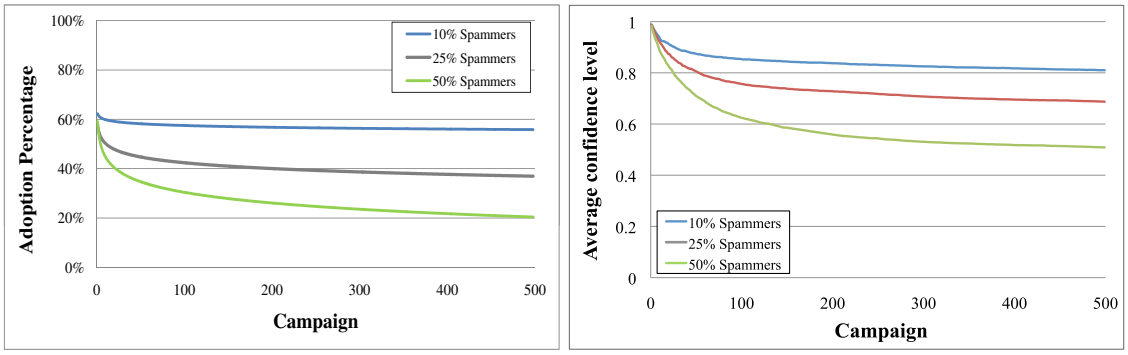
(a) Adoption Rate (b) Confidence Level

Figure 6.2: Learning preferences mode: Varying the conservation parameter α

their peers' preferences also takes into account the confidence levels, since the peers' response to recommendations account for both factors. This additional information is not contained in the direct observation of peers' preferences and since it is the composite of confidence and preference that determines actual adoption, the agents should be able to better predict their peers' adoptions. As shown in Figure 6.2, for moderate values of α, the performance of the proposed strategy is remarkably better than low and high levels of α, in terms of both product adoption and maintaining confidence levels in the network, which indicates that encouraging agents to target a small subset of their peers is the optimal strategy. This also shows that the adaptive rewards mechanism may work even better in contexts when individuals do not have perfect knowledge of their peers' preference but must infer them from observing past behavior.

6.4.3 Effect of Spammers

In order to test the robustness of the adaptive viral marketing model, I carried out another experiment where a set of spammers are manually inserted into



(a) Adoption Rate (b) Confidence Level

Figure 6.3: Varying the percentage of spammers at ($\alpha = 0.5$)

the network. A spammer is an agent that forwards recommendations for any product it adopts to all its peers, regardless of their preferences. I set ($\alpha = 0.5$) for the rest of the users, and examined the effect of various percentages of seeded spammers.

As illustrated in Figure 6.3, the agents in the network are able to identify the spamming agents after a relatively small number of campaigns, dropping their confidence in them. The effect of spamming behavior is obvious in this figure through the decreased adoption rate as the percentage of spammers present in the network is increased. However, the collective behavior of the non-spammer agents maintains the confidence level among trusted peers, while removing any confidence in spammers, which minimizes the effect of the spamming behavior on the adoption rates over time.

6.5 Conclusion

By analyzing the implications of the differential adaptive diffusion model on existing viral marketing strategies, I illustrated that most existing strategies focus on maximizing the product spread within each campaign, but fail to account for the long-term effects that spamming behavior can have on the underlying social network across campaigns. I introduced a new viral marketing strategy based on an adaptive incentive structure that accounts for the social network dynamics across different product campaigns. The experiments show that the proposed adaptive viral marketing strategy is able to account for the changes in peers' confidence across multiple campaigns, maintaining higher levels of product adoptions than those attained by classic strategies in the long term. I also showed that the proposed adaptive strategy is more robust to the existence of spammers in the network.

Chapter 7

Active Surveying: A Probabilistic Approach for Identifying Key Opinion Leaders

Opinion leaders play an important role in influencing people's beliefs, actions and behaviors. Although a number of methods have been proposed for identifying influentials using secondary sources of information, the use of primary sources, such as surveys, is still favored in many domains. In this chapter, I present a new surveying method which combines secondary data from different observable network modes, with partial knowledge from primary sources to guide the information gathering process. I apply the proposed active surveying method to the problem of identifying key opinion leaders in the medical field, and show how active surveying is able to accurately identify the opinion leaders while minimizing the amount of primary data required, which results in significant cost reduction in data acquisition without sacrificing its integrity.

7.1 Introduction

Studying influence in social networks is an important topic that has attracted the attention of a variety of researchers in different domains [87, 55]. People often seek the opinion and advice of their peers regarding various decisions, whether it is to try a new restaurant, buy a certain product or even to support a

particular politician [53]. This behavior gives rise to a certain set of individuals in the social network, referred to as *influentials* or *opinion leaders*, who have a huge impact on other people's opinions, actions and behavior.

In the commercial space, the question of how to identify opinion leaders within a given population of purchasers or decision makers is of great importance [77, 59]. Identifying these individuals properly leads to more effective and efficient sales and marketing initiatives [114]. This is true in multiple industries; here I begin my exploration in the medical domain, studying the influence networks of local physicians relative to the treatment of specific disease states. Key opinion leader identification has been the focus of multiple studies in the health care literature [103, 23].

Secondary data describing suggested influence is often easy to obtain; whereas primary data, representing surveys that measure trust and advice-seeking, is harder and much more expensive to acquire. For instance, citations are often used as an indirect indicator of influence in an academic settings, where influential authors' publications tend to receive higher citations than average. Obtaining a citation network between a set of authors in a certain field (e.g. infectious disease) can be easily constructed by looking at the publication record of each author. However, measuring the influence of each author directly requires more work, and often involves a labor-intensive process of interviewing subjects and extracting their "network of influence", e.g., who they turn to for advice and recommendations.

Methods for identifying opinion leaders can be classified into two categories

according to the type of data they use for drawing their conclusions. Primary methods rely on manually collecting information about peer-influence in a given population from the individuals themselves. One of the most commonly used primary methods is surveys, where the respondents are asked to report their opinion about who they perceive as opinion leaders. Although primary methods are considered to be the most informative about actual peer-influence, their main drawback is the high associated costs due to the time-intensive nature of the process: in many cases surveys are obtained through one-on-one interviews with the respondents, sometimes over the phone, but often in person.

On the other hand, secondary methods rely mainly on using an underlying interaction network as a "*proxy*" for influence, thus avoiding the manual aspect of primary methods. One of the most widely used techniques in this setting is relying on network centrality measures of these secondary networks (e.g., citation, co-authorship, etc.) to identify the opinion leaders. However, the major drawback of these methods is the fact that the correlation between peer-influence in the actual social network and the interactions occurring in the proxy networks cannot be verified. In a recent study on public opinion formation [117], the authors showed through a series of experiments that the customers who are critical in accelerating the speed of diffusion need not be the most connected in their corresponding social network.

In this work, I show how to combine the use of primary and secondary methods for leadership identification in the medical domain. I use primary data describing a physician nomination network in which physicians are surveyed to

nominate other physicians whom they turn to for professional advice. I augment this network with secondary data describing publication history (citation and co-authorship), as well as hospital affiliation information. I use ideas from the active learning literature to build a model that can use partial knowledge of primary data, together with secondary data, to guide the survey process. By targeting the most informative physicians for additional primary data collection, I minimize the amount of primary data needed for accurate leadership identification. As this type of primary data collection requires significant investment, this technique empowers organizations to tackle the task of accurate leadership identification in a much more cost effective and efficient manner.

The rest of the chapter is organized as follows. Section 7.2 provides a brief overview of the related work and background for both opinion leader identification and active learning. In Section 7.3, I give a detailed description of the problem and an outline of the proposed method. Section 7.3 describes the details of the active surveying algorithm. Section 7.5 discusses the experimental settings, the dataset and the results of using the proposed method compared to different baselines. Finally, Section 7.6 concludes my work and proposes future directions.

7.2 Background

7.2.1 Opinion Leader Identification

In the diffusion of innovation literature, there are two main methods for identifying opinion leaders from primary sources: self-designation and surveys

[88]. In the self-designation method, respondents are asked to report to what extent they perceive themselves to be influential. However, as can be expected, such methods are usually biased and often reflect self-confidence rather than actual influence. On the other hand, surveys are based on having selected individuals, referred to as respondents, report who they perceive as opinion leaders in a given domain [22]. Peer-identified opinion leaders are believed to be better sources of true influence compared to self-identified ones.

Due to the high costs associated with primary methods for leadership identification, there has also been a great deal of attention to methods that make use of secondary data sources. These methods rely mainly on using different structural measures for determining the importance of nodes in a proxy interaction network. In the sociology literature, various centrality measures [115] have been used to determine the most important individuals in a given social network. Among the most commonly used measures are degree centrality, indicating the most connected individuals in the network, and betweenness centrality, distinguishing the "brokers" in the network.

7.2.2 Active Learning

In this work, I build on ideas from the field of active learning, where the learner is able to acquire labels of additional examples to construct an accurate classifier or ranker while minimizing the number of labeled examples acquired. This is achieved by providing an intelligent, adaptive querying technique for ob-

taining new labels to attain a certain level of accuracy with minimal training instances. A generic algorithm for active learning is described in [90], where a learner is applied to an initial sample L of labeled examples, then each example in the remaining unlabeled pool is assigned an "*effectiveness score*," based on which the subsequent set of examples to be labeled is chosen until some predefined condition is met. The main difference between various active learning methods is how the effectiveness score of each example is computed; the score usually corresponds to the expected utility that the newly acquired example can add to the learning process.

One widely used method for active learning is uncertainty sampling [66], where the learner chooses the most uncertain data point to query, given the current model and parameters. Measuring the uncertainty depends on the underlying model used, but it usually translates to how close the data point is to the decision boundary. For instance, if a probabilistic classifier is used, the posterior probability can be used directly to guide the selection process. By acquiring the labels for the data points closer to the decision boundary, the model can be improved by better defining the existing margin. A variety of active learning methods have been proposed [94], with various ways to reduce the generalization error of the underlying model during learning. Active learning has proved to be useful in settings where acquiring labeled data is expensive. It has been applied successfully in numerous domains, such as image processing [112], speech recognition [113], and information extraction [111].

7.3 Problem Description

The problem can be formulated as determining the minimal set of respondents needed to correctly identify at least $k\%$ of the set of opinion leaders present in a given population. In order to achieve this goal, we need a method that can guide the surveying process for selecting the next respondent, such that the expected set of identified opinion leaders is maximized at each step. I apply a simple threshold model on the survey responses to identify opinion leaders; if a candidate receives more than α nominations, she is considered an opinion leader.

A key difference between this problem setting and the traditional active learning setting is that the acquisition of a survey response is more complex than that of a single label. A survey response is a structured object that includes a *set* of nominations $\{nominate(v, u) : u \in population\}$ made by a given respondent v; all of which should be accounted for in both the learning and inference phases. In some cases there may be weights associated with each nomination; although here I am assuming uniform weights, it is straightforward to extend the model to cases where weights vary.

I propose an active surveying approach that combines partial knowledge from primary sources along with secondary information to provide a dynamic framework for intelligently gathering additional primary data for opinion leader identification. In my approach, the next survey respondent is chosen to maximize the likelihood of identifying new opinion leaders. After the proposed respondent is surveyed, the survey results are incorporated back into the model to update

future predictions.

First, we need to define the conditions upon which the next respondent should be selected in order to maximize the set of identified opinion leaders. Suppose we are given an initial set of survey responses, and a threshold α that determines the minimum number of nominations an individual should obtain to be declared an opinion leader. Let the set of nominations received by a given nominee u be denoted as $nominations(u) = \{v : nominate(v, u) \wedge v \in respondents\}$. From the initial set of responses, we can generate the following two sets of individuals:

$$\begin{aligned} leaders = \quad & \{l : \quad |nominations(l)| \geq \alpha\} \\ candidates = \quad & \{c : \quad 0 < |nominations(c)| < \alpha\} \end{aligned}$$

where the *leaders* set represents the individuals who have received at least α nominations and are already identified as opinion leaders, while *candidates* is the set of individuals who have been nominated by at least one person, but have not yet received enough nominations to be declared opinion leaders. Figure 7.1 shows a toy example of how the *candidates* and *leaders* sets are generated.

Ideally, the best respondent to survey should be more likely to nominate new leaders, either from the ones already in the *candidates* set or introduce new individuals to expand it. In survey settings, there's typically a bound on the number of opinion leaders each respondent can nominate. Thus, I add a requirement that the respondent is also less likely to nominate individuals in the already

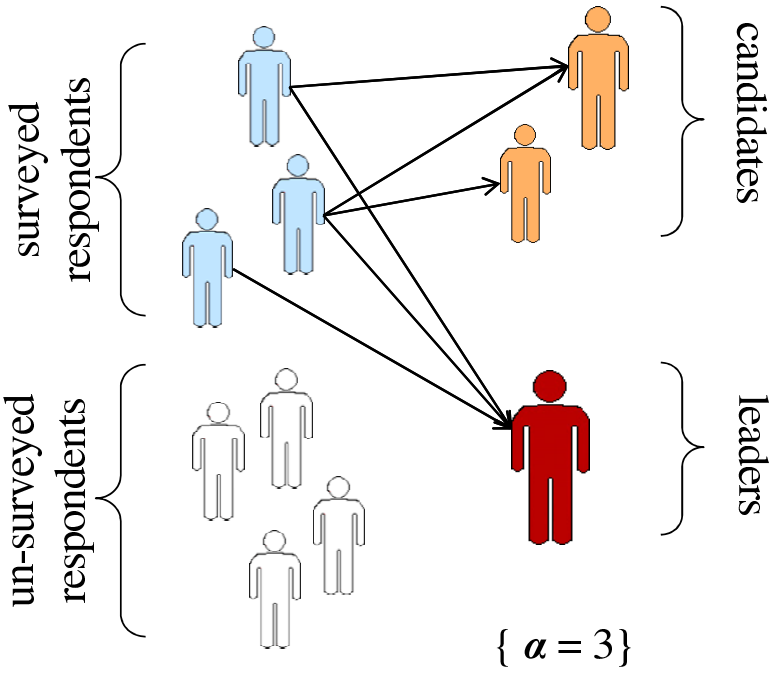


Figure 7.1: Example *candidates* and *leaders* sets

identified *leaders* set, in order to minimize the "non-informative" nominations to already identified opinion leaders. In order to estimate the likely nominations of a given respondent, I model the expected survey responses based on existing secondary sources, along with primary information from the current available surveys. By using this model to predict the nominations of the yet-to-be-surveyed respondents, we can then follow a greedy approach based on the above criterion to pick the respondent who is likely to expand the set of identified opinion leaders at each step.

The set of possible nominations in a given population can be viewed as a directed graph $G(V, E)$, where each node $v \in V$ in the network corresponds to an individual in the population, and a directed edge $e(u, v) \in E$ indicates that v is a possible nominee for respondent u. Generally, the set of potential edges

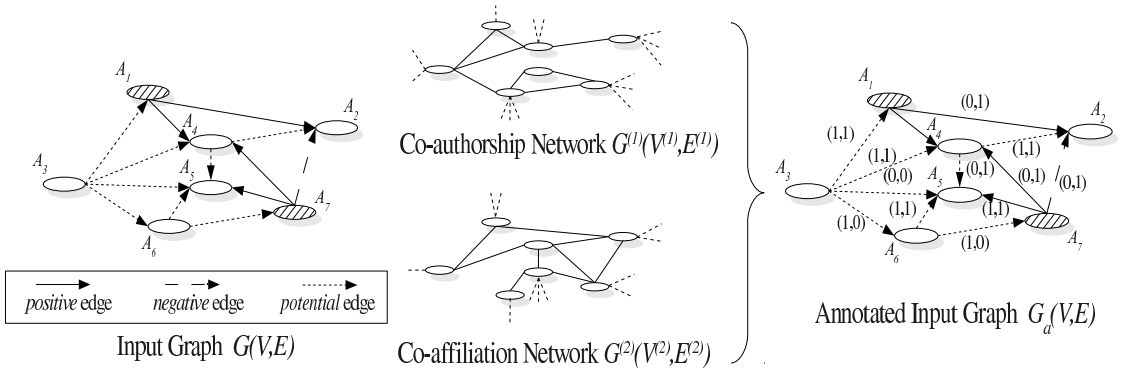


Figure 7.2: Feature generation for an example author network

in the network can be as large as $|V| \times |V|$, yielding a fully connected graph. However, in real scenarios, the number of potential edges can often be limited by using appropriate filters on the incident nodes, such as evidence from secondary sources, local proximity, similarity, or any other constraint imposed by the problem. I refer to the subset of potential edges that correspond to actual respondent nominations as "positive" edges, and the ones that are not realized through the survey as "negative" edges. I refer to the set of edges corresponding to the initial set of surveys as the "observed" edges.

The secondary sources of information are represented in the model as: a) a set of features $\mathcal{F}_v$ associated with the nodes V in G, and b) a set of secondary networks $G^{(1)}(V^{(1)}, E^{(1)}) \ldots G^{(n)}(V^{(n)}, E^{(n)})$ representing other types of interactions between the set of individuals in the target population (e.g. communication, co-authorship, co-affiliation, etc.). As these secondary networks may not necessarily align with the main graph G, I only consider the sub-networks comprising the nodes that overlap with the network of concern. Another set of edge features $\mathcal{F}_e$ is generated for the set of edges E in G, each representing a vector of the cor-

responding edge weights in each of the associated secondary networks. During this step, the set of node features $\mathcal{F}_v$ are also enriched by additional features from the secondary networks.

In Figure 7.2, the input graph G represents a partially observed author nomination network, where the shaded authors A_1 and A_7 are the ones who have already been surveyed. In this example, all of the potential nominations for author A_1 were realized (positive, denoted by solid edges), while for author A_7, although the nomination for A_2 was a potential edge, it was not realized (negative, denoted by a dashed outgoing edge). Each author in the primary nomination network G has a set of associated features, such as the geographical location, h-index, current academic position, etc. These features constitute the set of node features $\mathcal{F}_v$ in the model. In addition to the nomination network, we have two secondary sources of information in the example: a co-authorship network $G^{(1)}$ and a co-affiliation network $G^{(2)}$.

After aligning the secondary networks with the primary nomination network, the edge features generated are indicators of the edge existence in the corresponding secondary network. For instance, the edge in G corresponding to author A_1 nominating author A_2 does not have corresponding coauthorship evidence in network $G^{(1)}$, but the two authors do share the same affiliation as indicated in network $G^{(2)}$. Thus, the resulting feature vector for edge $e(A_1, A_2)$ in this simple example would be $\mathcal{F}_{e(A_1,A_2)} = (0, 1)$, as shown on the resulting annotated input graph G_a in Figure 7.2. In addition to the generated edge features, extra node features are derived from these secondary networks, such as the number

of publications from the co-authorship network, or the prestige of the affiliated organization from the co-affiliation network. These additional node features are then used to enrich the original set of author features obtained from the primary data.

7.4 Method

The proposed active surveying method uses a greedy probabilistic approach for solving the optimization problem. I use the current set of observed nominations as evidence for training a probabilistic classifier. The classifier is then used to infer how likely the potential nominations for each un-surveyed respondent are to be realized. Given the input graph G and the sets of node features $\mathcal{F}_v$ and edge features $\mathcal{F}_e$, a probabilistic classifier C is trained using the initial set of observed edges. For each un-observed potential edge $e(u, v) \in E$, the classifier outputs the posterior probability of that edge being positive, denoted as $p(+|e(u, v))$, or negative, denoted as $p(-|e(u, v))$.

Given the output probabilities from the classifier along with the initial sets of *leaders* and *candidates* determined by the observed edges in G, I define a score function $S(v)$ for each node $v \in V$ as:

$$S(v) = \sum_{c \in candidates} p(+|e(v, c)) - \sum_{l \in leaders} p(+|e(v, l))$$

The score function $S(u)$ represents the difference between the expected number of nominated *candidates* and the expected number of nominated *leaders* for a given

respondent u. Thus, following a greedy approach for finding the minimal set of respondents, the individual corresponding to node $v_S : \arg\max_v S(v)$ is then surveyed, and the resulting nominations are added to the training set and incorporated back into the model. Although there is an underlying independence assumption in predicting the respondents' choices of influential peers, I show in the experimental section that this approximation works well in practice.

One caveat with the above approach is the dependence between the quality of the decision of who to survey next with the accuracy of the underlying classifier. Therefore, a competing requirement is to choose respondents based on a criterion that will enhance the overall accuracy of the classifier. I rely on active learning to provide the necessary criterion for choosing the most informative respondents from the perspective of enhancing the overall accuracy of the underlying classifier.

In order to reduce the class probability estimation error, I use uncertainty sampling to select the respondents with the most uncertain responses, measured as the expected conditional classification error over their corresponding potential nominations. To choose the next respondent to survey, each nomination of a given respondent v is assigned a weight
$w(e(v,u)) = (0.5 - |0.5 - p(+|e(v,u))|)$ indicating the distance of the class probability estimate from 0.5, which is used to quantify the amount of uncertainty in the class prediction. Then, the weight of each respondent $W(v)$ is computed as the average of all the weights on her outgoing nominations. The respondents' weights are then used to make a probabilistic choice of the next respondent. This

weighted uncertainty sampling approach (WUS) has been shown to outperform traditional methods that pick the most uncertain sample [90].

To provide a robust mechanism, I incorporate the two objectives of maximizing the likelihood to identify a new opinion leader and minimizing the expected classification error for choosing the next respondent. For that, I quantify the amount of uncertainty in the classifier output over all un-observed edges E_u as:

$$H_{avg} = \frac{1}{H_{max} \times |E_u|} \sum_{e(u,v) \in E_u} H(e(u,v))$$

where the entropy of the classifier output with respect to a given edge $e(u,v)$ is defined as:

$$H(e(u,v)) = - \sum_{l \in \{+,-\}} p(l|e(u,v)) log(p(l|e(u,v)))$$

and H_{max} is a normalization factor, representing the maximum entropy of the classifier output, so that H_{avg} is a valid probability value between $[0,1]$

The next respondent to be surveyed v^* is then chosen via a probabilistic decision based on the current accuracy of the underlying classifier as follows:

$$v^* = \begin{cases} v \sim WUS & \text{with probability } p = H_{avg} \\ \arg\max_v S(v) & \text{with probability } p = (1 - H_{avg}) \end{cases}$$

Thus, the probability of choosing a respondent based on uncertainty sampling to enhance the classifier accuracy increases with higher uncertainty in the classifier output, while being more confident in the predictions yields a higher probability

of choosing a respondent that optimizes the objective function $S(v)$. The full details are presented in Algorithm 7.1.

Algorithm 7.1 Active Survey Algorithm

repeat
  Train classifier C using observed nominations
  for each un-surveyed respondent v **do**
    Compute the objective function $S(v)$
    Compute the weight $W(v)$ using uncertainty sampling
  end for
  Normalize uncertainty sampling weights $W(v)$
  Set $v_S \leftarrow \arg\max_v S(v)$
  Set $v_{WUS} \sim W(v)$
  With probability $p = H_{avg}$, set $v^* \leftarrow v_{WUS}$, otherwise set $v^* \leftarrow v_S$
  Survey respondent v^*, update *leaders* and *candidates* sets according to the resulting nominations
  Remove v^* from the un-surveyed respondents and add her survey results to the set of observed nominations.
until required number of opinion leaders is obtained

7.5 Experimental Evaluation

To test the proposed method, I use a health care dataset generously provided by Community Analytics, a social marketing research organization which specializes in analyzing influence networks and identifying opinion leaders through conducting surveys of their clients' target audiences. The data represents survey information for nominating influential local physicians, provided by their peers.

The dataset consists of 2004 physicians, with 899 actual survey respondents generating 1598 nominations. As the surveys are based on identifying locally influential physicians, I limit the potential edges for each respondent to the physicians whose locations are within a 150 mile radius, yielding a set of 127,420 po-

tential edges. By setting the nomination threshold (α = 2), 260 opinion leaders could be identified.

By using the physicians' lists of publications from PubMed[1], I constructed both a citation and a co-authorship network among the physicians in the primary network. I also used the physicians' affiliation information to construct a co-affiliation network as a third supplementary source to leverage the data. Finally, using these three secondary networks, I generated a set of 20 edge features on the primary physician network and enriched the node features with additional attributes from these networks. A sample of the features included in the augmented network as the input to the model are illustrated in Table 7.1.

Feature Name	**Source Network**
-Geographical Distance	$G_{nomination}$
-Respondent's current position (academic/non-academic)	$G_{nomination}$
-Nominee's current position	$G_{nomination}$
-Number of co-authored publications	$G_{co\text{-}authorship}$
-Nominee's publications count	$G_{co\text{-}authorship}$
-Number of respondent's citations of the nominee's publications	$G_{citation}$
-Nominee's h-index	$G_{citation}$
-Number of common affiliations	$G_{co\text{-}affiliation}$

Table 7.1: Sample features in the annotated physician nomination network

To conduct the experiments, I use a logistic regression classifier and vary the target percentage k of opinion leaders to be identified, showing the corresponding percentage of respondents required to reach this target using the proposed active survey method. I compare active surveying with a random baseline and

[1]http://www.ncbi.nlm.nih.gov/pubmed

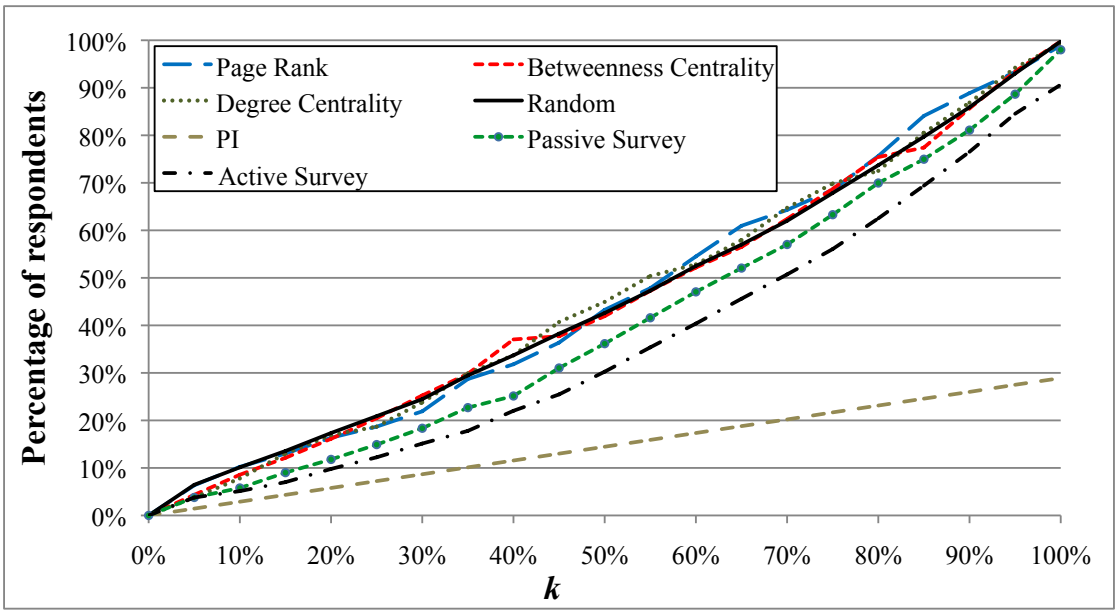


Figure 7.3: The percentage of respondents (y-axis) needed to identify $k\%$ of the opinion leaders (x-axis) at ($\alpha = 2$)

a set of other baselines based on various centrality measures for determining the most informative physicians. I use three widely used centrality measures for the structural baselines: degree centrality, betweenness centrality, and page rank. In order to understand the relative contribution of the classifier versus active learning, I compare the proposed approach to a "*passive*" surveying method, which follows the same procedure of active surveying for optimizing the score function $S(v)$ based solely on the classifier's output, but does not incorporate uncertainty sampling. Finally, I show the performance of a method referred to as perfect information (*PI*). *PI* uses the fully observed network and, at each step, greedily selects the survey respondent which identifies the maximum number of new opinion leaders. Note that the *PI* method represents a pseudo-optimal solution at each point, and hence the lower bound for the number of required respondents

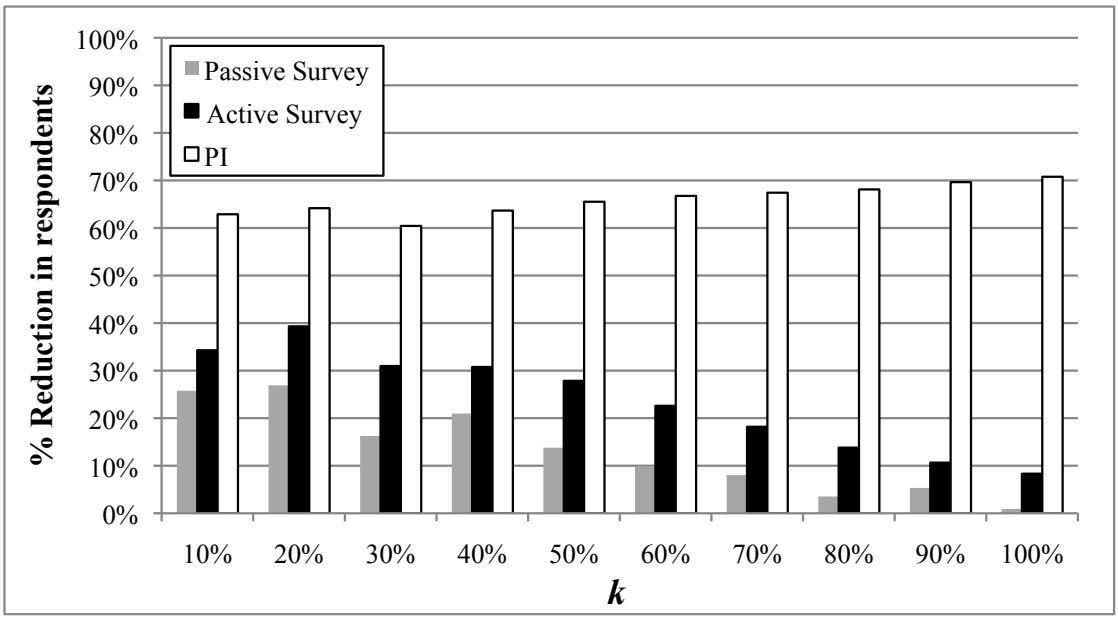


Figure 7.4: The percentage reduction in required respondents to identify $k\%$ of the opinion leaders at ($\alpha = 2$)

at each step.

As can be seen in Figure 7.3, while the performance of the centrality-based methods is indistinguishable from the random baseline, both the passive and the active surveying methods perform significantly better than the baselines. Furthermore, the proposed active surveying method outperforms passive surveying, showing that an intelligent acquisition strategy helps to improve the quality of the learned classifier. Figure 7.4 shows the actual percentage of reduction in the size of the respondent set of both the active survey method and the perfect information method, with respect to the minimum set obtained by the best performing baselines at the corresponding value of k. As can be noted from the figure, the proposed approach yields a 30% average reduction in the number of respondents required, as compared to a 19% average reduction by the passive

approach. The reduction attained by the active surveying method is reflected directly in surveying costs, thus helping survey conductors achieve their required goal at minimum cost. For instance, if a survey costs $500 per person, then in order to identify 50% of the opinion leaders in the used physician network, the active survey method needs only 270 surveys rather than 375 surveys required by the best performing baseline; this leads to a savings of $52,500.

7.6 Conclusion

In this chapter, I presented a novel, dynamic framework for prioritizing the acquisition of survey information, for the purpose of leadership identification. The approach enables intelligent integration of both primary and secondary data to identify which respondents to survey, based on both the likelihood of them expanding the set of identified opinion leaders and also the utility of the information for improving future predictions. I then validated the results on a real-world dataset describing a physician nomination network.

Although the algorithm is focused on opinion leadership identification, I believe the idea of exploration vs. exploitation behind active surveying can generally be applied in different settings for guiding the survey process to reduce the associated costs.

Chapter 8

Conclusion and Future Directions

In my dissertation, I have taken some initial steps towards reasoning in dynamic, multi-modal, multi-relational networks. I have shown the importance of considering the complex structure of these networks over time as opposed to limiting the analysis to a single entity or relationship type in different network analysis tasks. In this chapter, I first summarize the contributions of this dissertation, then I discuss the potential directions for future work and conclude.

8.1 Summary of Contributions

My dissertation focuses on analyzing and modeling different aspects associated with the dynamics of multi-modal, multi-relational networks, such as modeling network evolution, finding cohesive clusters in multi-relational domains, analyzing the dynamics of interactions among different entity types at both micro and macro-levels, and investigating the effect of these interactions on information diffusion and opinion leader identification.

8.1.1 Network Evolution

I started my research by analyzing how multi-modal networks evolve over time. To investigate this problem, I characterized the growth patterns of different

network modalities by analyzing the existing dependencies between the evolution of users and social groups in social and affiliation networks. Based on my findings, and other social network properties that were previously described in the literature, I proposed the first generative model that captures the statistical properties of multi-modal networks [122]. The proposed co-evolution model was able to mimic the evolution of real multi-modal networks, bringing new insights about the role of friendship in joining social groups. I believe that these insights not only affect the design of network evolution models but may have broader implications on mechanism design for tasks such as group recommendation, information diffusion and viral marketing strategies.

8.1.2 Multi-relational Clustering

Relational clustering algorithms try to bridge the gap between traditional, similarity-based data clustering and community detection algorithms for network data, by accounting for both feature and structure similarities. In multi-modal networks, there are more complex dependencies among different entity types that should be also taken into consideration to provide more meaningful clusterings. To address this aspect, I proposed a multi-relational affinity propagation clustering algorithm [96] to facilitates an exploration of a middle ground between feature-based similarity clustering, community detection, and block modeling in multi-relational networks. The proposed clustering approach extends the affinity propagation clustering algorithm to multi-relational domains by encod-

ing the dependencies across different node types in the form of soft structural constraints, with parametrized control over how they influence the final clustering. I showed empirically that the proposed algorithm outperforms previous approaches for multi-relational clustering.

8.1.3 Bi-modal Interaction Dynamics

In order to characterize the temporal interactions occurring in multi-modal networks, I started by investigating the user interactions with event-based groups in bi-modal networks from multiple domains: a political network of senator-bill sponsorship history, a publication network, and a dolphin observation network. By analyzing the participation patterns of users in groups over time, I was able to identify a set of factors that characterize the stability of these interactions: 1) *frequency*: stable relationships tend to have a larger number of associated interactions, 2) *recency*: up-to-date interactions indicate the liveliness of the corresponding relationship, and 3) *consistency*: stable relationships are usually associated with a consistent pattern of interactions among the corresponding entities over time. Taking these factors into consideration, I proposed a measure for quantifying users' loyalty to different groups as the degree of commitment a user shows to the group over time [99]. I showed that the proposed measure provides new insights about the temporal aspects of user-group interactions in dynamic affiliation networks that are not captured by existing centrality measures.

8.1.4 Information Diffusion

The next aspect I investigated was how information diffuses in multi-modal networks. To answer this question, I focused my analysis on the factors involved in product adoption in multi-modal networks, as opposed to relying only on the inferred influence across friendship links. I started by analyzing the shortcomings of existing information diffusion, such as 1) considering the diffusion of a single topic or type of information, neglecting the wealth of information in the existing user-topic preferences, and 2) the inherent assumption that the underlying social network among users is static, and thus failing to model the evolution of the aspects of individual relationships, such as trust, during the course of subsequent diffusion processes. To address these issues, I proposed an adaptive diffusion model that exploits both the dynamic social relationships among users, as well as their preferences for different topics or product types. Specifically, the proposed model was able to capture: 1) information across multiple diffusion processes (e.g. different product marketing campaigns), 2) the diversity in user preferences for different topics / product categories, and 3) the variation of confidence in peers recommendations over time [98]. The empirical evaluation of my proposed differential adaptive diffusion model on real-world network data showed that it outperforms earlier non-adaptive models in predicting future product adoptions. I also showed how the model can be utilized to identify influential users in the network for initial targeting.

8.1.5 Adaptive Viral Marketing

Next, I moved to discussing the implications of the proposed diffusion model on viral marketing strategies. I showed that existing techniques do not address the trust dynamics among users in the underlying social network, and thus might adversely affect the diffusion process over time. To address that, I proposed a novel adaptive viral marketing strategy that aligns the immediate utility of the users with the end goal of maintaining higher adoption levels across multiple campaigns. I showed that this objective can be achieved by redesigning the incentive structure of the viral marketing strategy. I validated the proposed adaptive viral marketing technique through a number of simulations over an agent-based model, and showed that the proposed strategy outperformed the classical ones in providing high adoption rates over time.

8.1.6 Active Surveying

In cases where historical interactions are unavailable, there's a need for alternate methods for estimating influence and identifying opinion leaders. To address this problem, I investigated how to leverage different network modalities in the process of identifying opinion leaders through primary sources, such as surveys. I proposed a new *active surveying* method for combining secondary data from different observable network modalities with partial knowledge from primary sources, to guide the information gathering process. I applied the proposed *active surveying* method to the problem of identifying key opinion leaders in the

medical domains [97], and showed that it could accurately identify the opinion leaders while minimizing the amount of primary data required. This results in significant cost reduction in data acquisition without sacrificing its integrity.

8.2 Future Directions

There are a number of promising future directions for the research on dynamic, multi-modal networks. Some of these potential avenues include:

- Combining structured information from networks with related unstructured data that is often under utilized in network models, such as text corpus, videos, music, etc. Incorporating these types of data into relational learning models offers a wealth of information that can be used in developing a holistic view of the associated network.

- Utilizing multi-modal network models in different application-oriented tasks from different domains, such as behavioral models, financial markets, recommender systems, and economic models.

- Scaling up network models to be able to handle large-scale problems. One potential approach for tackling the problem at the semantic level, is through analyzing the networks at successive levels of abstractions, as well as exploring semi-supervised techniques for network analysis.

8.3 Conclusion

With this thesis, I hope to motivate further research in developing new, scalable machine learning models that are capable of leveraging the temporal, multi-dimensional dependencies in complex networks. My work suggests that by incorporating additional network modalities, we can build rich models that provide a better understanding of the dynamics in complex networks, and thus help in providing highly accurate predictions about future events. I believe complex networks research constitutes an important pillar towards understanding and modeling how the local behavior of entities can affect the network on the global scale. By providing a holistic view of the corresponding network dynamics, this evolving line of research has the potential for considerably changing the systems and applications design for networks from different domains, which will in turn increase the value of the provided services for the users in these networks. It will also provide new means for researchers to uncover the behavioral patterns of the users, which enables accurate modeling of users activities and understanding the causal processes governing different social processes.

Bibliography

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[2] R. Albert, H. Jeong, and A. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378 – 382, 2000.

[3] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2007.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2006.

[5] A. Banerjee, S. Basu, and S. Merugu. Multi-way clustering on relation graphs. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2007.

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

[7] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 285(5439):509 – 512, 1999.

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

[9] H. S. Becker. Notes on the concept of commitment. *Journal of Sociology*, 66:32–42, 1960.

[10] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.

[11] T. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2006.

[12] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, October 2005.

[14] D. Centola and M. Macy. Complex contagion and the weakness of long ties. *American Journal of Sociology*, 113:702–734, 2005.

[15] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys (CSUR)*, 38(1), 2006.

[16] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357:370 – 379, 2007.

[17] R. M. Cunningham. Brand loyalty - what, where, how much? *Harvard Business Review*, 34:116–128, 1956.

[18] D. L. Davies and D. W. Bouldin. A clustering separation measure. *IEEE PAMI Trans.*, 1(2):224–227, 1979.

[19] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232:587 – 604, 2005.

[20] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80 – 82, 2005.

[21] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2001.

[22] S. Dorfman and J. Maynor. Under the influence. *Pharmaceutical Executive*, 26:148 – 150, 2006.

[23] G. Doumit, M. Gattellari, J. Grimshaw, and M. A. O'Brien. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database of Systematic Revs., 2007.

[24] M. Drehmann, J. Oechssler, and A. Roider. Herding and contrarian behavior in financial markets : An internet experiment. *American Economic Review*, 95(5):1403 – 1426, 2005.

[25] D. Druckman. Nationalism, patriotism, and group loyalty: A social psychological perspective. *Mershon International Studies Review*, 38(1):43–68, 1994.

[26] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[27] D. C. Dunphy. The social structure of urban adolescent peer groups. *Sociometry*, 26:230–246, 1963.

[28] M.G. Everett and S.P. Borgatti. Analyzing clique overlap connections. *Connections*, 21(1):49–61, 1998.

[29] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, September 1999.

[30] G. Flake, S. Lawrence, C. Giles, and F. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–71, 2002.

[31] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[32] L. Friedland and D. Jensen. Finding tribes: identifying close-knit individuals from employment patterns. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2007.

[33] A. Galstyan and P. Cohen. Influence propagation in modular networks. In *Proceedings of the AAAI Symposium on Social Information Processing*, 2008.

[34] L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3 – 12, December 2005.

[35] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

[36] R. Ghosh and K. Lerman. Leaders and negotiators: An influence-based metric for rank. In *Proceedings of the third International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[37] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[38] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the national academy of science (PNAS)*, 99(12):7821–7826, 2002.

[39] I. E. Givoni and B. J. Frey. A binary variable model for affinity propagation. *Neural Computation*, 21(6):1589–1600, 2009.

[40] J. Golbeck, editor. *Computing with Social Trust*. Springer, 2009.

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

[42] A. Goval, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)*, February 2010.

[43] Govtrack. Senate bill sponsorship data. website: http://www.govtrack.us, 2008.

[44] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420 – 1443, 1978.

[45] Habiba, T. Y. Berger-Wolf, Y. Yu, and J. Saia. Finding spread blockers in dynamic networks. In *Proceedings of the 2nd SNA-KDD workshop on Web mining and social network analysis (SNA-KDD)*, August 2008.

[46] S. Hill, D. Agarwal, R. Bell, and C. Volinsky. Building an effective representation of dynamic networks. *Journal of Computational and Graphical Statistics*, 25:584–608, 2006.

[47] P. Holme and M. E. J. Newman. Non-equilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74, 2006.

[48] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2003.

[49] J. Jacoby and D. B. Kyner. Brand loyalty vs. repeat purchasing behavior. *Journal of Marketing Research*, 10(1):1–9, 1973.

[50] S. Jegelka, S. Sra, and A. Banerjee. Approximation algorithms for tensor clustering. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 2009.

[51] S. Jurvetson. What exactly is viral marketing? *Red Herring*, pages 110 – 111, 2000.

[52] R. M. Kanter. Commitment and social organization: A study of commitment mechanisms in utopian communities. *American Sociological Review*, 33:499–517, 1968.

[53] E. Keller and J. Berry. *One American in ten tells the other nine how to vote, where to eat, and what to buy. They are the influentials*. Free Press, 2003.

[54] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, 2006.

[55] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence in a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2003.

[56] C. A. Kiesler. *The psychology of commitment: Experiments linking behavior to belief*. Academic Press, 1971.

[57] M. Kirsten and S. Wrobel. Relational distance-based clustering. In *Proceedings of the 8th International Conference on Inductive Logic Programming (ILP)*, 1998.

[58] S. Kok and P. Domingos. Statistical predicate invention. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007.

[59] D. Krackhardt. Structural leverage in marketing. *Networks in Marketing*, pages 50 – 59, 1996.

[60] K. Lerman. Social networks and social information filtering on digg. In *Proceedings of the first International Conference on Weblogs and Social Media (ICWSM)*, March 2007.

[61] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1, 2007.

[62] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2008.

[63] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. *ACM Transactions on Knowledge Discovery from Data*, 1(1):177–187, 2007.

[64] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web (WWW)*, April 2008.

[65] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, April 2006.

[66] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, July 1994.

[67] B. Long, Z. Zhang, X. Wú, and P. S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[68] B. Long, Z. Zhang, and P. S. Yu. A probabilistic framework for relational clustering. In *Proceedings of the thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[69] V. Mahajan, E. Muller, and F.M. Bass. New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing*, 54(1):1–26, 1990.

[70] J. Mann, R. C. Connor, L. M. Barre, and M. R. Heithaus. Female reproductive success in bottlenose dolphins (/tursiops/ sp.): Life history, habitat, provisioning, and group size effects. *Behavioral Ecology*, 11:210–219, 2000.

[71] J. Mann and B. Sargeant. *Like mother, like calf: The ontogeny of foraging traditions in wild Indian Ocean bottlenose dolphins*, pages 236–266. Cambridge University Press, 2003.

[72] J. Mann, B.L. Sargeant, J.J. Watson-Capps, Q.A. Gibson, M.R. Heithaus, R.C. Connor, and E. Patterson. Why do dolphins carry sponges. *PLoS One*, 3(12):e3868, 2008.

[73] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2008.

[74] J. P. Meyer, T. E. Becker, and C. Vandenberghe. Employee commitment and motivation: A conceptual analysis and integrative model. *Journal of Applied Psychology*, 89(6):991–1007, 2004.

[75] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC)*, October 2007.

[76] J. Moody and D. R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127, 2003.

[77] J. Myers and T. Robertson. Dimensions of opinion leadership. *Journal of Marketing Research*, 9:41 – 46, 1972.

[78] J. Neville, M. Adler, and D. Jensen. Clustering relational data using attribute and link information. In *Proceedings of the Text Mining and Link Analysis Workshop*, 2003.

[79] J. W. Newman and R. A. Werbel. Multivariate analysis of brand loyalty for major household appliances. *Journal of Marketing Research*, 10:404–409, 1973.

[80] M. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[81] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321–330, 2004.

[82] M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8696, 2006.

[83] R. L. Oliver. Loyalty: Whence consumer loyalty? *Journal of Marketing*, 63:33–44, 1999.

[84] A. Plangprasopchok, K. Lerman, and L. Getoor. A probabilistic approach for learning folksonomies from structured data. In *Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM)*, 2011.

[85] F. J. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, July 1998.

[86] M. Rattigan, M. Maier, and D. Jensen. Exploiting network structure for active inference in collective classification. Technical report, University of Massachusetts Amherst, 2007.

[87] B. H. Raven. Social influence and power. *Current studies in social psychology*, pages 371 – 382, 1965.

[88] E. M. Rogers and D. G. Cartano. Methods of measuring opinion leadership. *Public Opinion Quarterly*, 26:435 – 441, 1962.

[89] B. Ryan and N.C. Gross. The diffusion of hybrid seed corn in two Iowa communities. *Rural sociology*, 8(1):15–24, 1943.

[90] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153 – 178, 2004.

[91] R. J. Sampson and W. B. Grove. Community structure and crime: testing social disorganization theory. *American Journal of Sociology*, 94:774–802., 1989.

[92] B. L. Sargeant, A. J. Wirsing, M. R. Heithaus, and J. Mann. Can environmental heterogeneity explain individual foraging variation in wild bottlenose dolphins? *Behavioral Ecology and Sociobiology*, 61:679–688, 2007.

[93] T. C. Schelling. *Micromotives and Macrobehavior*. Norton, 1978.

[94] B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin - Madison, 2009.

[95] U. Sharan and J. Neville. Temporal-relational classifiers for prediction in evolving domains. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, December 2008.

[96] H. Sharara and L. Getoor. Multi-relational affinity propagation. Active submission, 2012.

[97] H. Sharara, L. Getoor, and M. Norton. Active surveying: A probabilistic approach for identifying key opinion leaders. In *Proceedings of the 22nd International Joint Conference On Artificial Intelligence (IJCAI)*, 2011.

[98] H. Sharara, W. Rand, and L. Getoor. Differential adaptive diffusion: Understanding diversity and learning whom to trust in viral marketing. In *Proceedings of the fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[99] H. Sharara, L. Singh, L. Getoor, and J. Mann. Understanding actor loyalty to event-based groups in affiliation networks. *Journal of Advances in Social Networks Analysis and Mining*, 1(2):115–126, 2011.

[100] M. E. Shaw. Group dynamics. *Annual Review of Psychology*, 12:129–156, 1961.

[101] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE PAMI Trans.*, 22(8):888–905, 2000.

[102] T. Snijders. *Models for Longitudinal Network Data*, page Chapter 11. New York: Cambridge University Press, 2005.

[103] S. Soumerai, T. McLaughlin, J. Gurwitz, E. Guadagnoli, P. Hauptman, C. Borbas, N. Morris, B. McLaughlin, X. Gao, D. Willison, R. Asinger, and F. Gobel. Effect of local medical opinion leaders on quality of care for acute myocardial infarction: A randomized controlled trial. *The Journal of the American Medical Association*, pages 1358 – 1363, 1998.

[104] F. Stonedahl, W. Rand, and U. Wilensky. Evolving viral marketing strategies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation (GECCO)*, July 2010.

[105] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining partitionings. In *AAAI*, 2002.

[106] J. Sun, C. Faloutsos, S. Papadimitriou, and P. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Jose, CA, August 2007.

[107] I. Sutskever, R. Salakhutdinov, and J. Tenenbaum. Modelling relational data using Bayesian clustered tensor factorization. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, 2009.

[108] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Jose, CA, August 2007.

[109] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[110] G. Tellis. Advertising exposure, loyalty, and brand purchase: A two-stage model of choice. *Journal of Marketing Research*, 25:134–144, 1988.

[111] C. A. Thompson, M. E. Califf, and R. J. Mooney. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, June 1999.

[112] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the 9th ACM international conference on Multimedia*, September 2001.

[113] G. Tür, D. Hakkani-Tür, and R.E. Schapire. Combining active and semisupervised learning for spoken language understanding. *Speech Communication*, 24(2):171 – 186, 2005.

[114] T. Valente and R. Davis. Accelerating the diffusion of innovations using opinion leaders. *The ANNALS of the American Academy of Political and Social Science*, 566(1):55 – 67, 1999.

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

[116] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. New York, Cambridge University Press, 1994.

[117] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441 – 458, 2007.

[118] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[119] F. Wu and B. A. Huberman. Social structure and opinion formation. *Computational Economics*, 2004.

[120] Z. Xu, V. Tresp, K. Yu, , and H. Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[121] H. J. Zeng, Z. Chen, and W. Y. Ma. A unified framework for clustering heterogeneous web objects. In *Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE)*, 2002.

[122] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD)*, 2009.