Critical Individuals in Dynamic Population Networks

BY

HABIBA HABIBA
B.S. (Computer Science) National University of Computer & Emerging Sciences, Pakistan, 2003

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense committee:

Tanya Berger-Wolf, Chair and Advisor
Barbara Di Eugenio
Piotr Gmytrasiewicz
Lev Reyzin
Vijay Subramanian, Northwestern University

ACKNOWLEDGMENTS

This work would not have been possible without the help of my advisor Tanya Berger-Wolf who has been very supportive and encouraging throughout regardless of my slow progress in research. I am thankful to the members of my thesis committee, Barbara Di Eugenio, Piotr Gmytrasiewicz, Lev Reyzin, and Vijay Subramanian, for their helpful feedback and advice. I am also grateful to a number of professors, Gyorgy Turan, Druv Mubayi, Robert Sloan, Joel Brown, and Dan Rubinstein for their mentoring, teaching, and guidance.

I am grateful to my lab mates and friends, Andrew Ring, Anushka Anand, Arun Maiya, Chayant Tantipathananandh, Heba Basiony, Marco Maggioni, Mayank Lahiri, Paul Varkey, Priya Govindan, Rajmonda Sulo Cáceres, and Saad Sheikh, for their help and encouragement.

Most importantly, this thesis would not have been possible if not for the love and support from my mom, my siblings, and my step father. I cannot express enough gratitude for their encouragement and faith in my capabilities even if I did not make sense to them all the time. Lastly, this thesis is indebted to Nauman Ahmad Shah for the innumerable occasions of patiently listening to whatever crisis I invented in the passage of its completion . Thank you!

H.

TABLE OF CONTENTS

CHAPTER		PAGE
1	**INTRODUCTION**	1
2	**PRELIMINARIES**	7
2.1	Types of networks	7
2.1.1	Aggregate network	8
2.1.2	Dynamic network	8
2.2	Static network structure and properties	9
2.2.1	Global properties	10
2.2.1.1	Density	10
2.2.1.2	Path	10
2.2.1.3	Diameter	10
2.2.1.4	Average path length	11
2.2.2	Local properties	11
2.2.2.1	Degree	11
2.2.2.2	Betweenness	11
2.2.2.3	Closeness	12
2.2.2.4	Clustering coefficient	12
2.3	Types of diffusion models	12
2.3.1	Independent cascade model	12
2.3.1.1	Independent cascade model and dynamic networks	13
2.3.2	Linear threshold model	14
2.3.2.1	Linear threshold model and dynamic networks	15
2.3.3	Independent cascade vs Linear threshold model	15
2.4	Sub–modularity and monotonicity	16
2.4.1	Maximizing sub–modular monotonic functions	17
2.4.1.1	Approximation	17
3	**RELATED WORK**	18
3.1	Diffusion of diseases and contagions	18
3.2	Diffusion of innovations	20
3.3	Algorithmic results in diffusion optimization	21
3.4	Computational network analysis	23
4	**NETWORK DATA AND MODELS**	26
4.1	Real world networks	26
4.2	Synthetic network models	31
4.3	Dynamic network generative model	33

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

	4.3.1	Model	34
	4.3.1.1	The dynamic generative process	34
	4.3.2	Experimental results	35
	4.3.3	Conclusion	38
5	**STRUCTURAL PROPERTIES OF DYNAMIC NETWORKS**		39
	5.1	Global properties	39
	5.1.1	Dynamic Density	40
	5.1.2	Temporal Paths	40
	5.1.2.1	Geodesic	41
	5.1.2.2	Shortest Simple Temporal Path	42
	5.1.2.3	Shortest Link Path	43
	5.1.2.4	Shortest Temporal Trails	43
	5.1.2.5	Dynamic diameter	44
	5.1.2.6	Average temporal path length	44
	5.2	Local properties	44
	5.2.1	Dynamic Degree	45
	5.2.1.1	Dynamic Average Degree	46
	5.2.2	Betweenness Centrality	46
	5.2.2.1	Temporal Betweenness Centrality	46
	5.2.2.2	Delay Betweenness Centrality	47
	5.2.3	Dynamic Closeness	47
	5.2.4	Dynamic Clustering Coefficient	48
6	**DIFFUSION MAXIMIZATION**		49
	6.1	Problem statement	49
	6.2	Maximizing the extent of diffusion in a dynamic network	50
	6.2.1	Problem Statement	51
	6.2.2	Complexity	52
	6.2.3	Approximation Algorithm	52
	6.2.3.1	The Overall Approach	52
	6.2.3.2	The Independent Cascade Model	53
	6.2.3.3	The Linear Threshold Model	55
	6.2.4	Experiments	58
	6.2.4.1	Types of experiments	58
	6.2.4.2	Experiment results	61
	6.2.4.2.1	Data sets	61
	6.2.4.2.2	Independent Cascade	61
	6.2.4.2.3	Linear Threshold	65
	6.2.5	Conclusions and Future Work	67
	6.3	The Impact of Structural Changes on Predictions of Spreading in Networks	69

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

6.3.1	Preliminaries	71
6.3.1.1	Static and Dynamic Networks	71
6.3.1.2	Network Diffusion	73
6.3.2	Methodology	73
6.3.2.1	Experimental Setup	75
6.3.3	Datasets	78
6.3.4	Results	78
6.3.4.1	The change in the relative spreading ability of individuals	78
6.3.4.2	The change in the identity of the top spreaders	81
6.3.4.3	The relative performance of the historical spreaders	82
6.3.5	Conclusions and Future Work	85
6.4	The effect of network structure regime on the extent of diffusion	87
6.4.1	Diffusion in Uniform random and Scale–free networks	89
6.4.1.1	Uniform random networks	90
6.4.1.2	Scale–free networks	91
6.4.1.2.1	Sparse scale–free networks	93
6.4.1.2.2	Dense scale–free networks	96
6.4.1.2.3	Intermediate density scale–free networks	97
6.4.2	Experimental methodology	102
6.4.2.1	Diffusion process	102
6.4.2.1.1	Extent of diffusion simulations	102
6.4.2.2	Datasets	104
6.4.2.2.1	Synthetic networks	105
6.4.3	Results and analysis	108
6.4.4	Optimal vs expected extent of diffusion	109
6.4.4.1	Optimal vs degree–based heuristic	111
6.4.4.2	Effect of network modularity on the extent of diffusion	116
6.4.5	Discussion and conclusion	118
6.5	Effect of dynamic network structure on diffusion in networks	120
6.5.1	Contributions	121
6.5.2	Methodology	122
6.5.2.1	Methods	122
6.5.2.2	Diffusion model	123
6.5.2.3	Parameter settings	123
6.5.3	Experimental results	124
6.5.4	Conclusions	125
7	**DIFFUSION MINIMIZATION**	129
7.1	Problem statement	130
7.2	Finding good blockers in dynamic networks	132
7.2.1	Contributions	132
7.2.2	Methodology	134

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

7.2.2.1 Algorithm . . . 134
7.2.2.2 Probability of activation . . . 135
7.2.2.3 Network properties . . . 136
7.2.2.4 Lower Bound: Best Blockers . . . 137
7.2.2.5 Upper Bound: Random Blockers . . . 137
7.3 Datasets . . . 138
7.4 Results and Discussion . . . 139
7.5 Conclusions and Future Work . . . 147

8 SOCIAL ORGANIZATION IN EQUIDS . . . 152
8.1 Motivation . . . 152
8.2 Network structure and functions . . . 153
8.2.1 Statistical physics . . . 154
8.2.2 Network theory . . . 154
8.3 Problem statement . . . 155
8.4 Diffusion maximization . . . 155
8.4.1 Independent cascade diffusion . . . 156
8.4.1.1 Extent of diffusion . . . 156
8.4.1.2 Rate of diffusion . . . 157
8.4.2 Linear threshold diffusion . . . 157
8.4.2.1 Extent of diffusion . . . 157
8.4.2.2 Rate of diffusion . . . 157
8.4.3 Grevys vs onagers . . . 160
8.5 Diffusion minimization . . . 162
8.5.1 Independent cascade . . . 162
8.5.1.1 Expected extent of diffusion . . . 162
8.5.1.2 Blocking set size . . . 163
8.5.2 Linear threshold . . . 163
8.5.2.1 Expected extent of diffusion . . . 163
8.5.2.2 Blocking set size . . . 163
8.5.3 Grevys vs Onagers . . . 168
8.6 Network measures as indicators of extent of diffusion . . . 168
8.7 Conclusions . . . 168

CITED LITERATURE . . . 170

VITA . . . 184

LIST OF TABLES

TABLE		PAGE
I	Real–world networks' statistics	31
II	Parameters for the generative model	35
III	Optimal Vs Approximate–grevys and onagers	63
IV	Dynamic vs Static: Quantitative and Qualitative	64
V	Optimal Vs Approximate–grevys and onagers	66
VI	Dynamic vs Static: Quantitative & Qualitative	67
VII	Extent of diffusion estimate: uniform and scale–free networks	98
VIII	Dynamic network dataset statistics.	138
IX	Best blockers	147

LIST OF FIGURES

FIGURE		PAGE
1	Modular structure of dynamic network	37
2	Degree distribution of skewed community structure	37
3	Comparison: Dynamic & Aggregate networks	73
4	Spearman's correlation: ranking of individuals across segments	79
5	Spearman's correlation: ranking of individuals VS network dissimilarity	80
6	Spearman's & Jaccard similarity in ranking of all individuals	81
7	Jaccard similarity: top 5 individuals across segments	83
8	Jaccard similarity: top 5 individuals VS. network dissimilarity	84
9	Relative performance of top 5 individuals across segments	84
10	Relative performance of top 5 individuals w.r.t. change in network	85
11	Reality Mining time series	105
12	Plains zebras series	106
13	Difference between the extent of diffusion	111
14	Reality Mining	112
15	plains zebras	113
16	Scale–free networks	114
17	Real-world networks	115
18	Difference in the greedy–optimal and the expected extent of diffusion	126
19	Difference in the greedy-optimal and Expected, Degree and, Betweenness heuristics	127
20	Reduction in the extent of diffusion - grevys	140
21	Reduction in the extent of diffusion - onagers	141
22	Reduction in the extent of diffusion - DBLP	142
23	Reduction in the extent of diffusion - Enron	143
24	Reduction in the extent of diffusion - Reality Mining	144
25	Reduction in the extent of diffusion - UMass	145
26	Enron-Best Measures Comparison	150
27	Onagers-Best Measures Comparison	151
28	Diffusion in grevys with independent cascade	158
29	Diffusion in onagers with independent cascade	159
30	Diffusion in grevys with linear threshold model	160
31	Diffusion in onagers with linear threshold model	161
32	Reduction in the extent of diffusion in grevys with independent cascade model	164
33	Reduction in the extent of diffusion in onagers with independent cascade model	165
34	Reduction in the extent of diffusion in grevys with linear threshold model	166

LIST OF FIGURES (Continued)

FIGURE **PAGE**

35 Reduction in the extent of diffusion in grevys with linear threshold model 167

SUMMARY

Diffusion of contaminants, diseases, rumors, fads, and many other dynamic processes typically take place through a network of interacting entities. Entities themselves can be as diverse as people, computers, organizations, and genes, among others. Similarly, interactions among the entities can be of many forms, including friendships, physical proximity, electronic and verbal communication, file sharing, physical molecular forces, and gene regulation. Network analysis is one way to mathematically model and understand relationships among entities and how a stochastic process, such as diffusion of a phenomenon taking place in the network, is affected by those relationships. One fundamental question in the context of diffusion, particularly in social networks is: which entities in a network are critical for a given diffusion process? For instance, these critical entities could be individuals to whom free products should be given in a network so that the adoption of the product is maximized, or individuals in a population who should be vaccinated so that the spread of a virus or a contaminant is minimized or, leaders in a network that are critical for initiating a mass movement. Mathematical forms of many variants of this question have been shown to be in the class of NP-hard problems. Moreover, strong inapproximability results exist for most of them. A number of heuristics have also been formulated in the last decade or so to produce effective solutions for certain network models. However, effective yet efficient solutions and better insights into effects of the network structure on diffusion have still eluded the scientific community.

SUMMARY (Continued)

In this research, I address the question of finding critical individuals for diffusion in networks in the context of network theory, graph mining, and social network analysis. My research focuses on two complementary optimization goals: maximization and minimization of the extent of the resulting extent of diffusion. I work with explicitly *dynamic* or time evolving networks instead of traditional static or aggregate representation of networks.

For diffusion *maximization*, I focus on finding the set of, what has been called, "influential individuals". That is, those individuals in the network who, when used as the initiators of the particular diffusion process, maximize the resulting extent of diffusion.

In my research I worked on the following problems in the context of diffusion maximization goal:

1. I extend the hardness results of diffusion maximization from traditional network models to explicitly dynamic networks. I show that the problem of finding the optimal set of influential individuals remains NP–hard for dynamic social networks. I provide a $(1-1/e)$-approximation algorithm for finding a set of most influential individuals.
2. I analyze the impact of structural changes in dynamic networks on the prediction of influential nodes in those networks. The motivation here is to measure the effectiveness of the best methods given the natural evolution of the underlying networks.
3. I examine the extent of diffusion problem by analyzing the global structural properties of real–world networks as indicators of diffusion trends. Specifically, my investigations use the *effective* density of the diffusion network as an indicator of: (a) when it is necessary to employ a sophisticated yet computationally expensive method, (b) when even a random

set of diffusion initiators performs as well as the best methods, in expectation; or (c) when and why certain heuristics, such as choosing high degree nodes for maximizing the diffusion in the network, work.

For diffusion *minimization* problem, I focus on finding individuals who should be vaccinated or blocked from the network in order to minimize the resulting extent of diffusion in the remaining network. I extend the standard network theoretic network measures to dynamic networks and investigate connections between local network structure and diffusion minimization. I develop simple, practical, and locally computable heuristics for identifying critical nodes for minimizing diffusion in those networks.

For rigorous analysis of the stochastic process of diffusion optimization, realistic network generative models are crucial. Unfortunately, there are not many such models for dynamic networks. To address this shortcoming, I present a statistical model for generating realistic dynamic networks over time. Modeling such networks is necessary for two reason. First, in order to better understand the underlying dynamics of the population. Second, in order to generate synthetic networks that emulate the real–world properties, for validation. I present a truly dynamic statistical generative network model that captures membership, formation, and fluidity of community membership and the resulting structure of interactions.

The application domain on which I primarily and extensively focus in my research, is behavioral ecology, especially, of *equids*. Specifically, I studied the relationship the effect of influence maximization on social organization in equids. My contribution in this area is the analysis of the relationship between social network structure of equids–specifically, grevys zebra and

onagers - to the functions (leadership formation, loyalties, protection, survival) performed by these networks. The goal of this work is to find clues in the global network structure of two types of equids that differentiate one network from the other for achieving various above mentioned goals. With this work I provide ecologists with sophisticated computational tools for understanding differences in social organization of various species of equids.

CHAPTER 1

INTRODUCTION

Many natural and artificial systems exhibit diffusion on or in network structures. This is a process by which information, viruses, ideas, and behavior spread over a network of individuals, computers, or webpages, among others. For example, adoption of a new technology begins on a small scale with a few "early adopters", then more and more people adopt it as they observe friends and neighbors using it. Eventually the adoption of the technology may spread through the social network as an epidemic "infecting" most of the network. Similarly, a process of infectious or viral spread starts with an initial set of infected individuals. The infection then spreads to susceptible individuals in close proximity or contact of the infected set of individuals. These diffusion processes then unfold over the network and potentially create cascades. These cascade processes have been the subject of studies in social sciences (80; 118; 139; 152), economics (76; 89; 144), epidemiology (34; 63; 143), statistical physics (124; 129; 133; 136), and recently in computer science (17; 30; 81; 92; 94; 100; 107; 111). Diffusion of information, innovation, viral marketing, and epidemic spread are some of the cascading phenomenon addressed in these research areas. Other than the population and the accompanying social networks, resource (such as water or air) distribution networks are another application of diffusion processes (27; 28; 29). Similarly, behavior propagation in animal proximity networks that results in mass movements in those networks is another interesting area of research for the

diffusion of influence phenomena. Hence, diffusion in networks is an important problem for a diverse set of applications.

One of the most crucial research problems in networks is the optimization of the extent of diffusion in those networks. For instance, which set of entities in a network facilitate or inhibit the diffusion process the most in a network? From a viral marketing perspective, which k individuals should be given free samples of a product so that the number of individuals who eventually adopt the product, can be maximized? Or, from an epidemiology perspective, which k individuals should be vaccinated so that spread of the disease in the population can be minimized? Both the diffusion maximization and many variants of the diffusion minimization objectives are NP–hard. Subsequently, approximation algorithms for these problems have been proposed (92; 107). However, most of these approximations rely on a stochastic simulation process that requires many iterations. These approaches then become infeasible for very large networks. To counteract this problem many effective and efficient heuristics have been proposed in the last decade or so (13; 51; 53; 63; 83; 100).

In my research, I investigate methods for finding nodes in a network that are critical for diffusion optimization in a network. *Given a network, a probabilistic diffusion model and the number k, the central question of my research is: which k individuals optimize the extent of diffusion the most when they are targeted to initiate the diffusion process.* In this research, I address the diffusion optimization objective for the two complementary goals of maximization and minimization.

Mathematical epidemiology, social science, and lately computer science has expended significant effort in developing and studying models of diffusion. Typically, these studies have assumed stationarity of the underlying network by aggregating the set of interactions (or edges) among the entities of a network over time into a fixed set of entities related to each other through the set of interactions. Such networks are called *aggregate* networks. An aggregate network is an accumulated and static representation of a population network. However, the real–world is not static and the diffusion process is inherently dynamic. Thus, to overcome these limitations of aggregate networks, some of my work is focused on explicitly dynamic or time–evolving networks.

Structural properties, such as, clustering coefficients, eigenvalues, expansion factors, and centrality measures are some of the graph theoretic properties that have been shown to relate to various social phenomena. In social sciences, as early as 1960s, Granovettor's insight about the strength of weak ties (79), and a large body of research on the diffusion of innovations (139; 152), has shown that an individual's ideas and behaviors are direct functions of the ideas and behaviors of the people the individual is connected to. Recently, many studies have shown a positive correlation between the network structure and various social phenomena such as community affiliations, communication pattern, as well as, diffusion process in networks. Based on the conclusions drawn from this vast body of research, in this work, I further explore the relationship of global and local network structure and diffusion in networks.

The goal of my research has been to develop efficient methods for diffusion optimization objectives while tapping into a graph theoretic framework of network structure. In addition,

I extend the study of diffusion processes from aggregate to explicitly dynamic networks. This research can be broadly divided into five parts.

1. **Network dynamics.** I formulated dynamic network structure properties by incorporating the time factor into the traditional network theoretic properties (Chapter 5). These properties provide the foundation for the analysis of diffusion in dynamic networks.

2. **Diffusion maximization.** Diffusion maximization is the problem of identifying a fixed set of k nodes in a network such that when a stochastic diffusion process in the network is initiated from them, the diffusing phenomenon reaches the maximal extent of the population (Section 6.1). For diffusion maximization, I first established why dynamics matter by comparing diffusion in dynamic and aggregate networks (Section 6.2). I further investigate how the evolution of network affects the prediction of the extent of diffusion and the identity of spreaders in a network (Section 6.3). I investigate the relationship between global network structure and the extent of diffusion (Section 6.4). Specifically, I investigate the effect of network density and modularity on the extent of diffusion. Moreover, I establish the connection between global network structure and the hardness of finding good diffusion initiators in networks. Lastly, I analyze the effect of dynamic network structure, such as dynamic communities, on diffusion in the network using certain stochastic diffusion models (Section 6.5).

3. **Diffusion minimization.** Diffusion minimization is the problem of identifying a fixed set of k nodes in a network that when blocked from the network, the effect of diffusing process is minimized the most in the network (Section 7.1). For diffusion minimization problem,

I devised heuristic methods for identifying critical individuals that minimize the extent of diffusion the most in dynamic networks. I approach this problem by identifying aggregate and dynamic network properties (Section 7.2) that are most effective in reducing the extent of diffusion. I compare the blocking ability of individuals in the network to their respective structural ranking based on the dynamic measures proposed in Chapter 5. I find that overall, simple ranking according to a node's static degree, or the dynamic version of a node's degree, performed consistently well. Surprisingly the dynamic clustering coefficient seems to be a good indicator, while its static version performs worse than the random ranking. This provides simple practical and locally computable algorithms for identifying key blockers in a network.

4. **Dynamic network generative model.** In this part of the thesis I present a statistical model for generating realistic dynamic networks over time (Section 4.3). Modeling such networks is necessary both to better understand the underlying dynamics of the population, as well as, for generating synthetic networks that emulate the real–world properties, for validating analysis. This model captures membership, formation, and fluidity of community membership, and the resulting structure of interactions. This model is designed to incorporate some of the most fundamental properties of the real world networks. I particularly use this model to analyze the effect of network structure on the extent of diffusion in dynamic networks in Section 6.5.

5. **Social organization in equids.** In this work I compare the relationship between social network structure of equids — specifically, grevys zebra and onagers — to the functions

(leadership formation, loyalties, protection, survival) these structures are optimized for to perform (Chapter 8). Equids display a variety of different forms of social organization; Plains zebra associate in small, closed harems, whereas onagers (wild asses) and grevys zebra are found in looser, more ephemeral associations. Horses appear to be somewhat intermediate. The goal of this part of my work is to find clues in the global network structure of two types of equids that differentiate one network from the other for achieving various diffusion goals. With this work I provide ecologists with sophisticated computational tools for understanding various risks and benefits for which these animals optimize for in their social organization.

The structure of the document is laid out in the following way. In Chapter 2, I briefly describe various concepts that are fundamental for this research. Chapter 3, provides a brief literature survey of applications of diffusion in networks, the computational complexity results, the approximation and the heuristic approaches for diffusion optimization as well as computational social network analysis. In Chapter 4, I briefly describe the dynamic and aggregate real–world networks used in the various experimental studies in this research. Moreover, I describe the network generative models employed in this research to generate synthetic networks. Lastly, I describe the statistical dynamic network model that I developed. In chapter 6, I address the diffusion maximization objective. Chapter 7 deals with the rigorous comparison of dynamic network structure with the goal of diffusion minimization. Lastly, in Chapter 8, I discuss the mathematical modeling of processes driving the social network organization in equids.

CHAPTER 2

PRELIMINARIES

Analysis of diffusion processes over networks has two separate components. One is the network that serves as a conduit to the diffusion process. Second is the stochastic diffusion process itself. In this chapter, I revisit the traditional definition of networks from the literature as well as state the time–incorporating extension to that representation of the network. Similarly, stochastic diffusion models come in various flavors. I discuss some of the models from mathematical epidemiology and social sciences that I use repeatedly in my work. Moreover, I state the network structural properties that are heavily used in the analysis of relationship between network structure and the extent of diffusion in those networks. These network properties are limited to static/aggregate network representations. In Chapter 5, I introduce the extension to these network properties for explicitly dynamic networks. Lastly, I state some of the functions and their properties that are used extensively as the basis for diffusion optimization.

2.1 Types of networks

In my work, I represent network structures in two distinct ways: the traditional aggregate network of connections and interactions and a non–traditional dynamic time–ordered series of interactions. Formal definitions of both are as follows:

2.1.1 Aggregate network

The aggregate network is the graph $G_A = (V, E)$ of individuals V and their interactions E observed over a period of time. In this representation an edge exists between pairs of individuals if they have ever interacted during the observed time period. Multiple interactions between a pair of individuals over time are represented as a single weighted (by frequency) edge between them or multiple edges between them (multigraph). This representation provides an *aggregate* view of the interactions where the timing and order of interactions is neglected. For the most part, my analysis is based on the multigraph representation of network of interactions.

Definition Let $\{1, \ldots, T\}$ be a finite set of discrete timesteps. Let V_t be the set of entities observed at time t and let E_t be the set of interactions among entities set V_t at t. The aggregate graph $G_A = (V, E)$ of such a network is the set of entities V and interactions E such that $V = \bigcup_t V_t$ and $(u, v) \in E$ if $\exists (u_t, v_t)$ at some timestep t between $1, \ldots, T$.

Many network problems are based on fundamental relationships involving time. Consider, for example, the problems of modeling the flow of information through a distributed network, studying the spread of a disease through a population, or analyzing the reachability properties of an airline timetable. In such settings, a natural model of a network is the one in which each interaction is annotated with a time label specifying the time at which the pair of entities involved actually interacted. I call such a network a *dynamic network*.

2.1.2 Dynamic network

I represent a dynamic network as a series $\langle G_1, \ldots, G_T \rangle$ of static networks where each G_t is a snapshot of individuals and their interactions at time t. For my work, I assume that the time

during which the individuals are observed is finite. For simplicity, I also assume that the time period is divided into discrete steps $\{1, \ldots, T\}$.

Definition Let $\{1, \ldots, T\}$ be a finite set of discrete timesteps. Let $V = \{1, \ldots, n\}$ be a set of individuals. Let $G_t = (V_t, E_t)$ be a graph representing a snapshot of a static network at time t. $V_t \subseteq V$, is a subset of entities V observed at time t. An edge $(u_t, v_t) \in E_t$ if entities u and v are connected at time t and for all $v \in V$ and $t \in \{1, \ldots, T-1\}$, $(v_t, v_{t+1}) \in E$ are directed self edges of entities across timesteps.

A *dynamic network* $G_D = \langle G_1, \ldots, G_T \rangle$ is the graph $G_D = (V, E)$ of the time series of graphs G_t such that $V = \bigcup_t V_t$ and $E = \bigcup_t E_t \cup \bigcup_{t-1} (v_t, v_{t+1})$.

Analysis of network structures and their local and global properties complement research about diffusion processes and how the global and local network configuration affect the extent or rate of spread. Diffusion of innovations research has been greatly enhanced by network analysis because it provides more exact specification of *critical individuals* like who influences whom during the diffusion process. Similarly, network analysis has benefited from diffusion related research by providing real-world applications to compare and validate many mathematical network models.

2.2 Static network structure and properties

Network theory provides an extensive vocabulary to label entities of a network. This vocabulary also provides a set of concepts by which various properties of the network and its entities can be evaluated. Thus, network theory gives us this variety of mathematical concepts

by which a network and its entities can be precisely quantified and measured. Following is a list of network related concepts that I refer to in the diffusion related work. They are divided into the macro–level, global network properties and the micro–level, local node level properties.

2.2.1 Global properties

Global properties of a network refer to the average properties of the entire network being studied. Although diffusion has been known to occur mostly at the local level the overall global structure has been shown to have a significant impact on the evolution of diffusion trends in a network (55). Some of the basic global properties of the network that have a significant impact on the extent of diffusion are the following.

2.2.1.1 Density

Density $D(G)$ is the proportion of the number of edges $|E|$ present in a network relative to the possible number of edges $\binom{|V|}{2}$.

$$D(G) = \frac{|E|}{\binom{|V|}{2}}.$$

2.2.1.2 Path

A simple path $P(u, v)$ between a pair of nodes u, v is a sequence of nodes $u = v_1, v_2, \ldots, v_p = v$ with every consecutive pair of nodes connected by an edge$(v_i, v_{i+1}) \in E$.

2.2.1.3 Diameter

Diameter $Dia(G)$ is the length of the longest shortest path.

$$Dia(G) = \max_{i,j}\{d(v_i, v_j) : v_i, v_j \in V\}.$$

2.2.1.4 Average path length

Average path length *AVG–PATH(G)* is the average of all shortest path lengths between all pairs of nodes in a network.

$$AVG\text{–}PATH(G) = \frac{\sum_{i,j \in V} d(v_i, v_j)}{\binom{|V|}{2}}.$$

2.2.2 Local properties

In network analysis various properties of the representing graph are studied as proxies of the properties of the entities and their interactions. For example, the degree, various centrality measures, clustering coefficients, or the Eigenvalues (PageRank) of the nodes have been used to determine the relative importance of the individuals (42; 90) in networks. Following are the local network properties I study in relation to diffusion in networks.

2.2.2.1 Degree

Degree of a node is the number of its unique neighbors. It is perhaps the simplest measure of the influence of an individual: the more neighbors one has, the higher the chances of reaching a larger proportion of a population.

2.2.2.2 Betweenness

Betweenness of an individual is the sum of fractions of all shortest paths between all pairs of individuals that pass through this individual. It is a parameter that measures the importance of individuals in a network based on their position on the shortest paths connecting pairs of non-adjacent individuals (11; 71; 72).

2.2.2.3 Closeness

Closeness of an individual is the average (geodesic) distance of the individual to any other individual in the network (72; 142).

2.2.2.4 Clustering coefficient

Clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together (125).

Next I state the stochastic diffusion models I use in my work for simulating diffusion in networks.

2.3 Types of diffusion models

Diffusion models are studied extensively in a wide range of disciplines. I focus on two of the general models from mathematical epidemiology and behavior/social sciences.

2.3.1 Independent cascade model

The independent cascade model was first introduced in (59; 75) in the context of word-of-mouth marketing. This is also the most commonly used simple model to study disease transmission in networks (57; 117; 119; 124; 133). In this model, transmission from one individual to another happens independent of interactions with all the other individuals.

The Independent Cascade model describes a diffusion process in terms of two types of individuals, active and inactive. The process unfolds in discrete synchronized timesteps. In each timestep, each active individual attempts to activate each of its inactive neighbors. The activation of each inactive neighbor is determined by a probability of success. If an active individual succeeds in affecting any of its neighbors, those neighbors become active in the next

time step. Each attempt of activation is independent of all previous attempts as well as the attempts of any other active individual to activate a common neighbor.

2.3.1.1 Independent cascade model and dynamic networks

Let $G_D = (V, E)$ be a dynamic network, $A_0 \subseteq V$ be a set of active individuals, and p_{uv} be the probability of influence of u on v[1]. An active individual $u_t \in A_0$ at timestep t tries to activate each of its currently inactive neighbors v_t with a probability p, independent of all the other neighbors. If u_t succeeds in activating v_t at timestep t, then v_t will be active in step $t+1$, whether or not $(u_{t+1}, v_{t+1}) \in E_{t+1}$. If u_t fails in activating v_t, and at any subsequent timestep u_{t+i} gets reconnected to the still inactive v_{t+i}, it will again try to activate v_{t+i}. The process runs for a finite number of timesteps T. I denote by $\sigma(A_0) = A_T$ the correspondence between the initial set A_0 and the resulting set of active individuals A_T. I call the size of the set A_T, $|A_T|$, the *extent of diffusion*.

The diffusion process in the independent cascade model in a dynamic network is different from the aggregate network in one important aspect. In the aggregate case, each individual u uses all its attempts of activating each of its inactive neighbors v with the same probability p in one timestep t. This is the timestep right after the individual u itself becomes active. After that single attempt the active individual becomes latent: that is, it is active but unable to activate others. However in the dynamic network model as defined above, the active individuals never

---

[1]For simplicity, I assume p is uniform for all V and remains fixed for the entire period of simulation. The uniform probability values also ensure that I test how the blocking ability of individuals depends solely on the structure of the network, controlling for other parameters that may affect this ability.

become latent during the spreading process. In my work, I only consider the progressive case in which an individual converts from inactive to active but never reverses (no recovery in the epidemiological model). It is a particularly important case in the context of identifying blockers since the blocking action is typically done before any recovery.

2.3.2 Linear threshold model

Granovetter first introduced threshold models of diffusion (80). He postulated that individuals differ in the degree they are influenced by the behavior of others in their social system. For example, some individuals instantly adopt a new product introduced in the market whereas others gradually adopt based on varying peer pressure. In general these sets of individuals are differentiated as early and late adopters. Thus, the degree by which each individual is influenced is not homogenous across a social system.

The linear threshold model also describes the diffusion over two sets of individuals, active and inactive. Each inactive individual has a certain susceptibility to become active, which is denoted by the individual's "threshold". Each active individual has a certain "weight" of influence over each of its inactive neighbors. An individual becomes active if the accumulated weight of all its active neighbors becomes larger than the individual's susceptibility threshold.

More formally, the linear threshold model is defined by two parameters. For each individual v, a threshold $\theta_v \leq 1$ indicates the latent tendency of this individual to be activated. For each edge $(w_t, v_t) \in E_t$ (for $0 \leq t \leq T$) the weight b_{w_t,v_t} is the influence of the individual w on v, that is, w's ability to activate v. For each v, $\sum b_{w,v} \leq 1$.

2.3.2.1 Linear threshold model and dynamic networks

The diffusion process described by linear threshold model in a dynamic network graph starts with a given set of thresholds θ_v assigned to each individual. The initial set of active individuals is A_0. The process unfolds in discrete timesteps, $1 \dots T$. At each step t, each inactive individual v_t is influenced by the set of its active neighbors. The inactive individual v_t becomes active at timestep $t+1$, if $\sum b_{w_t,v_t} \geq \theta_v$. If $\sum b_{w_t,v_t} < \theta_v$ then v remains inactive and at every subsequent timestep, a new attempt is made to activate v_{t+i} by the set of its neighbors active at time $t+i$. Each attempt is independent of any previously made attempts. The process continues for T steps or until no more activation are possible. The outputs, again, are the set of individuals active at time T, A_T and the size of that set $|A_T|$. I denote by $\sigma(A_0) = A_T$ the correspondence between the initial set A_0 and the resulting set of active individuals A_T.

The diffusion process in the dynamic network graph is different from the static network graph, where each active individual is given only one attempt to activate any of its inactive neighbor.

2.3.3 Independent cascade vs Linear threshold model

The difference between the independent cascade model and the linear threshold model is that in the former model each attempt of activation is independent of the attempts by all the other active individuals. However in the later model each inactive individual is influenced by the aggregated weight of all its active neighbors.

In my work I focus on discrete optimization functions only. In the following section I describe some of the fundamental properties of set functions that I heavily rely on for devising efficient solutions for various diffusion optimization objectives.

2.4 Sub–modularity and monotonicity

Sub–modularity is an extensively studied concept in theoretical computer science, especially for designing optimization algorithms. It appears naturally in many applications, either as a structural property of combinatorial problems or as a natural assumption on certain valuation function. The concept of sub–modularity has historically been known as *law of diminishing returns* in most economics literature. Sub–modular functions are set functions, which can be considered a discrete counterpart to convex functions.

Consider an arbitrary function $f(\cdot)$ that maps subsets of the finite set U to non-negative real numbers.

Definition $f(\cdot)$ is *submodular* if it satisfies the natural diminishing returns property. This property states that the marginal gain of adding an element v to a set S is at least as great as adding v to a set T, where $S \subseteq T$:

$$\forall v,\ S \subseteq T:\ \ f(S \cup v) - f(S) \geq f(T \cup v) - f(T)^1.$$

---

[1] I abuse the set notation slightly by denoting the singleton set $\{v\}$ as just v for simplicity.

One of the properties of the submodular functions is that if a non–negative function $f(\cdot)$ is submodular then it is monotonically increasing, i.e.

$$f(S \cup v) \geq f(S).$$

Definition (monotonicity.) A function $f : \{0,1\}^U \rightarrow R$ is monotone if for any $S \subseteq T \subseteq N$, $f(S) \leq f(T)$.

2.4.1 Maximizing sub–modular monotonic functions

Here I explain the hardness of maximizing a sub–modular function subject to a size constraint. That is, given a function $f(\cdot)$ that is monotone, submodular, normalized, and a $k \in \mathbb{N}$, find a set S with $|S| \leq k$ maximizing $f(S)$. This problem is NP-hard, since it contains set cover as a special case.

2.4.1.1 Approximation

A simple greedy algorithm by Nemhauser et al. permits a $(1 - 1/e)-$approximation for the maximization problem (122).

Algorithm 1: Greedy Approximation

Initialization: $S = \emptyset$;
while $|S| \leq k$ **do**
  $S := S \cup v : \arg\max_v \{f(S \cup v)\, \forall v \in N\}$
end while

CHAPTER 3

RELATED WORK

Dynamic phenomena such as opinions, information, fads, behavior, and disease spread through a network by contacts and interactions among the entities of the network. Such spreading phenomena have been studied in a number of domains including: social networks, epidemiology, electronic networks, and physical distribution networks. These diffusion processes can largely be divided into two forms of diffusion: Diffusion of innovation and diffusion of contamination. In the first two sections I touch upon the seminal work in these lines of research as well as briefly describe the subsequent progress in the last decade in these areas. Although it is natural to distinguish between the two classes of diffusion processes, the research in these areas have shown quite a lot of overlap (55). In a later section I describe the set of diffusion optimization objectives studied extensively in theoretical computer science. The last section deals with the research progress in network theory with respect to social network analysis.

3.1 Diffusion of diseases and contagions

Diffusion of communicable diseases, viruses, bacteria, contagions, and other contaminants is one of the most well studied class of diffusion processes. Other than the biological hazards, viruses and worms in computer networks and the Internet is another interesting application for processes spreading over the network. Traditionally, mathematical modeling of diffusion

of epidemics in networks has relied on the system of differential equations that makes the simplifying assumption of homogeneity among the interacting entities of the networks (9; 34).

In statistical physics, the concept of bond percolation has been applied to mathematical epidemic modeling (19; 124; 129; 143; 161). This line of research although started with the assumption of uniform random mixing among the entities of the network, the later work incorporates more complex network structures. Ball et al. in (19) provide a probabilistic framework that considers networks with "two levels of mixing". In this framework, each entity in the network belongs to the global network as well as a subgroup in the network. They conclude that the diffusion of diseases in the subgroups takes place much more easily than in the larger global network. Later, an alternative approach of calculating effect of clustering on SIR epidemic is given in (91). Kuperman and Abramson in (103) extend Keeling's work to SIRS model for diseases. In real–world, diseases do not die out completely, rather, they reappear periodically. This is typically known as the endemic state. Using more realistic network structures, Pastor-Satorras and Vespignani initiate the extent of diffusion of computer viruses on the Internet graph with scale–free degree distribution (133). Recently, Ferreira et al. quantitatively compare the epidemic thresholds for various network models using variants of mean–field theory (68). Claudio and Pastor-Satorras give a more nuanced relationship of hubs and network core to the extent of epidemics in heterogenous networks in (43).

Independently, there has been extensive research going on in studying diffusion of epidemics in game–theoretic settings (13). A remarkably different attempt in this line of work is the

analysis of competing activation mechanisms in epidemics on networks based on the network structural properties (44).

3.2 Diffusion of innovations

Diffusion of innovation is the spread of ideas, opinions, fads, and adoption of products through certain channels over time among the members of a social system. Models of diffusion of contagion have been used to explain the diffusion of ideas and opinions as early as nineteenth century (105; 115), *albeit* not in any scientific way. By the late 1960s, the conceptual link between adoption and diffusion had been greatly reinforced by their enshrinement in simple mathematical models, which were mostly derived from classical models of mathematical epidemiology (24). These various models, however, all embodied the core substantive assumption that new adopters are influenced by the proportion of the population that has adopted previously. For example, based on the theory of conditional decision models, social scientists have explained diffusion of various social behaviors like self–management (144; 146), segregation (145), ethnic norms (104), effects of social networks in diffusion of innovation (140), and rejection of innovation (1). These models vary in terms of both substance and details, but agree that individual decision–making can be formalized as the probability of any given individual acting is dependent on what his or her peers (friends, family, and coworkers, or other relevant social reference group) are willing to do. Granovetter later formalized this line of research by presenting "Threshold Models of Collective Behavior" (80). He postulated that individuals are not homogenous in the degree they are influenced by their social system. Dodds and Watts (58) later proposed a more general model of diffusion for "contagion" that contained both epidemic

as well as the threshold models as special cases. Initial modeling of diffusion in social networks heavily relied on uniform mixing models from mathematical epidemiology (9; 156).

However, in the two decades with the better understanding of network structures, more sophisticated and realistic model of diffusion of information have been developed (117; 124; 129; 160; 161). In recent years, a number of studies have made use of online data to study various social diffusion processes (4; 17; 18; 45; 106; 107).

3.3 Algorithmic results in diffusion optimization

The problem of identifying influential individuals in a social network has been most rigorously formulated and analyzed by Kempe et al. (92). They showed that this problem is NP-complete. Moreover, they provide a greedy approximation algorithm that guarantees a solution no worse than $(1 - 1/e)$ factor of the optimal for many general models of the spread of influence. Later, Mossel et al. (120) showed that the general case of finding a set of nodes with the largest "influence" is NP-hard, and has a $(1 - 1/e - \varepsilon)-$ approximation algorithm. Unfortunately, this approximation algorithm is computationally intensive. Strong inapproximability results for several variants of identifying nodes with high influence in interaction networks have been shown in (47). Given the variety of diffusion processes, a diverse range of applications for the diffusion problem, and the computational challenges involved, many techniques have been developed to address variants of the diffusion detection, maximization, and minimization problem. These techniques focus mainly on the diffusion function itself to solve the problem. Asur et al. presented an event–based characterization of critical behavior in interaction graphs for the purposes of modeling evolution, link prediction, and influence maximization (14).

Identifying key nodes for spread minimization objective has algorithmically and heuristically been dealt with in literature (63; 86; 93; 106). In (106), Leskovec et al. proposed an elegant algorithm that optimizes multiple criteria related to spreading processes such that the spreading phenomenon is detected as quickly as possible. Eubank et al. experimentally showed that global graph theoretic measures like expansion factor and overlap ratio are good indicators for devising vaccination strategies in static network models for populations (63). Another immunization strategy based on the aggregate network model was proposed in (51). They propose an efficient method of picking high degree nodes in a network to immunize thus inhibiting the spread of disease. Kempe et al in (93) showed that a variant of the blocker identification problem is NP-hard. In (10), Angluin et al. provide an approximation algorithm for diffusion maximization without resorting to learning the network structure.

For diffusion minimization (12) proposed a game theoretic approach for inoculation strategies for victims of viruses based on sum-of-squares partition problem. It has been shown that removing nodes of the highest eigenvalues reduces the diffusion the most (135; 151; 153). Expansion factor of a graph has been shown to be a good indicator for detecting the spread of a virus over the Internet (13). The core versus periphery structure of the network has been shown to facilitate the inference of network of influence (77). Immunization of hubs in a network (85) is a good strategy to minimize diffusion however, clustering coefficient is not an effective measure (83; 85) for such a goal. Immunizing neighbors of randomly selected nodes can result in efficient diffusion minimization (51). Degree centrality heuristic has been shown to work well

for containing spread of misinformation (38). Also, degree centrality has been shown to strongly correlate with rates of infection (89).

Other than optimization of information and contagion diffusion in population networks, a separate line of research has been going on to study diffusion of contaminants in distribution networks. (27; 28; 29) studied the optimal sensor placement sensors in water distribution networks in large cities. Recently, Brummitt et al. analyzed cascades of load shedding in electric grids coupled with other infrastructure (37). They provide a multi–type branching process framework for better prediction of cascading processes on modular random graphs and on multi–type networks in general.

3.4 Computational network analysis

Computational network analysis has attracted considerable interest in social and behavioral sciences, economics, and computer science communities among many others for a a while now. The development of faster and more sophisticated techniques for collecting and storing vast amount of data has resulted in improving our ability to analyze large social, biological, and computer networks. Computational techniques have started to play a central role in bridging the gap between network analysis methods and many real–world concepts and physical processes. Diffusion is one such phenomenon. Following is a brief literature survey of various network properties studied to analyze diffusion in networks.

The fundamental difference between network and non–network explanation of a process is the inclusion of concepts and information on *relationships* among entities in a study. Whether we want to study local actions in the context of global structure of relationships of entities or

the overall structure as it is, network analysis operationalizes structure in terms of networks of relationship among entities. Regularities and patterns in interactions give rise to structures or networks. The idea of representing societies as networks of interacting individuals dates back to Lewin's earlier work of group behavior (109). Typically, there is a single network representing all interactions that have happened during the entire observation period. Using this aggregate network model, the structure and properties of many social networks have been studied from different perspectives (23; 32; 33; 35; 87).

Interdependence between individual decisions alludes that the *structure of relations* between individuals is likely to impact the spread of conditional behaviors. Using systematic modeling of real–world networks, (2; 53; 113) model the diffusion process in networks by incorporating the network structure with the stochastic diffusion process. Social science shows that the location and other local structural characteristics of where the diffusion starts results in effectively different trends in diffusion (2). Size and density of the network affect the cascades of actions, depending on the structural position of a certain type of *critical individuals* (78). Recently, models of segregation have been extended to incorporate the real–world network structure like the *small world phenomenon* (66). Diffusion of innovation is more likely to fail in random networks than in a highly clustered network of consumers (48). "The strength of weak ties" is a highly influential work of Mark Granovetter (79) that highlights the social dynamics of a population. This study elucidates the cliquishness of the local community level structure and existence of infrequent links between densely connected subgroups. The significance of evolution of social populations for the diffusion of ideas, fads, traditions, riots, and other social

trends cannot be over stated. Thus, it is important to analyze the spread of knowledge in a population network while embedding its underlying structure in the overall equation of the spreading behavior. Network properties, such as, short average path length and high clustering coefficient correlate positively with diffusion maximization (95). Watts in (156) gives an exact solution for the threshold model in random graphs for the case where initially a low density of individuals have adopted a trend cascading through the network. Analysis of diffusion of knowledge in a fully regular lattice like network structure to completely random network, that is, from a local clique like structure to a completely irregular structure has shown that the diffusion of same types of knowledge results in different trends of prevalence in a population with different social structures (53). Recent works have focused on specific networks to co–relate spread of ideas and information in particular networks like the web (81; 101; 102).

In network analysis various local properties of the graph representing the population are studied as proxies of the properties of the individuals, their interactions, and the population itself. For example, the degree, various centrality measures, clustering coefficients, or the eigenvalues (PageRank) of the nodes have been used to determine the relative importance of the individuals, *e.g.,* (42; 90). Betweenness centrality has been used to identify cohesive communities (74) and the distributions of shortest path lengths employed to measure the "navigability" of the network (157). These and many other graph theoretic measures have been translated to many social properties (112; 123; 124). Network measures such as clustering and assortative mixing coefficients have been used to design local vaccination strategies (86).

CHAPTER 4

NETWORK DATA AND MODELS

This research work is based on the empirical study of many real–world and synthetic networks. The next section provide the summary of the real–world networks used in this work. In the later section, the description and generative process of the synthetic networks is provided. In the end, I present the dynamic network generative model that I developed for analysis on real time evolving networks.

4.1 Real world networks

The real–world networks used in this research, include a diverse range of samples from computer networks, human and animal population networks, co-authorship networks, social networks, and blogosphere among others. Following is a brief description of those networks.

Autonomous systems: The graph of routers comprising the Internet can be organized into sub–graphs called Autonomous Systems (AS). Each AS exchanges traffic flows with some neighbors (peers). A communication network of who–talks–to–whom from the Border Gateway Protocol logs constructed by (107) is used in our analysis. The data was collected from University of Oregon Route Views Project - Online data and reports. The dataset contains 735 daily instances which span an interval of 785 days from November 8, 1997 to January 2, 2000.

Co-authorship network: This data set is a sample of the *Digital Bibliography and Library Project* (110). This is a bibliographic dataset of publications in Computer Science. I use a subset of the data from 1984–2002. In this network each node represents an individual author and two authors are interacting if they are co-authors on a paper. The sample I used contains 1000 individuals and 1891 co-authorship links. This is one of the sparsest real networks used in this work.

Live journal: LiveJournal is a free on–line community with almost 10 million members; a significant fraction of these members are highly active. (For example, roughly 300,000 update their content in any given 24-hour period.) LiveJournal allows members to maintain journals, individual and group blogs, and it allows people to declare which other members are their friends (16; 108).

Gnutella P2P network: A sequence of snapshots of the Gnutella peer–to–peer file sharing network from August 2002. There are total of 9 snapshots of Gnutella network collected in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts (107; 138).

Epinions: This is who–trust–whom online social network of a general consumer review site Epinions.com. Members of the site can decide whether to "trust" each other or not. All the trust relationships interact and form the *Web of Trust* which is then combined with review ratings to determine which reviews are shown to the user (137).

Political Blogs: This is a directed network of 1224 web blogs connected by 16715 hyperlinks based on US politics before the 2004 elections. For our experiments I convert the directed graph to undirected graph (3).

Political Books: A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co–purchasing of books by the same buyers[1] (128).

Zachary Karate Dataset: This is an aggregate network of friendships between the 34 members of a karate club at a US university, as described by Wayne Zachary in 1977 (158). It is a sparse network with only 75 friendship links.

Enron: The Enron e-mail corpus is a publicly available database of e-mails sent by and to employees of the now defunct Enron corporation[2]. I used the version of the dataset restricted to the 150 employees of Enron organization who were actually subpoenaed. The raw Enron corpus contains 619,446 messages belonging to 158 users (5; 96).

Onagers: Populations of wild asses (*Equus hemionus*), also known as onagers, were observed by biologists (141; 150) in the Little Rann of Kutch, a desert in Gujarat, India, during January–May 2003. This data is recorded through visual scans. The dataset contains 29 individuals and 402 proximity relationships.

---

[1]The network was compiled by V. Krebs and is unpublished, but can found on Krebs' web site: http://www.orgnet.com

[2]Available with a full description at http://www.cs.cmu.edu/∼enron/

Grevys Zebras: Populations of grevys zebras (*Equus grevyi*) were observed by biologists (69; 70; 141; 150) from June–August 2002 in the Laikipia region of Kenya. Predetermined census loops were driven on a regular basis (approximately twice per week) and individuals were identified by unique stripe patterns. Upon sighting, an individual's GPS location was taken. In the resulting network, each node represents an individual animal and two animals are interacting if their GPS locations are the same. The dataset contains 28 individuals and 778 proximity links between them resulting in a very dense network.

MIT Reality Mining. The MIT Reality Mining dataset consists of social interactions among 100 students and faculty over a nine month period at the Massachusetts Institute of Technology (61). Interactions were inferred from recorded Bluetooth connections between Nokia 6600 smart phones distributed to the participants. Our processed dynamic network consists of 96 vertices. I quantized the data to 4 hours unit based on the analysis by Clauset and Eagle (50).

Haggle Infocomm. The Haggle Infocomm dataset consists of social interactions among attendees at an IEEE Infocomm conference in the Grand Hyatt Miami (148). There were 41 participants and the duration of the conference was 4 days. The time quantization period was 10 minutes.

Plains Zebra. Plains zebra (*Equus burchelli*) are another species of zebra. The data were collected in a similar fashion to that of the grevys dataset. The data were collected through visual scans (approximately once per day) over a period of several months (70). Each entity is a plains zebra and the interactions represent spatial proximity as determined

by ecologists based on GPS locations. It should be noted that this similarity between the plains Zebra dataset and the grevys Zebra dataset should *not* be taken to mean that the social interaction patterns will also be the same. There is evidence to indicate that different species of zebra can exhibit very different interaction patterns (150). The plains-1 dataset represents data from observations of 282 individuals from 12th July 2003 to 19th September 2006. The plains-2 dataset represents observations of a different population of 313 individuals from 5th January 2004 to 3rd July 2007.

Portland: The Portland dataset (64) is a statistically correct simulation of the movement of 1.6 million individuals of Portland on a daily basis. Each node in the network represents an individual and the two individuals are interacting if they are at the same "location", as defined by the simulation. I use a subset of 1588 individuals over 80 timesteps.

IMDB Photo Network The Internet Movie Database (IMDB)[1] maintains a large archive of tagged and dated photographs of individuals associated with the production of commercial entertainment, including actors, directors and musicians. One might reasonably assert that people tagged on a popular online movie information repository are 'recognizable' to the general public, and that a degree of social association exists between people photographed together. Thus, similar to the methodology of the plains Zebra sightings, IMDB Photo Network is a collection of metadata on 45,477 photos with two or more

[1]http://www.imdb.com

TABLE I: Real–world networks' statistics

Dataset	Nodes	Edges	Density
AS	16299	34157	.0002570
DBLP	964	1891	.0004074
Live Journal	15001	66286	.0005890
P2P	8114	26013	.0007903
Epinions	9997	216213	.0043270
Political Blogs	1224	16715	.0223320
Pol Books	105	441	.0809520
Karate	34	75	.1336900
Enron	147	3467	.3229890
plains	282	29050	.730606
Reality Mining	96	3625	.794956
onagers	29	402	.9901000
grevys	28	779	1

people, which collectively represents a partial structure of the social network of people associated with the entertainment industry. The quantization period was one day.

The summary basic statistics of the real–world networks is given in Table I.

4.2 Synthetic network models

Many empirical studies in this work also include synthetic networks based on real–world network models. The following two network models are primarily employed in this research.

Preferential attachment model The web, citation networks, and the network of film actors are among a few examples of social networks that have been shown to exhibit a skewed degree distribution (20; 21; 22; 157). Preferential attachment model (22) generates networks with skewed degree (specifically, power law) distributions. This is one of the first

network generative models. In this model the network evolves over time. Nodes are added to the network sequentially, each new node creates links to already existing nodes proportional to how well connected the other nodes are. Hence, a new node is much more likely to get connected to a high degree node than to a low degree node. The idea is based on the premise of "rich getting richer". Many real–world networks have been shown to exhibit this process of growth. The skewness of the degree distribution is determined by a parameter γ which is usually set between 2 and 3 based on the type of network being studied.

Block mixture model Nodes of real–world networks are often structured in tight relatively well–connected clusters (communities) (127). The stochastic block mixture model was proposed for this purpose in the context of social sciences, using a Bayesian approach (131). Further refinements, such as the assortative mixing (114) has made this model a natural choice to analyze real–world networks in controlled parameter settings. The block mixture model designates the nodes into C blocks. Given two parameters, the inter– and intra–block edge probabilities, the edges are generated uniformly at random, with the appropriate probabilities for each pair of nodes.

Apart from the progress in developing realistic network generative models that depict the static network properties relatively accurately, there is a need for realistic and tractable statistical models for the networks that evolve over time. For example, in epidemiology there is a need for data–driven modeling of human sexual relationship networks for the purpose of modeling and simulation of the diffusion of sexually transmitted disease. In the following section, I

present one such model that I developed. This is a probabilistic network generative model that given a fixed set of nodes, generates networks over time with skewed component distribution with varied sizes and tunable degree distribution within each component.

4.3 Dynamic network generative model

In this work I present a statistical model for generating realistic dynamic networks over time. Modeling such networks is necessary both to better understand the underlying dynamics of the population, as well as, for generating synthetic networks that emulate the real–world properties, for validating analysis.

Classically there are two main schools of network modeling. The approach primarily used in social sciences is to treat the entire network as a maximum likelihood object from a statistical distribution where some of the parameters, such as the number of dyads or triads, are fixed (84). A radically different approach originates from the random graph community, where generative models are designed to emulate large scale global statistical behaviors in networks, such as the degree distribution (8; 21) and average distance (114), among others. A glaring shortcoming of this second class of models is that these models are fundamentally evolutionary but not truly dynamic. That is, once a connection is established it cannot be removed. This intuitively contradicts how social links are formed, reinforced, and lost in real world. These models essentially give a "static" representation of the dynamics of interactions aggregated up to a certain point in time.

Here is a simple example of what the above mentioned models fail to capture. At given point in time, society exists as a collection of loosely formed communities (74). As an individual joins

a community, she/he forms relationships with its members. Overtime the individual updates those relationships by adding more links to already existing members or new–comers and by removing other links. Moreover, some individuals leave the communities all–together. Most of the models in the aforementioned two classes capture one or the other network properties but fail to incorporate many others. Especially, the existing models do not capture dynamic of forming and breaking relationships in a fluid community membership.

4.3.1 Model

In this work I present a truly dynamic statistical generative network model that captures membership, formation, and fluidity of community membership and the resulting structure of interactions. This model incorporates some of the most fundamental properties of the real–world networks. A time evolving network generative model that evaluate network community structure is presented in (121). However, the multiplicity of scale and relationships renders it harder to analyze in detail.

At a high level, the interactions in my model are driven by the individuals' membership in informal communities. Individuals tend to interact more within a community than accross communities however, over time they may update their affiliations.

4.3.1.1 The dynamic generative process

The network generative process works as follows. The model requires the parameters listed in Table II. Given the size of the network N and the number of timesteps T. In each timestep $t_i \in T$, I generate a set of communities according to the community distibution vector $\langle C_0, \ldots, C_m \rangle$. Nodes $1, \ldots, N$ are proportionally assigned to each community. Then, based

TABLE II: Parameters for the generative model

Parameter	Description
N	Size of the network
T	Number of timesteps
$\langle C_0, \ldots, C_m \rangle$	Distribution of community sizes.
P_{intra}	Intra–cluster link probability.
P_{inter}	Inter-cluster link probability.
Optional parameters to setup the dynamics of link formation within communities. For example, for *preferential attachment model*, skewness and average degree.	

on the P_{intra} probability, nodes within a community a linked to each other. In the next step, the nodes are connected across communities using the probability P_{inter}. The above steps are repeated for each timestep in T. Once the community distribution and node affiliations are fixed in each timestep, nodes are switched from one community to another community across timesteps based on the probability P_{sw}. Note that this process of community affiliation over time roughly follows the stochastic block model generative process with the evolution of the network with time incoporporated into it. Other than the basic uniform probability of interactions within a community P_{intra}, the model can take optional parameters to specify other individual level interactions pattern. For example, in the current version, this generative model can take *Preferential Attachement* model as the process by which nodes interact among themselves within a community.

4.3.2 Experimental results

I study the dynamic evolution of networks over a wide range of parameter values. For instance, to generate networks that have at most one giant community and other communities

that are trivially small, while maintaining a preferential attachment model of connectivity within a community, I sample thousands of networks by controlling for the relevant parameters. Figure 2 shows the degree distribution of a class of networks in which the giant component encompasses 40–50 % of the network and all the rest of the components are some constant fraction of log of the network size. For the preferential attachment model the exponent of skewness γ is set between 2 and 3, as has been shown for most real–world networks. For networks based on this model, the minimum average degree is set slightly above 1 to mimic the growth process. The resulting synthetic networks have a bimodal degree distribution. That is, there is a large frequency of smaller degrees but also a relatively significant number of nodes that have degrees closer to the size of the largest component. This instantiation of the network generative process also captures the modular structure of the underlying dynamic network, as is evident by the sample graph in Figure 1.

To verify how well this model imitates reality, I test it in two ways. First, I measure the global network properties of the dynamic network model. Real–world networks have been shown to belong to certain classes of degree distributions, have short geodesics, and/or high clustering coefficient. I estimate the parameter settings of our model that correlate to the properties observed in real world. Secondly, I use the maximum likelihood approach to measure the actual properties of the real–world dynamic networks. Such as, the probability of switching of individuals within communities, the probability of links within and across communities, the expected number of communities given an observed number of observations in time, and the sizes of the communities relative to the size of the population. I use those parameters to

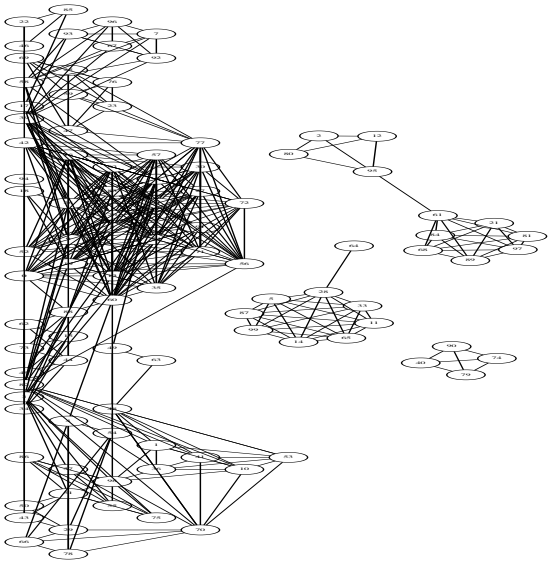


Figure 1: Modular structure of dynamic network

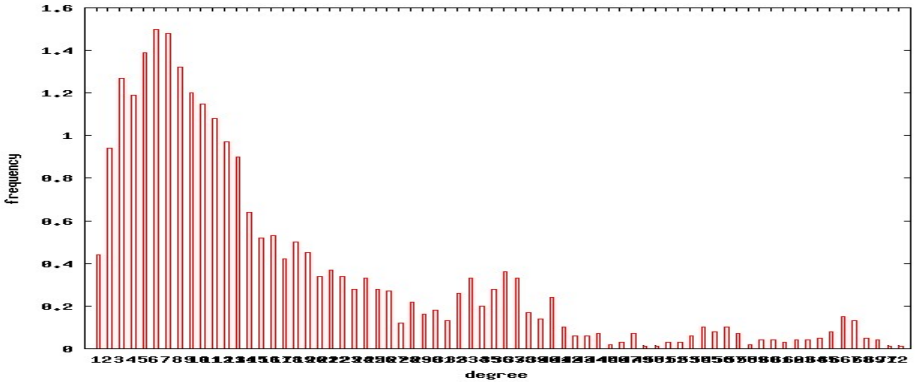


Figure 2: Degree distribution of skewed community structure

synthetically generate similar networks using our generative model and measure how well I can replicate the real process.

4.3.3 Conclusion

Research in computational generative modeling of networks has, over more than a decade, tried to build as realistic models as mathematically and statistically possible. However, for the most part they have failed to capture the complexity of multiplicity of properties exhibited by such networks. In this work, I propose a generative model for dynamic networks based on the notion of distribution of communities within the population that split and merge over time. Thus, this model not only emulates the dynamics of individual level interactions within communities but it also incorporates the structural changes of the communities themselves.

CHAPTER 5

STRUCTURAL PROPERTIES OF DYNAMIC NETWORKS

In this work, I extend the study of network structure measures to explicitly dynamic networks. Many network problems are based on fundamental relationship involving time. Consider, for example the problem of modeling the flow of information (40; 41; 54; 73), spreading of diseases in a population (57; 63; 91; 99; 124), viral marketing (59; 60), and transportation networks (7). For all such domains the evolution of the network over time plays a key role. Hence, it is vital to define the structural properties of such dynamic networks, so that I can compare such networks and better understand their differences.

In the following, I state the dynamic global as well as local structural properties that I developed.

5.1 Global properties

Global properties of a network refer to the average properties of the entire network. Although diffusion occur mostly at the individual level interactions, the overall global structure has been shown to have a significant impact on the evolution of diffusion trends in a network (55). Extensions of some of the basic global properties of the network that have a significant impact on the extent of diffusion are the following.

5.1.1 Dynamic Density

Dynamic Density is the average over all timesteps of the density of each time snapshot. Let $D(G_i)$ be the density of timestep i. Then, density of a dynamic network $D_T(G)$ over T timesteps is:

$$D_T(G) = \frac{1}{T} \sum_{1<t\leq T} D(G_t).$$

That is, the dynamic density averages the static density over the entire time series.

5.1.2 Temporal Paths

The notion of a path is fundamental to most of graph related measures, from connectivity problems and spanning trees to flows and cuts, all are based on paths. Paths that take into account the time labeling on edges are known as *temporal paths* (93). A temporal path is a *time respecting* path if the time labels of the sequence of edges on the path are nondecreasing. A path is *strictly time respecting* if the time labels of the edges are increasing.

Definition A *temporal path* $p(u, v)$ between a pair of individuals u and v, is a sequence of edges connecting them. Formally,

$$p(u, v) = \{(v_0 = u, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n), (v_n, v_{n+1} = v)\}$$

such that, $\forall (v_j, v_{j+1}), (v_i, v_{i+1}) \in p(u, v)$ and $\forall i < j, \lambda(v_i, v_{i+1}) < \lambda(v_j, v_{j+1})$.

5.1.2.1 Geodesic

Traditionally, geodesic is the shortest distance from one individual to another individual where distance is some function of edges on the shortest path. Usually, distance is the number of edges on the path when the graph is unweighted. In dynamic networks, geodesic is the length of the shortest temporal path. The length of temporal path is defined in two ways.

1. If the delays between any two consecutive interactions are ignored then length $d(u, v)$ of a path $p(u, v)$ is the number of edges on the path, assuming that each edge or interaction takes one timestep.

$$d(u, v) = |\{(v_0 = u, v_1), \ldots, (v_{n-1}, v_n = v)\}|$$

   where $(v_0 = u, v_1)$ is the first edge on the path starting at u and $(v_{n-1}, v_n = v)$ is the last edge ending at v.

2. Alternatively, the length of a path is the time it takes for an interaction to take place between a pair of non-adjacent individuals. Simply, the length is the time difference of the first and last interactions on the path. The delays are implicitly embedded in this distance measure between the two individuals.

$$d(u, v) = \lambda(v_{n-1}, v_n = v) - \lambda(v_0 = u, v_1) + 1$$

where $\lambda(v_{n-1}, v_n = v)$ and $\lambda(v_0 = u, v_1)$ are time labels of the last and first interaction respectively on the path $p(u, v)$ between u and v.

The first definition of length of a path is equivalent to the length of a path in a simple unweighted aggregate graph. However, unlike the aggregate graph, there is a path between any two individuals only if it is *time respecting*.

Hence, the geodesic $g(u, v)$ of two non-adjacent individuals u and v in a dynamic network is defined as:

$$g(u,v) = \begin{cases} 1 \quad if \quad (u,v) \in E \\ \lambda(v_{n-1}, v_n = v) - \lambda(v_0 = u, v_1) + 1 \quad if \quad (v_{n-1}, v_n), (v_0 = u, v_1) \in E \end{cases}$$

Geodesic between two individuals in a dynamic network can be interpreted differently based on the timing of interactions, the duration of interactions, and the number of individuals involved in the interaction. The value of the geodesic is the same for all three interpretations. Hence, any of them can be independently used for defining the shortest paths. All these factors have different relevance and meaning in various domains. Following are the three representations of the shortest paths in dynamic networks.

5.1.2.2 Shortest Simple Temporal Path

Shortest Simple Temporal Path $p_s(u, v)$, between a pair of individuals u and v is the shortest time respecting path between those individuals, with each intermediate individual present at most once. The length of the shortest simple temporal path is the geodesic $g(u, v)$.

$$p_s(u,v) = \begin{cases} (u,v) \quad if \quad (u,v) \in E \\ \{(v_0 = u, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n), (v_n, v_{n+1} = v)\} \quad if \quad (v_i, v_{i+1}) \in E \end{cases}$$

5.1.2.3 Shortest Link Path

Shortest Link Path $p_l(u,v)$, is the shortest simple temporal path with minimum number of individuals on the path. The significance of this type of path is that it reduces the dependency of the source and destination on intermediate individuals.

$$p_l(u,v) = \min |p_s(u,v)|$$

5.1.2.4 Shortest Temporal Trails

Another way of measuring the significance of intermediate individuals on the shortest paths of non–adjacent individuals is in terms of the ratio of the time spent on an intermediate node to the total length of the path. Note, that the length of the path is defined in terms of time for the temporal paths. It is obvious that an individual that retains the information for the longest time in a shortest path is more significant than others in some ways. Then, by accounting for delay at each intermediate individual, the paths are no longer simple. The delay on each intermediate individual is essentially a self loop around that individual. These routes in temporal networks are called temporal trails[1]. A temporal trail $p_d(u,v)$ between u and v is

[1]A trail may repeat vertices but must not repeat edges

a sequence of edges $p_d(u,v) = \langle (v_0 = u, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n), (v_n, v_{n+1} = v) \rangle$, such that $\lambda(v_{t-1}, v_t) < \lambda(v_t, v_{t+1})$ for $t = 1, \ldots, n-1$.

The shortest such trails are the one with the smallest geodesic. I call them the *Shortest Temporal Trails*. The important thing is that these shortest trails have the same geodesic as the shortest simple paths and shortest link paths between the same pair of individuals.

5.1.2.5 Dynamic diameter

Dynamic diameter is the length of the longest shortest *temporal* path. Since the geodesic is the same for all three types of temporal paths, the value of the diameter is consistent across all three path measures. However, the actual diameter can be interpreted differently based on the definition used.

$$Dia_T(G) = \max\{d_T(v_i, v_j) : v_i, v_j \in V\}.$$

5.1.2.6 Average temporal path length

Average temporal path length is the average of all shortest *temporal* path lengths between all pairs of nodes in a network.

$$AVG\text{–}PATH_T = \frac{\sum_{i,j \in V} d_T(v_i, v_j)}{\binom{V}{2}}.$$

5.2 Local properties

In network analysis various local properties of the graph are studied as proxies of the properties of the nodes/entities and their interactions. For example, the degree, various centrality measures, clustering coefficients, or the Eigenvalues (PageRank) of the nodes have been used

to determine the relative importance of the individuals (42; 90) in networks. Following are the extensions of some of the local network properties I study in relation to diffusion in networks, to dynamic networks.

5.2.1 Dynamic Degree

Dynamic Degree is the change in the neighborhood of an individual over time. Informally, it is the rate at which new friends are gained over time. Let $N(u_t)$ be the neighborhood of individual u at timestep t. The relative change in the neighborhood is then[1] :

$$\frac{|N(u_{t-1}) \triangle N(u_t)|}{|N(u_{t-1}) \cup N(u_t)|}|N(u_t)|.$$

The Dynamic Degree DEG_T of u is the total accumulated rate of friend addition.

$$DEG_T(u) = \sum_{1<t\leq T} \frac{|N(u_{t-1}) \triangle N(u_t)|}{|N(u_{t-1}) \cup N(u_t)|}|N(u_t)|.$$

Note, that here I consider a friend to be "new" if it was not a friend in the previous timestep. The definition is easily extended to incorporate a longer term memory of friendship. The dynamic degree captures the gregariousness of an individual, an important quality from a spreading perspective.

---

[1]Here $\triangle$ denotes the symmetric difference of the sets

The dynamic degree, unlike its standard aggregate version, carries the information of the timing of interactions and is sensitive to the order, concurrency, and delay among the interactions.

5.2.1.1 Dynamic Average Degree

In the dynamic network the interactions among the individuals change with time. The dynamic average degree is the average over all timesteps of the interactions of an individuals in each timestep:

$$AVG\text{-}DEG(u) = \frac{1}{T} \sum_{1 \leq t \leq T} DEG(u_t).$$

where, $DEG(u_t)$ is the size of the neighborhood of u at timestep t.

5.2.2 Betweenness Centrality

Based on the above definition of temporal paths, I define betweenness centrality in dynamic networks in two different ways. Temporal betweenness centrality is based on the shortest simple or link paths. Whereas, delay betweenness centrality is based on the shortest simple trails definition of temporal paths.

5.2.2.1 Temporal Betweenness Centrality

Temporal betweenness centrality measures the importance of individuals based on their position in the shortest temporal paths of all other nodes. Temporal betweenness centrality $B_T(v)$ of node v is defined as:

Let g_{st} be the number of shortest temporal paths between s and t. Let $g_{st}(v)$ be the number of shortest temporal paths between s and t that pass through v. Let $B_{T(st)}(v)$ be the fraction

of shortest temporal (s,t) paths passing through v. Then, the temporal betweenness centrality $B_T(v)$, of a node v is defined as the sum of fraction of all shortest temporal paths passing through the node v between all pairs of nodes. Formally, the temporal betweenness centrality is:

$$B_T(v) = \sum_{s \neq t \neq v} B_{T(st)}(v) = \sum_{s \neq t \neq v} \frac{g_{st}(v)}{g_{st}}.$$

5.2.2.2 Delay Betweenness Centrality

The delay betweenness centrality, $B_D(v)$, of individual v, is defined as:

Let nst_{st} be the number of shortest trails from s to t. Let $nst_{st}(v)$ be the number of time steps of delay of v that all shortest trails from s to t. Let $B_{D(st)}(v)$ denote the delay-dependency of (s,t) on v. The delay-betweenness centrality $B_D(v)$ of a vertex v is the sum of all delay-dependencies $B_{D(st)}(v)$ of all other node pairs (s,t). Formally, the delay betweenness centrality is:

$$B_D(v) = \sum_{s,t:s \neq t \neq v} B_{D(st)}(v) = \sum_{s,t:s \neq t \neq v} \frac{nst_{st}(v)}{nst_{st}}.$$

5.2.3 Dynamic Closeness

Dynamic Closeness of a node is the average *time* it takes from that individual to reach any other individual in the network. Dynamic closeness is based on shortest temporal paths and the geodesic is defined as the time duration of such paths.

Let $d_T(u, v)$ be the length of the shortest temporal path from u to v. Following the definition in (142) I define dynamic closeness as follows.

$$C_T(u) = \frac{1}{\sum_{v \in V \setminus \{u\}} d_T(u, v)}.$$

5.2.4 Dynamic Clustering Coefficient

Dynamic Clustering Coefficient is the sum of the fractions of an individual's neighbors who have been friends among themselves in previous timesteps. That is, the dynamic clustering coefficient measures how many of your friends are already friends.

Let $CF(u_t)$ be the number of friends of u that are already friends among themselves by timestep t. Let $N(u_t)$ be the neighborhood of individual u at timestep t. Then,

$$CF(u_t) = \sum_{i,j \in N(u_t)} \sum_{k=0}^{t-1} E_k(i, j).$$

The dynamic clustering coefficient is then the fraction of u's neighbors who are already friends among themselves.

$$CC_T(u) = \sum_{0 \leq t < T} \frac{CF(u_t)}{|N(u_t)|(|N(u_t)| - 1)}.$$

CHAPTER 6

DIFFUSION MAXIMIZATION

Diffusion maximization is the process by which diffusion of a phenomenon takes place in a network through the entities of the network with the goal of maximizing the diffusing process. Finding entities critical for maximizing the extent of diffusion in a network is one of the fundamental questions related to diffusion in networks. I call this set of critical individuals as *spreaders*. A multitude of definitions and variants of the corresponding function for spreaders and diffusion maximization exist in literature. Following is some of the general to specific interpretations of diffusion and spreaders in literature.

6.1 Problem statement

Definition (Diffusion) $\mathit{Diffusion}(\cdot)$ is a function that gives the overall average extent of diffusion of a spreading process in a network. That is, the expected number of individuals affected by a stochastic diffusion process after a specified number of timesteps. The estimate of the diffusion is dependent on the underlying diffusion model and the structure of the network. $\mathit{Diffusion}_v(\cdot)$ is the expected extent of diffusion in a network, when the diffusion process is initiated by the individual v. Given a diffusion model $\mathcal{M}$ and a distribution of the probability of activation $\mathcal{X}$, I define the diffusion functions as follows:

$$\mathit{Diffusion}_v : \{G, \mathcal{M}, \mathcal{X}\} \rightarrow \mathbf{R}^+$$

$$Diffusion(G) = \frac{1}{|V|} \sum_{v \in V} Diffusion_v(G, \mathcal{M}, \mathcal{X})$$

The extent of diffusion is the commonly used criterion for diffusion maximization objective but other variants of this function have been introduced in the literature with respect to the kind of applications of diffusion processes being studied. Lescovec et al. in (106) define the optimization goal in terms of *detection of a diffusion process* taking place in a network. In this case, the goal is to minimize the time to detection of a diffusion process, the density of population unaffected by the diffusion process, or the size of the population affected before detection. The last goal is exactly the extent of spread objective defined above.

6.2 Maximizing the extent of diffusion in a dynamic network

In this part of my research, I address the problem of maximizing the extent of diffusion in dynamic networks (Section 2.1.2). I show that the problem of finding the optimal set of influential individuals remains NP–hard for dynamic social networks. I provide a $(1 - 1/e)$-approximation algorithm for finding a set of most influential individuals (Section 6.2.2). I experimentally compare the optimal solution with the proposed approximation algorithm using real–world network data (Section 4.1) and show that the approximation algorithm performs comparable to the optimal in practice.

In this work, diffusion process is simulated using two of the most common diffusion models: the Independent Cascade model (75; 76) and the Linear Threshold model (80). The details of the two models are provided in Section 2.3. I compare the resulting extent of diffusion, quantitatively and qualitatively, for both the aggregate and the dynamic networks. I experimentally

show that for independent cascade model, diffusion in the aggregate network overestimates the actual extent of diffusion taking place in the dynamic networks. Whereas, for the linear threshold model, the aggregate network underestimates the actual extent of diffusion measured in the dynamic networks. Moreover, I show that the set of influential individuals derived from the aggregate network model may be completely different from the one obtained from the explicitly dynamic analysis. Hence, it shows that ignoring the time order of interactions in general gives results that do not relate to the actual spreading process taking place in the population. Algorithms that explicitly address the dynamic aspects of the data are necessary and must be used where the data on the timeline of interactions is available and I provide such an algorithm.

6.2.1 Problem Statement

I now formally define the *Dynamic Influence Maximization* problem:

Given: a dynamic social network $G = \langle G_1, \ldots, G_T \rangle$ over T timesteps and an integer k and a diffusion model $\mathcal{M}$.

Find: the initial set of individuals A_0 of size $|A_0| = k$

Objective: maximize the extent of spread initiated by A_0 as measured by the expected size of the set of active individuals $f(A_0) = |A_T|$ in G after T timesteps under the diffusion model.

I consider this problem for both the independent cascade model (Section 2.3.1) and the linear threshold model (Section 2.3.2). I show that the Dynamic Influence Maximization problem is NP–hard for both models but it can be efficiently approximated using a simple greedy approach.

6.2.2 Complexity

Kempe *et al.* (92) show that the static Influence Maximization problem is NP-hard using reductions from the Set Cover problem for the Independent Cascade model and from the Vertex Cover problem for the Linear Threshold model. Both reductions are to instances of the Influence Maximization problem where the diffusion happens in one timestep. Thus, the static instance of the Influence Maximization problem is a special cases of the Dynamic Influence Maximization problem with $T = 1$ timestep. Therefore, the Dynamic Influence Maximization problem for both spread models remains NP–hard.

6.2.3 Approximation Algorithm

Here I present an approximation algorithm for both diffusion models. I use the notion of *submodular functions* described in 2.4.

6.2.3.1 The Overall Approach

For the Dynamic Influence Maximization problem the goal in the terminology of the submodular functions is to find a set S of individuals such that $f(S)$ is maximal, that is adding any other individual to S will not increase the value of $f(S)$. This is an NP–hard problem for any submodular function by reduction from the Hitting Set problem (52; 122). A simple greedy hill climbing algorithm (122) gives a $(1 - 1/e)$-approximation of the optimal solution. I use a similar greedy approach to achieve a $(1 - 1/e)$-approximation and show that you cannot do any better for the specific case of the Dynamic Influence Maximization problem. I need to show that the expected number of active individuals is a monotonic submodular function of the set of initiating individuals.

Theorem 1. *If $f(.)$ is a submodular and monotonic function, then the greedy hill climbing algorithm gives a $(1-1/e)$–approximation of the optimal solution (122).*

The definition of the approximation ratio assumes that the function $f(\cdot)$ is either already known or is easily computable. However, for the stochastic diffusion process there is no such known function. The best that can be done is to simulate the stochastic diffusion process on sample inputs several times and to estimate the expected set of active individuals A_T at time T using the Monte Carlo method. Kempe *et al.* (92) show that for some $\epsilon > 0$, there is a $\gamma > 0$ such that using $(1+\gamma)$-accurate estimates of f results in a $(1-1/(e-\epsilon))$-approximation of the static Influence Maximization problem. This provides an approximation scheme that achieves an approximation ratio arbitrarily close to the $1-1/e$. I use this approximation approach for the Dynamic Influence Maximization problem with the two diffusion models. In the next two sections I prove that the influence maximization function $f(\cdot)$ is submodular and monotonic for the two diffusion models. With the following results, it naturally follows that the above approximation algorithm gives within $(1-1/e)$ of the optimal solution.

6.2.3.2 The Independent Cascade Model

Using the above approach I prove that for an arbitrary instance of the independent cascade model, the resulting influence function $f(\cdot)$ is submodular as well as monotonic. That is, the increase in the expected number of activated individuals for any set $S \subseteq T$ and an individual v_i is $f(S \cup v) - f(v) \geq f(T \cup v) - f(v)$. The dynamic network graph is defined by the natural ordering of individual to individual connections based on the timesteps. For each timestep t, to find whether an individual v_t is successful in activating an individual w_t I flip a coin for

the edge (v_t, w_t) with the bias p_{v_t,w_t}. The probability $p_{v,v}$ for the edges of the form (v_t, v_{t+1}) for all timesteps t and all individuals is set to 1. Also, I only consider the progressive case in this work, therefore, when an individual v_t becomes active at time t, it will remain active for all subsequent timesteps. The coin flips can occur earlier than the processing of a given edge. Thus, I flip all the coins for G at the beginning of the diffusion process and use the results when the corresponding edge needs to be processed. I call the edges for which the coin flips indicate success "live" and the rest are "blocked". All edges of type (v_t, v_{t+1}) have probability 1 of activation, therefore, they are live by default. Note, that an individual u will be active if only if there is a path from some individual in A_0 to u consisting entirely of live edges. By fixing the outcomes of coin flips and selecting a set A_0 at the beginning, I can easily find the set of active individuals at the end of the spreading process by tracing a path of live edges from each individual in A_0.

Let X be a sample set of outcomes of coin flips for all edges. I define the activation outcome of an active individual v by $R(v, X)$, which is the set of individuals activated by v. Given a sample set of coin flips X and an initial set A_0, I denote the number of active individuals at the end of the spreading process by $f_X(A_0)$. Note, that $f_X(A_0)$ is the size of the union of all $R(v, X)$ where $v \in A_0$. The overall (incomputable) influence function $f(A_0)$ is the expectation of $f_X(A_0)$ over all possible sets of coin flips X. I show that $f_X(A_0)$ is submodular.

Lemma 2. *$f_X(A_0) = \sum_{v \in A_0} R(v, X)$ is a submodular function.*

Proof. Let S and T be two sets such that $S \subseteq T$. Consider the quantity $f_X(S \cup v) - f_X(S)$. By definition

$$\begin{aligned} f_X(S \cup v) &- f_X(S) \\ &= |\bigcup_{u \in S \cup v} R(u, X) - \bigcup_{u \in S} R(u, X)| \\ &= |R(v, X) - \bigcup_{u \in S} R(u, X)| \\ &\geq |R(v, X) - \bigcup_{u \in T} R(u, X)| \\ &= f_X(T \cup v) - f_X(T) \end{aligned}$$

That is, the number of elements in $R(v, X)$ that are not already in $\cup_{u \in S} R(u, X)$ is at least as big as the number of elements not in the larger set T. This inequality shows that $f_X(A_0)$ is submodular for the independent cascade model. Moreover, $f_X(\cdot)$ is monotone, since $f_X(S \cup v \notin S) \geq f_X(S) + 1$. □

6.2.3.3 The Linear Threshold Model

For the linear threshold model as well, I show that the resulting diffusion function $f(\cdot)$ is submodular. Unlike the independent cascade model, a general influence function for the linear threshold model based on fixed choices of thresholds θ_v for each individual is not submodular. This is due to the *accumulated* effect of individuals on other individuals. As defined above, each individual u_t has an influence $b_{v,u} \geq 0$ on each of its neighbor v_t, such that the total influence on v_t at time t is $\sum_{(u_t, v_t) \in E_t} b_{v,u}$. Here, I use the approach of (92), to resolve the non–submodularity of the spreading function. At time t, for each individual v_t I pick at most one edge incident

on v_t from any neighbor individual u_t with probability $b_{v,u}$, and do not pick any other incident edge with the probability $1 - \sum_{(u_t,v_t)\in E_t} b_{v,u}$. If u_t is not connected to v_t then $b_{v,u} = 0$. An edge (u_t, v_t) is called "live" if it has been picked for activation and all other edges connected to v_t are called "blocked". Note, that unlike in the independent cascade model, where each edge is live independent of the other edges and an individual can have multiple live edge paths, in this process, each active individual belongs to a unique live edge path.

In this modified version of the linear threshold model, the submodularity property of the function $f(\cdot)$ holds as in the independent cascade model. I pick a sample distribution X of live/blocked edges, and define $R(v, X)$ as the set of individuals reachable through live edge paths from v. Thus, for the initial set A_0, it follows that $f_X(A_0)$ is the cardinality of $\cup_{v\in A_0} R(v, X)$, which is the submodular function of A_0. Finally, since $f(\cdot)$ is an expectation of $f_X(\cdot)$ over all X, it is a non–negative linear combination of the functions $f_X(\cdot)$ and hence is also submodular. Moreover, $f_X(\cdot)$ is monotone, since $f_X(S \cup v \notin S) \geq f_X(S) + 1$.

Now, I show that the original linear threshold model is equivalent to the modified version.

Lemma 3. *The distribution over active sets obtained by running the linear threshold process to completion starting from A_0, is equivalent to the distribution of active sets reachable by A_0 under the random selection of live edges defined above.*

Proof. Consider the active set A_t at time t, and let v be an individual which is not yet active. This individual becomes active at $t+1$ if the set of individuals $A_t - A_{t-1}$ pushes it over the threshold θ_v. Thus, the probability of activation of each individual can be defined as:

$$Pr(\text{activation of } v) = \frac{\sum_{u \in A_t - A_{t-1}} b_{v_t,u_t}}{1 - \sum_{u \in A_{t-1}} b_{v_{t-1},u_{t-1}}}.$$

Similarly, for the redefined linear threshold model using random live–edge pattern, I run the diffusion process by revealing the identities of live edges gradually as follows. At time t, A_t is the set of active individuals. I pick an individual v_t outside of A_t and check whether v_t has a live edge connecting it to any individual in A_t. If yes, then v_t becomes active in time $t+1$. The activation of v_t at $t+1$ shows that the live edge of v_t comes from the set $A_t - A_{t-1}$. This is equivalent to the probability

$$\frac{\sum_{u \in A_t - A_{t-1}} b_{v_t,u_t}}{1 - \sum_{u \in A_{t-1}} b_{v_{t-1},u_{t-1}}}.$$

Thus, the probability of activation through the live edge path is equivalent to the probability of activation of an individual obtained by the original diffusion process. Hence, by induction over the time steps, the distribution over active sets obtained by linear threshold model is the same as that of the live edge path process. □

This completes the proof of the submodularity of $f(\cdot)$ for both diffusion models. Thus, as outlined in Section 6.2.3, a simple greedy heuristic of selecting one initially active individual, at a time, that provides the largest marginal gain, gives a $(1 - 1/e)$-approximation.

In the next section, I show the experimental results of evaluating the performance of the approximation algorithm in practice, as well as compare the static and the dynamic influence maximization problem solutions.

6.2.4 Experiments

6.2.4.1 Types of experiments

I perform three sets of simulations using each diffusion model.

1. I compare the optimal solution to the results of the greedy approximation algorithm for the dynamic network.
2. I compare quantitatively the extent of diffusion given by the approximation algorithm in the dynamic versus static/aggregate version of the dynamic networks.
3. I qualitatively compare the extent of diffusion in the dynamic networks using the initial set obtained by solving for diffusion maximization in static networks.

Note, for both the optimal and greedy approximation, the function $f(\cdot)$ needs to be computed. It is an open question to compute this quantity exactly by an efficient method, but very good estimates can be obtained by simulating the random process. More specifically, I simulate both the process for $T = 10000$ trials for each initial set. This value is chosen after simulating the diffusion process for several thousands runs and finding the least T beyond which the extend of spreads obtained are comparable to T.

Following is the sketch of the the **optimal algorithm**.

1. Set the size of the initial set to k.

2. For each possible subset of individuals of the initial set size:

   (a) Simulate the spreading process for T trials with each spread model.

   (b) Mark each individual not in the initial set as activated if that individual was active in more than half of the trials.

   (c) Pick the initial set that results in the maximum number of active individuals. This is the optimal set that maximizes the spread.

Algorithm 2 formally explains the optimal method.

Algorithm 2: Optimal Influence Maximization

Set size of the initial set: $|S| = k$;
Initialization: $S = \emptyset$;
for all $S \in \binom{N}{k}$ **do**
  Initialize: $A_T = \emptyset$;
  repeat
    Estimate the extent of diffusion A_T in $G(V, E)$ using S as the initially active set.
  until T *trials*
  if $v \in V$ active for more than half of the Trials **then**
    $A_T = A_T \cup \{v\}$
  end if
end for
return S with $\max_{\binom{N}{k}} |A_T|$.

Following is the sketch of the **greedy approximation algorithm**.

1. Simulate the diffusion process independently with each individual i.

2. Calculate the expected extent of diffusion from each individual i.

3. Pick the individual i that gives the maximum extent of diffusion.

4. Remove that individual and all those individuals that were activated by it in more than half of the trials from the population.

5. Repeat steps (1–4) until k most influential individuals have been picked, where k is the initial set size.

Algorithm 3 formally explains the greedy approximation method.

Algorithm 3: Greedy Approximation

Initialization: $S = \emptyset$;
while $|S| \leq k$ **do**
  $S := S \cup v : \arg\max_v \{f(S \cup v)\, \forall v \in N\}$
end while
return S

Following are the details of the experiments.

Optimal vs Approximate: For small datasets I simulate the approximate as well as the optimal diffusion processes using the two models. The optimal solution is computed through exhaustive search of k most influential individuals out of the N in the network. I compare the difference in the extent of diffusion for the initial set sizes of $k = 3$, 5, and at

most 10 individuals. Exhaustive search is not feasible beyond $k = 10$ even for very small networks ($N \leq 30$).

Dynamic vs Static - Quantitative: For the second set of experiments I compare the extent of diffusion in the static and the dynamic networks.

Dynamic vs Static - Qualitative: For the third set of experiments I compare the diffusion in static and dynamic networks qualitative. I compare the sets of activated individuals in the static and the dynamic networks when the same initial set of individuals is used. I take the initial sets obtained by the approximation algorithm on the static network. I use those those sets as the initial sets for the spreading processes in the dynamic network.

6.2.4.2 Experiment results

6.2.4.2.1 Data sets

The experimental results are based on the grevys, onagers, Portland, and Co–authorship networks described in Section 4.1.

Following are the results for each diffusion model.

6.2.4.2.2 Independent Cascade

Parameter settings I compare the extent of diffusion for the initial set size k of 3 to at most 10. It is important to mention here that for $k = 1$ the optimal and approximate methods have identical results. For $k = 2$ the two solutions are too close to compare. Hence, I start with $k = 3$ as the size of the initial set. For $k > 10$ the exhaustive optimal method is infeasible to evaluate. Moreover, by $k = 10$, the initial set is too large already and all the individuals

get activated very easily from both the methods, making the distinction between the two hard to compared. The probability of activation on each edge is set uniformly at random. These experiments are repeated for a range of probability values. After rigorously and exhaustively trying a range of parameters, I find that at the upper threshold of activation, when the entire population gets activated is at $pr > 0.2$. While for $pr \leq 0.05$ the extent of diffusion is trivially small and is not very interesting to analyze. Here I show the results for $pr = 0.2$. Each stochastic diffusion simulation is run for over a 1000 trials. The number of iterations 1000 was chosen as the one to be sufficient for the convergence of the Monte Carlo simulations.

Optimal vs Approximate For this set of experiments I use small data sets, as it is infeasible to compute the optimal solution on large networks. These experiments are performed on grevys and onagers datasets, as each has less than 30 individuals. Both optimal and approximate(greedy) methods are simulated on the two networks. The first set of simulations are on the grevys dataset. I compare the results of initial set sizes of 3, 5, 7 and 10 individuals. I could not get the optimal solution for the initial set size of 10 since the number of subsets of size 10 for the dataset, $\binom{28}{10} > 2.8^{10}$, is too big and the exhaustive search is not feasible. However, the value of the optimal solution for 10 initial individuals is at least 27 (which is the value of the optimal solution for 7 initial individuals) and is more likely to be 28, the entire population.

The second set of simulations were conducted on the onagers data. Initial set sizes are 3, 5, and 6. I did not need to consider larger initial sets since 6 initial individuals activate the entire population. The results for both datasets are shown in Table III. The results are based on the Independent cascade model with probability of diffusion $pr = 0.2$.

The objective in animal populations is to test the hypothesis of how behavior spreads. Thus, I can then compute predictions with the observed behavior.

TABLE III: Optimal Vs Approximate–grevys and onagers

Data Set	Pop.size	Init set size	Opt	Approx
grevy's	28	3	20	18
		5	23	20
		7	27	22
		10	–	25
onagers	29	3	26	26
		5	28	28
		6	29	29

The experiments show that the results obtained using the approximation algorithm are within 80% accuracy of the optimal solution. This is better than the $(1-1/e) \approx 0.632$ accuracy predicted by the theoretical analysis.

Dynamic vs Static: Quantitative Comparison In the second set of experiments, I compare the extent of diffusion in the dynamic and the static networks using the approximation algorithm only. I use the Portland and DBLP data sets for this set of experiments. The size of these networks are 211 and 1374, respectively. The initial sets are of the sizes 3, 5, and 10 individuals. The probability of activation is set to be the same for all edges and is 0.2. The results for the Portland and DBLP datasets are given in Table IV. I observe that the extent of

spread resulting from the same size of the initial set of individuals is much greater in the static network than in the dynamic network. The difference is much more pronounced in the DBLP dataset, the sparser of the two.

TABLE IV: Dynamic vs Static: Quantitative and Qualitative

Data set	Pop.size	Init set size	Static	Dynamic	Dynamic using Static init. set
Portland	211	3	186	96	40
		5	211	125	65
		10	211	181	-
DBLP	1374	3	58	7	4
		5	65	11	9
		10	76	22	18

Dynamic vs Static: Qualitative Comparison In the last set of experiments I compare the extent of diffusion in the dynamic network obtained using the initial set resulting from running the approximation algorithm on the static network. I use the Portland and the DBLP datasets for this set of experiments as well. The initial set sizes are 3 and 5 for Portland, since 5 initial individuals activate the entire network, and 3, 5, and 10 for DBLP. The probability of activation is set to 0.2. The results are shown in Table IV.

I take the best (approximate) set obtained in the static network and use it to initiate the spread in the dynamic network. However, the resulting extent of diffusion is much less than

the extent of diffusion obtained using the approximation algorithm directly on the dynamic network. This experiments demonstrates that the optimal set of initial individuals is not the same in the static and the dynamic networks. Moreover, a set that performs well in the static network may perform very poorly in the dynamic network.

6.2.4.2.3 Linear Threshold

Parameter settings For all the experiments I used initial sets of sizes up to 10 or the size that resulted in the entire population being activated. The threshold probabilities were set randomly for each individual. Each stochastic diffusion simulation is run for over a 1000 trials. The number of iterations 1000 was chosen as the one to be sufficient for the convergence of the Monte Carlo simulations.

I perform the same three sets of the experiments using the linear threshold model of spread.

Optimal vs Approximate I calculate the extent of diffusion using the optimal exhaustive search and the greedy approximation algorithm on the grevys and the onagers datasets. The results for this set of experiments are presented in Table V. The results demonstrate that the approximate algorithm is 100% accurate on this dataset for the linear threshold model and, thus, can be used in practice for a good estimate of the extent of diffusion.

Using the same data used in the experiments for the independent cascade model, I observe that the extent of spread in linear threshold model is greater with the same size of the initial set. For grevys, population size of 28, starting with 6 initial individuals results in the activation of the entire population. For onagers, with the population size of 29, having only 3 individuals

TABLE V: Optimal Vs Approximate–grevys and onagers

Data Set	Pop.size	Init set	Opt	Approx
grevys	28	3	25	25
		5	27	27
		6	28	28
onagers	29	3	29	29

in the initial set results in the activation of the complete population, both in the optimal and the approximate solutions.

Dynamic vs Static: Quantitative Comparison To compare the extent of diffusion obtained by the approximation algorithm in the dynamic and the static networks, I use the complete Portland data set of 1588 individuals and DBLP data of 1374 individuals. The initial set sizes are 3, 5, and 10 individuals. None of the initial sets resulted in activation of the entire population but the 10 individuals' set came very close to it. The results are shown in Table VI. As in the independent cascade case, the extent of diffusion in the dynamic and the static networks is very different. However, unlike in the independent cascade, here the extent of diffusion in the dynamic network is greater than in the static.

Dynamic vs Static: Qualitative comparison In the last set of experiments with the linear threshold model I compare the extent of diffusion in the dynamic network obtained using the initial set resulting from running the approximation algorithm on the static network. Again,

TABLE VI: Dynamic vs Static: Quantitative & Qualitative

Data set	Pop.size	Init set size	Static	Dynamic	Dynamic using Static init. set
Portland	1588	3	665	1569	25
		5	724	1571	30
		10	1396	1576	71
DBLP	1374	3	961	1165	62
		5	971	1167	70
		10	993	1172	84

I use the Portland and the DBLP datasets with the initial set sizes 3, 5, and 10 individuals. The results are shown in Table VI.

As in the independent cascade case, the best initial set obtained for the static network performs extremely poorly in the dynamic network. Thus, also for the linear threshold model, to find the right set of initial individuals we must perform the analysis on the dynamic network.

6.2.5 Conclusions and Future Work

In this work I defined the problem of finding the set of individuals that maximizes the extent of diffusion in an explicitly dynamic social network. I show that this Dynamic Influence Maximization is NP–hard for two common stochastic diffusion models. I propose a $(1 - 1/e)$-approximation algorithm that performs well in practice.

I compare the results of Influence Maximization in the dynamic and the static representations of social networks. The extent of diffusion in the two types of representations is different and the difference depends on the diffusion model and the dynamic topology of the networks.

Moreover, the individuals that maximize the extent of diffusion in the dynamic and static graphs are different.

Thus, the aggregate static representation not only gives an inaccurate information about networks, using it to make predictions about the diffusion of ideas, behavior, or diseases in a social network may lead to absolutely incorrect conclusions. Hence, it is crucial to use dynamic network graphs where the interactions among individuals is sensitive to the time order.

However, in many instances, we must provide an estimate of the extent of spread and some idea of the individuals that might be its facilitators even when the explicit dynamic information about the interactions is not unavailable. Thus, we need to come up with some measures that give theoretical estimates of the spread in dynamic networks based on static graphs.

There are many other possible extensions of this research. In this work I focused on identifying the facilitators of a spreading process. A very important but computationally a much harder problem is to identify the set of individuals that effectively prevent diffusion in a network. Such individuals would be the best candidates for vaccination in case of a spread of a disease. In this work, I simplified the models by considering only the progressive case, where an individual remains active once it has been activated. This work naturally extends to non-progressive models. Finally, diffusion models other than the two discussed here can be used to get a better comparison between the dynamic and the static network representations.

In conclusion, understanding how processes spread in networks is an important and growing area within many application domains. We must find the right models, measures, and algorithms to study explicitly dynamic networks. My work demonstrates the importance of

dynamic information and takes the first steps in analyzing the behavior of spreading processes in explicitly dynamic data. In the next section I explore just how much dynamic change affects the estimates of spread in dynamic networks.

6.3 The Impact of Structural Changes on Predictions of Spreading in Networks

In a typical realistic scenario, there exist some past data about the structure of the network which are analyzed with respect to some possibly future spreading process, such as behavior, opinion, disease, or computer malware. How sensitive are the predictions made about spread and spreaders to the changes in the structure of the network? I investigate the answer to this question by considering seven real–world networks that have an explicit timeline and span a range of social interactions, from celebrity sightings to animal movement. For each dataset, I examine the results of the spread analysis with respect to the changes that occur in the network as the time unfolds as well as introduced random perturbations. I show that neither the estimates of the extent of spread for each individual nor the set of the top spreaders are robust to structural changes. Thus, analysis performed on historic data may not be relevant by the time it is acted upon.

Prediction of the course and extent of processes spreading in social networks and identification of the top spreading individuals have become important issues in many contexts, from epidemiology to viral marketing. In a typical realistic scenario, there exists some past data about the structure of the network which is analyzed with respect to the future spread of some process, such as behavior, opinion, disease, or computer malware. The important tasks are (a)

estimating the possible number of affected individuals once the process starts, (b) predicting who those individuals may be, (c) identifying the most effective spread initiators, and (d) identifying individuals that can effectively block the spread of the process. However, by the time the outcomes of such analysis are acted upon, such as by selecting marketing targets or vaccination candidates, time has elapsed and the network structure may have changed significantly from what was used for the initial analysis. The effectiveness of the marketing scheme (97) or epidemiological response may be sabotaged if analysis results are sensitive to such structural changes.

In this work, I focus mainly on the tasks of estimating the extent of spread and identifying the top spreaders. These are the individuals that, when used as the start of a spread, affect the largest proportion of the population. I ask how sensitive the predictions made about spread and spreaders are to changes in the structure of the network. To answer this overall question, I formulate three specific questions:

1. **How much does the relative spreading ability of individuals change?** Most algorithms for estimating the extent of spread and for identifying the top spreaders fundamentally rely on estimates of the spreading ability of each individual. Thus, it is important to know how reliable those estimates are both in terms of actual numbers and in the ranking they impose on individuals.

2. **How much does the identity of the top spreaders change?** While the first question asks whether our predictions hold for all the individuals in the population, this question

focuses only on the top spreaders. The set of top spreaders may be more or less robust than the rest of the individuals, yet it is typically more critical to the impeding action.

3. **How does the spreading ability of the top spreaders from the past compare with that of the top spreaders after the change?** While the identity of the top spreaders may change as the network changes, the previous set of top spreaders may still perform well. Although it may not be the best set of top spreaders in the new network, I ask whether it is good enough.

I investigate the answers to these questions by considering seven real–world networks (Section 4.1) that have an explicit timeline and span a range of social interactions, from celebrity sightings to animal movement. For each dataset, I examine the results of the spread analysis with respect to the changes that occur in the network as time unfolds, as well as introduced random perturbations (Section 6.3.2). I show that neither the estimates of the extent of spread for each individual nor the set of the top spreaders are robust to structural changes (Section 7.4). Thus, analyses performed on historic data may not be relevant by the time they are acted upon if the network changes substantially in the meantime.

6.3.1 Preliminaries

6.3.1.1 Static and Dynamic Networks

Recall from Section 2.1 the definition of the aggregate and the dynamic networks.

Definition A *dynamic network* is a time-series of labeled graphs $\mathcal{G} = \langle G_1, \ldots, G_T \rangle$, where $G_t = (V_t, E_t)$ is the graph of interactions taking place at timestep t. $V_t \subseteq V$ is the set of

individuals observed at timestep t, and an edge (v_1, v_2) exists in E_t if v_1 and v_2 were observed interacting in that time period.

The question of how much actual time should be quantized into a 'timestep' is beyond the focus of this work. However, I note that many types of social systems have natural time quantizations such as hours or days. Recently, Caceres et al. gave a rigorous formulation for choosing scales of quantization of temporal data (39). Figure 3 shows a dynamic network of interactions between four individuals.

One can aggregate a range of timesteps in a dynamic network into a single *static* graph, or an *aggregate network*. This is done by accumulating vertices and edges present in a given range of timesteps. Note that aggregating the entire range of timesteps results in a traditional social network, i.e. a single graph of all edges without any temporal information.

Definition Given a dynamic network $\mathcal{G}$, a *static*, or an *aggregate network* $G_{[i,j]}$ is the accumulation of vertices and edges of the range $[i, j]$ of timesteps of $\mathcal{G}$:

$$V(G_{[i,j]}) = \bigcup_{i \leq t \leq j} V_t \qquad E(G_{[i,j]}) = \bigcup_{i \leq t \leq j} E_t$$

If $i = 1$ and $j = T$, then the resultant aggregate network $G_{[1,T]}$ is equivalent to the traditional social network G.

In traditional approaches, network analysis is performed on the aggregate network rather than the original explicitly dynamic network. One might immediately recognize a problem here: paths in an aggregate network might not correspond to valid propagation paths in the

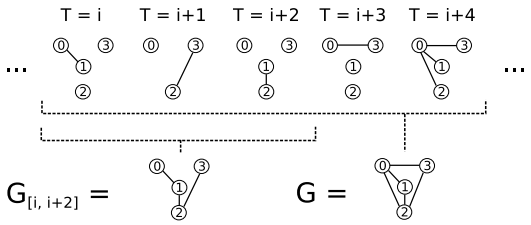


Figure 3: A dynamic network (top), an aggregated network (bottom left), a traditional social network (bottom right).

original dynamic network. Since the propagation of a rumor or virus must proceed along a sequence of edges that are increasing in time, and since an aggregate network has no temporal information, modeling the spread of a process without considering time can lead to grossly inaccurate results (82).

6.3.1.2 Network Diffusion

A process spreading in a network can be described formally using many models of transmission. For this paper I use a model that has been extensively studied in the context of social networks and viral marketing, the *Linear Threshold* model (80). This model is extensively described in Section 2.3.2.

6.3.2 Methodology

To answer the three questions posed in Section 6.3, I recall a typical scenario for network analysis: the network is observed for some time, then is analyzed as one aggregated social network. The results of the analysis are then deployed in the network, which has changed in

the meantime. In order to determine the effect of the changes on the results of the historical analysis, I use the following overall experimental template:

1. Consider (part of) the dynamic network as an aggregate "historical data" network G_h.
2. Perform analysis on G_h: estimate the spreading ability of each individual and identify the top spreaders.
3. Extract the changed network G_f. I do this in two ways, both by considering the actual future segment of the network and by randomly perturbing the network to introduce changes.
4. Perform analysis on G_f: estimate the spreading ability of each individual and identify the top spreaders. Compare the results of the analysis on G_h with the results on G_f. This will answer questions 1 and 2: how much does the relative spreading ability of each individual and the identify of the top spreaders changes.
5. Recall that the third question was how well do the top spreaders from the past perform in the future relative to the best spreaders of the future. To answer this question, I simulate the spread in G_f (changed network) starting both from the top spreaders of G_f and the top spreader of G_h and compare their performance.

As I have pointed out, there are at least two ways to consider the changes that may happen in the structure of the network as the network evolves with time. First, I may look at the actual dynamic network and aggregate a portion of it into an initial "historical" segment used for analysis. Subsequent segments are designated "future" data and used to validate the results

of the analysis. The changes in the structure of the network then are the actual changes that are recorded in the data. This is the approach of *temporal cross-validation*, which is an adaptation of the well-known statistical technique.

Temporal cross-validation involves dividing the timeline of the dynamic network into several segments and aggregating each segment into a single graph. Any analysis technique performed on the graph of one segment should then produce similar results in another segment, given that it is the same underlying network (and presumably the same dynamics) being modeled. If this is not the case, then I can conclude that either the analysis technique is not robust, or that the underlying dynamics of the network are changing. In either case, the particular analysis technique is then unlikely to produce actionable results. For this study, for temporal cross-validation I divide each dynamic network into five segments of equal duration.

While the temporal cross-validation approach examines the robustness of the analysis with respect to actual recorded network changes, these changes may not be representative of the changes that may, *in principle,* happen in the network. Thus, for the second way to introduce changes into the structure of the network, I take the aggregate network and randomly remove and add edges to introduce possible perturbations and provide the answer to the *expected* robustness of the analysis.

6.3.2.1 Experimental Setup

I initiate the experiments with the "historical" data network G_h. This is the aggregate network of either a segment of or the entire dynamic network. That is, $G_h = (V_h, E_h)$, where

V_h are the nodes present in timesteps $i \leq t \leq j$, $E_h = \bigcup_{t=i}^{j} E_t$. Here i and j are either the first and the last timestep of a particular network segment or $i = 0$ and $j = T$.

I define an *objective function* for each vertex v, denoted $spread(v)$, which is the proportion of the population that v eventually activates if it starts as the only active node in the network. This is determined, as in earlier approaches (92), by Monte Carlo simulations. I simulate the linear threshold spreading process on the network starting with v as the only active node, for each node $v \in V$. The threshold values Θ_v are chosen randomly for each iteration, and all three variants of the linear threshold spreading process are simulated – aggregate spread on G_h, and dynamic spread with and without memory on the underlying dynamic network $\langle G_i, \ldots, G_j \rangle$.

In all cases, I simulate the spreading process for $j - i + 1$ timesteps. I used 500 iterations of each spread simulation, which was sufficient to produce consistent results. After simulating the spread from each individual, in each iteration, I note the number of activated individuals $\sigma(v) = |A_f|$. The overall spreading capacity of v is then the average over all iterations of the size of the active set proportionally to V_h: $spread(v) = \frac{1}{500} \sum \sigma(v)/|V_h|$.

I then rank the individuals in the order of their spreading ability: $spread(v_1) \geq spread(v_2) \geq \cdots \geq spread(v_{|v_h|})$. I call the first k individuals in this order the "top k spreaders". While as a set, this may not be the best *set* of k individuals from which to start a spreading process, individually they are the top k performers. I investigate whether they remain in the top k as the network changes.

As mentioned earlier, I obtain the changed network of "future" data G_f in two ways. In the temporal cross-validation setting this is one of the segments that follows the segment of G_h.

That is, $G_f = (V_f, E_f)$ is the aggregate network of a latter segment. For random perturbations, I remove a fraction p of existing edges uniformly at random and add the same number of edges that were not in the network, preserving the overall number of edges. I use the range of $p = \{0.05, 0.1, 0.3\}$, that is, changing $5, 10$ and 30 percent of the edges. Note, that if the density of a network is d, one cannot change more than $1 - d$ fraction of the edges in this scheme.

To answer the first question about the change in the relative spreading ability of individuals, I measure the correlation between the orderings imposed on the individuals by their spreading ability. That is, given the ordering imposed by $spread(v)$ function in G_h and G_f, I measure the difference in the orderings using Spearman's rank correlation coefficient (149). For the second question, I measure the difference in the identity of the top k spreaders for $k = \{5, 10\}$ by measuring the Jaccard similarity (88) of those sets in networks G_h and G_f.

Finally, I answer the last question by taking the top k spreaders from G_h and using them as the set of initially active individuals in G_f. I denote the average proportion of activated individuals as APX. I compare that number to the same process repeated with the initial set being the top k spreaders from G_f itself. I denote the average proportion of activated individuals in the latter case by OPT. I then measure the performance of the historical top spreaders in the changed network as the fraction APX/OPT.

I perform our analysis across different datasets representing a wide range of types of interactions. A summary of results obtained from my analysis is presented in Section 7.4.

6.3.3 Datasets

For these experiments, I used real dynamic networks spanning the range of interactions from animal behavior to celebrity sightings. Following are the dynamic networks used in this analysis. Animal Social Networks: grevys zebra and plains zebra animal networks, Mobile P2P: MIT Reality Mining and Haggle InfoComm network, Enron Email Network, IMDB Photo network. Section 4.1 provides the description of all these networks. Table I summarizes the basic statistics of the datasets.

6.3.4 Results

I now describe the results of the experiments and analysis that address each of the three questions, in turn, posed in Section 6.3. In all the figures, the datasets are shown in the order of increasing network density.

6.3.4.1 The change in the relative spreading ability of individuals

Recall that to answer the question of how valid the predictions about the spreading capacity of each individual in the network are as the network changes over time, I compare the rankings of the individual's spreading capacity. I calculated the Spearman correlation coefficient between those rankings in the original and changed networks. Figure 4 shows the correlations of the ranking by static linear threshold spread in the aggregate network of a given segment versus the rankings by the static and the two dynamic linear threshold spread within the same and all future segments of a dynamic network. For example, the bottom row of the plots shows the static spread ranking in the first segment versus all the rankings in each of the five segments, while the top row shows the ranking of the fourth segment versus the rankings of the fourth and

fifth segments. As expected, the only perfect correlation is between the static spread ranking with itself within the same segment. What is unexpected, however, is how little correlation there exists between the dynamic and the static spread models and how quickly the correlation deteriorates as time unfolds.

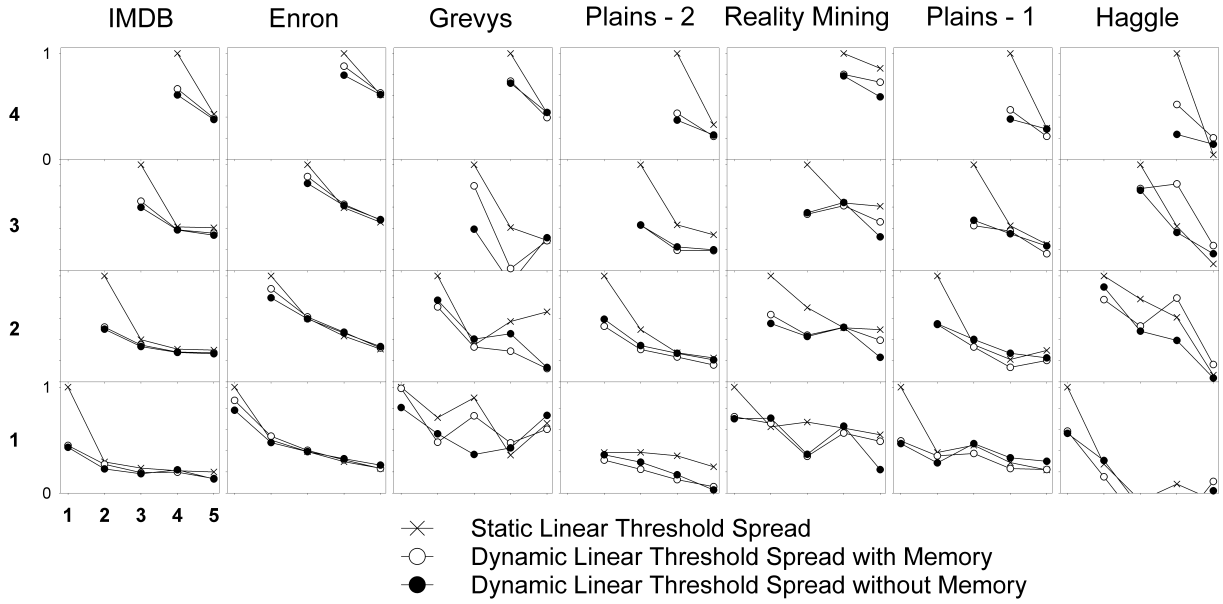


Figure 4: Spearman's correlation coefficient comparing the ranking of individuals (by estimated spreading capacity) across different segments. Comparisons are made both within the same segment and between current and future segments. Results for three different spread models are shown.

To measure how much the network actually changes with time, I calculated the distance between the aggregate networks of every two segments. I measure this distance as the complement

of the Jaccard similarity of the edge sets of the two networks. Figure 5 shows the scatter plot of the distance between segment networks versus the Spearman correlation coefficient between the spread rankings of the individuals in the two networks. The surprising feature of the plots is how little correspondence there is between the network similarity and the consistency of the rankings. Moreover, note that in most datasets the distance between any two segments is at least .4 and often reaches .8, which means networks typically change very fast with time.

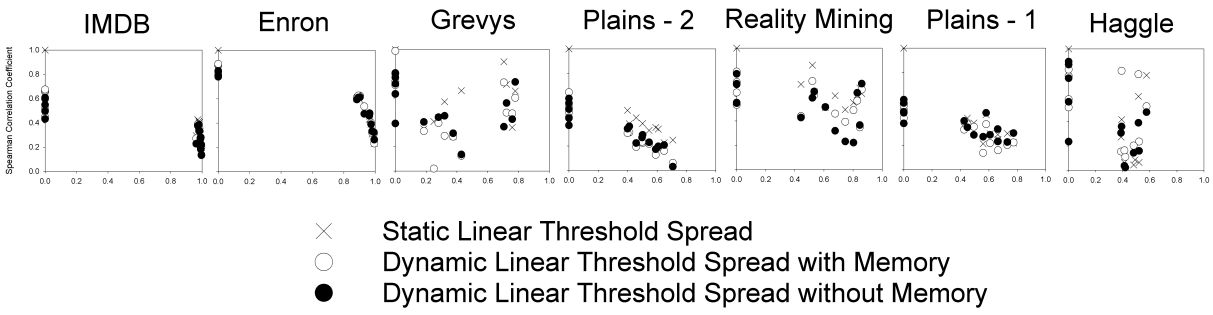


Figure 5: Spearman correlation coefficient between the ranking of individuals (by their estimated spreading capacity) as a function of the dissimilarity between the underlying networks. Dissimilarity (x-axis) is measured as the complement of the Jaccard similarity on the edge sets of the networks.

The result of random perturbations of the edges in networks are shown in Figure 6. Here I only compare the rankings of the static linear threshold spread on the aggregate networks before and after the perturbation. I do not consider the dynamic spread models since the perturbations are not explicitly dynamic and I do not control the timesteps in which the random edges are perturbed. These results also show that the quality of the predictions of the spreading capacity of the individuals deteriorates rapidly with the increase in the amount of perturbation.

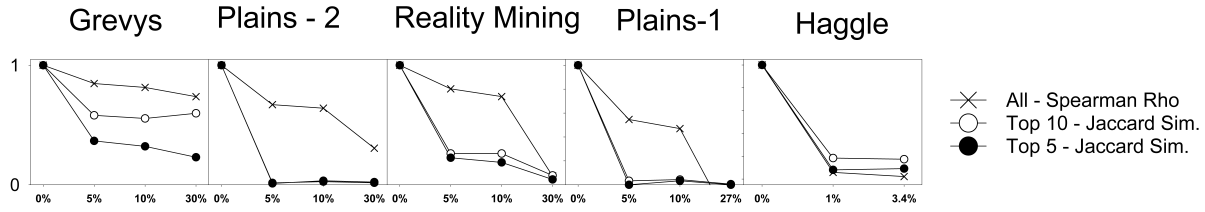


Figure 6: Spearman's correlation coefficient between the ranking of all individuals and the Jaccard similarity between the 5 and 10 top ranked individuals in the original and perturbed networks, both as functions of the percent of randomly perturbed edges in the network. The datasets are ordered by increasing density. The amount of perturbation cannot exceed the complement of the density.

6.3.4.2 The change in the identity of the top spreaders

The second question I asked was whether, despite the fact that overall the relative predictions about the spreading capacity of individuals may not be robust, the identities of the top spreaders remain relatively constant. I compare the sets of the top five and top ten ranked

individuals in the network before and after the spread. I measure the Jaccard similarity of the sets of the top spreaders. Figure 7 shows the similarity of the sets of the top five spreaders between the first segment and all subsequent segments. All other pairs of segments show a similar trend. As the results show, the identity of the top five spreaders changes drastically as time unfolds. Figure 8 shows the scatter plot of the similarity between the top five sets versus the amount of change in the network.

Even in the identity of the top ranked individuals there is little correspondence between the amount of change in the network and the consistency of the results. Recall that the datasets are ordered by their density. There is a possible trend that in sparser networks the top ranked individuals remain more consistently in the top, but before I draw any conclusions, this trend must be investigated further. Figure 6 shows the similarity of the top spreaders for the randomly perturbed networks. Even at 5% perturbation these sets are almost entirely different. Thus, for example, even small changes in the network may invalidate the predictions about the potentially good marketing targets.

6.3.4.3 The relative performance of the historical spreaders

Finally, despite the fact that, as I saw, the identity of the top spreaders may have changed, I asked whether the old top spreaders would still "perform" sufficiently well in the new, changed network. That is, if the top spreaders from the original network are used to initiate the spread in the new, changed network, how would the number of individuals affected by this spread compare to the extent of spread initiated by the new set of the top spreaders in the new

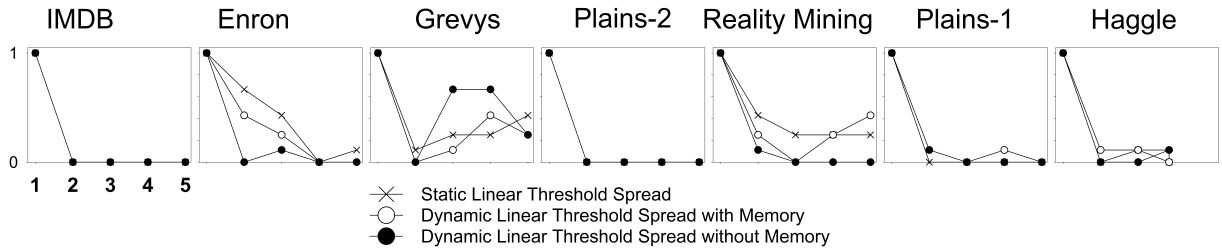


Figure 7: Jaccard similarity comparing the top 5 individuals (ranked by spreading capacity) in segment 1 with the top 5 individuals in subsequent segments. The x-axis is the current segment to which the top five set from segment 1 is being compared.

network? The answer to this question directly implies the answer to whether the actions based on past predictions are valid as the network changes and lead to sufficiently good results.

Figure 9 shows the relative performance of the top five ranked individuals in segment 1 compared to the top five individuals in each subsequent segment. Despite the fact that the identity of the top individuals changes, I see that the old top spreaders perform as well as the new top spreaders in many cases. However, this is true only for the networks where spread saturation is easily achieved and spread initiated from almost any set of individuals reaches everybody in the population. In sparser networks like IMDB and Enron, the performance of the old top spreaders deteriorates with time. Figure 10 shows the relationship between the performance of the top ranked sets and the amount of change between the old and the new networks which, again, demonstrates that there is little correspondence between the amount of perturbation and the performance of the top individuals.

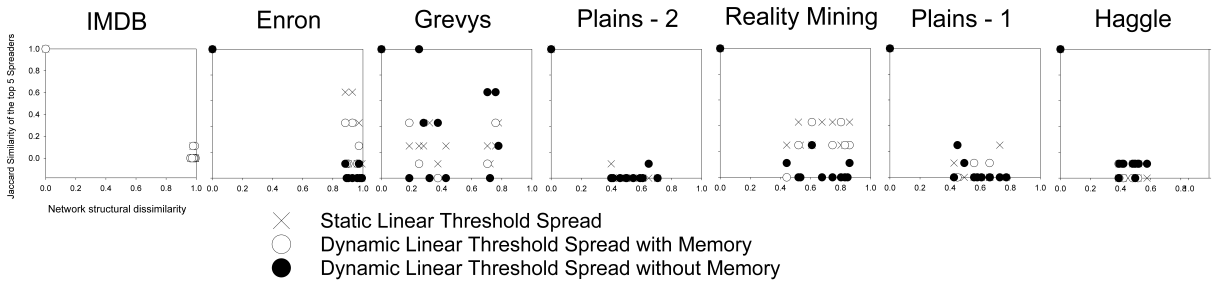


Figure 8: Jaccard similarity between the sets of top 5 individuals (ranked by spreading capacity) as a function of the dissimilarity between the underlying networks.

Surprisingly, on randomly perturbed networks, the original top 5 set of individuals performed always nearly as well as the top set from the perturbed network. This is despite the fact that the sets themselves have few individuals in common. Thus, random perturbations may not be representative of changes in real networks and it is, then, particularly important to distinguish true patterns of network evolution and noise.

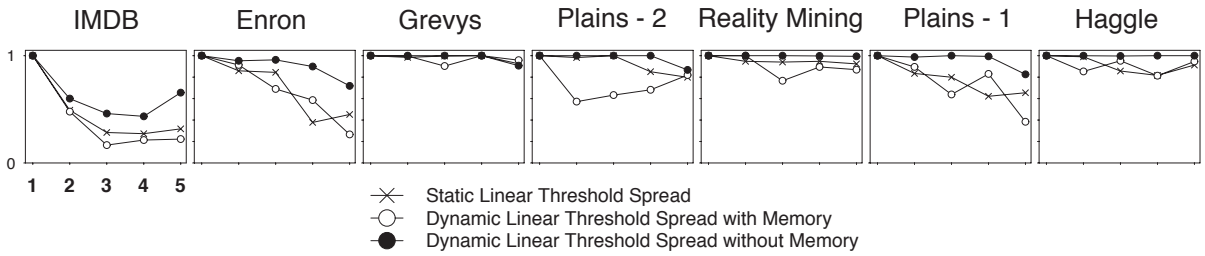


Figure 9: Relative performance of the top five sets of individuals from each segment in subsequent segments. Each datapoint with the x-coordinate value x is the ratio of the estimated extent of spread in segment x initiated by the top five individuals from segment 1 to that initiated by the top five individuals from segment x itself.

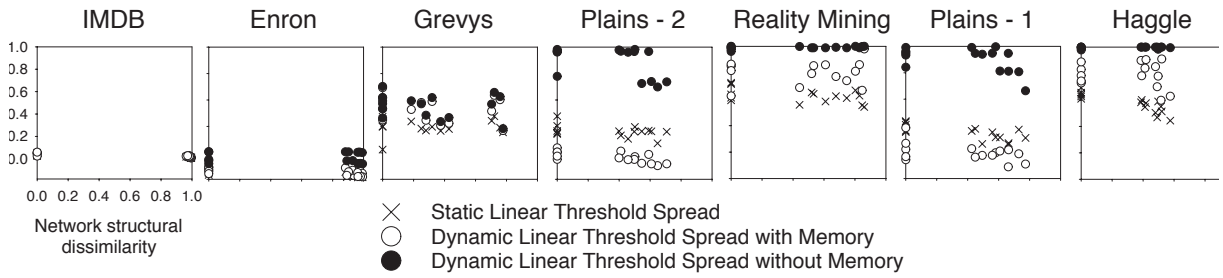


Figure 10: Relative performance of the top five sets of individuals as a function of the amount of change in the network. Each datapoint with x-coordinate x is the ratio of the estimated extent of spread initiated in the changed network by the top five individuals from the original network to that initiated by the top five individuals from the changed network itself, where the fraction of edges differing between the original and the changed networks is x.

6.3.5 Conclusions and Future Work

Most social network analysis is performed on historical and typically aggregate data, and the possible structural changes that happen as the network evolves are not taken into consideration. Thus, by the time the analysis is completed and acted upon, its results may not be valid if the network indeed has changed in the meantime. In this paper, I asked how much such changes can affect the results of network analysis in the context of diffusion in networks. Specifically, I asked three questions: (a) whether the predictions about the relative spreading capacity of each individual are robust; (b) whether the sets of the top spreaders are relatively unaffected; and (c) whether the performance of the top spreaders in terms of the extent of spread they may cause remains good enough even after the changes. In the process of answering these questions,

I also compared the predictions made on an the traditional aggregate, static representation of a network to the explicitly dynamic view of social interactions.

I found that in real dynamic networks the predictions about the relative spreading capacity of individuals and the identity of the top spreaders are sensitive even to minimal changes in the network. Moreover, I found that networks change significantly with time, often by as much as 40% of edges in a short time period. Surprisingly, I also found that there is little correspondence between the amount of change in the network and the robustness of the predictions. Finally, while in the real timeline, the performance of the top spreaders from the past did not compare well with the performance of the current top spreaders, in randomly perturbed networks past top spreaders typically did well even after the perturbations.

Thus, overall, I found that not only do predictions from the past not hold well into the future, these predictions do not deteriorate gracefully either. This implies that I cannot estimate the robustness of our predictions by measuring the amount of structural change in the network. Moreover, since random changes do not diminish the relative spreading ability of the top spreaders as much as the changes with real passage of time, I conclude that a few critical edges can make a big difference. Thus, we need methods for identifying edges that are critical to the robustness of the predictions. We must also develop analysis techniques that take possible future network changes into consideration.

Finally, in almost all experiments, the analysis performed in an aggregate network using a static diffusion model had little correspondence to the explicitly dynamic models of spreading

processes simulated on dynamic networks. Thus, in explicitly dynamic networks we must use analysis methods that explicitly take the dynamic nature of interactions into consideration.

6.4 The effect of network structure regime on the extent of diffusion

Who are the most influential individuals in a (social) network and can we efficiently find them? Computationally this and similar questions have been shown to be in the class of NP-hard problems (92; 106; 153). Moreover, even the best approximation algorithms are infeasible for massive network data available on today's ubiquitous large platforms, such as social networking websites, blogosphere, and communication networks. However, is it really necessary to expend this intensive computational effort in order to identify those individuals? While theoretically the answer seems to be 'yes', in practice many efficient heuristics have been demonstrated to be effective. In this work I show that the hardness of computationally finding the most influential individuals in a network depends on the structural properties of that network, such as its density (or, rather, the effective density, which is the combination of density and the probability of activation, that is, the density within the spread network) and modular structure.

The goal of this work is to find clues in the global network structure that make a particular heuristic for identifying influential individuals work better than others and to figure out when heuristics work at all and when a serious computational effort is unavoidable. Thus, I take a step back from devising yet another method that works for a certain set of networks, to answer a more general question of *what makes a certain heuristic effective, and not another, for a given network?* Specifically, I show that it is possible to use effective density and modular structure of a network as indicators of when it is necessary to employ a sophisticated yet computationally

expensive method versus when even a random set of spread initiators will perform as well as the best, in expectation, for maximizing the diffusion of influence in the network. I find that the effective density, which is the product of the density and the rate of diffusion, is a better indicator of the extent of diffusion than the density on its own. A diffusion process with very low rate of diffusion in denser networks behaves similar to the diffusion process in sparse networks at high rates of diffusion. I show that trends in diffusion follow two regimes.

1. Density regime: networks with (effective) densities above and below a certain threshold are amenable to simple heuristics. In dense networks, in fact, there is no differentiation between influence of individuals and any random set of individuals is good, given a high enough rate of spread. In effectively sparse networks simple heuristics like highest degree nodes, perform well.

2. Modularity regime: In between the two extremes, the difference between the best and a random set of individuals may be significant. That difference, in fact, depends on how modular (clumped, clustered, non–uniform) the network is. The more modular, non-uniform, the network is, the bigger that difference is. Thus, it is for those intermediate density networks with rich complicated structure that we need to use sophisticated and computationally intensive approaches to find influential individuals.

This result supports the findings that simple heuristics perform well in practice since most real world networks are very sparse, with few exceptions that are very dense (that represent small single communities).

To demonstrate my results, I systematically empirically evaluate the difference between the expected and the optimal extent of diffusion on a variety of synthetic and real–world networks, over a range densities and degrees of modularity. My results consistently tell the same story: density and modularity matter and can tell us when to use a simple heuristic and when to put the effort and use more sophisticated yet computationally expensive methods. Indeed, once stated this way, it is not surprising and quite intuitive: in dense networks everybody is connected to almost everybody and any set of individuals will do well; and in sparse networks nothing spreads well beyond the immediate neighbors so high degree nodes do best. However, my work for the first time test this assertion in a systematic way, quantifies the critical network properties, identifying the transition thresholds and lay the groundwork for a more rigorous theoretical analysis.

6.4.1 Diffusion in Uniform random and Scale–free networks

To understand how the extent of diffusion is affected by the global network structure we estimate the difference in the best and worst extents theoretically in two network models extensively studied in social network analysis. I first estimate the expected extent of diffusion in the most tractable network model, uniform random model. This model is defined as a set of N nodes connected to each other with a uniform probability p (62). Next, I give estimates for the more realistic real–world model, scale–free networks. For estimates on diffusion, we assume the underlying stochastic process is of the monotonic *susceptible-infected* form of diffusion (92), defined by one parameter $p \in (0, 1]$: the rate of diffusion. Without loss of generality, I as-

sume that the number of diffusion initiators $k = 1$ for the theoretical estimates. We denote by $E(N, p)$, the expected extent of diffusion, with rate of diffusion p, in a network of size N.

6.4.1.1 Uniform random networks

A uniform random network $G(N, d)$ is a set of N nodes connected to each other with probability d. Thus, a uniform random network is fully defined by the probability (density) d of the network. I use the bounds on network connectivity in such networks from (62) to estimate the expected extent of diffusion in these networks.

Theorem 4. ***Sparse:*** *Expected extent of diffusion,* $E(N, p) \sim 0$*, with high probability, as* $N \to \infty$*, when* $q \times d < 1/N$*.*

Intermediate density: *Expected extent of diffusion,* $E(N, p) \in (p \times n^{2/3}, \log(N)/N]$*, with high probability, when,* $1/N < d \times p < \log(N)/N$*.*

Dense: *Expected extent of diffusion,* $E(N, p) \sim N$*, with high probability, when* $d \times p \geq \log N/N$*.*

Proof. **Sparse:** In uniform random graph, if $d < 1/N$, then the graph $G(N, d)$ almost surely has no connected component of size larger than $O(\log(N))$ (62). Assuming each activation happens independently in the graph, from the above result, it immediately follows that if $d < 1/N$ and $p \leq 1$, then, the expected extent of diffusion $E(N, p)$ is bounded above by $d \times p < d \leq 1/N$.

Intermediate density: In uniform random graph, if $d = 1/N$, then the graph $G(N, d)$ almost surely has a *giant* component whose size is of order $n^{2/3}$. If $d \to c/N$, for some

constant $c > 1$, then the graph $G(N, d)$ almost surely has a unique *giant* component containing a positive fraction of the N nodes. No other component has more than $O(\log(N))$ nodes (62). Using these results, the expected extent of diffusion in such a network is bounded below by $pn^{2/3}$ yet it is no more than $\frac{\log(N)}{N}$. The lower bound comes from the expected extent of diffusion $E(n^{2/3}, p)$ in the giant component. Assuming each edge in the giant is activated independently with probability p from the diffusion process, the expected extent in this component is at least $pn^{2/3}$. On the other extreme, when the density $d \to \frac{c}{N}$ of $G(N, d)$, for some constant $c > 1$, the expected extent of diffusion is at most $p\frac{c}{N} < \frac{\log(N)}{N}$ when $N \to \infty$.

Dense: In uniform random graphs, if $p \geq \log(N)/N$, the graph is almost surely connected (62). It follows immediately from this result that for $q \sim 1$, the diffusing process affects almost the entire population.

□

6.4.1.2 Scale–free networks

Real–world networks have been shown to have skewed degree distribution, approximated by power law distribution with the exponent $\gamma \in [2, 3]$, resulting in what is called a scale–free network. In such networks, it has been shown that there is no "threshold" for diffusion (132). This is because the second moment of the distribution diverges so that the ratio of the second and first moment is zero (31). In other words, a diffusion process with an arbitrarily small rate of diffusion can potentially affect the entire population. Hence, *the extent of diffusion is dictated by the structure of the network more than the rate of diffusion.* To understand the

effect of scale–free structure on the extent of diffusion, I tie the skewness γ to the densities of such networks. I use the estimates of scaling of maximum degree and the number of minimum degree nodes in general scale–free networks from (56) to formulate various regimes of densities in this type of networks.

Theorem 5. *In scale–free networks with exponent of skewness γ, the maximum expected degree $\hat{d} \to N^{\frac{1}{\gamma-1}}$ for any γ value as $N \to \infty$. In addition, the number of nodes A with degree $O(1)$ is of order N when $\gamma > 0$. However, $A \to N^{\gamma}$ when $\gamma < 0$ as $N \to \infty$.*

Proof. Using the generalized harmonic mean $H_{N-1,\gamma} = \left(\frac{1}{N}\sum_{k=1}^{N-1} k^{\gamma}\right)^{-1}$, the expected maximum degree $\hat{d}$ in a scale–free graph with exponent γ can be estimated as follows (56).

$$\hat{d} = max\{x : \frac{N}{H_{N-1,\gamma}} \sum_{k=x}^{N-1} k^{-\gamma} > 1\}. \tag{6.1}$$

Where, $H_{N-1,\gamma} = \sum_{k=1}^{N-1} 1/k^{\gamma}$. The above summation converges to Riemann zeta function [1] $\zeta(\gamma)$ for $N \to \infty$. That is, $\sum_{k=x}^{N-1} k^{-\gamma} = \zeta(\gamma)$. The above maximum expected degree estimate can be rewritten as the integral between the interval $[x, N-1]$. Moreover, when $\gamma > 1$ and $N \gg 1$, $\zeta(\gamma)$ converges to 1. Thus,

$$\frac{N}{H_{N-1,\gamma}} \int_{x}^{N-1} k^{-\gamma} dk = 1. \tag{6.2}$$

---

[1]The Riemann zeta function or EulerRiemann zeta function, $\zeta(s)$, is a function of a complex variable s that analytically continues the sum of the infinite series which converges when the real part of s is greater than 1.

Integrating,

$$\frac{N}{H_{N-1,\gamma}(1-\gamma)}[(N-1)^{1-\gamma} - x^{1-\gamma}] = 1. \tag{6.3}$$

Now, solving for x, the maximum expected degree, we get:

$$\hat{d} = x = [\frac{N}{(\gamma-1)H_{N-1,\gamma}}] \to N^{\frac{1}{\gamma-1}}, \quad as N \to \infty. \tag{6.4}$$

The number of nodes A with unitary degree $O(1)$ in scale–free networks, can be given as:

$$A = \frac{N}{H_{N-1,\gamma}}. \tag{6.5}$$

When $\gamma > 0$, the harmonic series diverges, hence, $A \to N$ as $N \to \infty$. On the other hand, when $\gamma < 0$, the harmonic series converges, hence, $A \to N^{\gamma}$ as $N \to \infty$. □

Corollary 1. *From Theorem 5, in scale–free networks when $\gamma \leq 2$, the maximum expected degree $\hat{d} \to N^{1/\gamma}$ as $N \to \infty$. When $0 < \gamma < 2$, the maximum expected degree $\hat{d} \to N$ as $N \to \infty$. When $\gamma < 0$, the maximum expected degree $\hat{d} \to 1/N$ as $N \to \infty$ by substituting in Equation 6.3.*

6.4.1.2.1 Sparse scale–free networks

Here I show that scale–free network with exponent of skewness $2 < \gamma \leq 3$ are sparse.

Theorem 6. *Scale–free networks with power law exponent $\gamma \in [2, 3]$ are sparse, that is, the order of the edges $\|E\|$ is $O(N)$.*

Proof. Sparsity of scale–free networks with $2 < \gamma \leq 3$ follows immediately from Theorem 5. Since the number of nodes with degree $O(1)$ is near N it implies that most of the nodes have degree $O(1)$, that is all but $c_1 N$ nodes, for some constant $c_1 \propto 1/N$, have very small degree. Then the number of edges in the network can be approximated as $|E| \sim c_1 N\hat{d} + (1-c_1)NO(1) \sim O(N)$. □

Extent of diffusion. Sparseness and skewness of such scale–free networks with $2 < \gamma \leq 3$ render degree–based heuristics (such as choosing nodes with highest degrees) very effective in maximizing the extent of diffusion. From Theorem 5, sparse scale–free networks are characterized by the existence of "hubs" – nodes with a very high degree compared to the average. Selectively initiating the diffusion process through these nodes maximizes the extent of diffusion in expectation.

Theorem 7. *In scale–free network $G(N, E)$ with $2 < \gamma \leq 3$, the expected extent of diffusion $E(N, p)$ at the rate of diffusion p, is optimal in expectation, when initiated by the node with the maximum expected degree $\hat{d}$.*

Proof. I show that the worst case extent of diffusion from a maximal degree node is comparable to the best case extent of diffusion from the lowest degree node, asymptotically. Hence, even in the worst case, the nodes with maximum expected degree are good candidates for maximizing the extent of diffusion in sparse scale–free networks.

Recall, from Theorem 5, we know that all but a small fraction of nodes have degree $O(1)$. Suppose $c_1 N$ is the number of nodes that have the maximum expected degree $\hat{d} \sim N$. From the

definition of power law distribution, $c_1 \propto \frac{1}{N}$. That is, only a small fraction of nodes have high degrees. Similarly, the number of nodes A with degree $O(1)$ tends to N. Let c_2 be the fraction of nodes with degree $O(1)$. Then, $0 \ll c_2 < 1/O(1)$. For the proof, I assume the diffusion process to be a two step process. Note, in a stochastic diffusion process, each progressive step of diffusion propagation, is at least a factor $p(< 1)$ times less than the previous step. That is the probability of activating nodes in each subsequent step reduces by a multiplicative factor of p. Hence, the expected extent of diffusion in the network is dominated by the first few steps.

The best case network for starting from a node with a low degree is an instance of a scale–free network where a node of degree $O(1)$ is connected to all nodes with maximum expected degree $\hat{d} \sim N$. The expected extent of diffusion in such case can be approximated as: $E(N,p) \leq pc_2[p\hat{d}+p(c_2-c_1)O(1)] \leq p^2c_2\hat{d}^2+p^2(c_2-c_1)$. Since $c_1 \sim 1/N$ and $c_2 \sim N$, then $E(N,p) \approx p^2N$.

The worst case network for starting from a node with a high degree is an instance of a scale–free network where a node with maximum expected degree is $\hat{d} \sim N$ is connected to all nodes that have some small constant $O(1)$ degree themselves. The expected extent of diffusion in such case can be approximated as: $E(N,p) \leq p\hat{d}[pc_1\hat{d} + p(c_2 - c_1)O(1)] \leq p^2c_2\hat{d}^2 + p^2N$. Since $\hat{d} \to N$, then $c_1 \sim 1/N$ and $c_2 \sim N$, then $E(N,p) \approx p^2N$.

That is, asymptotically, the expected extent of diffusion from a node with the maximum expected degree in the worst case scenario is comparable to the expected extent of diffusion from a node with degree $O(1)$. □

Most real–world networks are sparse (Table I). Hence, even though degree based heuristics may not provide provable guarantees to the extent of diffusion, now we have a better under-

standing of why such heuristics work in real world examples. In Section 6.4.3, I experimentally show the effectiveness of degree heuristics in sparse real–world and synthetic scale–free networks.

While sparseness is a common property, which is regularly exploited in devising efficient heuristics for diffusion, there are many examples of dense networks in real–world. Therefore, it is essential to understand the structure of such networks for estimating the extent of diffusion. Next, I show some analysis of denseness of scale–free networks and its affect on the extent of diffusion.

6.4.1.2.2 Dense scale–free networks

Theoretically, a network is dense if the number of edges, $E \sim N^2$. From Theorem 5, we know that when $\gamma < 0$, the maximum expected degree is of order N, whereas the number of nodes with O(1) degree decrease in the order of N^{γ}. That is, most of the nodes have uniformly high degrees.

Extent of diffusion. The above result shows that dense scale–free networks have uniformly high degrees. That is, the resulting network is almost a clique and any edge can be activated with probability p of diffusion. Thus, the resulting diffusion network is equivalent to a diffusion process in a uniform random graph and is completely determined by the diffusion probability p.

Theorem 8. *In scale–free network $G(N, E)$ with $\gamma < 0$, the expected extent of diffusion $E(N, p)$ with rate of diffusion p is $\sim pN$.*

Proof. The proof of this theorem follows directly from Newman's results on the scaling of moments of degree distribution in scale-free networks with $\gamma < 0$ (126). When $\gamma < 0$, the

first moment of degree distribution diverges as $N \to \infty$. That is, the number of highest degree nodes is no longer small but rather of the order of N. Hence, asymptotically almost all nodes have uniformly high degrees. In other words, the resulting network has *dense* clique–like structure. Thus, the resulting diffusion network is essentially a uniform random graph with edge probability p and the extent of diffusion depends solely on p and not on the identity of the initiating node. Hence, for a dense network of N nodes, $E(N, p) \sim pN$. □

Table VII summarizes some of the significant results about the uniform and scale–free network structures that are useful for estimating the extent of diffusion in such networks.

For $0 \leq \gamma < 1$, Genio et al. show that scale–free networks are either not realizable due to the extreme difference in the highest degrees ($\sim N$) and the number of nodes with the $O(1)$ degree ($\sim N$) or the degree distribution is not continuous (that is, the degree distribution exists with cut-offs) (56). Such a distribution results in a wide range of complex structures. For networks within such range of γ values, I devise a heuristic that exploits the modularity of the network to maximize the extent of diffusion in such networks. I analytically show in Section 6.4.1.2.3 and experimentally in Section 6.4.3 the improvement in extent of diffusion using this heuristic.

6.4.1.2.3 Intermediate density scale–free networks

From the results in the above two sections, it is clear that for $\gamma < 0$, scale-free networks are dense whereas, for $2 < \gamma < 3$ the scale-free networks are very sparse. The networks with $\gamma \in [0, 2]$, are neither very sparse nor dense. Aiello et al. (6) show that in scale–free networks when $0 < \gamma < 1$, with high probability the graph is connected and when $1 < \gamma < 2$, with high probability the second largest components are of size $\Theta(1)$. That is, the graph is not

Density	sparse	intermediate density	dense
No. of edges	$\|E\| < \log N$	$\sim \log N \leq \|E\| \leq N$	$\|E\| > O(N)$
density	$d < \frac{1}{N}$	$d \in [\frac{1}{N}, \frac{\log N}{N})$	$d > \frac{\log N}{N}$
Uniform	No connected component larger than $O(\log N)$ (62)	Connected component of order $O(n^{2/3})$. Unique giant component. No other component contains more than $O(\log N)$ node (62).	Connected graph (62).
Diffusion estimates	Very little connectivity, hence, extent of diffusion 0 with probabilty $O(1)$.	No more than the size of largest connected component.	With high probability $O(N)p$ network affected by the diffusion process.
skewness	$2 < \gamma < 3$	$0 < \gamma < 2$	$\gamma < 0$
Scale-free	Small number of high degree nodes, and large number of low degree nodes. (56)	Scale–free graph not realizable with high probability. (56)	Connected graph. With high probability almost all nodes have uniformly high degrees (56).
Diffusion estimates	Extent of diffusion $\sim$ maximum expected degree.	Depends on modularity structure of the network.	With high probability the extent of diffusion similar to uniform random networks.

TABLE VII: Summary: Structure and estimates of diffusion in uniform and scale–free networks

connected with high probability. Hence, in this range of γ values, we find that both loosely connected as well as unconnected structures of scale–free networks can be realized. Thus, such networks have large connected components yet not necessarily very dense. In this intermediate density range, one way to estimate the extent of diffusion is by incorporating various structure characteristics estimated in such networks in literature. Here, I specifically use the modular nature of the network to devise practical and effective heuristic for estimating near optimal extent of diffusion.

Extent of diffusion The optimal diffusion maximization method requires exhaustive search for the best k nodes from the $\binom{n}{k}$ possible choices as the initiators of the diffusion process. On the other extreme is the naive approach of picking any uniformly random set of k nodes from the network as diffusion initiator. In the previous sections I showed that the structure of the network simplifies this diffusion maximization problem, such that simple heuristics and even the uniform random choice of diffusion initiators give estimates close to the optimal. Here I extend that analysis to networks with complex modular structures. Hence, in scale–free networks which do not have simple structures such as regular graphs or very skewed connectivity, the complexity and non-uniformity of the structure has to be incorporated in devising effective strategies for diffusion maximization. As I show in the previous sections and later verify with experimental results, that once we have the knowledge of the expected network structure, the optimal extent of diffusion can be estimated by simple structural measures. Hence, in this intermediate density region with complex structures we must analyze various global structural properties to find properties that estimate the extent of diffusion in those network structures

the best. For instance, there is substantial evidence in literature that population networks organize themselves in clusters (74; 114; 123; 127). In such modular or clustered networks, an improvement over choosing a uniformly random set of diffusion initiators would be to choose the diffusion initiator by taking into consideration the modular structure of the network. Here I propose a heuristic that selects nodes to initiate the diffusion process from the "boundaries" of clusters within a network, that is, those nodes that connect disparate clusters to each other. Intuitively, such nodes are structurally well placed to be effective diffusion maximizer *within* as well as *across* modules. In networks with such modular structure I show that this simple heuristic that favors the boundary nodes, as diffusion initiators, between various modules gives the extent of diffusion very close to the optimal.

The optimal extent of diffusion is the exhaustive method of finding the k initiators from $\binom{N}{k}$ choices.

Let B_i be the i^{th} cluster in graph G with N_i nodes and E_i edges. The expected extent of diffusion $E(N, k, p)$ in such a block mixture network when k initiators are chosen uniformly at random from among all the nodes can be approximated as:

$$E(N, k, p) = E(N = \cup_i B_i, k, p) = \sum_i E(B_i, k_i, p).$$

Where p is the rate of diffusion and k_i are the spread initiators picked uniformly at random from block B_i.

On the other hand, the expected boundary extent of diffusion $B(N,k,p)$ is given by:

$$B(N,k,p) = \sum_i E(B_i, k_i, p_i) + \sum_{ij} E(\{B_i, B_j\}, k_i + k_j, p_{ij})$$

Where, $E(\{B_i, B_j\}, k_i + k_j, p_{ij})$ is the expected extent of diffusion across blocks B_i, B_j. Hence, $B(N,p,k) > E(N,p,k)$ by a factor of $\sum_{ij} E(\{B_i, B_j\}, k_i + k_j, p_{ij})$. (The optimal spread is, of course, at least as large as the heuristic estimate.) Hence, I conclude that the modular network structure calls for better methods of approximating the optimal extent of diffusion. It is also worth noting that such boundary nodes have above average degrees. Since the expected degree of a node within the block B_i is d_i, the boundary node by definition has an expected degree of $d_i + \sum_{j \in \{\cup B_j\}} d_{ij}$. Thus, with relatively high degrees, high betweenness, and high vertex expansion factor such nodes are better candidates for diffusion maximization. While using boundary nodes as initiators may be a good heuristic, it is clear that, theoretically, they do not always constitute an optimal choice. The NP-completeness and approximability gap of $(1 - 1/e)$ factor holds for these intermediate density networks.

I now present the results of an experimental investigation of the performance of the proposed boundary heuristic, as well as the high degree and random approaches for diffusion maximization.

6.4.2 Experimental methodology

6.4.2.1 Diffusion process

I simulate spread in the real–world and synthetic networks using the Independent cascade spreading model. This model is explained in full detail in Section 2.3.1.

6.4.2.1.1 Extent of diffusion simulations

The extent of spread in a network is the number of individuals affected at the end of a diffusion process (simulation), initiated by a set of individuals. I measure this extent based on three types of diffusion initiators.

1. Optimal diffusion: The k spreaders that maximize the extent of diffusion in the network. Since for large networks it is NP-hard to find the exact k best spreaders, I use the greedy approximation described in Section 6.2.4.1.

2. Random spreaders: Any k spread initiators picked uniformly at random from the network. This set of k individuals is picked randomly overly multiple runs of the stochastic simulation. The resulting extent of diffusion is the average over all runs. The number of runs to simulate the diffusion process from each random set of diffusion initiators is set to $c\log(N)$. Where $c > 0$ is a constant, that is set to 10 for this set of experiments. I tried higher values of c, however, the extra runs did not improve the variance in the results by any significant margin. N is the size of the network population. This method is formally described in Algorithm 4.

3. Degree and boundary heuristics: k highest degree or highest conductance boundary nodes chosen from the network as diffusion initiators. In this approach, I first calculate the structural property M (degree or boundary node value) for the network. Rank the nodes in decreasing value of the property M. Then pick the top k highest degree or most well connected boundary nodes from the network. The diffusion process is then simulated in the network through this set of top k nodes. Algorithm 5 formally states this method.

Algorithm 4: Expected extent of diffusion with random set of initiators

Initialize: $\mathit{Total\ Spread} := 0$;
while $\mathit{Run} \leq c\log(N)$ **do**
  Initialize: $S = \mathit{Random}(k)$ from N;
  $f(S) := \mathit{Spread}(S)$;
  $\mathit{Total\ Spread} := \mathit{Total\ Spread} + f(S)$;
end while
$\mathit{Expected\ Spread} := \frac{\mathit{Total\ Spread}}{c\log(N)}$;
return *Expected Spread*;

Algorithm 5: Extent of diffusion with network property M

Initialize: Set $S = Top(k, M[])$ from N;
return $f(S) := \mathit{Spread}(S)$;

In each network I find the extent of diffusion using the three types of spreaders. I compare the the optimal extent of diffusion (greedy approximation) to the expected extent in the network - based on randomly selected sample of diffusion initiators. I also compare the optimal extent to the two heuristic methods. The difference in the extent of diffusion, for each pair of methods, highlights the disparity in the network structure and how it affects the resulting extent of spread. I focus on how this difference varies with network density.

6.4.2.2 Datasets

The experimental analysis in this work is based on a large set of real–world networks as well as synthetic networks generated using well studied generative network models. I examine numerous real–world networks including blogosphere, online social networks, email exchange networks, router networks, co-authorship, human, and animal proximity networks. Following are the the network datasets that I analyze in this work: Autonomous Systems (AS), Co–authorship network (DBLP), Live Journal, Peer 2 Peer (P2P), Epinions, Political Blogs, Political Books, Karate, Enron, onagers, and grevys. The details of these networks are presented in Section 4.1. The list of networks and their basic statistics is given in Table I. It is important to note here that most of the real–world networks are very sparse. The results of these experiments, shown in Figure 17, are consistent with the trends I deduce in Section 6.4.1.2.

Further for Reality mining, and plains zebra networks I analyze the *densification* of these networks over time and its effect on the estimates of the extent of diffusion. Note that of the list of the datasets I worked with here, only these two datasets have the actual evolution of network recorded over time 4.1. I aggregate the interactions in each of these networks over

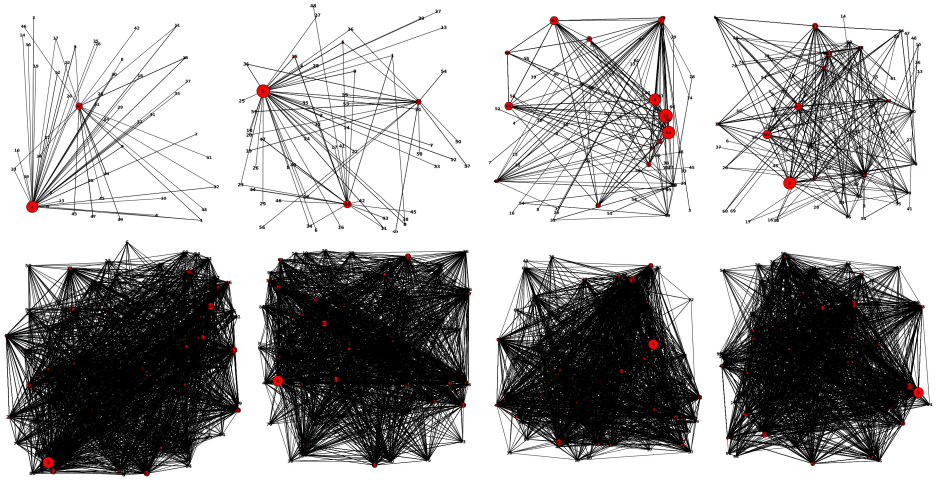


Figure 11: Reality Mining time series

time. In this way I get samples of the same network that are "evolving" with time. Hence, with this aggregation process I get a set of networks from the same domain and interaction dynamics but varying (increasing) densities. Figure 11 and Figure 12 show the series of aggregated Reality Mining and plains zebra networks. These time series of networks is the progressively (successively) accumulated interactions in each network. The size of the node represents the relative degree of the node.

6.4.2.2.1 Synthetic networks

To evaluate how the theoretical analysis the extent of diffusion in networks is manifested in practice, I use two generative network models that have been shown to represent real–world networks closely.

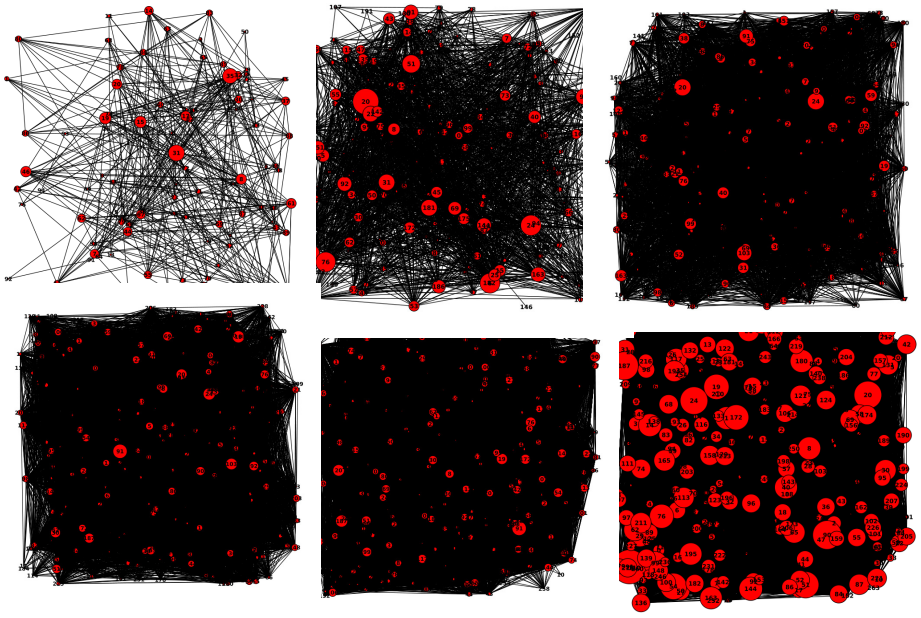


Figure 12: Plains zebras series

I use the Preferential Attachment model for generating scale-free networks (22). This model is explained in detail in Section 4.2. To generate networks from low to high density using this model I take the following steps. This model is completely defined by two parameters. The exponent γ of the skewness of degree and the minimum average degree m of each node. The γ parameter dictates the degree distribution whereas the m parameter the process of edge formation among the nodes. The generative process takes three inputs, a fixed network size N, γ, and m. In these experiments I set $N = 500$. (For networks with larger N, especially at high density, estimating even the approximate extent of diffusion becomes infeasible.) Initially, an empty network $G(0, 0)$ is created with 0 nodes and 0 edges. In each subsequent steps, one

node is added to the network until all N nodes are exhausted. Each node $v \in N$ is added to the network and connected to m other nodes proportional to the degree of the existing nodes. At the end of N steps I get a sample of a scale–free network with a given γ and m. For the same parameter setting I generate multiple networks. For this work, I sample 50 networks for each parameter settings. I estimate the extent of diffusion using the diffusion model and the methods described in Section 6.4.2.1. The exponent γ is set in the range $[-1, 3]$. The parameter m controls for the density of each network.

For finer insights into how network structure, other than its density, contributes to the disparity in the optimal and the expected or the heuristic spreads, I use the block mixture model (131) to generate another set of synthetic networks. The stochastic block mixture model was proposed for this purpose in the context of social sciences, using a Bayesian approach (131). Further refinements, such as the assortative mixing (114) has made this model a natural choice to analyze real world networks in controlled parameter settings. The details of this model are presented in Section 4.2. The block mixture model designates the nodes into m blocks. Given two parameters, the inter– and intra–block edge probabilities ($d_i, d_{i,j}$ respectively), the edges are generated uniformly at random, with the appropriate probabilities for each pair of nodes. For this generative process, I need four parameters: N–size of the network, m–the number of modules or clusters, d_i–the intra–block connectivity probability, $d_{i,j}$–the inter–block connectivity probability. In the first step, N are distributed proportional to the sizes of the cluster. (Here the sizes of the clusters are assumed to be uniform.) In the second step, nodes within each cluster are connected with probability d_i. In the last step nodes across clusters

are connected with probability $d_{i,j}$. To instantiate networks for the entire range of overall network density $d \in (0, 1]$ and the fixed $N = 500$, for density d I sample $0.5d \geq p_i < d$ and $0 \geq p_{i,j} < 0.5d$. This process results in networks of various densities and various levels of modularities. I sample d at the scale of 0.01. The results shown in Figure 13 are at a less granular scale as all intermediate densities resulted in approximately similar estimates of diffusion for similar parametric values. Again, I sample 50 networks for each parameter settings. I estimate the extent of diffusion using the diffusion model and the methods described in Section 6.4.2.1.

The resulting set of synthetic networks not only provide me with data to compare density with the extent of diffusion but also gives a better insight into the modular structure of networks for finer analysis. The process of adding (but not removing) interactions to create increasingly denser networks is similar to the way most network generative models are defined. Moreover, assuming the underlying generative process which is evidenced by the network is stable, those series represent networks with similar dynamics.

6.4.3 Results and analysis

I observe the following three trends in my experimental analysis. (a) The optimal extent of diffusion is well approximated by the expected extent of diffusion in networks of high densities but not in the intermediate density networks; (b) degree–based heuristics result in optimal or near–optimal diffusion in sparse networks, but these heuristics may not perform consistently well for denser networks; And (c) modularity of a network can be exploited for designing better heuristics that approximate the optimal extent of diffusion well especially for networks of intermediate density range.

Following are the experimental results and more detailed observations of the three trends.

6.4.4 Optimal vs expected extent of diffusion

Figure 14, Figure 15, and Figure 16 show the difference between the (greedy) optimal and the expected extent of diffusion as a function of the *effective* density (product of density and rate of diffusion: $d \times p$) in Reality mining, plains zebra, and synthetic scale–free networks, respectively. The plots show the following three trends in the extent of diffusion as the *effective* densities of the networks progressively increase.

First, in low effective density (sparse) networks ($\leq .004$ for real–world networks and $\leq .001$ for synthetic networks), the extent of diffusion is very low, irrespective of the method used. Sparse networks lack a "well defined" structure, by definition, most nodes in such networks are minimally connected. Such real–world networks have extremely skewed degree distribution: a very small number of nodes have disproportionately high degrees and the remainder of nodes are sparsely connected. Hence, in these networks, the extent of diffusion is very low relative to the size of the network even with high rates of diffusion since most nodes have few or no neighbors to whom they can propagate. Only high degree nodes can influence many others but there are so few of them, and their neighbors have such low degrees, that they hardly make a dent in the overall extent of diffusion.

The second trend we observe is that at high effective densities (≥ 0.2 for real–world networks and $\geq .003$ for synthetic networks), most nodes are uniformly well connected. In expectation, the extent of diffusion initiated by any random node is high and is comparable to the optimal extent of diffusion in these dense networks due to the high uniformity in the network structure.

In high effective density networks, I find that diffusion processes behave almost deterministically, that is, in expectation, the entire connected component is affected by the diffusion process, irrespective of who initiates it. Hence, similar to sparse effective density, I find that in such networks, the optimal approach does not outperform the expected by much.

Finally, and most interestingly, in networks of intermediate effective densities I find a clear phase transition in the difference between the extent of diffusion resulting from the optimal and random initiators. This difference in the optimal and the random methods increases until it peaks and starts to decline gradually, as network densities progressively increase. In this intermediate region, the extent of diffusion is very sensitive to the identity of the initiator. Clearly, in this region it is worth while to devise near–optimal yet better performing diffusion maximization techniques.

Figure 17, shows the difference in the optimal and expected extent of diffusion for all the real–world networks from Table I. I observe that in sparse networks, the difference in the optimal and random method exists but is very small and is not consistent with increasing density. One important thing to highlight here is that depending on the size of the network, the set of initially activated nodes had to be adjusted proportional to the network size (in this work I pick no more than 5%) of the network nodes as the initial set. In intermediate density region, I notice that the difference in the two methods peaks. However, in very dense networks (these are mostly proximity networks(clique–like structures), the difference is almost zero. I observe that in these high density networks, the connectivity among the nodes is uniformly high.

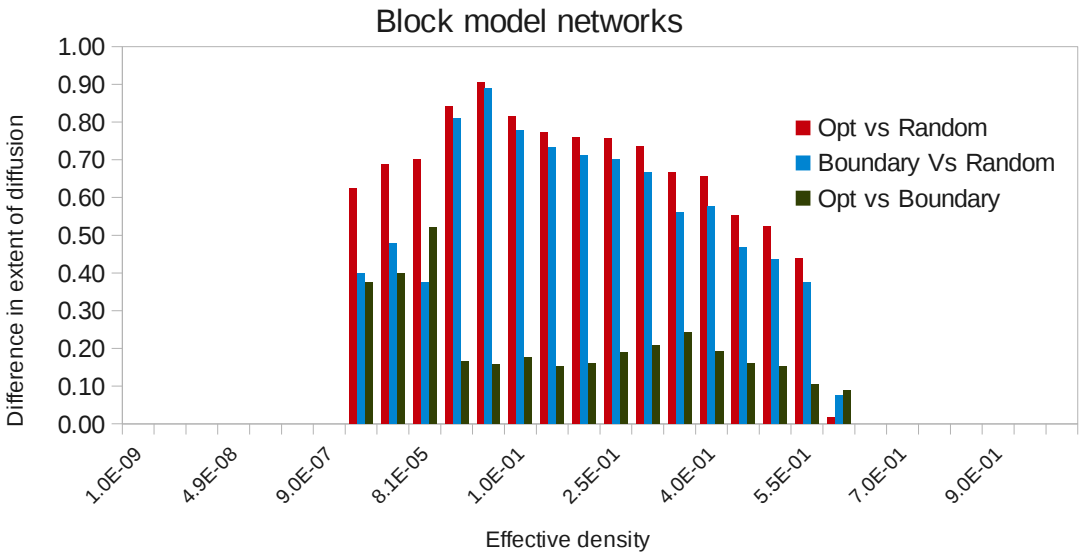


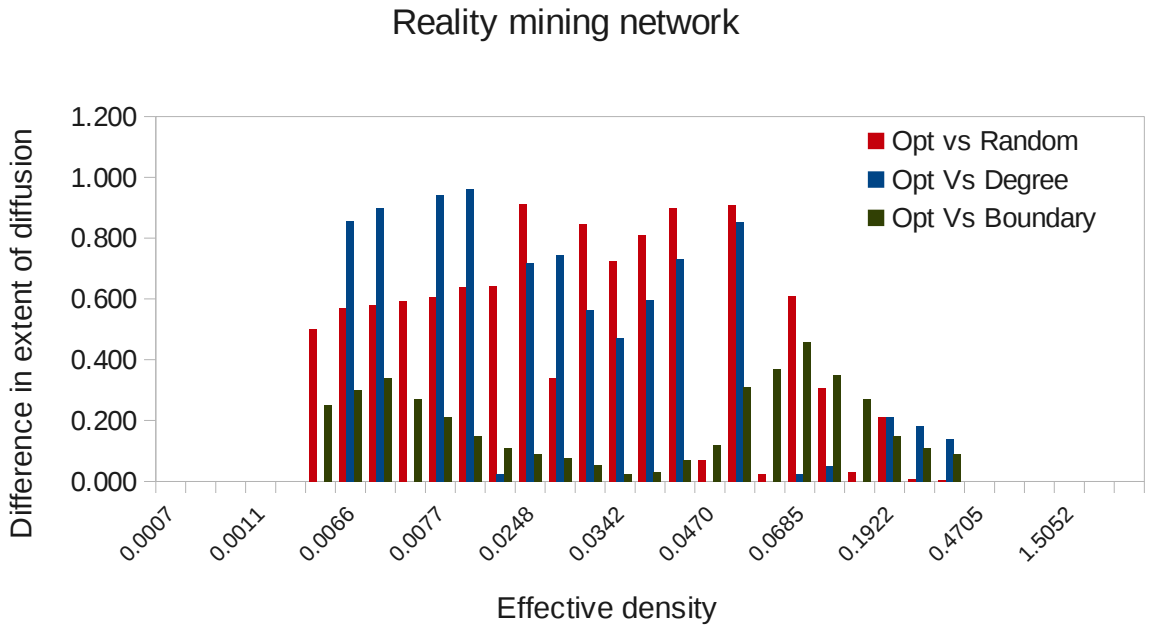
Figure 13: The difference in extent of diffusion, w.r.t increasing effective density of networks, between the optimal, random, and boundary heuristic methods.

Note one difference between the scale–free and the real–world networks: compared to the real–world networks, the effective density region of the high difference between (greedy) optimal versus random methods is confined to a smaller range of densities. Clearly, network characteristics other than density and skewness in degree distribution implicitly contribute to the extent of diffusion, that are not captured by the scale–free networks. Before I delve further into this issue I analyze the behavior of some of the most well studied heuristics for diffusion maximization in the same effective density settings.

6.4.4.1 Optimal vs degree–based heuristic

I compare the extent of diffusion from the optimal, as well as, the random method to the extent of diffusion given by one of the very well studied heuristics for diffusion maximization,

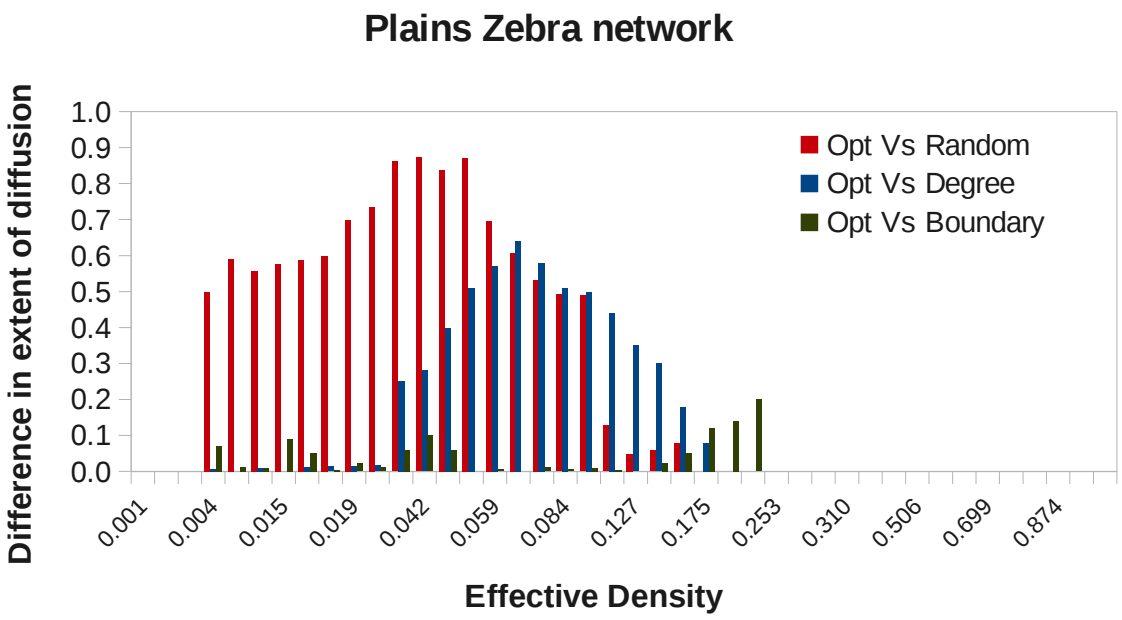
Figure 14: Reality Mining



the degree–based heuristic. Degree centrality has frequently been shown to correlate with the influence maximization objective (38; 89) and I have proved similar results in Section 6.4.1.2.

I simulate diffusion in Reality Mining, plains zebra, scale-free synthetic networks by using k highest degree nodes. I compare the extent of diffusion from those high degree nodes with the optimal and the expected extent of diffusion. The difference in the extent of diffusion between the optimal and the degree heuristic for $k = 1$ are shown in Figure 14 and Figure 15 for Reality Mining and plains zebra networks, respectively. For other values of k the difference in the extent of diffusion between the optimal and the degree heuristic is almost the same. I find that the difference in the degree heuristic and greedy–optimal extent of diffusion is negligible

Figure 15: plains zebras



at low densities. In fact, at low effective density (below 0.005) the optimal diffusion initiators are indeed precisely the nodes with the highest degrees, in concordance with the theory.

At high effective density (above 0.5), even though the results from heuristics are comparable to the optimal, the identities of the optimal spread initiators and the heuristically chosen nodes are not necessarily the same. Recall, that at those densities, I proved earlier that any randomly chosen set of initiators performs as well as the optimal. Thus, the good performance of the heuristics here is not due to the identity of the chosen initiator but due to the fact that any

Figure 16: Scale–free networks

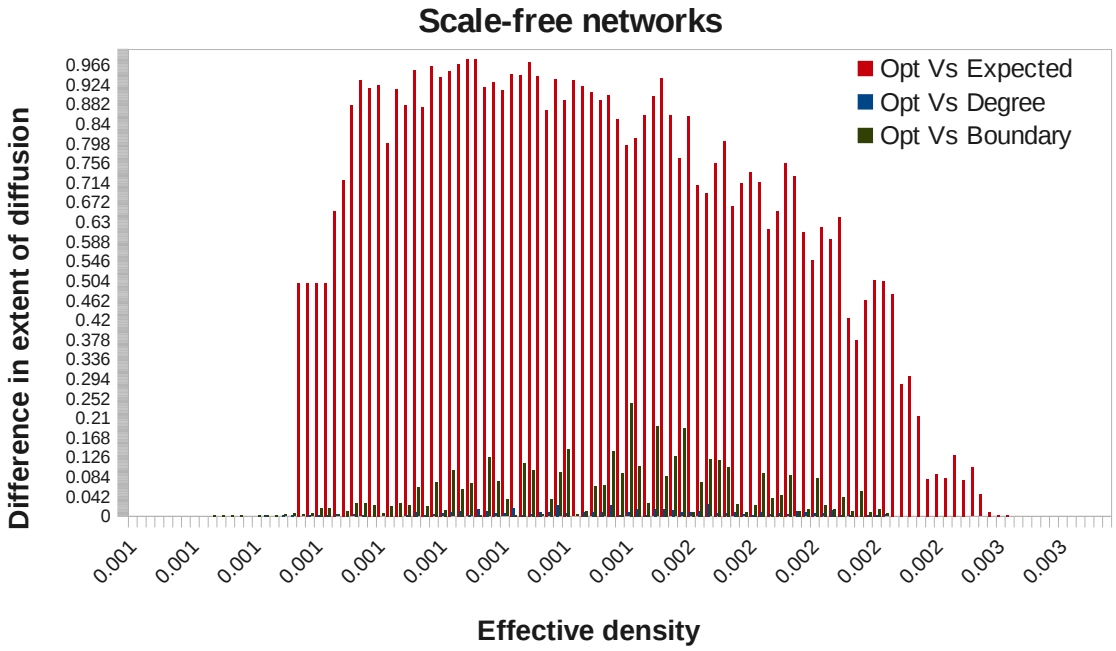
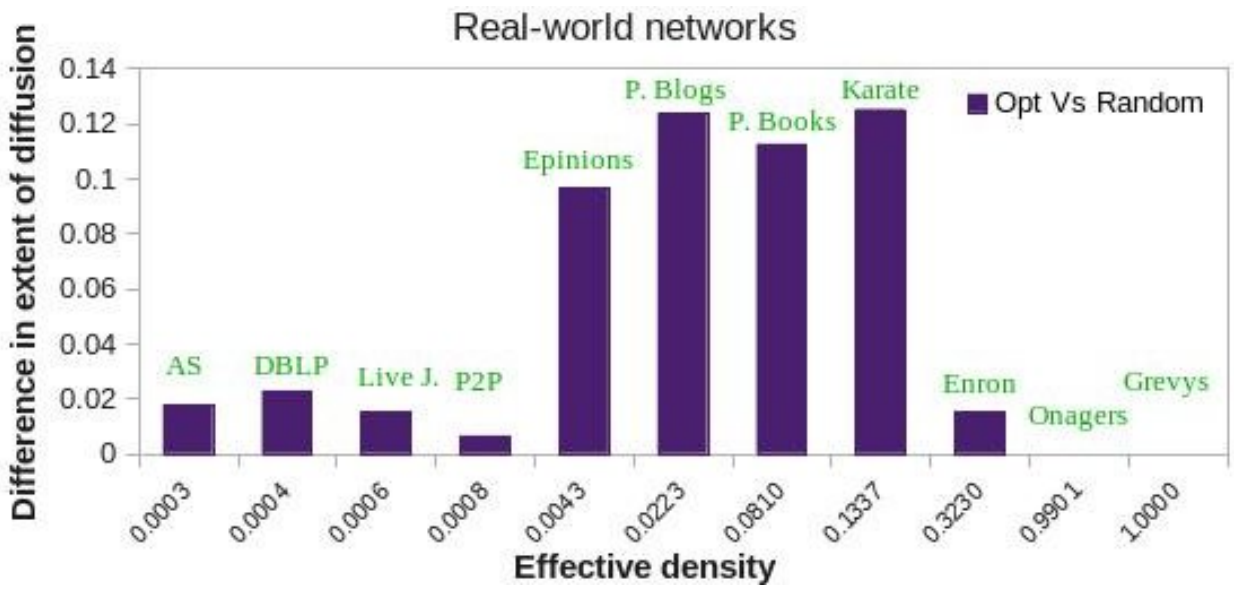


Figure 17: Real-world networks



node serves well. This eliminates the significance of a heuristic or even the optimal approach over a random one in high density networks.

In the intermediate density region, although the degree heuristic still works better than random, the differences in the extent of diffusion between the greedy–optimal and degree heuristic methods are *inconsistent*. Thus, the degree heuristic may work for some networks but not others, and there are no theoretical or empirical guarantees for its performance in this region. This is typically the case of any other heuristic for estimating extent of diffusion in networks. These fluctuations in effectiveness of various methods have not been taken into consideration in formulating better methods for diffusion maximization. This study gives us a better under-

standing of the effects of network structure on diffusion and it shows where we need to put more efforts, by exploiting the structure of networks, for devising better methods for diffusion maximization.

For better insights into the effect of network structure (especially in networks that fall within the intermediate effective density region) on the extent of diffusion, I use stochastic block mixture model process to generate networks in the full range of density from 0 to 1. I compare the extent of diffusion of the greedy–optimal with the randomly selected initiators, as well as the heuristic of selecting the boundary nodes, proposed above. I show that this heuristic outperforms the random by a large margin. Moreover, in the intermediate density networks, this heuristic performs very close to the greedy–optimal method.

6.4.4.2 Effect of network modularity on the extent of diffusion

Recall from Section 6.4.2.2.1, that a network can be represented as a set of blocks, with *inter*– and *intra*–block probabilities for connectivity. For a well–defined clustering of nodes to exist it is assumed that the inter–block probabilities (e.g. $p_{i,j}$) of forming links (between blocks B_i, B_j) are much lower than the intra–block probabilities (e.g. p_i within a block B_i). I generate a large number of networks with density between $[0, 1]$ having a range of inter– and intra–block probabilities. In these networks, I estimate the greedy–optimal, random, and boundary heuristic based extents of diffusion. I compare the boundary heuristic with the optimal and expected extent of diffusion. I observe essentially the same three trends in the difference in extent of optimal and random extent of diffusion across the range of density $[0, 1]$. That is, at the two extremes of the density regime, the difference between the optimal and random is

negligible. In between the two extremes the two methods do not compare well. Here, with the supplementary knowledge of the modular structure of the network, I experimentally observe that selecting nodes at the boundary of the blocks in the network, results in the extent of diffusion remarkably close to the greedy-optimal.

In this intermediate density region the optimal extent of diffusion differs significantly from the expectation, where the diffusion initiators are chosen uniformly at random. This result supports what I have shown in real–world and scale–free networks in Section 6.4.4. However, the choice of the stochastic block mixture model provides us with a better understanding of the effect of network modularity on the extent of diffusion. Figure 13 shows that at the effective density of 3×10^{-5} and above (much lower than networks that do not exhibit significant level of modularity, example, real–world sparse networks and scale–free networks) the random method performs very poorly compared to optimal (70% difference). This difference peaks at the effective density of .0001. I observe that the difference between the inter–and intra–block probabilities is maximal here (50%). This is the point at which, in the scale–free networks I observe significant differences between the greedy–optimal and the expected (Figure 16). However, surprisingly, the boundary heuristic that takes into consideration the modular structure of the network gives close to optimal result. Notice, that the difference between the greedy–optimal and the boundary heuristic is minimum where the inter–and intra–block probabilities difference is maximum. This reinforces the fact that network modular structure plays a deciding role in diffusion maximization.

Towards the other extreme, when networks become very dense, the random method is comparable to optimal (as proved). The boundary heuristic, although performs reasonably close to the greedy–optimal, at such high density the difference between the inter–and intra–block probabilities is very low (for the sampled networks reported here, it is < 0.007).

Hence, I show that the additional knowledge of the modular structure was helpful in devising a better heuristic for this specific model for networks with intermediate density. Consequently, I conclude, that the global network properties, density and degree of modularity, of real–world networks are natural parameters for designing efficient and accurate methods for influence maximization.

6.4.5 Discussion and conclusion

This work is a systematic exploration of the connection between network structure, specifically, density, modularity, and the extent of diffusion in those networks. My results show a strong effect of density on the extent of diffusion and the ease of finding effective diffusion initiators. I show that with respect to the extent of diffusion, networks can be demarcated into three broad classes: sparse, intermediate density, and dense networks. I show that networks with densities above and below certain thresholds are amenable to simple heuristics or even simple random choices, whereas in between the two extremes is a region that require better and efficient methods for effective solutions of spread maximization.

In sparse networks, simple, degree based heuristics perform as well as the the optimal methods. Moreover, real–world networks are mostly sparse and generally have skewed degree

distribution. Hence, in most real–world networks, we do not need to employ computationally expensive methods for diffusion maximization.

In dense networks, I find that, due to the uniformity of local structure of nodes, extents of diffusion by any method, including the optimal, are very similar. I find that proximity networks, such as, animal social networks are usually very dense. Hence, a random set of spread initiators, in expectation, gives close to the optimal result for diffusion maximization.

In between the low and high effective density regions is an intermediate range of densities where the behavior of the optimal is markedly different from random diffusion method. Hence, it is this intermediate range of densities for which application of sophisticated and computationally intense methods is necessary for finding good diffusion initiators efficiently. However, even in this intermediate range of densities we find that the particular defining structure of the networks make local structural measures like degree and variants of degree heuristics to be good estimates of the optimal extent of diffusion. Moreover, observing the relatively well defined modularity of the network structure in this intermediate region for block mixture model we get a better understanding of what makes certain heuristics more effective than others. Thus, taking advantage of the insights provided by this rigorous experimental analysis, better structural based heuristics can be devised that are efficient and easier to evaluate.

To summarize, I experimentally showed that the optimal extent of diffusion in sparse networks is achieved by the high degree nodes and in very dense networks it is achieved by any choice of diffusion initiators for a given diffusion model. Hence, for networks of these effective density ranges we do not need to use computationally challenging methods to find the most

effective diffusion initiators. This result leads to the discovery of an intermediate effective density range where diffusion in networks is sensitive to the identity of the initiators. I further verify that the general degree–based heuristics for finding good diffusion initiators work very well for most networks due to their inherent sparsity, or boundary nodes–based heuristics due to the modular structure of the underlying network. Hence, we can simplify a computationally hard problem of finding critical individuals for maximizing diffusion by limiting the application of computationally expensive methods to within a certain range of densities and certain structures. This work also gives us the basis to further explore the effect of network structure on the extent of diffusion, both theoretically and empirically, to design algorithms that take advantage of this connection of extent of diffusion to the network structure.

6.5 Effect of dynamic network structure on diffusion in networks

In this work I examine how density and modularity of *dynamic networks* affect the choice of influence maximization heuristics and the optimality of the resulting solutions. In recent years, online social networks, such as Facebook and Twitter, have provided real examples for understanding diffusion of information, ideas, and adoption of products in fast evolving networks (15; 159). Other than social networks, applications such as contamination detection in water distribution networks or spread of diseases, also inherently rely on the dynamic process of diffusion (130; 134). In the past, almost all efforts focused on "static" networks (106) in which the ties among nodes remain fixed over time. In recent years, there has been some work done in studying diffusion in "evolutionary" networks, such as co–authorship networks. However, this is also a limiting example of social networks as the assumption here is that links are never lost.

Yet, most networks are highly dynamic. In social networks, individuals maintain different connections in different contexts, such as friends or work colleagues and these relationships change with time. Similarly, the animal networks change with age, migration, and risk factors in the environment. Despite all the real–world examples, there has been little work done in incorporating these relationship dynamics while analyzing the extent of influence in such networks with a few exceptions such as (10). And most of the work that has been done focused on designing algorithms or heuristics that work for any general dynamic networks.

6.5.1 Contributions

In this part of my work, analyze the global structure of explicitly dynamic networks for understanding when to employ a sophisticated algorithm vs a simple heuristic. Moreover, similar to the static network analysis for diffusion maximization (Section 6.4), I employ simple dynamic structural measures that can guide this choice. Specifically, I analyze networks in term of their changing density and community structure to investigate the trends in the extent of diffusion of influence.

In this work, I use density of a dynamic network as an indicator of influence maximization. I ask the following three questions:

- When it is necessary to employ a sophisticated yet computationally expensive method?
- When even a random set of spread initiators perform as well as the best in expectation for maximizing the spread in the network?
- Why certain heuristics, such as, degree centrality can be used as good initiator of the diffusion process in certain networks and not for others?

I show experimentally that for network densities above and below a certain threshold, the difference between the optimal and expected spread is negligible. In between the two extremes, the difference between the two approaches is markedly large. This region, rich with non–trivial and complicated structure, requires further work to devise efficient techniques for finding optimal spreaders.

6.5.2 Methodology

I use the statistical dynamic network generative model introduced in Section 4.3 to generate dynamic networks with a wide range of properties. This generating model provides us with an extensive set of parameters through which networks of varying densities and community structures can be generated. For this analysis, I use this dynamic network generative process to instantiate networks with skewed community size distribution. I use the Independent cascade diffusion model (92) to simulate diffusion in networks. I estimate three main types of extent of diffusion in these networks.

6.5.2.1 Methods

Following are the three methods used for estimating the extent of diffusion in dynamic networks.

1. Optimal (or the greedy approximation) extent of diffusion. Note that the greedy approximation gives the same result as the optimal when the initial set size $k = 1$. The details of both methods are provided in Section 6.2.4.1.
2. Expected extent of diffusion by a uniformly random set of diffusion initiators. Algorithm 4 formally describes this algorithm.

3. Extent of diffusion by degree and betweenness heuristics. This method is described in Algorithm 5.

6.5.2.2 Diffusion model

I simulate spread in the real–world and synthetic networks using the Independent cascade spreading model. This model is explained in full detail in Section 2.3.1.

6.5.2.3 Parameter settings

I perform extensive Independent cascade model based diffusion simulation over a set of networks sampled from the network generative process explained in Section 4.3. I sample thousands of networks by controlling for network modularity over time while controlling for the static network density. I sample networks with a skewed cluster size distribution based on the conclusions by Chung et al. for networks with skewed degree distributions (49). Thus, in this set of synthetic networks the giant component encompasses about 40–50 % of the network and all the rest of the components are some constant fraction of log of the network size. The connectivity of nodes within each cluster is set using the preferential attachment model. For the preferential attachment model the exponent of skewness γ is set between 2 and 3, as has been shown for most real–world networks (123). For networks based on this model, the minimum average degree is set slightly above 1 to mimic the growth process. The resulting synthetic networks have a bimodal degree distribution. That is, there is a large frequency of smaller degree nodes but also a relatively significant number of nodes that have degrees closer to the size of the largest component. This instantiation of the network generative process also captures

the modular structure of the underlying dynamic network, as is evident by the sample graph in Figure 1. I sample over 500 such networks of various densities and levels of modularity.

6.5.3 Experimental results

I compare the optimal extent of diffusion with the expected extent of diffusion in the synthetic dynamic networks. I find that networks with very low static effective density have highly skewed static structure. This renders heuristics like highest degree very effective for diffusion maximization. On the other hand, networks that are very dense (static density), the best methods for estimating the extent of diffusion are comparable to the baseline, that is a uniformly random set of spreaders. However, there is an in–between region of density for which the difference between the optimal and random is very significant. This is meaningful since with only the basic knowledge about the underlying network structure, I observe, that it is easier to determine when we need to employ a sophisticated yet computationally expensive method and when something as simple as random set of spread initiators will work for diffusion maximization. I also compare the optimal spreaders to the heuristic methods in that intermediate density region to evaluate their effectiveness. I find that although degree heuristic gives result closer to the optimal, it is the nodes that connect communities that consistently perform well in this region. I further estimate the extent of diffusion in this region of density using the dynamic betweenness centrality measure. Recall, the dynamic betweenness centrality presented in Section 5.2.2, ranks each node in the network based on their structural as well as temporal position.

Figure 18 shows results of difference in optimal and expected for various seed set = 5, 10, and 15. From this result I observe the two phase shifts in the difference in the extent of diffusion at effective density of 0.01 and 0.3. Again, this shows that in dynamic networks, just as in static network, above and below certain densities it is easier to estimate the greedy–optimal extent of diffusion by something as simple as a random set of diffusion initiators. However, in between the low and high density region, the difference in the greedy–optimal and the expected is significantly large. I use the degree and the dynamic betweenness heuristic to show that within this region simple heuristics give estimates close to the greedy–optimal.

Figure 19 shows the difference in optimal and random, degree, and betweenness. Clearly, there are two phase shifts for when the expected performs as well as optimal. As is observed in the results degree gives near optimal extent of diffusion in low density networks. Dynamic betweenness heuristic gives diffusion estimates remarkably close to optimal in dynamic networks with intermediate density. Moreover, degree as well as random expected extent of diffusion perform much worse as compared to the optimal. Towards the dense network extreme, all methods converge to the optimal, including the uniform random expected extent of diffusion.

6.5.4 Conclusions

I show that dynamic networks with densities above and below a certain threshold are amenable to simple heuristics. In dense networks, in fact, there is no differentiation between optimal influence of individuals and any random set of individuals is good, given a high enough rate of diffusion. In sparse networks simple heuristics like highest degree nodes perform well. In between the two extremes, the difference between the optimal and a random seed set is signifi-

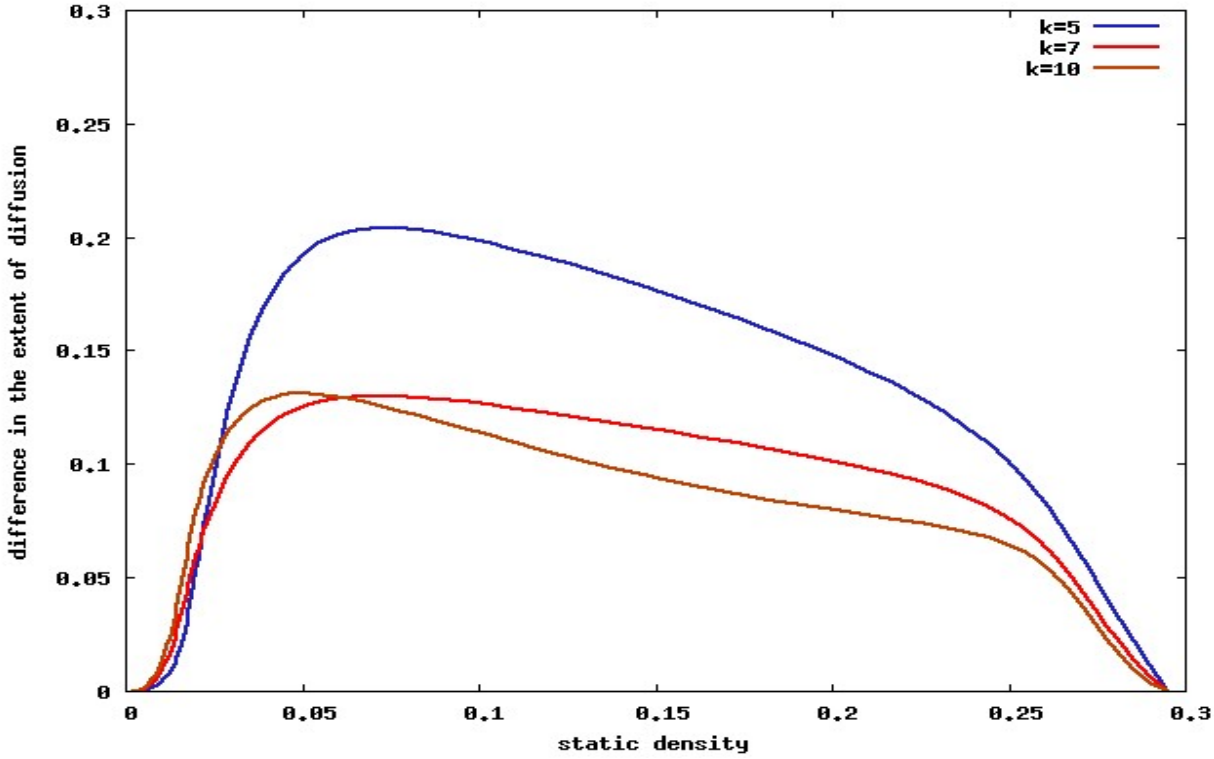


Figure 18: Difference in the greedy–optimal and the expected extent of diffusion

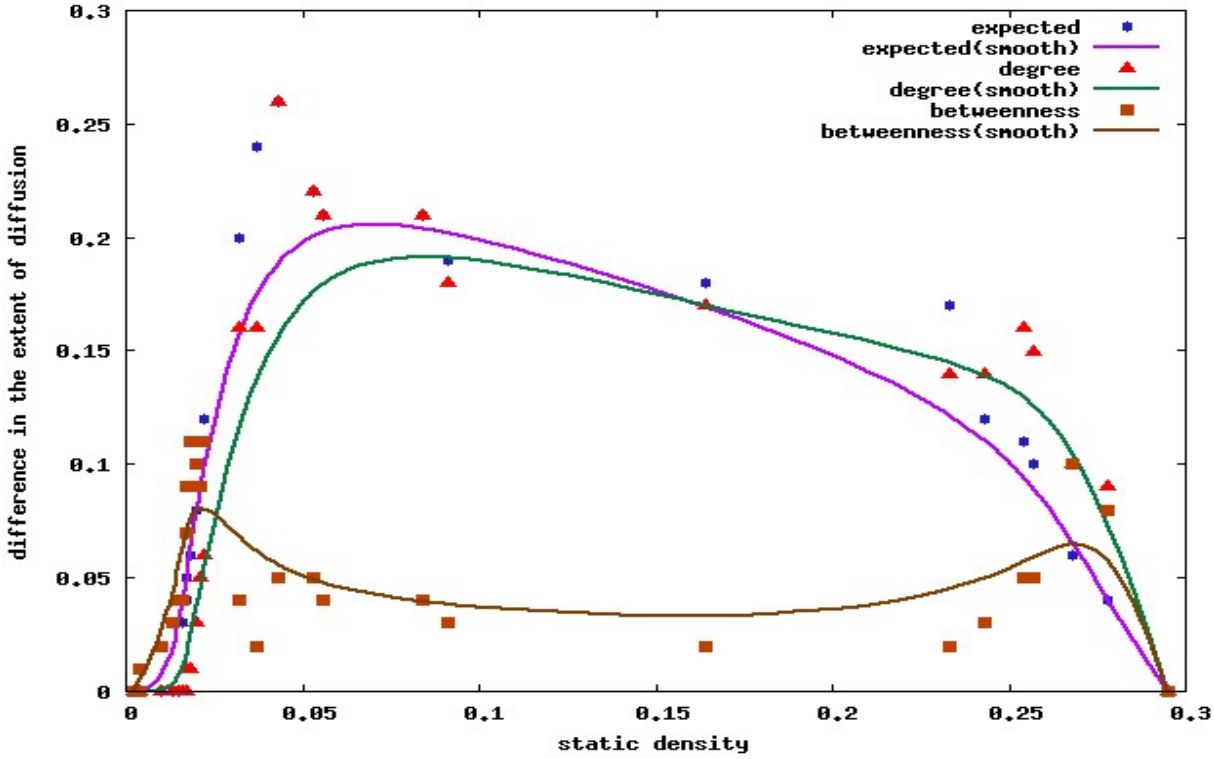


Figure 19: Difference in the greedy-optimal and Expected, Degree and, Betweenness heuristics

cant. That difference, in fact, depends on how modular (clumped, clustered, non uniform) the network is. Networks with such densities have complex structures that require sophisticated methods for influence maximization.

CHAPTER 7

DIFFUSION MINIMIZATION

In the context of the propagation of diseases or undesirable behavior, it is important to *minimize* extent of diffusion in a network by identifying a set of *blockers*: individuals that are most effective in stopping or slowing down the spread of a process through the population. This problem of diffusion minimization is complementary to the diffusion maximization problem discussed in Chapter 6.

How can we stop the diffusion of a dynamic process through a social network? This problem has applications to many diverse areas such as preventing or inhibiting the spread of diseases (25; 63; 86), computer viruses[1] (26; 57), rumors, and undesirable fads or behaviors (59; 60; 80; 81). A common approach to inhibit the diffusion of such processes is to identify key individuals whose removal will most dampen the diffusion. In the context of the spread of a disease, it is a question of finding individuals to be quarantined, inoculated, or vaccinated so that the disease is prevented from becoming an epidemic. I call this set of key individuals the *blockers* of the spreading process.

---

[1]In particular, we are concerned with computer malware that spreads through social networks, such as email viruses and worms, cell-phone viruses, and other related malware such as the MySpace worm.

7.1 Problem statement

Diffusion minimization is the problem of finding a set of nodes in a network, whose removal from the network reduces the extent of diffusion in the network the most. I call such a set of nodes as "blockers" of the diffusion process and the function that finds such a set of nodes as "blocking function". Informally, blockers are a set of entities in a network that minimizes the extent of diffusion in the network the most. In case of a virus or disease propagation in a network, it is the set of individuals that once quarantined or vaccinated will result in containing the extent of diffusion of the disease the most. In case of contaminant flowing through a network of pipes and junction, it is the set of junction that when equipped with proper sensors, minimizes the time of detection of the contaminant or the source of contaminant.

I formally define the *Diffusion Minimization* problem in terms of the blocking function as follows:

Definition (Blocking) A blocking funcion $kBl(G)$ is a minimization function that finds a set of k nodes whose removal from the network *minimizes* the expected extent of diffusion the most in that network. Let $Diffusion_v(\cdot)$ be the expected extent of diffusion in the network. Then, the expected minimization in the extent of diffusion after removing k nodes from the network G is:

$$kBl(G) = \min_{X \subseteq V, |X|=k} Spread(G \setminus X).$$

An alternative but equivalent definition is based on maximization:

Definition (Blocking) Blocking $kBl(\cdot)$ is the function that finds k individuals in the network that results in the maximum reduction in the extent of spread in the network when that set is removed from the network. That is, this function finds the best blocker(s) in the network. Let $Bl_X(\cdot)$ is a function that measures the reduction in the expected extent of diffusion *after removing* the set X of individuals from the network.

$$Bl_X : G \to \mathbf{R}^+, \qquad Bl_X(G) = Spread(G) - Spread(G \setminus X).$$

Then Blocking is:

$$kBl(G) = \max_{X \subseteq V, |X|=k} Bl_X(G).$$

This definition of the individuals' blocking capacity by its removal, in the context of diffusion of a contagion, corresponds to the quarantine action. Vaccination or inoculation leave the node in the network but deactivate its ability to propagate the contagion. A subtle difference between removal of a node and vaccinating a node is that the former changes the network structure whereas the later only blocks the flow. For *susceptible infected removed* model from mathematical epidemiology the two actions are equivalent at the abstract level for estimating the extent of diffusion in networks. Another interpretation of diffusion minimization is to limit the diffusion of a contaminant by detecting the source of contaminant as quickly as possible. Here the goal of optimization is to minimize the time of detection. Depending on the application, this goal is different from minimizing the *impact* of diffusion from the above definitions.

7.2 Finding good blockers in dynamic networks

A comprehensive survey of work in this area is given in Section 3.3. Although diffusion minimization by finding critical junctions, points, or nodes in a network is an extremely important problem for many applications such as spread of viruses, diseases, contamination, among others. However, most of the scientific work in this regard is limited to either some applications or some simple functions. This problem generally has so far resisted systematic algorithmic solutions. In an effort to formulate practical solutions, in the following work we ask: *Are there structural network measures that are indicative of the best blockers in dynamic social networks?*

I compare the blocking ability of individuals to the values of their structural properties proposed in Chapter 5. For instance, I analyze whether a high degree node is a good blocker. I find that overall, simple ranking according to a node's static degree, or the dynamic version of a node's degree, performed consistently well. Surprisingly the dynamic clustering coefficient seems to be a good indicator, while its static version performs worse than the random ranking. This provides simple practical and locally computable algorithms for identifying key blockers in a network.

7.2.1 Contributions

There has been significant previous work related to studying and controlling the spread of dynamic processes in a network (27; 28; 29; 40; 46; 57; 59; 63; 75; 86; 92; 93; 106; 107; 117; 124; 133; 139). Unfortunately, these results have three properties rendering them ineffective for identifying good blockers in large networks.

1. Many proposed algorithms focus on identifying nodes that will be most effective in *starting* the spread of a process rather than blocking it (92; 106); or alternatively, focus on identifying nodes in the network that will be most effective in sensing that a process has started to spread, and where the process initiated (27; 28; 29). In this work, my focus is specifically on identifying those nodes that are good blockers.

2. Algorithms proposed in previous work all require computationally expensive calculations of some global properties over the entire network, or rely on expensive, repeated stochastic simulations of the diffusion of a dynamic process. In this work, I present algorithms that identify good blockers quickly, based only on local information.

3. Perhaps the most critical problem in previous work is the elision of the dynamic nature of social interactions. The very nature of a diffusion process implies an explicit time axis. For example, the flow of information through a social network is dependent on who has the information at what point in time and who are the individuals in contact at that moment with the information carrier that are likely to acquire the information next (93). In this work, I focus on explicitly dynamic networks, defined in Section 2.1.2. In these networks, I study the social interactions over a finite period of time, measured in discrete timesteps.

The main contributions of this work are summarized below.

- I formally define the problem of identifying key diffusion blockers in networks in Section 7.1.

- I measure the network properties defined in Section 2.2 of each node and rank the nodes based on those values.
- I compare the reduction in the extent of diffusion based on removing individuals from a network in the ranking order imposed by various network measures. I identify measures that consistently give a good approximation of the best spread blockers.
- I compare the difference in the sets of top blockers identified by various measures.
- I extensively evaluate our methods on the following real dynamic networks: Enron email network, Bluetooth networks of MIT Reality Mining and UMass proximity, DBLP co–authorship network, and animal population networks of grevys zebras and onagers. The details of these networks are provided in Section 4.1.

Ultimately, I show that the dynamics of interactions matters and simple local measures, such as degree, are highly indicative of an individual's capacity to prevent the spread of a phenomenon in a population. The implication of our results are that there are practical scalable heuristics for identifying quarantine and vaccination targets in order to prevent an epidemic.

7.2.2 Methodology

I evaluate the effectiveness of each of the structural properties as indicator of node's blocking capacity under the Independent Cascade spreading model. The model is explained in Section 2.3.1. Following is the algorithm for estimating the extent of diffusion in each dynamic network $G = (V, E)$ using each network property/measure.

7.2.2.1 Algorithm

For each measure and for each dynamic network dataset, we follow the following steps:

1. Order the individuals $0, \ldots, |V-1|$ according to the ranking imposed by the measure.
2. For $i = 0$ to $|V-1|$ do:
   (a) Remove node i from $G = (V, E)$.
   (b) Estimate the extent of spread in $G \setminus i$ by averaging over stochastic simulations of Independent Cascade model initiated at each node in turn, 3000 iterations for each starting node.
   (c) If the extent of spread is less than 10% of the nodes in the network then Stop.

   Set of the removed individuals are the best blockers according to their ranks.

7.2.2.2 Probability of activation

We conducted the Independent Cascade spreading experiments on a variety of networks with diverse structural properties like density/sparsity, diameter and average path length etc. In each network, we assigned a different probability of activation based on the structure of the network. Following is the procedure we used to find a meaning full probability of activation to work with for a given network.

1. For a given $G = (V, E)$, run the Independent Cascade Spreading process with $p = 1$. Notice that this is a deterministic process.
2. Calculate the average extent of spread S in $G = (V, E)$. Notice, this will give us roughly the size of the largest connected component of the network G.
3. Rerun the spreading process while setting $p < 1$. Calculate the average extent of spread in the network.

4. Repeat the last step while reducing p until the average extent of spread is only as large as 50% of S.

5. Set probability of activation for G equal to p.

7.2.2.3 Network properties

I use the following network properties as heuristics: dynamic and aggregate versions of degree, betweenness, closeness centralities, and clustering coefficient, as well as the average dynamic degree (turnover rate). For the global measures of betweenness and closeness we locally approximate them within 1, 2, and 3 hops neighborhoods. I also rank individuals based on their neighbors within 1, 2, and 3 hops of nodes and edges. The static network properties are defined in Section 2.2.2 and their dynamic counterparts are introduced in Section 5.2. Since the diffusion occurs primarily due to local level interactions, I introduce two properties based on the local connectivity of each node. Following are the two measures that exploit the local neighborhoods of individuals.

Nodes in Neighborhood is the number of nodes in the local neighborhood of an individual. The number of nodes in the 1-neighborhood is precisely the degree of an individual. I extend the measure by considering the 2- and 3-neighborhoods of each individual.

Edges in Neighborhood is the number of edges in the local neighborhood of an individual. I compute the edges in neighborhood for 1, 2 and 3 hop neighborhoods of each individual. This measure loosely captures the local density of the neighborhood of an individual.

Overall, I use 26 different measures.

I compare the extent of diffusion using the above structural properties heuristic to the optimal exhaustive search of the best blockers or greedy local search algorithm. Moreover, I compare the heuristics to the random selection of k nodes as blockers of diffusion in the network.

7.2.2.4 Lower Bound: Best Blockers

I identify the best blockers one at a time using exhaustive search over all the individuals. To find one best blocker, I remove each individual, in turn, from the network and estimate the extent of spread using stochastic simulations of the Independent Cascade model in the remaining network. The best blocker, then, is the individual whose removal results in the minimum extent of diffusion. I then repeat the process with the remaining individuals. This process imposes another ranking on the nodes.

Optimally, one needs to identify the *set* of top k blockers. However, this problem is computationally hard and an exhaustive search is infeasible. I have conducted limited experiments on the datasets considered in this paper and in all cases the set of iterative best k blockers (estimated using the above procedure) equals to the set of top k (optimal using exhaustive search) blockers. This preliminary result warrants future investigation and rigorous evaluation.

7.2.2.5 Upper Bound: Random Blockers

To estimate the worst blocking scenario in a network I pick individuals at random to block. I estimate the reduction in spreads after removing each individual. I repeat the process until only 10% of the remaining population is infected by the spreading process. This method is repeated for several runs R and eventually the average extent of diffusion is calculated over all R runs. For these experiments R is set to $c\log(N)$, where c is a constant greater than 1 and N

is the network size. This Upper bound gives us an insight into how hard blocking the spread of a phenomenon in certain network structure can be.

7.3 Datasets

In this section I briefly describe the datasets used in the experiments. I use the following datasets in these experiments: grevys, onagers, DBLP, Reality Mining, Enron, and UMass. These network datasets are explained in extensive detail in Section 4.1. Table VIII provides a summary of the statistics of the networks used in these experiments. Here V is the number of individuals, E is the number of edges, T is the number of timesteps, D is density and D_T is dynamic density, d is the diameter within a connected component and d_T is the dynamic diameter, and p is average shortest path length and p_T is the average temporal shortest path length.

	grevy	onagers	DBLP	Enron	MIT	UMass
V	28	29	1374	147	96	20
E	779	402	2262	7406	67107	2664
T	44	82	38	701	2940	693
D	0.30	0.36	0.002	0.04	0.68	0.72
D_T	0.52	0.24	0.09	0.14	0.18	0.35
d	4	3	15	6	2	
d_T	36	74	37	618	315	8
p	1.84	1.66	5.5.4	2.66	1.32	
p_T	4.81	7.51	5.12	461.24	4.21	3.71

TABLE VIII: Dynamic network dataset statistics.

7.4 Results and Discussion

For each of the datasets I have evaluate all the measures to determine how effectively they identify good blockers. Figure 20, Figure 21, Figure 22, Figure 23, Figure 24, and Figure 25 show reduction in the extent of diffusion from the best measures on all datasets. For all the plots, the x–axis is the number of individuals removed and the y–axis shows the corresponding extent of spread. The lower the extent of diffusion, the better is the blocking capacity of the individuals removed. Thus, curves lower on the plot correspond to measures that are better indicators of blocking power.

The comparison of all the measures showed that four measures performed consistently well as blocker indicators: degree in aggregate network, the number of edges in the immediate aggregate neighborhood (local density), dynamic average degree, and dynamic clustering coefficient. This is good news from the practical point of view of designing epidemic response strategies since all the measures are simple, local, and easily scalable. Surprisingly, while the local density and the dynamic clustering coefficients seem to be good indicators, the aggregate clustering coefficient turned out to be the worst, often performing worse than a random ordering. Betweenness and closeness performed inconsistently. Page Rank did not perform well in the only dataset with directed interactions (Enron)[1].

As seen in Figure 24 and Figure 25, the ease of blocking the spread depends very much on the structure of the dynamic network. In the two bluetooth datasets, MIT Reality Mining

---

[1]On undirected graphs, Page Rank is equivalent to degree in aggregate network

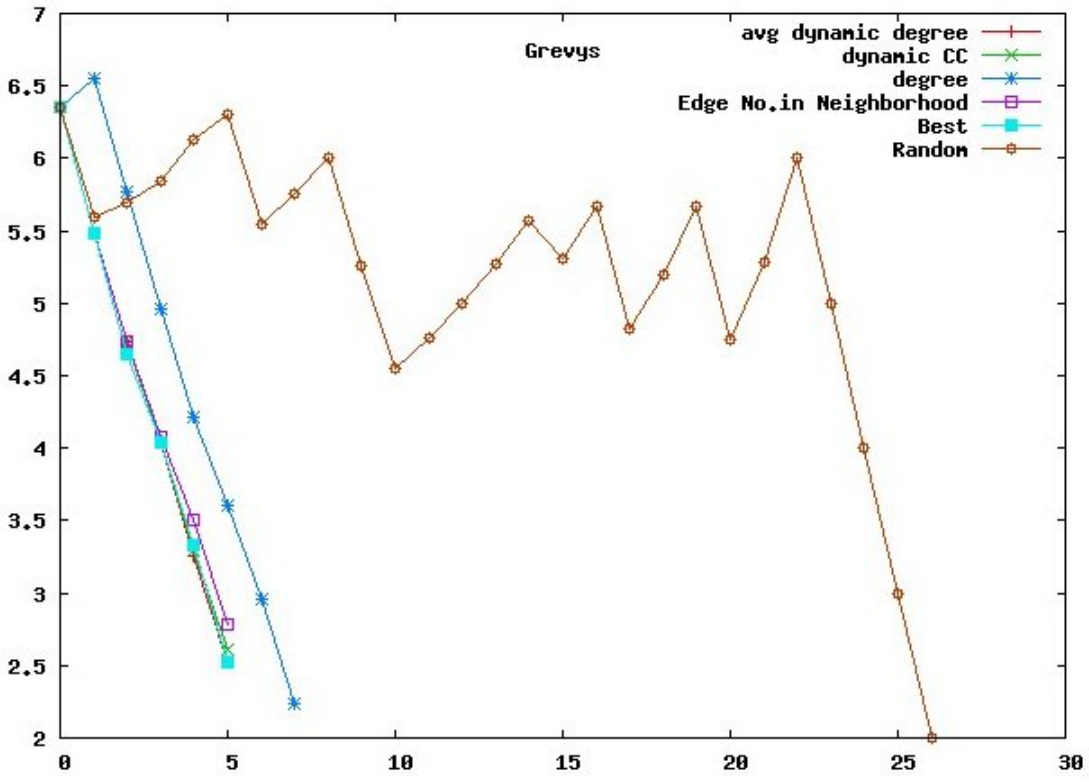


Figure 20: Reduction in the extent of diffusion - grevys

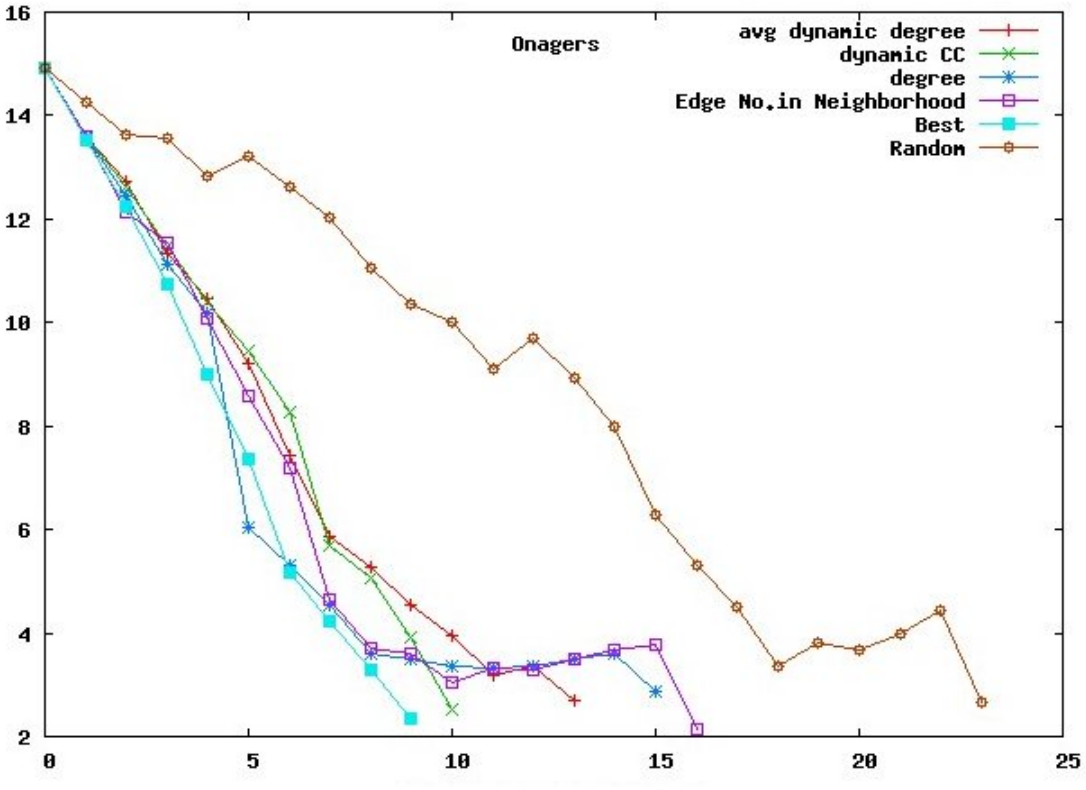


Figure 21: Reduction in the extent of diffusion - onagers

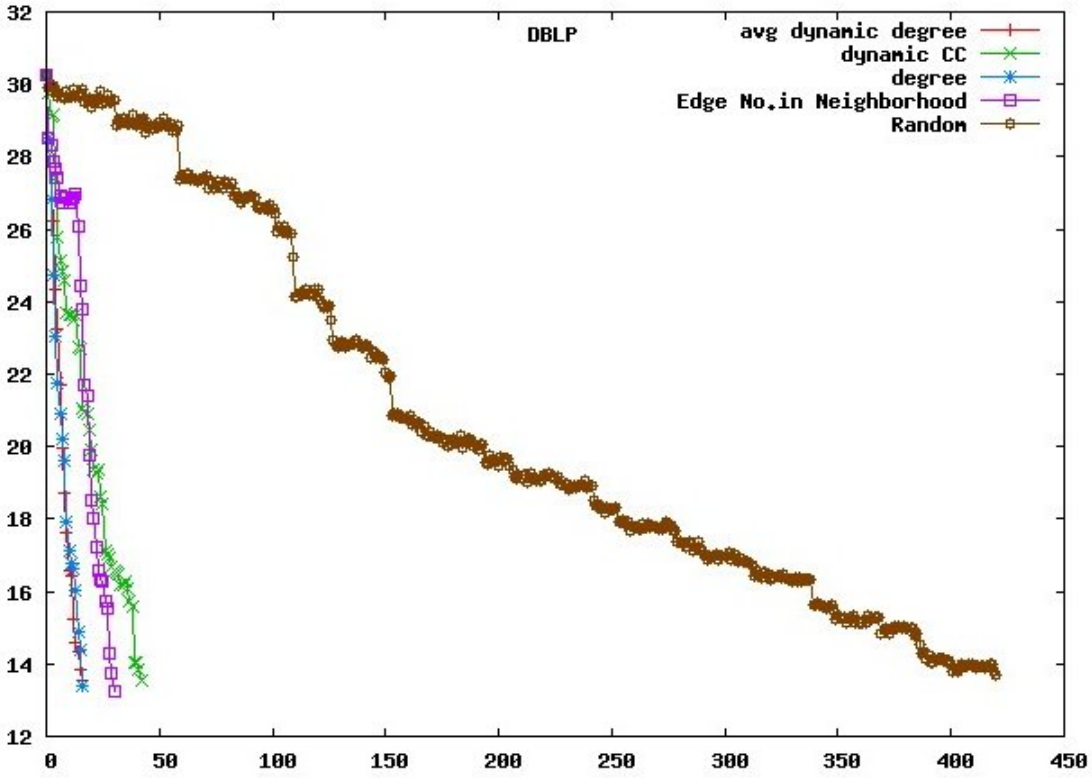


Figure 22: Reduction in the extent of diffusion - DBLP

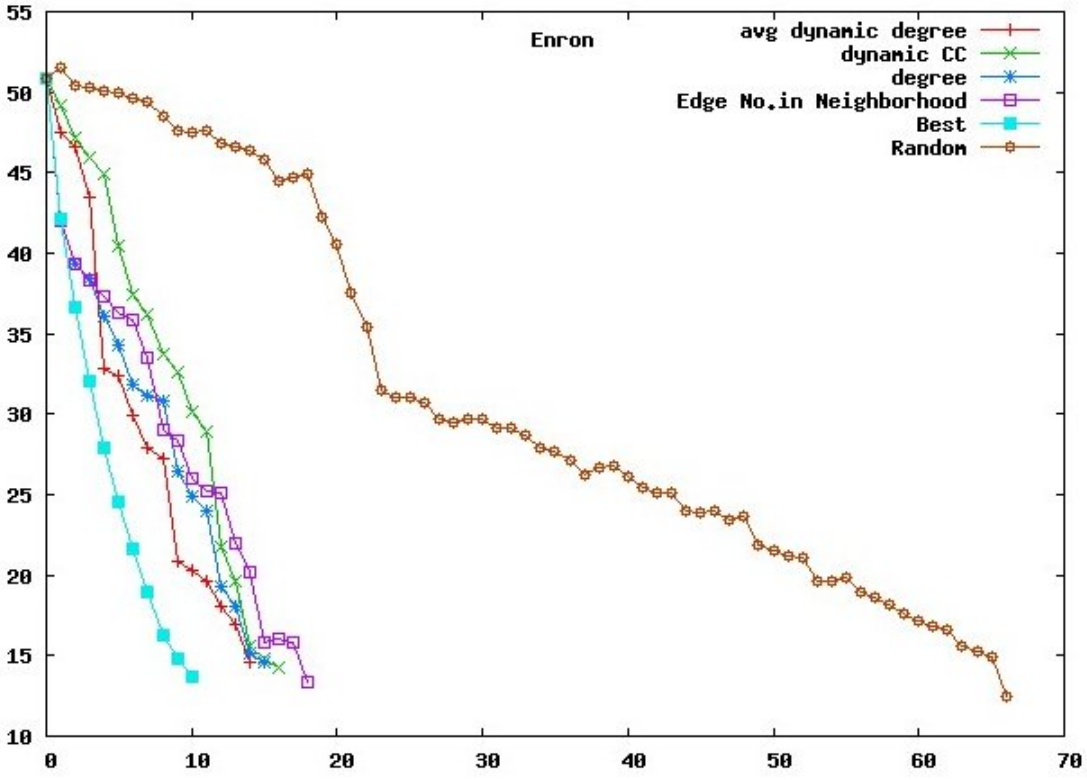


Figure 23: Reduction in the extent of diffusion - Enron

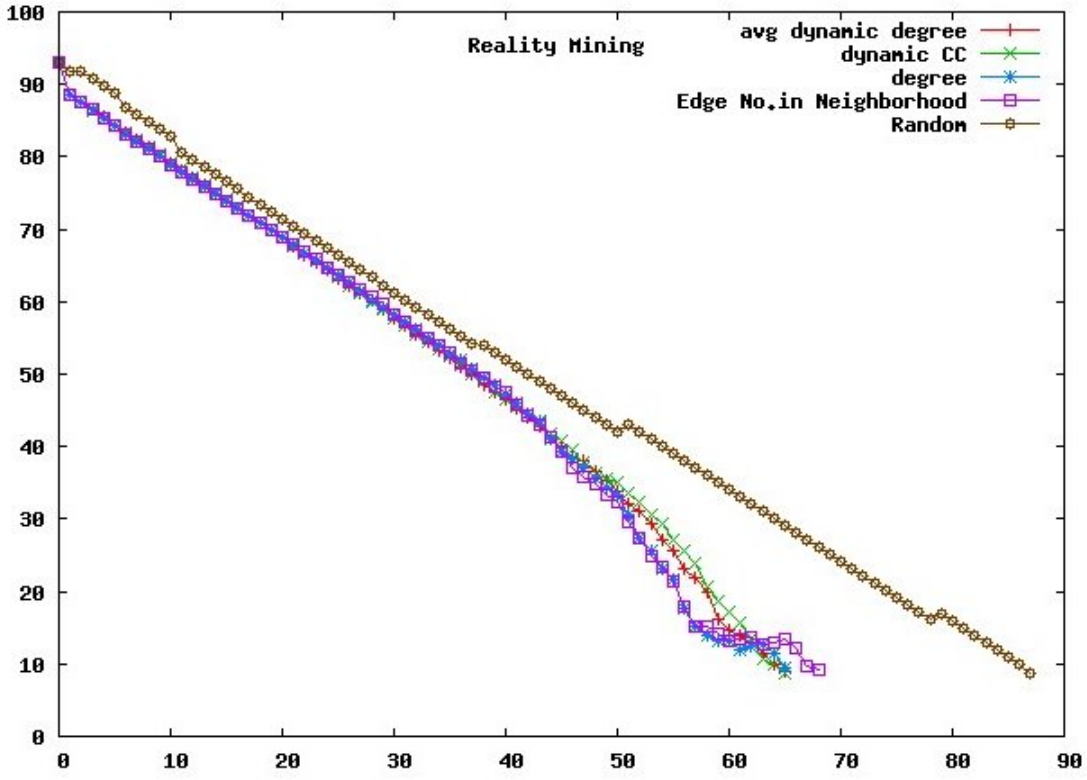


Figure 24: Reduction in the extent of diffusion - Reality Mining

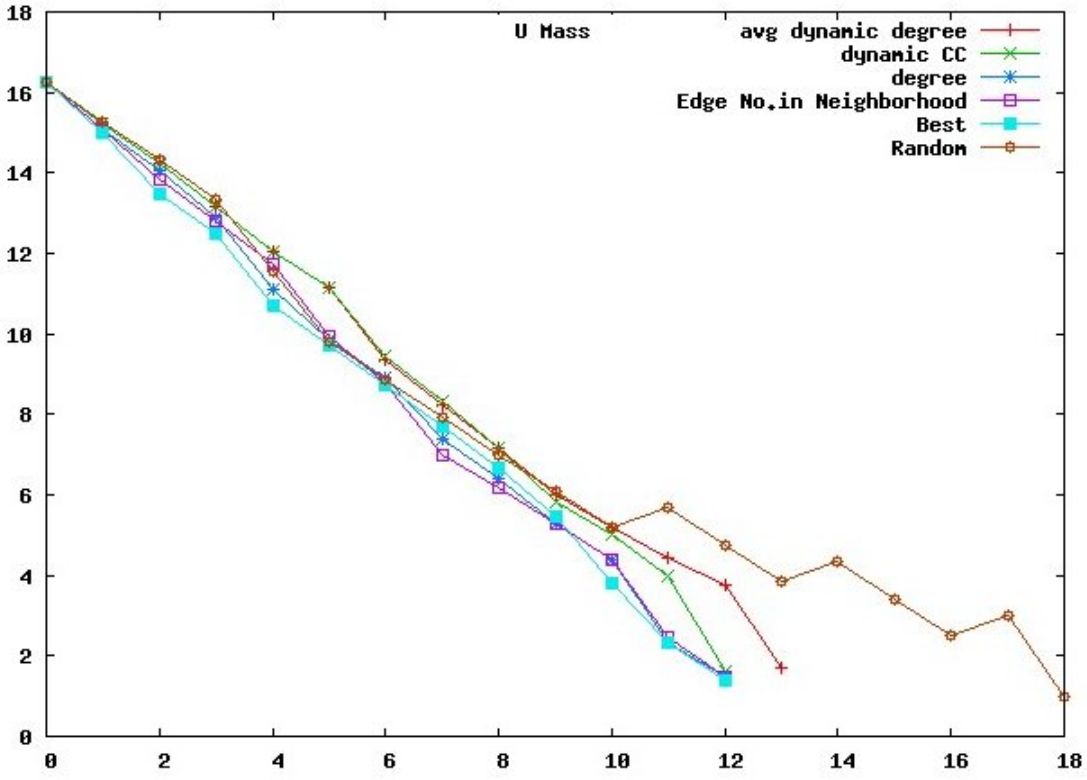


Figure 25: Reduction in the extent of diffusion - UMass

and UMass, all orderings, including the random, performed similarly. Those are well connected networks, as evident by the large difference between the dynamic diameter and the average shortest temporal path. The only way to reduce the extent of spread to below 10% of the original population seems to be trivially removing nearly 90% of the individual population. On the other hand, Enron and DBLP show the opposite trend of being easily blockable by a good ranking measure.

In addition to comparing the extent of spread based on the ranking by various measures, I compare the sets of the top ranked blockers identified by the four best measures as well as the best possible ordering. Figure 26 and Figure 27 show the scatter plots of the pairwise comparisons of rankings induced by the four measures. The scatter plots show that, in general, there is little correspondence between the rankings imposed by various measures. The only strong relationship, as expected, is between the number of edges in the neighborhood of a node and its degree in the aggregate network. Each plot shows the ranking of the individuals according to two measures. Individuals ranked top by both measures are in the upper right corner of each plot. Nodes with the same value for a particular measure have the same rank for that measure.

I further explore the difference in the sets of the top ranked individuals by computing the size of the common intersection of all the top sets ranked by the four measures and the best possible ranking. I use the size of the set determined by the best possible ordering as the set size for all measures. Table IX shows the size of the common intersection for all datasets. Again, I see a strong effect of the structure of the network. The MIT Reality Mining and the UMass

Dataset	Set size	Inter. size	Inter. frac
grevys	5	2	.40
Onagres	9	3	.33
DBLP	16	0	0
Enron	13	4	.31
Reality Mining	60	48	.80
UMass	12	10	.83

TABLE IX: The size of the common intersection of all the top sets ranked by the four measures and the best ranking. Set size is the size of the sets determined by the best blocking ordering. The size of the intersection is the number of the individuals in the intersection and the Intersection fraction is the fraction of the intersection of the size of the set.

datasets have most of the same nodes ranked as top by all measures. On the other hand, in DBLP the four measures produced very different top ranked sets, yet all four measures were extremely good indicators of the blockers. In other networks, while there are some individuals that are clearly good blockers according to all measures, there is a significant difference among the measures. Overall, these results lead to two future directions: 1) investigating the effect of the overall network structure on the "blockability" of the network; and 2) designing consensus techniques that combine rankings by various measures into a possible better list of blockers.

7.5 Conclusions and Future Work

In this work I investigated the task of preventing a dynamic process, such as disease, from spreading through a network of social interactions. I formulated the problem of identifying good blockers: nodes whose removal results in the maximum reduction in the extent of spread in the network. I focused on identifying structural network measures that are indicative of whether or not a node is a good blocker. Since the timing and order of interactions is critical

in propagating many spreading phenomena, I focused on explicitly dynamic networks. I used 26 different measures as candidates for identifying good blockers in dynamic networks.

I conducted experiments on six dynamic network datasets spanning a range of contexts, sizes, density, and other parameters. I compared the extent of diffusion while removing one node at a time according to the ranking of nodes imposed by each measure. Overall, four measures performed consistently well in all datasets and were close to identifying the overall best blockers. These four measures were node degree, number of edges in node's neighborhood, dynamic average degree, and dynamic clustering coefficient. The traditional aggregate clustering coefficient and dynamic closeness performed the worst, often worse than a random blocking process. All four best measures are local, simple, and scalable measures and, thus, potentially lead to good practical epidemic prevention strategies. However, before such policy decisions are made, we need to verify that my results hold true in other, larger and more complete datasets and are not an artifact of the particular types of dynamic data sets that are currently available to the research community.

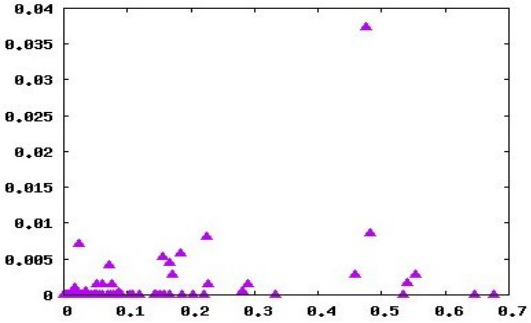
The striking disparity between the performance of the dynamic and aggregate clustering coefficient indicates the necessity of taking the dynamic nature of interactions explicitly into consideration in network analysis. Moreover, this disparity justifies the extension of traditional network measures and methods to the dynamic setting. In future, the informativeness of a range of dynamic network measures in various application contexts needs to be further investigated.

I also compared the sets of nodes ranked at the top by various measures. Interestingly, in the networks in which it was difficult to block the diffusion process, all the measures resulted

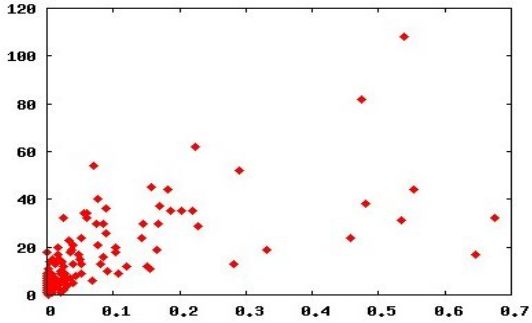
in very similar rankings of individuals. In contrast, in the networks where the removal of a small set of individuals was sufficient to reduce the spread significantly, the best measures gave very different rankings of individuals. Thus, there seems to be a dichotomy in the real–world networks I studied. On one hand, there are certain types of networks (*e.g.* MIT Reality Mining and UMass datasets) that are hard to block in that it is inherently challenging to find good blockers, no matter what measures are used. On the other hand, there are certain types of networks that are inherently easy in the sense that many different sets of nodes are good blockers. In future, the specific *structural* attributes of a network that delineate this difference between networks for which it is hard or easy to identify good blockers needs to be further investigated.

The comparison of the top ranked sets also shows that while there may be some common nodes ranked high by all measures, there is a significant difference among the measures. Yet, all the rankings perform comparably well. Thus, there is a need to test a consensus approach that combines the sets ranked top by various measures into one set of good candidate blockers. This is similar to combining the top k lists returned as a web search result (65).

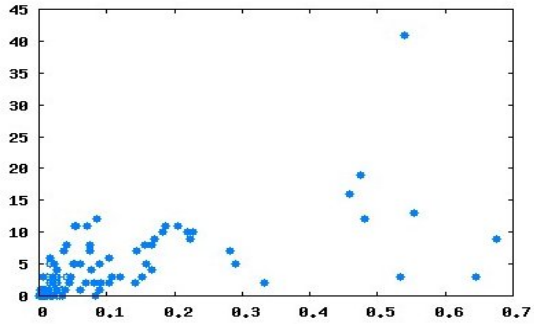
This work focused on the practical approaches to identifying good blockers. However, the theoretical structure of the problem is not well understood. While the blocking function is submodular, it is non–monotonic and, thus, so far has defied good approximation algorithms. Recent developments in the analysis of non–monotonic submodular functions (67; 154) may be applicable in this context and may result in good approximation guarantees.



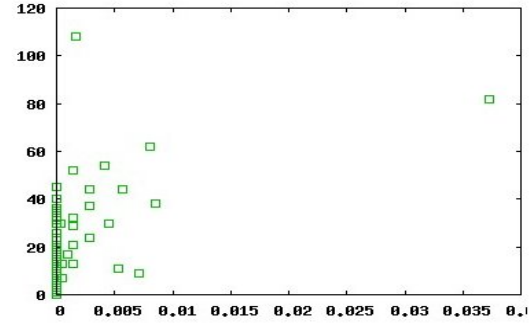
Dyn Avg Deg vs Dyn CC



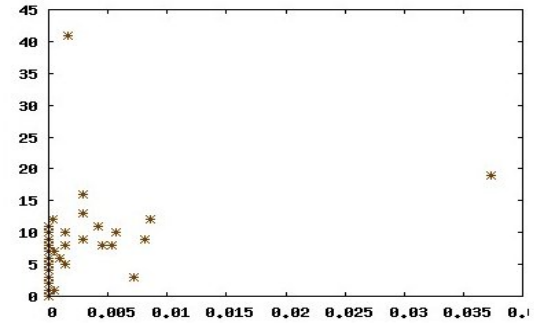
Dyn Avg Deg vs ENN1



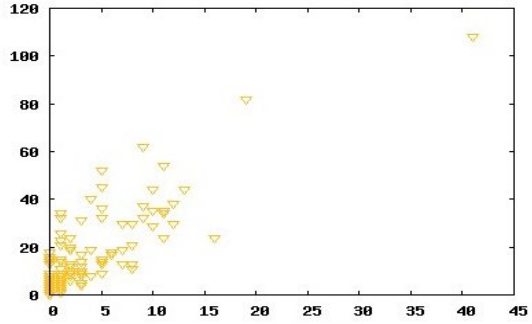
Dyn Avg Deg vs Agg. Deg



Dyn CC vs ENN1

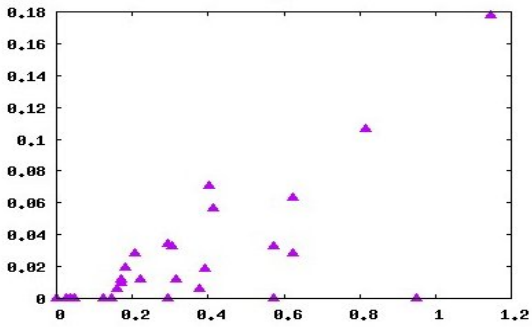


Dyn CC vs Agg Deg

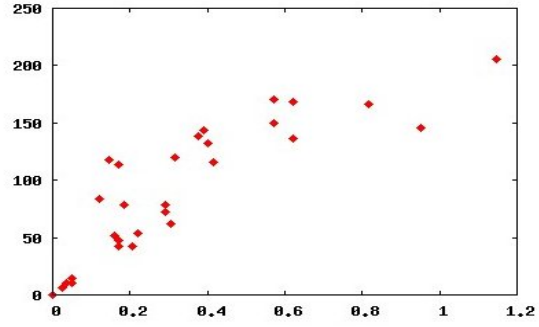


Agg Deg vs ENN1

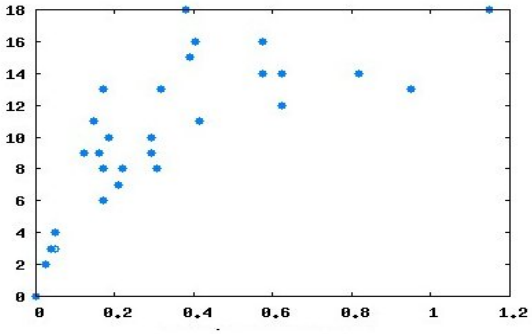
Figure 26: Comparison of best measures in Enron dataset



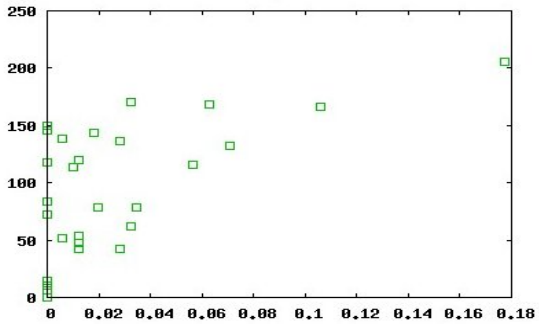
Dyn Avg Deg vs Dyn CC



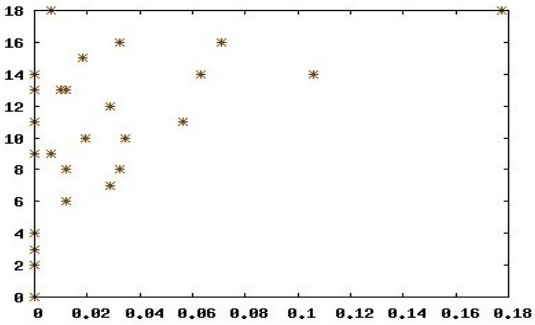
Dyn Avg Deg vs ENN1



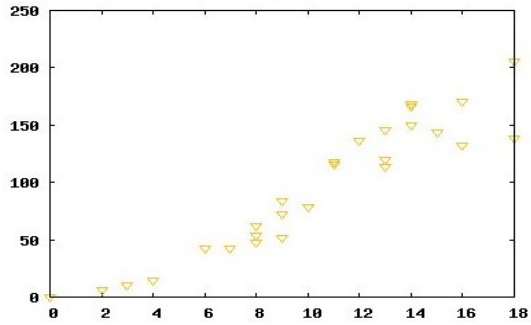
Dyn Avg Deg vs Agg. Deg



Dyn CC vs ENN1



Dyn CC vs Agg Deg



Agg Deg vs ENN1

Figure 27: Comparison of best measures in onagers dataset

CHAPTER 8

SOCIAL ORGANIZATION IN EQUIDS

Social interactions among various *equids* (zebras, horses, asses) populations is one of the ecological applications of my research. Across various types of equids, ecologists have observed an immense diversity in social behavior. Social interactions differ in their type, frequency, and duration. For example, interactions among these animals can be cooperative, antagonistic, or territorial, as well as interactions or social bonds may last for years or minutes or just seconds (98). Which type of interaction occurs and with what frequency and duration will depend on factors such as foraging resources, risks of predators, population size, among others (98). This raises the question for ecologists as to how to comprehend multiple and complex interaction patterns that can arise even if the number of participants is relatively small.

8.1 Motivation

Equids display many different forms of social organization; plains zebra associate in small, closed harems, whereas onagers (asses) and grevys zebra are found in looser, more ephemeral associations. Horses appear to be somewhat intermediate (69; 70; 150). The goal of this work is to find links between the population network structure and the various functions these animals tend to optimize for in their interactions. A by–product of this analysis is that the ecologists get a better sense of salient differentiating features in the network structures of the two types of equids based on the vitalness of the functions each population is attempting to optimize for.

With this work I provide ecologist sophisticated computational tools for understanding various social functions and organization of these animals. I perform extensive experiments on these networks to analyze the network structures and their properties. I identify critical network properties that characterize a certain network structure for optimizing a certain diffusion goal. With these methods ecologists are better equipped to answer questions about different social structures in equids.

8.2 Network structure and functions

Sociologists started addressing the above question more than sixty years ago by observing human interaction patterns and how those interactions were shaped by various functions performed by humans. In recent years, models of interactions and the processes (functions) taking place on them, have been extensively formalized in the fields of statistical physics and mathematical epidemiology, among others. Network theory has long been used to investigate human social organization. Its main strengths being the potential to address population–level problems by building up complex social structures from individual level interactions (147; 155). Furthermore, network statistics can be used to derive novel quantitative measures that characterize social structure at the level of the individual, and these measures can then be used alongside other standard statistical variables (*e.g.* measures of relatedness or reproductive success). These advances provide us with a powerful set of tools for computational analysis of animal social networks.

8.2.1 Statistical physics

Physicists have long been hypothesizing that biological systems can be understood in the same way as collective behavior in physics, where statistical mechanics provides a bridge between microscopic rules and a macroscopic phenomena. Most of the work in this line of research aims to show how individual level attributes can help to explain and predict patterns at the level of populations that can propagate to higher –more abstract– levels of organization in the population (116). For example, power–law has been shown to explain the fractal nature of nature (36). In the same vein, I study the behavior propagation of equids using probabilistic diffusion models. Overtime, this propagation results in stable network structures that is analogous to a certain network signature that signifies the social dynamics of various species and how they differ from one another.

8.2.2 Network theory

Network theory provides a quantitative framework that can be used to characterize social structure both at the level of the individual and the population. These quantitative variables provide a new tool in addressing key questions in behavioral ecology particularly in relation to the evolution of social organization and the impact of social structure on evolutionary processes. For example, network measures could be used to compare social networks of different species or populations making full use of the comparative approach. However, the networks approach can in principle go beyond identifying structural patterns and also can help with the understanding of processes within animal populations such as disease transmission and information transfer.

Finally, understanding the pattern of interactions in the network (i.e. who is connected to whom) can also shed some light on the evolution of behavioral strategies.

These techniques from stochastic modeling as well as network theory, make it possible to calculate quantitative metrics describing social structure across different scales of organization, from the individual to the population with respect to the functions performed by the individuals in these networks.

8.3 Problem statement

In this work I compare the relationship between social network structure of equids - specifically, grevys zebra and onagers - to the functions (leadership formation, loyalties, protection, survival) these structures are optimized for, using Independent Cascade (Section 2.3.1) and Linear Threshold (Section 2.3.2) diffusion models.

8.4 Diffusion maximization

I study the differences, between grevys and onagers populations, in maximizing various trends of diffusion (such as, awareness of a predator, knowledge of water and food resources) using the Independent cascade and Linear threshold models. I compare the following two trends in the two networks:

Extent of diffusion: The expected size of the population affected by the diffusing process given a constant rate of diffusion and a fixed size population initiating the diffusion process.

Rate of diffusion: The expected extent of diffusion relative to the size of population initiating the diffusion process.

Following sections provide the description of the conclusion drawn for maximizing the above two goals using the independent cascade and linear threshold model of diffusion in the two populations.

8.4.1 Independent cascade diffusion

8.4.1.1 Extent of diffusion

I find that in grevys, with high expectation diffusion starts from the largest component, that is composed of almost 55% of the population, but it does not go easily beyond that component Figure 28. However, at the same rate of diffusion, the diffusion process extends to almost the entire population in onagers Figure 29.The diffusion progresses linearly beyond the the largest component when best diffusion initiator is picked using the greedy approximation approach stated in Section 2.4.1.1.

After rigorously simulating the independent cascade process on the two networks with various parametric values, I find that at very low rates of diffusion, it is relatively easier to spread in grevys than in onagers. Since the onagers network is sparsely connected the low rates render it harder to reach tenuously connected individuals. On the other hand, the tight, temporally persistent, cohesive structure of the grevys network facilitates the diffusing process to reach a large extent of population. At the diffusion rate of as low as 0.05 the best extent with just a single active individual reaches as high as 50% of the population. Whereas, for the same rate of infection, in onagers, only 25% of the population gets activated while any additional activation occurs linearly with the size of the initially active individuals. Hence, at a rate high enough, the distributiveness of the onagers network gives it an edge over more structured networks.

8.4.1.2 Rate of diffusion

I find that the rate of diffusion is higher in onagers than in grevys. As shown in the results in Figure 29 for the same rate of infection of .25, it takes 3 individuals to reach the entire population in onagers. In grevys, on the other hand, for the same extent of diffusion it takes about 10 individuals (Figure 28). (Extent of diffusion in grevys for 0.05 and .25 are not much different from each other.)

8.4.2 Linear threshold diffusion

8.4.2.1 Extent of diffusion

I find the extent of diffusion using the linear threshold model is quantitatively similar to that of the independent cascade model for both grevys and onagers networks. That is, extent of diffusion is generally much higher in onagers than in grevys, as evident from Figure 30 and Figure 31. The similar extent of diffusion for the two models indicates, that even with completely opposite dynamics of the diffusion processes, the structure of the population significantly influences the extent of diffusion in the population.

8.4.2.2 Rate of diffusion

Figure 30 and Figure 31 show that the rate of diffusion in onagers is much higher than in grevys using the linear threshold model. Even in the worst case (uniform random) it takes no more than 4 individuals to activate the entire population of onagers. Since at any given time each individual is close to very few other individuals, it becomes easier for the set of active individual to activate their inactive neighbor. Thus, a diffusion process initiated with a

Figure 28: Diffusion in grevys with independent cascade

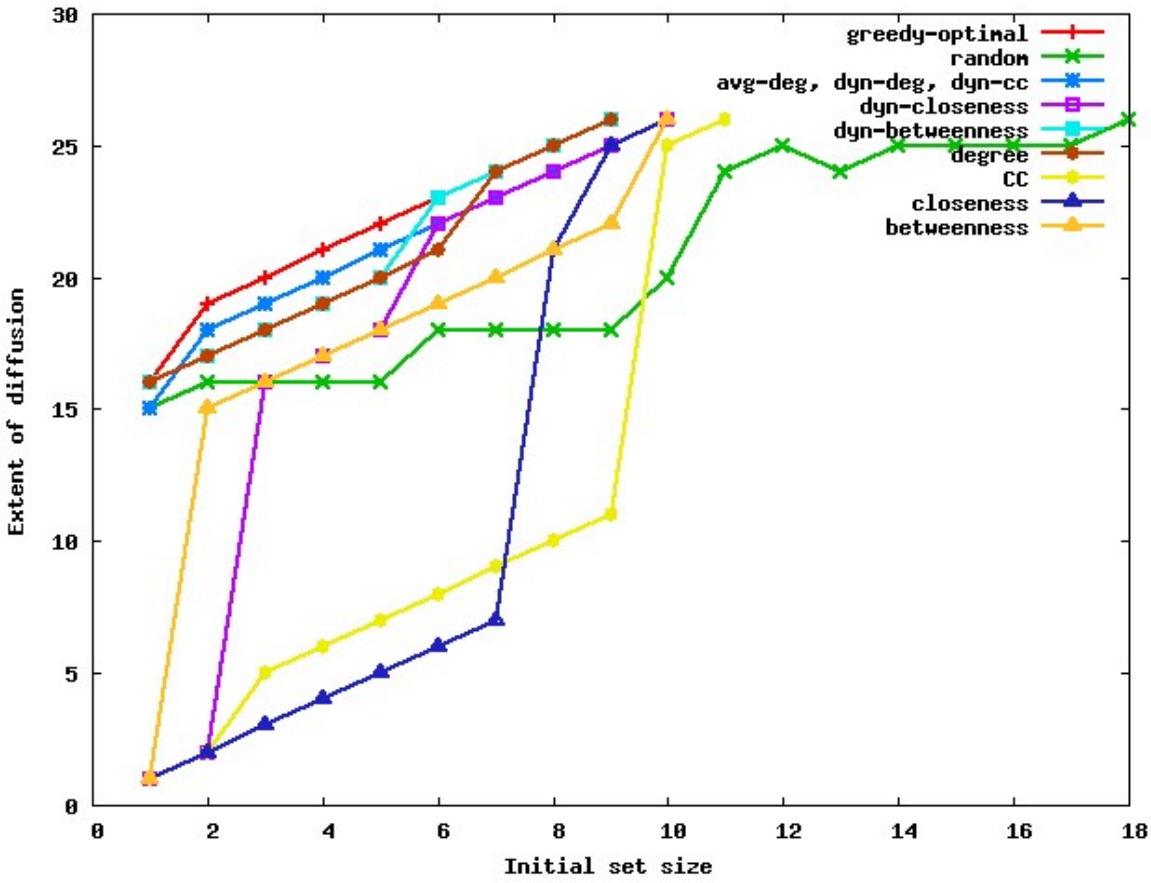


Figure 29: Diffusion in onagers with independent cascade

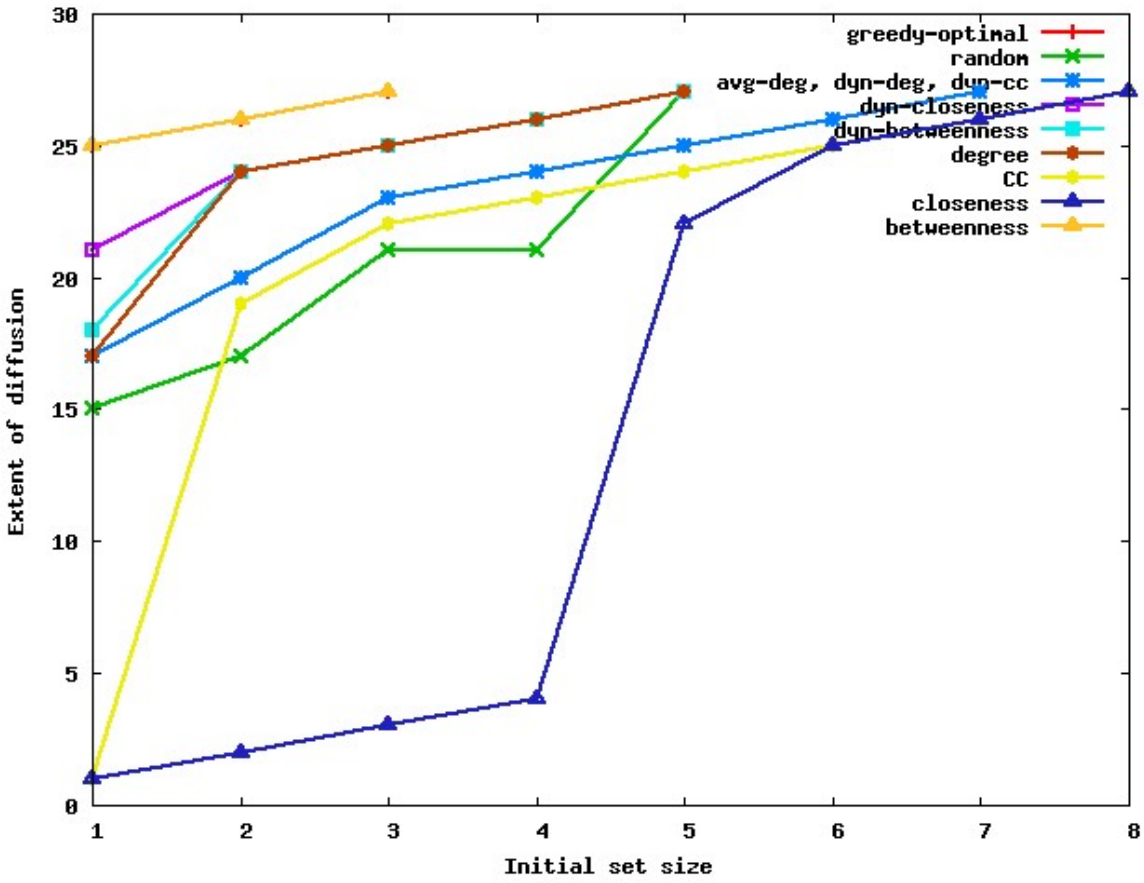
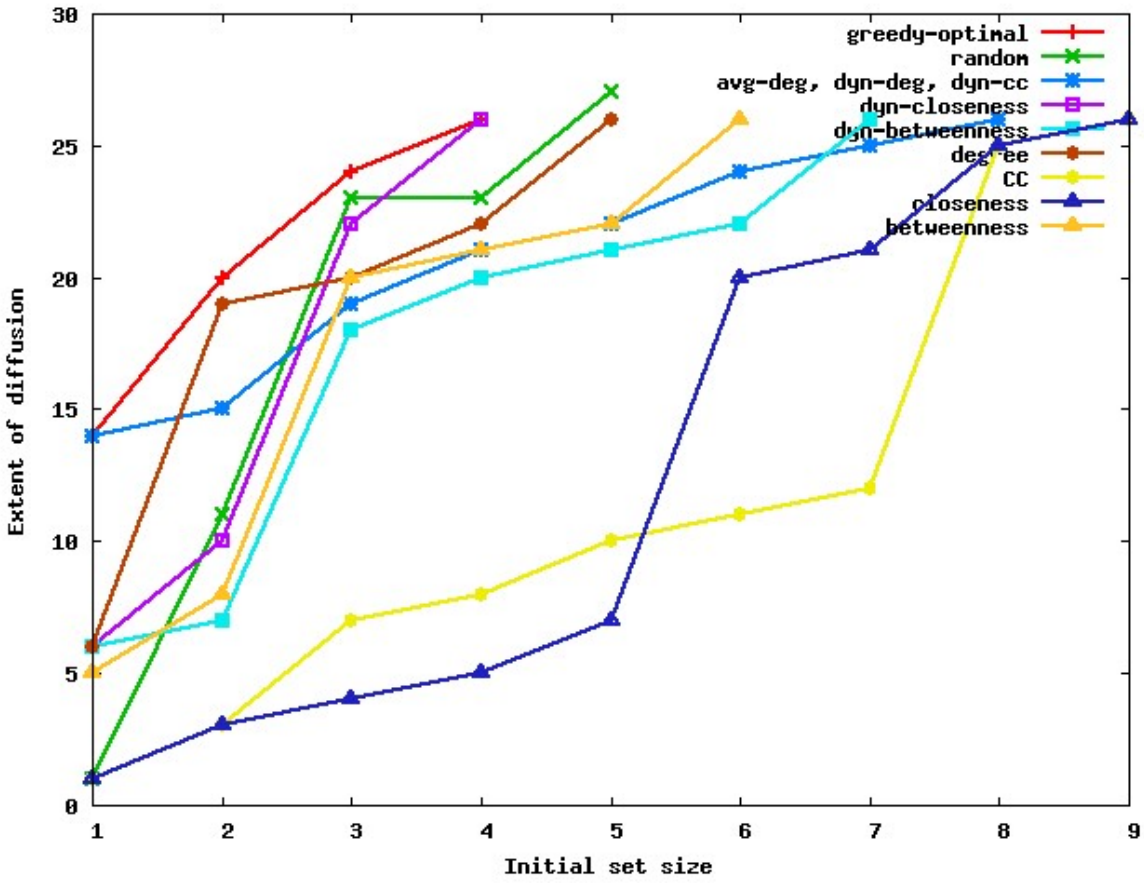


Figure 30: Diffusion in grevys with linear threshold model

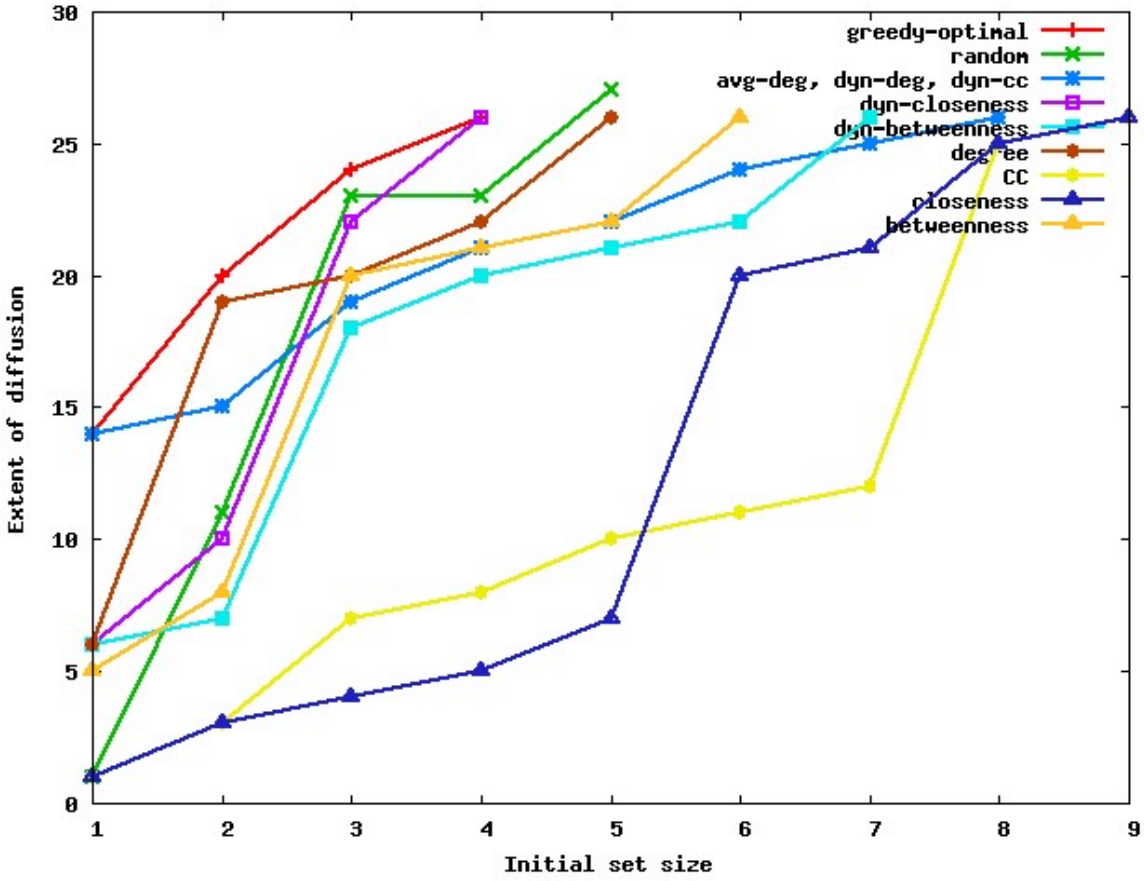


strategically "good" set of individuals, can reach the entire population. Whereas, in grevys, it takes 4 or more individuals to result in the same extent of diffusion.

8.4.3 Grevys vs onagers

The overall trend observed for the two population is that diffusion happens relatively easily in onagers than in grevys. The random and distributed nature of the onagers network facilitate the diffusion of processes such as diseases and information. Grevys, on the other hand have

Figure 31: Diffusion in onagers with linear threshold model



a relatively more cohesive structure. From Section 6.4 we know that in more clustered the network structures, the extent of diffusion across the cluster is dependent on the rate of diffusion. Therefore, in grevys its easier to spread within a cluster but harder to propagate the spreadin process across clusters.

8.5 Diffusion minimization

I focus on the following two measures for diffusion minimization, using the independent cascade and linear threshold diffusion models.

Expected extent of diffusion: This is the average extent of diffusion in the network over all nodes. That is, it is the expected value of the extent of diffusion, when the diffusion process is initiated by each node independently.

Blocking set size: A blocking set is a group of individuals that when removed (blocked) from the network brings the extent diffusion in the network to below some minimum threshold. Blocking set size is then the minimal such set of individuals which when blocked from the network will result in the extent of diffusion no more than the threshold.

8.5.1 Independent cascade

8.5.1.1 Expected extent of diffusion

The expected extent of diffusion in onagers is significantly more than in grevys. This trend can be observed in Figure 32 and Figure 33. For the two population of the same size, we see that the expected extent of diffusion in onagers and grevys are about 55% and 33% of

the population, respectively. Hence, for onagers the network with higher expected extent of diffusion, it is harder to block the diffusion process in this network.

8.5.1.2 Blocking set size

Secondly, although the best blocking in both grevys and onagers require almost identical number of blockers to limit the extent of diffusion to about 10% of the population. I find that in most cases (of measures) it requires somewhere between 12 and 17 individuals to achieve such low extent in onagers whereas in grevys most measure ranking enable effective blocking with top 10 individuals (Figure 32 and Figure 33).

8.5.2 Linear threshold

8.5.2.1 Expected extent of diffusion

Figure 34 and Figure 35 show that the expected extent of diffusion in onagers is significantly more than in grevys. For the two populations of the same size, I see that the expected extent of diffusion in onagers is 80% of the population unlike grevys where the diffusing process reaches only 30% of the population. Hence, for onagers the network with higher expected extent of diffusion, it is harder to block the process.

8.5.2.2 Blocking set size

Secondly, although the best blocking in both grevys and onagers require almost identical number of blockers to contain the spread to about 10% of the population. I find that in most cases (of measures) it requires around 20 of the 28 individuals to contain the extent of diffusion to within 10% of the onagers population whereas in grevys most measure ranking enable effective blocking with top 15 individuals. These observations are summarized in Figure 34 and Figure 35.

Figure 32: Reduction in the extent of diffusion in grevys with independent cascade model

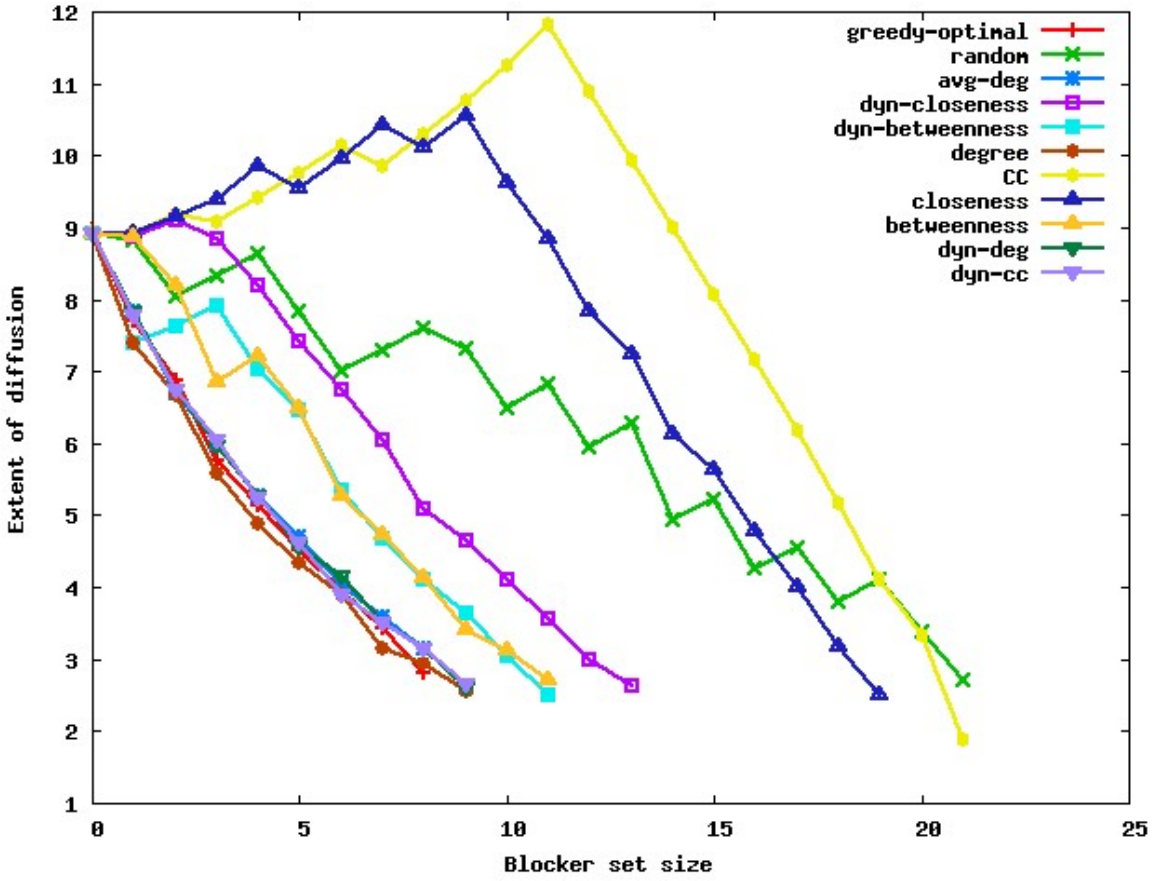


Figure 33: Reduction in the extent of diffusion in onagers with independent cascade model

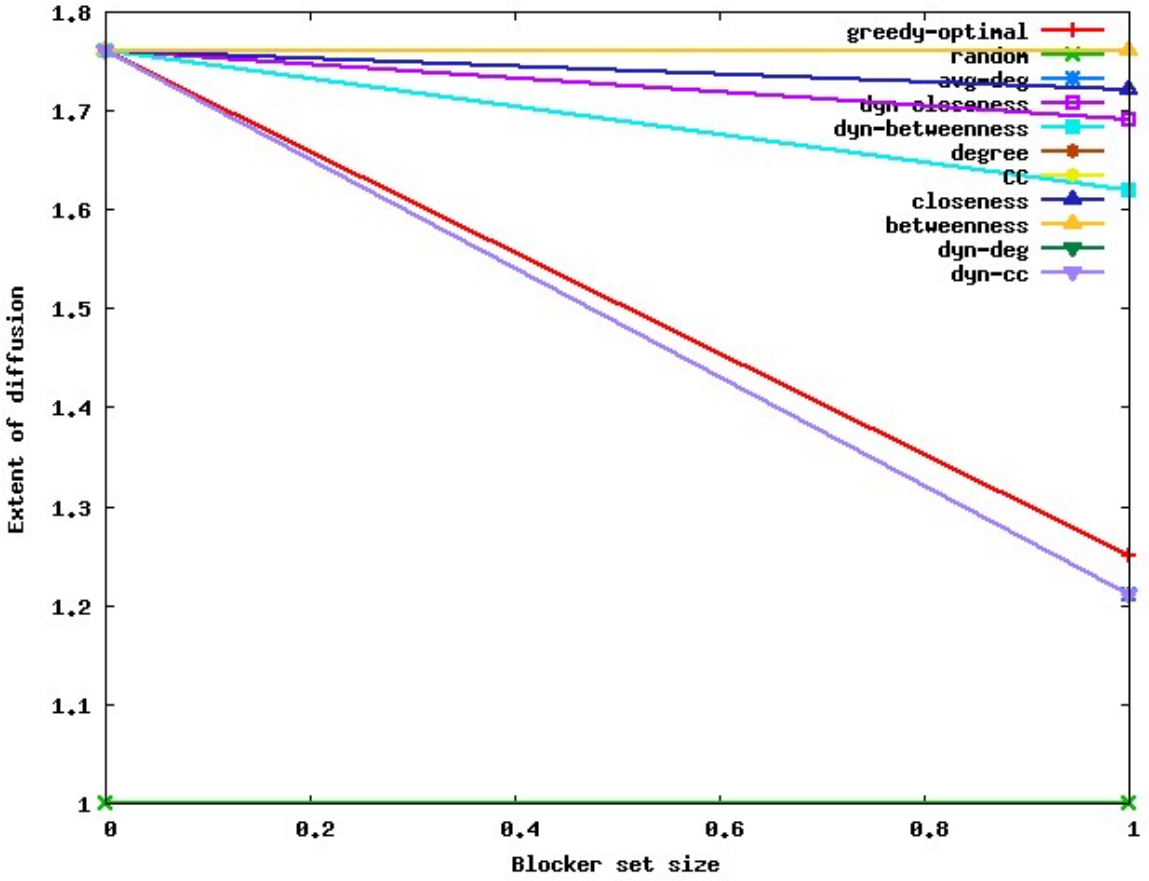


Figure 34: Reduction in the extent of diffusion in grevys with linear threshold model

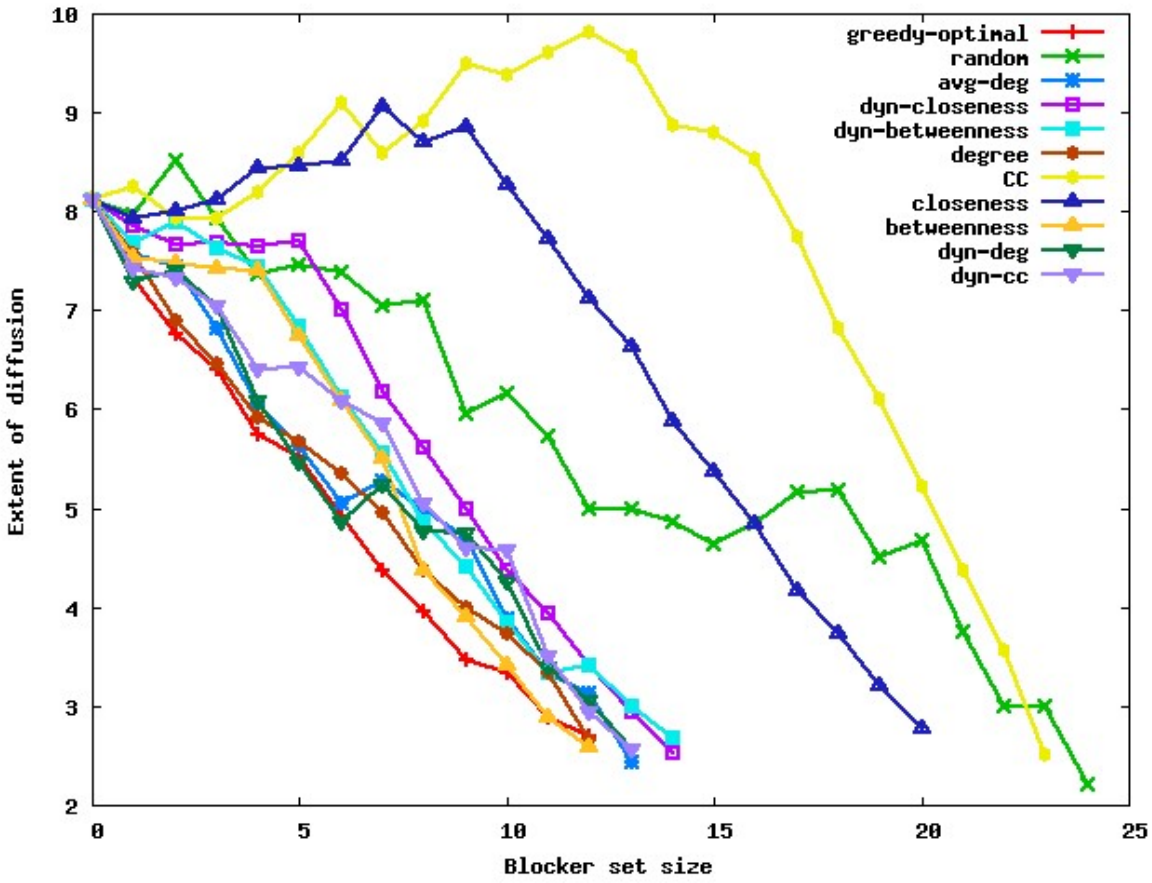
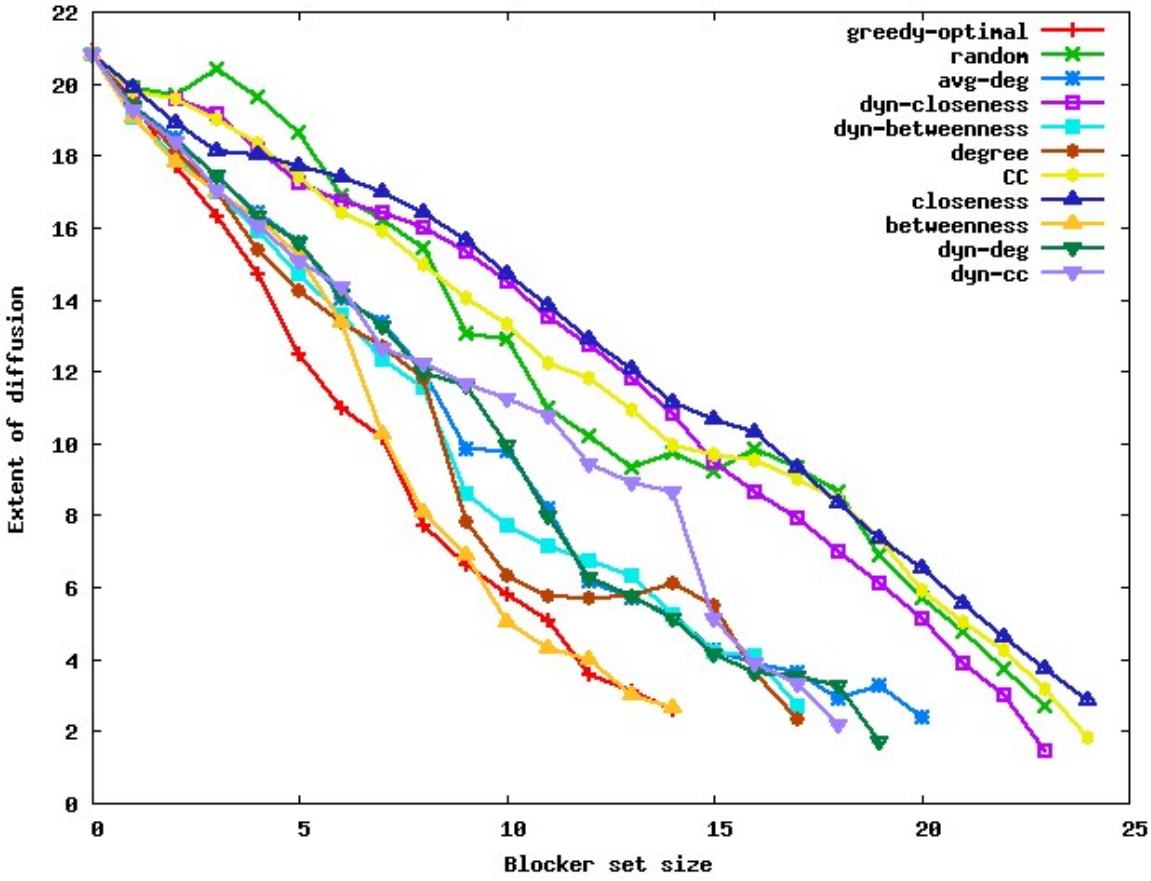


Figure 35: Reduction in the extent of diffusion in grevys with linear threshold model



8.5.3 Grevys vs Onagers

It is easier to block the diffusion of a process in grevys than in onagers. This trend is consistent with what is generally observed in distributed (non–centralized) networks. Distributed networks are more robust to random failures. Whereas, more centralized and structured networks are reliant on its cores and its easier to disrupt them by removing the core or "hubs" in the network.

8.6 Network measures as indicators of extent of diffusion

Following is the summary of my preliminary observation about the effectiveness of various network measures as indicators of diffusion in the two networks.

- Individuals with higher betweenness centrality are effective in blocking the diffusion process in both grevys and onagers (Figure 34 and Figure 35).
- In onagers it is easy to reach a large set of population with almost all the network measures (Figure 30 and Figure 31).
- In grevys, it is relatively easy to reach all the individuals within a group since this animal tend to live in cohesive groups. However, it is hard to reach across the groups (**??** and Figure 31).

8.7 Conclusions

With the help of these preliminary observations I provide ecologists with systematic interpretation of animal network structure and its effect on their behavior. The goal of this work is to find links between the population network structure and the various functions these animals

tend to optimize for in their interactions. A by–product of this analysis is that the ecologists get a better sense of salient differentiating features in the network structures of the two types of equids based on the vitalness of the functions each population is attempting to optimize for. With this work I provide ecologist sophisticated computational tools for understanding various social functions and organization of these animals. I perform extensive experiments on these networks to analyze the network structures and their properties. I identify critical network properties that characterize a certain network structure for optimizing a certain diffusion goal. With these methods ecologists are better equipped to answer questions about different social structures in equids.

CITED LITERATURE

1. Abrahamson, E.: Managerial fads and fashions: The diffusion and rejection of innovations. The Academy of Management Review, 16(3):586–612, 1991.

2. Abrahamson, E. and Rosenkopf, L.: Social Network Effects on the Extent of Innovation Diffusion: A Computer Simulation. ORGANIZATION SCIENCE, 8(3):289–309, 1997.

3. Adamic, L. A. and Glance, N.: The political blogosphere and the 2004 u.s. election: divided they blog. In LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery, pages 36–43, New York, NY, USA, 2005. ACM.

4. Adar, E. and Adamic, L. A.: Tracking information epidemics in blogspace. In Web Intelligence, pages 207–214, 2005.

5. Adibi, J.: Enron email dataset. [ONLINE]. http://www.isi.edu/?adibi/Enron/Enron.htm.

6. Aiello, W., Chung, F., and Lu, L.: A random graph model for power law graphs. Experimental Math, 10:53–66, 2000.

7. Aizen, J., Huttenlocher, D., Kleinberg, J., and Novak, A.: Traffic-based feedback on the web. Proceedings of the National Academy of Sciences, 101(Suppl.1):5254–5260, 2004.

8. Albert, R. and Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys., 74:47–97, 2002.

9. Anderson, R. M. and May, R. M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, 1992.

10. Angluin, D., Aspnes, J., and Reyzin, L.: Optimally learning social networks with activations and suppressions. Theor. Comput. Sci., 411(29-30):2729–2740, 2010.

11. Anthonisse, J.: The rush in a graph. Amsterdam: Mathematische Centrum, 1971.

12. Aspnes, J., Chang, K., and Yampolskiy, A.: Inoculation strategies for victims of viruses and the sum-of-squares partition problem. Journal of Computer and System Sciences, 72(6):1077–1093, September 2006.

13. Aspnes, J., Rustagi, N., and Saia, J.: Worm versus alert: Who wins in a battle for control of a large-scale network?

14. Asur, S., Parthasarathy, S., and Ucar, D.: An event-based framework of characterizing the evolutionary behavior of interaction graphs. In Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.

15. Asur, S. and Huberman, B. A.: Predicting the future with social media. CoRR, abs/1003.5699, 2010.

16. Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X.: Group formation in large social networks: Membership, growth, and evolution. In The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

17. Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J.: Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

18. Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L.: The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web, WWW '12, pages 519–528, New York, NY, USA, 2012. ACM.

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

20. Barabási, A.-L.: Linked. Perseus Books Group, 2002.

21. Barabási, A.-L.: The origin of bursts and heavy tails in human dynamics. Nature, 435:207–211, 2005.

22. Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks. Science, 286:509–512, 1999.

23. Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, 311(3-4):590–614, August 2002.

24. Bass, F. M.: Comments on "A New Product Growth for Model Consumer Durables": The Bass Model. Management Science, 50(12):1833–1840, 2004.

25. Berger, E.: Dynamic monopolies of constant size. Journal of Combinatorial Theory Series B, 83:191–200, 2001.

26. Berger, N., Borgs, C., Chayes, J. T., and Saberi, A.: On the spread of viruses on the internet. In SODA '05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms, pages 301–310, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.

27. Berger-Wolf, T., Hart, W., and Saia, J.: Discrete sensor placement problems in distribution networks. Mathematical and Computer Modelling, 2005.

28. Berry, J., Fleischer, L., Hart, W., Phillips, C., and Watson, J.: Sensor placement in municipal water networks. Journal of Water Resources Planning and Management, 131(3), 2005a.

29. Berry, J., Hart, W., Phillips, C., Uber, J. G., and Watson, J.: Sensor placement in municipal water networks with temporal integer programming models. Journal of Water Resources Planning and Management, 132(4):218–224, 2006.

30. Bharathi, S., Kempe, D., and Salek, M.: Competitive influence maximization in social networks. In Proceedings of the 3rd international conference on Internet and network economics, WINE'07, pages 306–311, Berlin, Heidelberg, 2007. Springer-Verlag.

31. Boguñá, M., Pastor-Satorras, R., and Vespignani, A.: Absence of epidemic threshold in scale-free networks with degree correlations. Phys. Rev. Lett., 90:028701, Jan 2003.

32. Börner, K., DallAsta, L., Ke, W., and Vespignani, A.: Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. In Complexity, Special issue on Understanding Complex Systems, 2006. in press.

33. Börner, K., Maru, J., and Goldstone, R.: The simultaneous evolution of author and paper networks. PNAS, 101(Suppl 1):5266–5273, 2004.

34. Brauer, F. and Castillo-Chávez, C.: Mathematical Models in Population Biology and Epidemiology. Springer, 2000.

35. Broido, A. and Claffy, K.: Internet topology: connectivity of ip graphs. In Proceedings of SPIE ITCom, 2001.

36. Brown, J. H., Gupta, V. K., Li, B.-L. L., Milne, B. T., Restrepo, C., and West, G. B.: The fractal nature of nature: power laws, ecological complexity and biodiversity. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 357(1421):619–626, May 2002.

37. Brummitt, C. D., D'Souza, R. M., and Leicht, E. A.: Suppressing cascades of load in interdependent networks. Proceedings of the National Academy of Sciences, 109(12):E680–E689, March 2012.

38. Budak, C., Agrawal, D., and Abbadi, A. E.: Limiting the spread of misinformation in social networks.

39. Caceres, R. S., Berger-Wolf, T. Y., and Grossman, R.: Temporal scale of processes in dynamic networks. In ICDM Workshops, eds, M. Spiliopoulou, H. Wang, D. J. Cook, J. Pei, W. W. 0010, O. R. Zaane, and X. Wu, pages 925–932. IEEE, 2011.

40. Carley, K.: Communicating new ideas: The potential impact of information and telecommunication technology. Technology in Society, 18(2):219–230, 1996.

41. Carley, K.: Dynamic network analysis. In Dynamic Social Network Modeling and Analysis, eds, R. Breiger, K. Carley, and P. Pattison, pages 133–145. Washington, D.C., The National Academic Press, 2003.

42. Carreras, I., Miorandi, D., Canright, G., and Engøo-Monsen, K.: Eigenvector centrality in highly partitioned mobile networks: Principles and applications. Studies in Computational Intelligence(SCI), 69:123–145, 2007.

43. Castellano, C. and Pastor-Satorras, R.: Competing activation mechanisms in epidemics on networks. CoRR, abs/1105.5545, 2011.

44. Castellano, C. and Pastor-Satorras, R.: Competing activation mechanisms in epidemics on networks. CoRR, abs/1105.5545, 2011.

45. Chaoji, V., Ranu, S., Rastogi, R., and Bhatt, R.: Recommendations to boost content spread in social networks. In Proceedings of the 21st international conference on World Wide Web, WWW '12, pages 529–538, New York, NY, USA, 2012. ACM.

46. Chen, L. and Carley, K.: The impact of social networks in the propagation of computer viruses and countermeasures. IEEE Trasactions on Systems, Man and Cybernetics, forthcoming.

47. Chen, N.: On the approximability of influence in social networks. ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1029–1037, 2008.

48. Choi, H., Kim, S., and Lee, J.: Role of network structure and network effects in diffusion of innovations. Industrial Marketing Management, 2008.

49. Chung, F. and Lu, L.: Connected components in random graphs with given expected degree sequences. Annals of Combinatorics, 6:125–145, 2002. 10.1007/PL00012580.

50. Clauset, A. and Eagle, N.: Persistence and periodicity in a dynamic proximity network. Unpublished manuscript.

51. Cohen, R., Havlin, S., and ben Avraham, D.: Efficient immunization strategies for computer networks and populations. Physical Review Letters, 2003.

52. Cornuejols, G., Fisher, M., and Nemhauser, G.: Location of bank accounts to optimize float. Management Science, 23, 1977.

53. Cowan, R. and Jonard, N.: Network structure and the diffusion of knowledge. Journal of Economic Dynamics and Control, 28(8):1557 – 1575, 2004.

54. Croft, D. P., James, R., Thomas, P., Hathaway, C., Mawdsley, D., Laland, K., and Krause, J.: Social structure and co-operative interactions in a wild population of guppies (*poecilia reticulata*). Behavioural Ecology and Sociobiology, 2006.

55. David, E. and Jon, K.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. New York, NY, USA, Cambridge University Press, 2010.

56. Del Genio, C. I., Gross, T., and Bassler, K. E.: All scale-free networks are sparse. Phys. Rev. Lett., 107:178701, Oct 2011.

57. Dezsö, Z. and Barabási, A.-L.: Halting viruses in scale-free networks. Physical Review E, 65(055103(R)), 2002. DOI: 10.1103/PhysRevE.65.055103.

58. Dodds, P. S. and Watts, D. J.: A generalized model of social and biological contagion. Journal of Theoretical Biology, 232(4):587–604, February 2005.

59. Domingos, P.: Mining social networks for viral marketing. IEEE Intelligent Systems, 20:80–82, 2005.

60. Domingos, P. and Richardson, M.: Mining the network value of customers. In Seventh International Conference on Knowledge Discovery and Data Mining, 2001.

61. Eagle, N. and Pentland, A.: Reality mining: Sensing complex social systems. Journal of Personal and Ubiquitous Computing, 2006.

62. Erd́os, P. and Rényi, A.: On random graphs i. Publ. Math. Debrecen, 6:290297, 1959.

63. Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z., and Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature, 429:429:180–184., Nov 2004. Supplement material.

64. Eubank, S., Kumar, V., Marathe, M., Srinivasan, A., , and Wang, N.: Structural and algorithmic aspects of massive social networks. In Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, pages 718–727, 2004.

65. Fagin, R., Kumar, R., and Sivakumar, D.: Comparing top k lists. In SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, pages 28–36, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

66. Fagiolo, G., Valente, M., and Vriend, N. J.: A dynamic model of segregation in small-world networks. Lecture Notes in Economics and Mathematical Systems, 613:111–126, 2009.

67. Feige, U., Mirrokni, V., and Vondrák: Maximizing non-monotone submodular functions. In Foundations of Computer Science(FOCS), 2007.

68. Ferreira, S. C., Castellano, C., and Pastor-Satorras, R.: Epidemic thresholds of the. CoRR, abs/1206.6728, 2012.

69. Fischhoff, I. R., Sundaresan, S. R., Cordingley, J., Larkin, H. M., Sellier, M.-J., and Rubenstein, D. I.: Social relationships and reproductive state influence leadership roles in movements of plains zebra (*equus burchellii*). Animal Behaviour, 2006.

70. Fischhoff, I. R., Sundaresan, S. R., Cordingley, J., and Rubenstein, D.: Habitat use and movements of plains zebra (equus burchelli) in response to predation danger from lions. Behavioral Ecology, 18(4):725–729, 2007.

71. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry, 40:35–41, 1977.

72. Freeman, L. C.: Centrality in social networks: I. conceptual clarification. Social Networks, 1:215–239, 1979.

73. Freeman, L. C.: Finding social groups: A meta-analysis of the southern women data. In Dynamic Social Network Modeling and Analysis, eds, R. Breiger, K. Carley, and P. Pattison. Washington, D.C., The National Academies Press, 2003.

74. Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci., 99:8271–8276, 2002.

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

76. Goldenberg, J., Libai, B., and Muller, E.: Using complex systems analysis to advance marketing theory development. Academy of Marketing Science Review, 2001.

77. Gomez-Rodriguez, M., Leskovec, J., and Krause, A.: Inferring networks of diffusion and influence. CoRR, abs/1006.0234, 2010.

78. Gould, R. V.: Collective action and network structure. American Sociological Review, 58(2):182–196, 1993.

79. Granovetter, M.: The strength of weak ties. American Journal of Sociology, 78(6):1360–1380, 1973.

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

81. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A.: Information diffusion through blogspace. In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 491–501, New York, NY, USA, 2004. ACM Press.

82. Habiba and Berger-Wolf, T. Y.: Influence maximization in dynamic networks. Technical Report 2007-20, DIMACS, 2007.

83. Habiba, Yu, Y., Berger-wolf, T. Y., and Saia, J.: Finding spread blockers in dynamic networks, 2010.

84. Handcock, M. S. and Morris, M.: A curved exponential family model for complex networks. Comput. Math. Organ. Theory, 15(4), December 2009.

85. Hartvigsen, G., Dresch, J., Zielinski, A., Macula, A., and Leary, C.: Network structure, and vaccination strategy and effort interact to affect the dynamics of influenza epidemics. Journal of Theoretical Biology, 246(2):205 – 213, 2007.

86. Holme, P.: Efficient local strategies for vaccination and network attack. Europhys. Lett., 68(6):908–914, 2004.

87. Hopcroft, J., Khan, O., Kulis, B., and Selman, B.: Natural communities in large linked networks. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 541–546, 2003.

88. Jaccard, P.: The distribution of flora in the alpine zone. The New Phytologist, 11(2):37–50, 1912.

89. Jackson, M. O. and Rogers, B. W.: Relating network structure to diffusion properties through stochastic dominance. The B.E. Journal of Theoretical Economics, 2007.

90. Jordán, F., Benedek, Z., and Podani, J.: Quantifying positional importance in food webs: A comparison of centrality indices. Ecological Modelling, 205:270–275, 2007.

91. Keeling, M.: The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Lond. B, 266:859–867, 1999.

92. Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

93. Kempe, D., Kleinberg, J., and Kumar, A.: Connectivity and inference problems for temporal networks. J. Comput. Syst. Sci., 64(4):820–842, 2002.

94. Kleinberg, J.: Temporal dynamics of on-line information streams. Draft chapter for the forthcoming book Data Stream Management: Processing High-Speed Data Streams (M. Garofalakis, J. Gehrke, R. Rastogi, eds.), Springer.

95. Kleinberg, J.: Small-world phenomena and the dynamics of information. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence,, Morgan Kaufman, 2001.

96. Klimt, B. and Yang, Y.: The enron corpus: A new dataset for email classification research. In Proceedings of the European Conference on Machine Learning, 2004.

97. Korkki, P.: For marketers, viruses just won't cooperate. The New York Times, July 6 2008.

98. Krebs, J. R.: Behavioural Ecology: An Evolutionary Approach. 4th ed. Oxford/Malden, MA, Blackwell, 1997 1997.

99. Kretzschmar, M. and Morris, M.: Measures of concurrency in networks and the spread of infectious disease. Math. Biosci., 133:165–195, 1996.

100. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A.: On the bursty evolution of blogspace. In Proc. International WWW Conference, 2003.

101. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A.: On the Bursty Evolution of Blogspace. World Wide Web, 8(2):159–178, 2005.

102. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.: Trawling the web for emerging cyber-communities. In Proceedings of the 8th International World Wide Web Conference, May 1999.

103. Kuperman, M. and Abramson, G.: Small world effect in an epidemiological model. Physical Review Letters, 2001.

104. Kuran, T.: Ethnic norms and their transformation through reputational cascades. Journal of Legal Studies, 27(2):623–59, June 1998.

105. Le Bon, G.: The Crowd a Study of the Popular Mind. Macmillan, 1896.

106. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., and VanBriesen, J.: Cost-effective outbreak detection in networks. In Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.

107. Leskovec, J., Adamic, L. A., and Huberman, B. A.: The dynamics of viral marketing. In EC '06: Proceedings of the 7th ACM conference on Electronic commerce, pages 228–237, New York, NY, USA, 2006. ACM Press.

108. Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. CoRR, abs/0810.1355, 2008.

109. Lewin, K.: Principles of Topological Psychology. New York: McGraw Hill, 1936.

110. Ley, M.: Digital bibliography & library project (DBLP). `http://dblp.uni-trier.de/`, December 2005. A digital copy of the databse has been provided by the author.

111. Liben-Nowell, D. and Kleinberg, J.: Tracing the flow of information on a global scale using Internet chain-letter data. Proceedings of the National Academy of Sciences, 105(12):4633–4638, 25 2008.

112. Liljeros, F., Edling, C., and Amaral, L. N.: Sexual networks: Implication for the transmission of sexually transmitted infection. Microbes and Infection, 2003.

113. Liu, B. S., Madhavan, R., and Sudharshan, D.: Diffunet: The impact of network structure on diffusion of innovation. European Journal of Innovation Management, 8(2):240–262, February 2005.

114. M., N. and M., G.: Finding and evaluating community structure in networks. Phys. Rev., 69, 2004.

115. Mackay, C.: Memoirs of Extraordinary Popular Delusions and madness of crowds. G. Routledge & Company, January 1856.

116. Marquet, P. A., Quiñones, R. A., Abades, S., Labra, F., Tognelli, M., Arim, M., and Rivadeneira, M.: Scaling and power-laws in ecological systems. Journal of Experimental Biology, 208:1749–1769, 2005.

117. May, R. M. and Lloyd, A. L.: Infection dynamics on scale-free networks. Physical Review E, 64(066112), 2001. DOI: 10.1103/PhysRevE.64.066112.

118. Mercken, L., Snijders, T. A., Steglich, C., and de Vries, H.: Dynamics of adolescent friendship networks and smoking behavior: Social network analyses in six european countries. Social Science & Medicine, 69(10):1506–1514, 2009.

119. Morris, M.: Epidemiology and social networks:modeling structured diffusion. Sociological Methods and Research, 22(1):99–126, 1993.

120. Mossel, E. and Roch, S.: On the submodularity of influence in social networks. In The Annual ACM Symposium on Theory of Computing(STOC), 2007.

121. Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P.: Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science, 328(5980):876–878, May 2010.

122. Nemhauser, G., Wolsey, L., and Fisher, M.: An analysis of the approximations for maximizing submodular set functions. Management Science, 14:265–294, 1978.

123. Newman, M.: The structure and function of complex networks. SIAM Review, 45:167–256, 2003.

124. Newman, M. E.: Spread of epidemic disease on networks. Physical Review E, 66(016128), 2002. DOI: 10.1103/PhysRevE.66.016128.

125. Newman, M. E. J.: Scientific collaboration networks. i. network construction and fundamental results. Physical Review E, 64:016131, 2001.

126. Newman, M. E. J.: Power laws, pareto distributions and zipf's law. Contemporary Physics, 46(5):323–351, May 2005.

127. Newman, M.: Networks: An Introduction. New York, NY, USA, Oxford University Press, Inc., 2010.

128. Newman, M.: Network data, 2011.

129. Newman, M. and Watts, D.: Scaling and percolation in the small-world network model. Physical Review E, 1999.

130. Nicolaides, C., Cueto-Felgueroso, L., Gonzlez, M. C., and Juanes, R.: A metric of influential spreading during contagion dynamics through the air transportation network. PLoS ONE, 7(7):e40961, 07 2012.

131. Nowicki, K. and Snijders, T. A.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96:1077–1087, 2001.

132. Pastor-Satorras, R. and Vespignani, A.: Evolution and structure of the Internet. Cambridge, Cambridge University Press, 2004.

133. Pastor-Satorras, R. and Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett., 86(14):3200–3203, Apr 2001.

134. paul Watson, J., Hart, W. E., and Greenberg, H. J.: An analysis of multiple contaminant warning system design objectives for sensor placement optimization in water distribution networks.

135. Prakash, B. A., Tong, H., Valler, N., Faloutsos, M., and Faloutsos, C.: Virus propagation on time-varying networks: Theory and immunization algorithms. In ECML/PKDD (3), pages 99–114, 2010.

136. Read, J. M. and Keeling, M. J.: Disease evolution on networks: the role of contact structure. Proc. R. Soc. Lond. B, 270:699–708, 2003.

137. Richardson, M., Agrawal, R., and Domingos, P.: Trust management for the semantic web. In 2nd International Semantic Web Conference, 2003.

138. Ripeanu, M., Foster, I., and Iamnitchi, A.: Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Computing Journal, 2002.

139. Rogers, E. M.: Diffusion of Innovations. Simon & Shuster, Inc., 5th edition, 2003.

140. Rosenkopf, L. and Abrahamson, E.: Modeling reputational and informational influences in threshold models of bandwagon innovation diffusion. Comput. Math. Organ. Theory, 5(4):361–384, 1999.

141. Rubenstein, D. I., Sundaresan, S., Fischhoff, I., and Saltz, D.: Social networks in wild asses: Comparing patterns and processes among populations. In Exploration into the Biological Resources of Mongolia, eds, A. Stubbe, P. Kaczensky, R. Samjaa, K. Wesche, and M. Stubbe, volume 10. Martin-Luther-University Halle-Wittenberg, 2007. In press.

142. Sabidussi, G.: The centrality index of a graph. Psychometrika, 31:581–603, 1966.

143. Sander, L. M., Warren, C. P., Sokolov, I. M., Simon, C., and Koopman, J.: Percolation on heterogeneous networks as a model for epidemics. Mathematical Biosciences, 180(1-2):293 – 305, 2002.

144. Schelling, T.: Micromotives and Macrobehavior. Norton, 1978.

145. Schelling, T. C.: Models of segregation. American Economic Review, 59(2):488–93, May 1969.

146. Schelling, T. C.: Egonomics, or the art of self-management. American Economic Review, 68(2):290–94, May 1978.

147. Scott, J.: Social Network Analysis. SAGE Publications, 2012.

148. Scott, J., Gass, R., Crowcroft, J., Hui, P., Diot, C., and Chaintreau, A.: CRAWDAD trace cambridge/haggle/imote/infocom (v. 2006-01-31). Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom, January 2006.

149. Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology, 15:72–101, 1904.

150. Sundaresan, S. R., Fischhoff, I. R., Dushoff, J., and Rubenstein, D. I.: Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. Oecologia, 2006. doi 10.1007/s00442-006-0553-6.

151. Tong, H., Prakash, B. A., Tsourakakis, C. E., Eliassi-Rad, T., Faloutsos, C., and Chau, D. H.: On the vulnerability of large graphs. In ICDM, pages 1091–1096, 2010.

152. Valente, T. W. and Saba, W.: Campaign recognition and interpersonal communication as factors in contraceptive use in bolivia. Journal of Health Communication, 2001.

153. Valler, N., Prakash, B. A., Tong, H., Faloutsos, M., and Faloutsos, C.: Epidemic spread in mobile ad hoc networks: Determining the tipping point. In Networking (1), pages 266–280, 2011.

154. Vredeveld, T. and Lenstra, J.: On local search for the generalized graph coloring problem. Operations Research Letters, 31:28–34, 2003.

155. Wasserman, S. and K., F.: Social Network Analysis.. Cambridge, MA, Cambridge University Press, 1994.

156. Watts, D.: A simple model of global cascades on random networks. PNAS, 99:5766–5771, 2002.

157. Watts, D. and Strogatz, S.: Collective dynamics of small-world networks. Nature, 393:440–442, 1998.

158. Zachary, W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977.

159. Zaman, T. R., Herbrich, R., Van Gael, J., and Stern, D.: Predicting information spreading in twitter. 2010.

160. Zanette, D. H.: Dynamics of rumor propagation on small-world networks. Phys. Rev. E, 65(4):041908, Mar 2002.

161. Zekri, N., Porterie, B., Clerc, J.-P., and Loraud, J.-C.: Propagation in a two-dimensional weighted local small-world network. Phys. Rev. E, 71(4):046121, Apr 2005.

VITA

NAME:	Habiba
EDUCATION:	Ph.D., Computer Science, University of Illinois at Chicago, Chicago, Illinois, 2013.
	B.S., Computer Science, National University of Computer & Emerging Sciences, Islamabad, Pakistan, 2003.
ACADEMIC EXPERIENCE:	Research Assistant, Computational Population Biology Lab, Department of Computer Science, University of Illinois at Chicago, 2006–2013.
	Teaching Assistant, Department of Computer Science, University of Illinois at Chicago: • Introduction to Computing and Programming, 2011. • Introduction to Computing, 2012.
PROFESSIONAL EXPERIENCE:	Research Intern, Microsoft Research, Redmond, Washington, 2012.
	Software Engineer, FAST, Islamabad, Pakistan, 2004-2005.
PROFESSIONAL MEMBERSHIP:	Institute of Electrical and Electronics Engineers (IEEE)
	Women In Engineering (WIE)
HONORS:	Chancellor's student service & leadership, University of Illinois at Chicago, 2008, 2009.
	Fifty for the Future award, 2011
	Fulbright student, 2005–2010.
PUBLICATIONS:	Habiba, T. Berger-Wolf. Affect of network structure on influence maximization in dynamic networks. SIAM workshop on network science, 2013.

Habiba, C. Tantipathananandh, T. Berger-Wolf. Dynamic networks generative model for skewed component distribution. SIAM workshop on network science, 2013.

Habiba, T. Berger-Wolf. Working for influence: effect of network density and modularity on diffusion in networks. IEEE ICDM 2011 Workshop on Data Mining in Networks, 2011.

Ipek Kulachi, Habiba, Caitlin L. Barale Rajmonda Sulo, Tanya Berger-Wolf, Dan I. Rubinstein. Influence of ecological and social factors on group dynamics in sheep. Animal Behavior Society Annual Meeting, 2010.

Caitlin L. Barale, Ipek Kulachi, Habiba, Rajmonda Sulo, Tanya Berger-Wolf, Dan I. Rubinstein. Social Networks Approach to Sheep Movement and Leadership. The 7th international conference on Applications of Social Network Analysis ASNA 2010.

Habiba, Yintao Yu, T. Berger-Wolf, Jared Saia. Book chapter: Finding Spread Blockers in Dynamic Networks. Lecture Notes in Computer Science journal (LNCS).2009.

Mayank Lahiri, Arun S. Maiya, Rajmonda Sulo, Habiba, Tanya Y. Berger-Wolf. The Impact of Structural Changes on Predictions of Diffusion in Networks. ICDM Analysis of Dynamic Networks (ADN) workshop. 2008.

Habiba, Yintao Yu, T. Berger-Wolf, Jared Saia. Finding Spread Blockers in Dynamic Networks. KDD Social Network Analysis (SNA) workshop. 2008.

Habiba, T. Berger-Wolf. Graph Theoretic Measures for Identifying Effective Blockers of Spreading Processes in Dynamic Networks. Mining Large Graphs (MLG) workshop. 2008.

Habiba, C. Tantipathananandh, T. Berger-Wolf. Betweenness Centrality Measure in Dynamic Networks. DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction Networks. 2007.