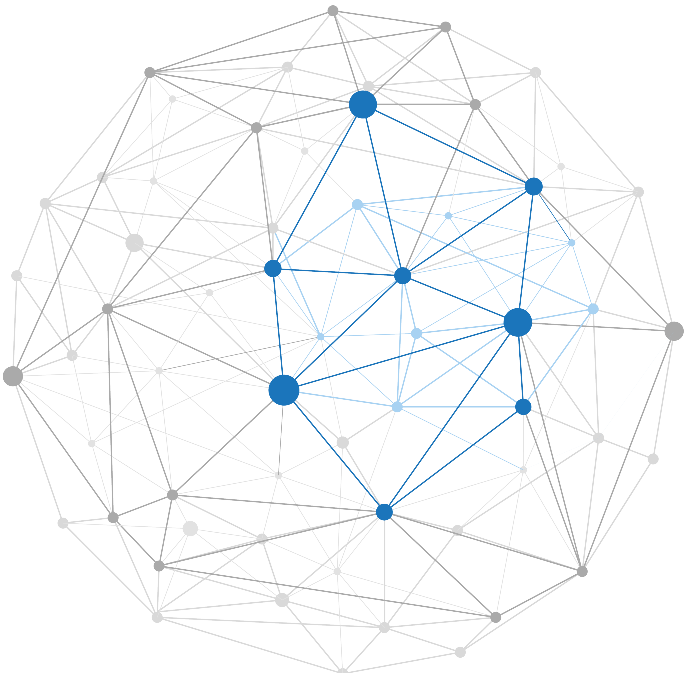
University of Zurich[UZH]

Quantifying and Correcting the Majority Illusion in Social Networks

Thesis December 1, 2016



Markus Göckeritz
of Spreitenbach AG,
Switzerland

Student-ID: 13-750-633
markus.goeckeritz@uzh.ch

Advisor: **Daniel Spicar**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

Acknowledgements

I would like to express my gratitude to Professor Abraham Bernstein, Ph.D., for giving me the opportunity to write my bachelor thesis. Also, I would like to thank my supervisor, Daniel Spicar, who has provided me with valuable substantive help and thoughtful advice and who has guided me through the process of writing this thesis.

Zusammenfassung

Das Phänomen der Majority Illusion, welches ursprünglich von Lerman et al. (2015) entdeckt wurde, versetzt Personen in den Irrglauben, dass ein Verhalten, oder Attribut, sehr populär sei, obwohl es in der Realität nur selten vorkommt. Dadurch, dass sehr stark vernetzte Personen ein bestimmtes Verhalten aufweisen, wird weniger stark vernetzen Personen, die mit ihnen verbunden sind, vorgetäuscht, dass dieses Verhalten sehr populär wäre. Wir zeigen in unserer Arbeit, wie die Existenz und das Ausmass der Majority Illusion für ein Netzwerk auf Knotenebene, wie auch auf Netzwerkebene berechnet werden kann. Im Kontext der Adoption sozialen Verhaltens ist die Majority Illusion ein interessanter Fall von falschen Wahrnehmungen. Wir präsentieren eine Methode, wie die Majority Illusion ausgenutzt werden kann, um die Verbreitung eines binären Attributes in Netzwerken in einem Threshold Modell nach Granovetter (1978) künstlich zu verstärken. Wir vergleichen unsere Methode mit einem naiven Knotengrad-Ansatz, der nach stark vernetzen Knoten sucht. Unsere Methode findet Knotenmengen, die garantiert eine Kaskade von Adoptionen auslösen, die das ganze Netzwerk aktivieren. In skalenfreien Netzwerken, die Strukturen aufweisen, wie sie von Barabási et al. (2000) und Dorogovtsev und Mendes (2002) beschrieben wurden, liefert unsere Methode bessere Resultate, als der Knotengrad-Ansatz. In Netzwerken die Strukturen aufweisen, wie sie von Watts und Strogatz (1998) beschrieben wurden, liefert sie Resultate, welche im Durchschnitt doppelt so viele Knoten beinhalten, wie die des Knotengrad-Ansatzes. Darüber hinaus präsentieren wir eine Variante eines dynamischen Adoptionsmodells, welches den Zeitaspekt berücksichtigt und Annahmen, die wir über das menschliche Verhalten in der realen Welt treffen, in das Modell integriert. Es konnte keine Beziehung zwischen dem Ausmass und der Geschwindigkeit der Adoption und dem Clustering innerhalb des Netzwerkes, wie es von Centola (2010) und Centola und Baronchelli (2015) vorgeschlagen wird, belegt werden. Ebenso war es uns nicht möglich, einen solchen Zusammenhang zu widerlegen.

Abstract

The majority illusion that was discovered by Lerman et al. (2015) tricks individuals into perceiving a social behavior to be popular when in reality, it is not. That is, vertices in a network overestimate the presence of an attribute as highly connected vertices skew the perception of their neighbors. We show how the majority illusion can be quantified on a vertex-centric and a global perspective for binary as well as for continuous attributes. In the context of social contagion, the majority illusion is an interesting case of disproportionate experiences that can cause a false truth to propagate through a network. We propose an approach to exploit the majority illusion in order to artificially promote the diffusion of a binary attribute in a network in a threshold model as introduced by Granovetter (1978). Our approach returns target vertex sets that are guaranteed to cause an influence cascade that eventually activates the entire network. Our approach out-performs a naive highest-degree approach in scale-free networks that exhibit network structures as described by Barabási et al. (2000) and Dorogovtsev and Mendes (2002). In small-word networks as described by Watts and Strogatz (1998) our approach returns target vertex sets that, on average, have twice the size of target vertex sets retrieved with a highest-degree approach. Additionally, we introduce an alternative dynamic diffusion model that considers the time dimension and incorporates assumptions we make about human behavior in the real world. In the diffusion model we introduce, we were unable to confirm or to disprove that the extent and speed at which a social behavior propagates in a diffusion process profits from highly clustered network structures as suggested by Centola (2010) and Centola and Baronchelli (2015).

Table of Contents

1

Introduction

People live in networks. Private networks of other people they are connected with. Ranging from friendships, work and advice networks to information transfer, online and offline social networks and many more. Within these networks, people are in constant interaction with other individuals, which exposes them to external social or behavioral information and influence. Within these networks, various counterintuitive phenomena and paradoxes were discovered. One of the most famous network paradoxes is the friendship paradox (Feld, 1991). The friendship paradox states that on average, your friends are more popular than you. Similar paradoxes that apply to various other characteristics such as the number of followers Twitter users have and the number of citations authors have were since identified (Hodas et al., 2013; Eom & Jo, 2014). The so called *majority illusion* is a special case of the friendship paradox. It describes how under certain naturally occurring network configurations individuals in a network may perceive a social behavior to be popular when in in reality, it is not. This creates situations in which individuals systematically misperceive the extent to which their peers exhibit a certain attribute. Researchers have found that individuals condition their own social behavior to a large degree on the social behaviors of their peers (Schelling, 1973; Granovetter, 1978; Bettencourt et al., 2006; Rogers, 2010; Young, 2011; Salganik et al., 2006). Social behavior encompasses not only physical interactions but also the adoption of conventions, innovations and ideas. Misperceptions that cause people to overestimate the extent at which their friends engage in risky and unhealthy behavior may lead someone to approximate the misperception with his or her own behavior (Baer, Stacy, & Larimer, 1991; Berkowitz, 2005; Bearak, 2014). Furthermore, social behavior is often modelled akin to communicable diseases (Granovetter, 1978; Goldenberg et al., 2001a, 2001b; Bettencourt et al., 2006; Centola, 2010; Centola & Baronchelli, 2015). These models try to simulate the spread of a social behavior in a network of interacting individuals. In these networks, individuals interact only with a small subset of the entire population. The structure of these networks and the direct neighborhood of an individual play a significant role for the propagation of a social behavior. The risk of contagion depends on the distribution of contagion among an individual's direct neighborhood as well as the strength and frequency of social influence. Researchers have studied the problem of identifying influential individuals in a network that play an important role for such diffusion processes in order to exploit aspects of network paradoxes and the contagiousness of social behavior to artificially promote the spread of a social behavior (Domingos & Richardson, 2001; Domingos, 2005; Kempe et al., 2015; Goyal et al., 2010; Chen et al., 2009).

We present an overview of related scientific research that elaborates on network phenomena and paradoxes, the contagiousness of social behavior, influence maximization and networks and their structures in Chapter 2. Chapter 3 is dedicated to an introduction into the majority illusion and create the context for this work. In Chapter 4 we propose an approach to quantify the existence and the magnitude of the majority illusion for a given network on a vertex-centric as well as a global perspective. We introduce and evaluate an approach to estimate sets of influential vertices in order to artificially promote the spread of a social behavior in a network in Chapter 5. In Chapter 6 we introduce an alternative approach to simulate the spread of a social behavior that considers the time dimension and incorporates assumptions that we make on human behavior in the real world. We show the results of an experiment we conducted using the diffusion model that we introduce. We discuss limitations to our work in Chapter 7 and discuss possible further scientific contributions in Chapter 8. The thesis concludes with the general results of our evaluations as well as our thoughts on the contributions that we could make.

2

Related Work

In this chapter we want to give the reader a general overview of scientific research that is related to this work. In Section 2.1 we introduce various phenomena and paradoxes related to the majority illusion that emerge in networks. The paradoxes and phenomena explain misperceptions that arise and lead individuals into misperceiving the world around them. Additionally, we introduce the work of scientists that successfully used implications drawn from these network phenomena to improve the early detection of contagious outbreaks. In Section 2.2 we introduce a number of scientific papers that discuss the contagiousness of social behavior, including examples in which individuals were shown to adopt social behaviors from their peers, as well as two models that are popularly used to model the diffusion of social behavior and two experiments in which the relationship between network structure and the extent and speed of a diffusion process were studied. In Section 2.3 we introduce various studies that investigate finding highly influential individuals in networks. In Section 2.4 we introduce two major findings of the structure of real-world networks including recent research on large real-world online social networks as well as models to generate networks that exhibit specific properties that were found in real-world networks.

2.1 A World Full of Illusions

2.1.1 Cases of Disproportionate Experiences - The Class Size Paradox

The class size paradox describes the phenomenon that students perceive average class sizes to be different than the actual average class size, i.e. the average that the faculty or university sets and perceives. The paradox was discovered by Feld and Grofman (1977).

The phenomenon occurs because students experience the class size of the classes they are in, and the size of the class they are in accounts to the average class size proportionally to the number of students in it, i.e. the higher the size of a class, the larger will be its contribution to calculating the average class size among students. That is, if a class comprises a high number of students, a high number of students will experience the corresponding class size simultaneously, whereas if a class is constituted of only a small number of students, only a small number of students will experience the lower class size, which leads to the biased perception of average class sizes.

Feld and Grofman (1977) have shown that the variation of class sizes is responsible for this contrariety in perceptions and that if there is any variation in class sizes at all, the paradox is guaranteed to occur.

A class size paradox arises when individuals disproportionately experience classes of different sizes. Similarly, disproportionate experiences may emerge in cases other than classes, such as the average size of corporations, average size of cities, population density, or how crowded public places, restaurants and roads on average are (Hemenway, 1982; Granovetter, 1984).

2.1.2 Your Friends Are More Popular Than You

"On average, your friends have more friends than you", states the friendship paradox discovered by Feld (1991), one of the best known network paradoxes. The friendship paradox describes the seemingly counterintuitive observation that the majority of individuals in a social network has fewer friends than their friends have on average. That is, the average person has less friends than the average friend of a person.

Let us illustrate the friendship paradox based on a friendship network that was originally used by Feld (1991), built with data on friendships among students collected at different high schools.

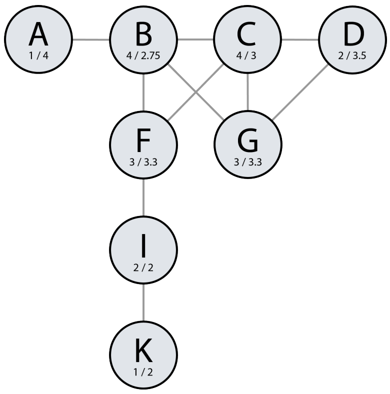


Figure 2.1: Friendships among eight students (Feld, 1991)

Figure 2.1 shows each student's number of friends as well as the average number of friends among his or her friends. For example, F has three friends and F's friends, on average, have slighty more friends. There are eight individuals depicted in figure 2.1 out of which five have less friends, i.e. are less popular, than their friends on average. One individual has as many friends as his or her friends do and only two have more friends than their friends have on average. Looking at the entire network, on average, each person has 2.5 friends, while on average, their friends have 2.98 friends, which means that the average person in the friendship network in Figure 2.1 is on average less popular than his or her friends.

The distortion of the paradox arises because highly connected individuals show up disproportionately more often as friends because they are observed more often by others and thus significantly skew the mean number of friends of friends. Very similar to the class size paradox discovered by Feld and Grofman (1977), people are overrepresented according to their popularity. That is, people with a high number of friends, appear on a high number of lists of friends, whereas people with only a few friends will appear on fewer lists of friends.

In his work, Feld (1991) has shown that for a social network with a degree distribution that has non-zero variance, i.e. where individuals do not all have the same number of friends, the mean number of friends of friends is always greater than the mean number of friends of individuals.

2.1.3 Your Friends Are More Interesting Than You

Hodas et al. (2013) have shown in their work that, using a sample of the Twitter firehose, the friendship paradox holds for more than 98% of Twitter users. For most of the users, everyone that they follow or who follows them has more friends and followers than they themselves have on average. Furthermore, they have discovered two more detailed variations of the friendship paradox, namely the virality paradox that states *"your friends receive and send out more viral content than you"* and the activity paradox which states *"your friends are more active than you"*.

Hodas et al. (2013) have shown a large correlation between the activity and the connectivity, i.e. followers and friends, in the follower graph on Twitter. Therefore, Hodas et al. (2013) propose the hypothesis that activity causes connectivity, leading to the two variations of the friendship paradox that they have discovered.

2.1.4 Generalized Friendship Paradox

Using two coauthorship networks of physicists and network scientists, Eom and Jo (2014) have shown that the friendship paradox holds for coauthorship characteristics such as the number of coauthors, number of citations and number of publications. That is, your coauthors, on average, have more coauthors, citations and publications than you.

To Eom and Jo (2014), this was an indication that the friendship paradox might hold for node characteristics other than degree. Thus, they have investigated a generalized form of the friendship paradox. Their research focused on answering the question, whether the friendship paradox holds for any other arbitrary node characteristic such as income or happiness.

Eom and Jo (2014) were able to show that the origin of the generalized friendship paradox at network level is rooted in a positive correlation between degree and characteristic, suggesting that the friendship paradox holds for any arbitrary node characteristic that is positively correlated with degree.

2.1.5 Exploring the Origin of Network Paradoxes

Kooti et al. (2014) examined whether the origin of network paradoxes such as the friendship paradox is rooted in mathematical properties of networks and their structures, or whether they have a behavioral origin, i.e. whether they originate from individuals connecting themselves to other individuals with certain characteristics. Using two social networks drawn from Digg and Twitter, Kooti et al. (2014) confirmed the friendship paradox as well as variations of the friendship paradox such as the activity paradox and the virality paradox. Furthermore, Kooti et al. (2014) have shown that the friendship paradox, as well as most variations, also hold when user attributes are compared to the median of the attributes among their neighbors, as opposed to the mean, which strengthens network paradoxes for several attributes. That is, not only are your friends *on average* more popular than you, but *the majority* of your friends are more popular than you.

Similar to Eom and Jo (2014), Kooti et al. (2014) explored whether the presence and magnitude of the friendship paradox changes depending on different degree-attribute correlations and assortativity settings. They conclude that, although both correlations appear to play a significant role in the presence and magnitude of certain variations of the friendship paradox, individuals seem to dynamically position themselves in the network to remain subject to the paradox. Kooti et al. (2014) propose that this is because behavioral factors that create those correlations are often related to desirability of the attribute of the paradox. This suggests that besides the mathematical nature of the friendship paradox and related paradoxes, behavioral factors play a significant role in the cause of the correlations and the paradoxes.

2.1.6 Early Detection of Contagious Outbreaks

Christakis and Fowler (2010) and Garcia-Herranz et al. (2014) have explored a novel strategy for placing sensors in social networks to detect contagious outbreaks which employs findings from the friendship paradox (Feld, 1991). Christakis and Fowler (2010) and Garcia-Herranz et al. (2014) propose the following approach: monitoring the friends of randomly sampled individuals. The idea is simple, friends of randomly sampled individuals are expected to be more popular, e.g. have higher degree, than the randomly sampled individuals and are therefore more likely to interact with a higher number of people, making them better targets for detecting contagious outbreaks.

Christakis and Fowler (2010) present the results of their strategy, tested on 744 students from Harvard College, consisting of 319 randomly sampled students and 425 friends of students in the random sample. Using the group of friends, they were able to identify outbreaks of flu a significant number of days earlier than solely with the random sample, indicating that their approach could be an effective way to detect contagious outbreaks at early stages of an epidemic.

Garcia-Herranz et al. (2014) used this method to detect contagious message spreading in Twitter and present their results as follows. Using Twitter data of 40 million users, 1.5 billion directed relationships and nearly half a billion messages, they were able to identify outbreaks of contagious message spreading about seven days earlier than with traditional approaches.

2.2 Contagiousness of Social Behavior

2.2.1 The Threshold Model

Granovetter (1978) introduced a threshold model of collective behavior to describe the relation between collective behavior and individual motives, which previous models did not include but implicitly assume. A threshold model is used in situations in which individuals are given two distinct and mutually exclusive alternatives. Individuals condition their decisions on a simple threshold rule. Each individual has a threshold, a number or proportion of other individuals a given individual is connected to, that must make the same decision before a given individual adopts it. That is, at each time step, an individual observes the decisions other individuals have made and if the proportion of individuals that have adopted a certain behavior is equal to or larger than the threshold of the individual, the individual will adopt the same behavior. Threshold models assume that individuals are rational and possess complete information, i.e. an individual observes the decisions of all other individuals.

2.2.2 The Cellular Automata Model

Goldenberg et al. (2001b, 2001a) studied the effects of word-of-mouth communication in the context of marketing. In their research, they used a cellular automata model to simulate the effect word-of-mouth communication has on the promotion and distribution of a new product. In a cellular automata model, individuals influence each other following a stochastic model. Initially, a subset of individuals that have adopted the social behavior convince their peers with a predefined probability to also adopt the social behavior. Each individual that could be convinced tries to convince its peers with the same probability. An individual that fails to convince its peers does not retry. Eventually, either all individuals could or could not be convinced and the diffusion process halts.

2.2.3 The Social Norms Theory

Berkowitz (2005) investigated the theory of social norms. The social norms theory describes situations in which an individual perceives the extent at which its peers exhibit a certain social behavior to be different from their own when in reality, this is not the case. Berkowitz (2005) discusses previous research that suggests that an overestimation of certain social behaviors among an individual's peers influences an individual to increase the extent to which it exhibits the same social behavior. That is, individuals that overestimate the use of alcohol, cigarettes, drugs and other behaviors and attitudes among their peers increase their own use to approximate the misperceived extent.

2.2.4 The Spread of Ideas

Bettencourt et al. (2006) studied the spread of ideas using several generic population models inspired by epidemiology. Particularly, they studied the spread of a specific scientific idea, Feynman diagrams. Their results suggest that adapted epidemic models are able to accurately capture aspects of the spread of ideas. With this, Bettencourt et al. (2006) have shown that the diffusion of ideas shares similarities to the diffusion of diseases. Furthermore, they have shown that the spread of Feynman diagrams shared similarities to the spread of communicable diseases.

2.2.5 The Spread of Popular Songs

Salganik et al. (2006) investigated the social influence in cultural markets, studying what qualitative difference tremendously successful songs, books and movies exhibit in comparison to their unsuccessful alternatives. Salganik et al. (2006) conducted an experiment in which they created an artificial music market and assigned $14,341$ participants to two experimental groups. The participants were asked to listen to various songs and assign a rating to each song they listened to. Participants in one group were only given information about the music that they listened to, e.g. song and band name, whereas participants in the other group were additionally given information about the choices of other participants. The results of their experiment suggest the choices people make on what music they listen to is affected by the music their peers listen to. That is, the social behavior of individuals is affected by the social influence of their peers.

2.2.6 The Spread of Social Innovation

Young (2011) modelled the dynamics of social innovation as a coordination game played on a network of interacting individuals. In his research, Young (2011) studied how a particular set of conventions and rules establish, such that they become common practice in a population of individuals. Young (2011) suggests that the speed at which conventions emerge depends on the relative advantage individuals obtain from adopting the convention and the existence of autonomous groups that adopt the innovation in early stages of the diffusion process.

2.2.7 The Spread of Behavior in an Online Social Network Experiment

Centola (2010) investigated the effects of network structure on the spread of behavior by studying the spread of behavioral adoption in artificially structured online communities. His results show that local clustering in a social network can tremendously improve behavioral adoption on an individual level by reinforcing signals, but also that large scale diffusion can reach more people and spread more quickly in clustered networks.

2.2.8 The Spontaneous Emergence of Conventions

Centola and Baronchelli (2015) have studied the effects of network structure on the spontaneous emergence and evolution of social conventions, without any institutional mechanism guiding or coordinating the process. That is, Centola and Baronchelli (2015) investigated how purely local interactions and conventions can produce global coordination and trigger the emergence of a universal convention.

Their results show that depending on the underlying network structure, local groups of coordinated individuals may emerge and start competing for global dominance, i.e. compete to produce a globally accepted convention. Usually, if local groups of coordinated individuals emerge, particularly if multiple groups emerge quickly, the population is not able to generate collective agreement on a global convention as neighboring groups compete and never persuade each other. In their experiments, this was the case for random graphs as well as spatially embedded networks. Furthermore, Centola and Baronchelli (2015) present in their results that homogeneously mixing populations leads to coordination dynamics in which individuals quickly generate a dominant convention which, in all of their trials, resulted in a global convention.

2.3 Influence Maximization

2.3.1 The Network Value of a Customer

Domingos and Richardson (2001), Richardson and Domingos (2002) and Domingos (2005) have studied the influence among customers in the context of viral marketing. Traditionally, customer value was defined as the expected profit from sales to a customer, leaving out the fact that customers might influence their friends to buy the same product. Domingos (2005) included this fact and introducted the customer network value. The network value of a customer is the expected profit from sales to other people that the customer might influence to buy a product. Domingos and Richardson (2001), Richardson and Domingos (2002) and Domingos (2005) studied the problem of influence maximization, to be able to find the most influential set of customers, using a probabilistic model of interactions.

2.3.2 Influence Maximization: A Discrete Optimization Problem

Kempe et al. (2015) studied the problem of influence maximization as a problem in discrete optimization. Kempe et al. (2015) used two different diffusion models. A threshold model, as described by Granovetter (1978), and a cellular automata model, as described by Goldenberg et al. (2001a, 2001b). Kempe et al. (2015) introduced approximation algorithms that solve the problem of finding good sets of target vertices in the context of the two models. Furthermore, Kempe et al. (2015) have shown that the optimization problem is NP-hard.

2.3.3 Influence Maximization Problems

Goyal et al. (2010) investigated alternative influence maximization problems motivated by Kempe et al. (2015). The problems include finding sets of target vertices such that the number of vertices that will be activated covers at least n vertices and finding such a coverage in minimal time. Goyal et al. (2010) have introduced approximation algorithms to solve the problems they proposed and have shown that both problems are NP-hard.

2.3.4 Efficient Influence Maximization

Chen et al. (2009) introduced several improvements to the approximation algorithms introduced by Kempe et al. (2015) and Goyal et al. (2010). Furthermore, Chen et al. (2009) introduced a new vertex selection heuristic based on degree discounts. The idea of degree discount is that if a vertex v_i was already selected to be in the target vertex set, the degree of a new vertex v_j that is considered as potential target vertex is discounted by one if it is connected to vertex v_i, i.e. the edge between the two vertices is not counted towards the degree of v_j. The same discount is applied to v_j for all vertices it shares an edge with that are already selected as target vertices. Chen et al. (2009) have shown that their ehancements on the approximation algorithms introduced by Kempe et al. (2015) and Goyal et al. (2010) could improve the running time by 15% to 34%, while matching the quality of the result. Furthermore, the degree discount heuristic that they introduced was able to improve the running time of the approximation algorithm by more than six orders of magnitude.

2.4 Networks and Network Structures

2.4.1 Price's Model

Researcher de Solla Price (1965) has studied a network of citations between scientific papers and found that the distribution of references follows a power-law. In a later paper, Price (1976) introduced a theory applied to the growth of a network that explains the power-law distribution. Price (1976) called the effect cumulative advantage, the effect that *"the rich get richer"*. Price (1976) proposed that the power-law follows because scientific papers that are frequently cited are more likely to be cited than papers that are less known. Also, Price (1976) suggested that authors who published many papers are more likely to publish additional papers than authors that have published fewer papers.

2.4.2 Preferential Attachment

Barabási and Albert (1999) re-introduced the idea of cumulative advantage and reported their existence in three different networks: an actor collaboration network, the world-wide web and a power grid network. Furthermore, Barabási and Albert (1999) have shown that in these three networks, the probability $P(k)$ that a vertex in a network has degree k, decays as a power-law following $P(k) \sim k^{-\gamma}$, whereby $\gamma \in [2.1, 4.0]$. Barabási et al. (2000) introduced a model to generate networks, in which vertices that are added to the network preferentially attach to vertices with higher degrees, which naturally leads to a degree distribution that follows a power-law.

2.4.3 Natural Preferential Attachment

Dorogovtsev and Mendes (2002) introduced a model to generate networks similar to the model that Barabási et al. (2000) proposed. Networks generated by both models have degree distributions that follow a power-law. The preferential linking in the model that Dorogovtsev and Mendes (2002) introduced, arises not from a probability $P(k)$ to connect to high degree vertices, but naturally from the underlying process. In their model, when a new vertex is added to a network, it is connected to both ends of a randomly chosen edge in the network. Since high degree vertices have a high number of edges, it is more likely to find high degree vertices on one of the endpoints of a randomly chosen edge.

2.4.4 Small-World Network Model

Watts and Strogatz (1998) introduced a model to generate networks that share the property of the small-world phenomenon, the phenomenon of *six degrees of separation*. In their model, vertices are initially arranged in a ring and connected to their k nearest neighbors. Conditioned upon a parameter β, the edges of a vertex are randomly reconnected to a vertex chosen uniformly at random. This creates networks that can be highly clustered yet have small characteristic path lengths.

2.4.5 Structure and Evolution of Online Social Networks

Kumar et al. (2006) studied the structure and evolution of online social networks such as Flickr and Yahoo! 360. They have shown that the degree distribution of the two online social networks follows a power-law. Furthermore, Kumar et al. (2006) classify members of the social networks into three groups: singletons, giant component and middle region. Singletons are zero-degree vertices that joined but never participated in the social network. The giant component represents the large group of vertices who are directly and indirectly connected to each other. The middle region consists of various isolated communities that are not part of the giant component. Kumar et al. (2006) show that almost all isolated communities in the middle region are structured like stars. Furthermore, Kumar et al. (2006) show that the average diameter in the giant component is close to six.

2.4.6 The Anatomy of the Facebook Social Graph

Ugander et al. (2011) studied the structure of the social graph of active facebook users. They have shown that the degree distribution of the Facebook social graph is monotonically decreasing, but does not follow a strict power-law. Furthermore, Ugander et al. (2011) have shown that almost the entire Facebook social graph consists of one giant connected component with an average distance between pairs of users of 4.7, an average local clustering coefficient of 0.14 and degree assortativity of 0.226. Ugander et al. (2011) confirm the friendship paradox in the Facebook social graph and show that the friendship paradox holds for 92.7% of all users.

2.4.7 Similarity Breeds Connection

"Birds of a feather flock together" - the principle that individuals tend to connect to individuals they share similarities with. McPherson et al. (2001) studied the occurrence and extent of homophily in social networks. That is, McPherson et al. (2001) studied the extent to which individuals in a social network share sociodemographic, behavioral or intrapersonal characteristics and show that people's personal networks are homogeneous with regard to those characteristics.

3

The Phenomenon That Tricks Our Minds

Social networks have many complex properties that can lead to counterintuitive phenomena. A well known example is the class size paradox, which was initially discovered by Feld and Grofman (1977). The class size paradox shows that students of a university perceive the average class size very differently from the actual average class size. The class size paradox is a case, where students disproportionately experience high numbers and thus they think that the average class size is larger, than it actually is. Hemenway (1982) and Granovetter (1984) have found other cases of disproportionate experiences, where individuals perceive the average sizes of cities to be larger than they actually are, or they perceive public places and roads to be more crowded than they are in reality. Another famous example of a network paradox is the friendship paradox discovered by Feld (1991), which states that on average, your friends are more popular than you. Hodas et al. (2013) have shown that the friendship paradox is true for nearly all twitter users and Eom and Jo (2014) and Kooti et al. (2014) discovered and showed that the friendship paradox also emerges for network characteristics other than degree. Lerman et al. (2015) have discovered a novel variation of the friendship paradox, the *majority illusion*, an example of disproportionate experiences, a paradox that affects an individuals perception and the collective social phenomena that emerge, a paradox that might explain why some messages, informations or ideas can spread rapidly and widely while others just perish in the sheer mass of a social network.

3.1 The Majority Illusion

Lerman et al. (2015) have discovered what they call the majority illusion, a phenomenon that can trick people into thinking that something is common, when in reality, it is actually rare. The phenomenon describes how individuals in a social network, under certain naturally occurring network configurations, may perceive the majority of their neighbors to exhibit a certain characteristic or attribute, even though, among the entire network, only a minority of individuals actually carries the characteristic. That is, individuals in a social network may perceive a certain attribute to be popular, when in fact, it is globally rare.

The majority illusion applies to networks in which vertices have attributes. Attributes can be as simple as a binary characteristic such as “has a smartphone” versus “does not have a smartphone" or "is a vegetarian" versus "is not a vegetarian". Attributes can also be more complex and describe behavioral characteristics, personality traits or opinions such as "is adventurous" versus "is unadventurous" or "agrees with abc's political views" versus "disagrees with abc's political views". In such cases, attributes may be continuous and describe the extent to which an individual agrees with a certain opinion. Additionally, attributes may also be either implicit or explicit, which we will see in Chapter 4. For binary attributes, we say that a vertex is active, if it carries the attribute and likewise, we say that a vertex is inactive, if it does not carry the attribute.

Let us illustrate the illusion with the example that Lerman et al. (2015) already used to familiarize with the paradox.

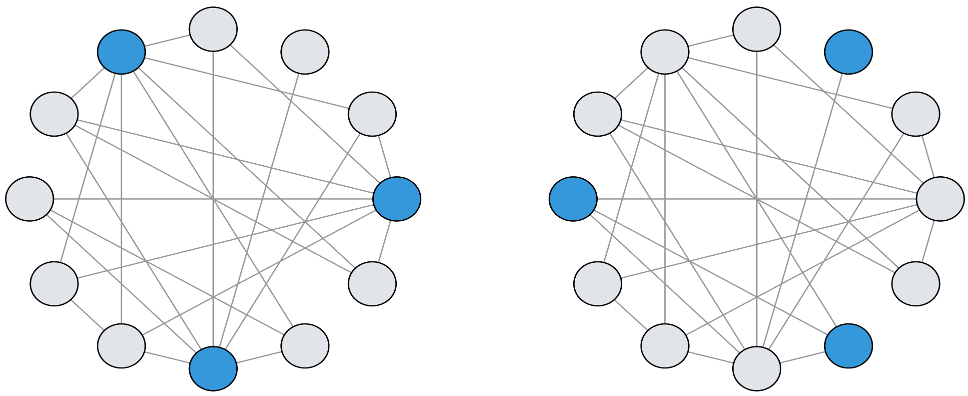


Figure 3.1: Illustration of the majority illusion paradox (Lerman et al., 2015)

The two networks in Figure 3.1 are identical, except for which three vertices are colored. Suppose that the color of a vertex represents whether a vertex is active or not, whereby grey vertices are inactive and blue vertices are active. In the network on the left, all grey vertices observe the majority of their neighbors to be active, whereas in the network on the right, not a single inactive vertex has a majority of active neighbors. Even though the blue vertices are clearly a minority, when looking at the entire network, it appears to *all* vertices, except for the active vertices themselves, in the left network, that most of the their friends are active.

An individual's social behavior appears to be affected to a large degree by its perception of the behaviors and actions of others. Social behavior is exhibited not only through physical or explicit interactions between individuals, but also through the adoption of new ideas, innovations and preferences in the form of cognitive or emotional behavior. Bettencourt et al. (2006), Rogers (2010) and Young (2011) have shown that individuals are more likely to adopt a new technology or innovation, the more of their peers also adopt the same new technology or innovation. Salganik et al. (2006), Schelling (1973), Granovetter (1978) and Bearak (2014) have shown that the strength of social influence also affects the decisions people make to listen to music, join social movements or engage in risky behavior. Social behavior is contagious and the dynamics underlying the diffusion of social behavior have many similarities to those involved in the spread of communicable diseases. Therefore, the spread of social behavior is often modelled akin to contagious diseases. In these models, social interactions are modeled as networks of interacting individuals whose decisions are determined by the actions of their neighbors. Often in these models, the way an individual is influenced by its neighbors follows a threshold rule. That is, an individual has some threshold $T \in [0, 1]$, which is a fraction of neighbors, and at each time step, if the fraction of neighbors which exhibits a certain social behavior is at least T, the individual adopts the social behavior itself. The threshold represents the social evidence it takes to persuade an individual to adopt a certain social behavior.

Let us again use the illustration of the majority illusion in Figure 3.1. Recall that 'active' means that an individual carries an attribute and that attributes encompass social behavior. In the network on the left, all inactive vertices experience a majority of active neighbors. That is, every inactive vertex experiences a fraction of active neighbors that is at least 0.5. In a threshold setting where an individual would switch from inactive to active if at least half of its neighbors are active, i.e. individuals have threshold $T = 0.5$, in the network on the left in Figure 3.1, every inactive individual would eventually become active, whereas in the network on the right, not a single individual would become active.

In the context of social contagion, the majority illusion is a very interesting case of disproportionate experiences. Similar to the examples of Feld and Grofman (1977) and Feld (1991), the majority illusion creates a false truth, something that is believed by many individuals to be true but is not. In the case of the majority illusion however, depending on what social behavior is promoted, the collective social phenomena that might emerge can have a tremendous impact on the individuals that are affected. The majority illusion is able to promote the spread of social behaviors as simple as what music will be perceived as popular or what products people will preferably buy, but also social behaviors as complex as revolutionary political movements. The majority illusion potentially explains why some messages, pictures, videos, opinions or ideas can reach a tremendous number of people in a very short time, while others do not; simply because those individuals that promote the spread play a more important, hidden role for the propagation dynamics.

4

Quantifying the Majority Illusion

In this chapter, we show how the presence and magnitude of the majority illusion in a social network can be quantified both on a network level as well as an individual vertex level. Furthermore, we discuss how the magnitude of the illusion could be approximated in cases when networks exceed sizes for which exact calculation is feasible.

4.1 Preliminaries and Definitions

For a given network with vertices V, edges E and an attribute a, we want to derive for each vertex, whether the vertex is subject to the majority illusion. Furthermore, we want to be able to tell what proportion of a network experiences the majority illusion as well as what proportion of vertices causes the misperception. We say, every vertex in a network has an attribute a, where a_i denotes the attribute of vertex v_i, such that i is a generic index and v_i is a member of the network, i.e. $v_i \in V$. Attribute a_i indicates, whether vertex v_i exhibits a certain characteristic a or not, or simply, whether it agrees with opinion a, such that a_i is either binary, i.e. $a_i \in \{0, 1\}$, indicating whether a vertex does or does not exhibit a, or continuous, i.e. $a_i \in [0, 1]$, indicating to what degree vertex v_i agrees to a. Furthermore, a_i can be implicit, i.e. $a_i \in \{0, 1\}$ or $a_i \in [0, 1]$ as well as explicit, i.e. $a_i \in \{-1, 1\}$ or $a_i \in [-1, 1]$. That is, if a_i is explicit, additionally to what extent vertex v_i agrees to a, a_i is able to more naturally describe disagreement. For completeness, we will quantify the majority illusion for attributes a as described in this section, i.e. binary, continuous as well as implicit and explicit. In later chapters, we will work with binary and implicit attributes and continuous where noted.

We define the majority illusion to be existent in a network, if there is at least one vertex $v_i \in V$, for which the majority illusion holds. We call such a vertex to be affected by or to experience the majority illusion. Additionally, we call a network to be affected by or experience the majority illusion, if the majority of vertices experiences the majority illusion.

4.2 Quantifying the Majority Illusion

A vertex v_i experiences the majority illusion if what is popular among the neighbors of v_i and appears to v_i to be common, is globally rare. Therefore, to be able to tell whether some vertex v_i experiences the majority illusion, we compare what is locally popular to what is globally popular. That is, compare the most common attribute among the neighborhood of v_i to the most common attribute in the entire network.

We first want to identify the attribute that occurs most often among the entire network, i.e. the *majority attribute* of the network. In a first step, we take the sum of all attributes of all vertices

$$S_a = \sum_{i=1}^{n} a_i \tag{4.1}$$

where n is the number of vertices in the network. The sum S_a already tells us whether the majority of people in the network are in favour of or against a. Suppose that attributes are explicit and binary, i.e. $a_i \in \{-1, 1\}$. If the majority of people is against a, which means that the majority of vertices has attribute $a_i = -1$, the sum of all attributes of all vertices S_a will be negative. If the sum of all attributes of all vertices S_a is positive, the majority of people agree with a and if S_a equals zero, attributes are distributed evenly and there are as many vertices with $a_i = -1$ as there are vertices with $a_i = 1$. Respectively, suppose that attributes are implicit and binary, i.e. $a_i \in \{0, 1\}$. If the majority of people is against a and thus carry attribute $a_i = 0$, the sum S_a will be lower than $\frac{n}{2}$. This time, we check whether S_a is greater than, less than or equal to $\frac{n}{2}$ instead of 0.0 and can tell what the majority attribute will be, depending on which inequality will be true.

We divide the sum of all attributes over all vertices S_a by the number of vertices in the network, n,

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{4.2}$$

which gives the *average attribute* $\bar{a}$. The average attribute has the same properties as S_a. Namely, whether it is below or above 0.0, or 0.5 respectively, gives us the majority attribute. Additionally, $\bar{a}$ represents the weighted sum of the attributes, such that each attribute accounts to the sum only to the extent of the fraction of attributes that it makes up. The average attribute $\bar{a}$ thus includes the proportion of vertices that actually carry the majority attribute, which indicates how popular the majority attribute really is. For example, suppose there is a network with $n = 500$ and each vertex v_i in this network carries an attribute $a_i \in \{0, 1\}$ such that $\bar{a} = 0.8$. What is the number of vertices that carry attribute $a_i = 1$ in this network?[1]

When dealing with continuous attributes, the nature of the average attribute $\bar{a}$ and the sum of all attributes of all vertices S_a is very much the same, except that we cannot infer from $\bar{a}$ the proportion of people in the network that have some attribute a_i, while still giving us a good estimate of the degree to which people in the network agree with a.

Generally, the average attribute represents the opinion of the entire network. Therefore, if we round the average attribute to the closest attribute a_i, such that $a_i \in \{0, 1\}$ or $a_i \in \{-1, 1\}$, we obtain the majority attribute $[\bar{a}]$.[2] Rounding the average attribute

[1] The number of vertices with $a_i = 1$ must equal $\bar{a} * n = 0.8 * 500 = 400$.

[2] Note that if the average attribute $\bar{a} = 0.5$, or 0.0 respectively, there is no majority attribute.

to the closest integer if attributes are continuous classifies the average attribute into a binary category.

In order to quantify the majority illusion in a social network, we want to identify whether there are vertices that perceive an attribute other than $[\bar{a}]$ to be the most popular attribute among its neighbors. That is, we compare the majority attribute among the neighborhood of a vertex to the actual majority attribute.

To do this, we follow the same principle. We sum up the attributes among the neighbors of a vertex v_i and divide this number by the number of neighbors v_i has, i.e. its degree k_i, which gives us

$$\bar{a_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} a_j \tag{4.3}$$

the average attribute among the neighbors of v_i, which then can be rounded to the closest integer to obtain $[\bar{a_i}]$, the majority attribute among the neighbors of v_i. If for any vertex v_i the majority attribute among its neighbors $[\bar{a_i}]$ does not equal the actual majority attribute $[\bar{a}]$, then the vertex v_i is affected by the majority illusion and the vertices that trick v_i into the illusion are all neighboring vertices v_j that have $a_j = \bar{a_i}$.

After that we have obtained the majority attribute for every vertex $v_i \in V$, we want to find out whether there are vertices in the network that experience the majority illusion, how many there are, by whom it is caused and whether the entire network itself is affected by the illusion. The proportion of vertices that experience the majority illusion is the number of vertices for which $[\bar{a_i}] \neq \bar{a}$, caused by the number of vertices for which $a_i \neq [\bar{a}]$ and if the proportion of vertices that experience the majority illusion is greater than $\frac{n}{2}$, or 0.0 respectively, then the entire network is affected by the majority illusion.

Note that when working with continuous attributes, the described calculation of the majority attribute might not necessarily be suitable. The calculation of the average attribute gives the opinion among the entire network, rounding the average attribute to the nearest integer classifies it into a binary statement, which *does* represent what attribute vertices in the network on average exhibit, but not whether the majority of vertices, if expressed in binary in terms, are in favour of or against the attribute, i.e. agree or disagree with a. Consider the isolated view of a vertex and its neighbors with depicted attribute distribution in Figure 4.1.

The average attribute $\bar{a_i} = 0.525$, which is slightly above the neutral state and means that on average, v_i's neighbors agree with a to a small degree and thus, $[\bar{a_i}] = 1$. However, depending on the perception model one is working with, this might not give the anticipated result. One could argue that a vertex does not interact with all its neighbors, or that an individual only has a limited view on the attributes among its neighbors, i.e. it only sees whether a neighbor agrees or disagrees with a[3], such that it would naturally perceive the majority of its neighbors to *disagree* with a, because in fact, the majority of its neighbors actually do. Nevertheless, quantifying the majority illusion as we have described in this section certainly yields a proper result and quantification for the illusion, particularly in a setting where vertices are aware of the attribute distribution among

[3]Which would be very much the same, as if we would transform a continuous attribute to its binary representation.

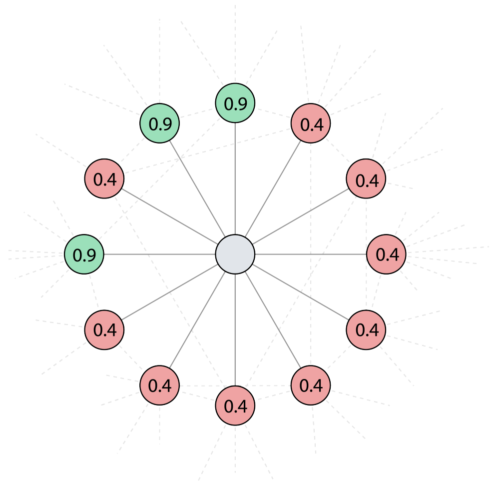


Figure 4.1: Illustration of vertex neighborhood with continuous attributes

their neighbors and where we make no assumptions on how individuals perceive their neighbors. The provided example arises simply from the nature of continuous attributes. Further elaboration on different models are described in Chapter 6.

4.3 Approximating the Majority Illusion

The approach shown in Section 4.2 requires knowledge of the entire network as well as every single edge in the network, which is not suitable for networks of large size as, depending on the network structure, this might require intensive computation. In this section we want to give our thoughts on how the magnitude of the majority illusion could be approximated on a network level as opposed to the vertex-centric fashion shown in Section 4.2.

Lerman et al. (2015) introduced a statistical model to calculate the magnitude of the majority illusion and demonstrated empirically how different network structures affect the magnitude of the illusion. The proposed model uses network properties such as the degree distribution $p(k)$, the probability that an arbitrary vertex v_i has degree k, the joint degree distribution $e(k, k^{'})$, the probability that an arbitrary vertex v_i with degree k is connected to a vertex v_j that has degree $k^{'}$, the joint probability distribution $P(k, x)$, the probability that an arbitrary vertex v_i with degree k has attribute x, as well as inferred properties, such as the assortativity coefficient, or the average degree. Given that the magnitude of the majority illusion can be calculated using these specific network properties, one should be able to approximate the magnitude of the illusion using approximate values for the required network properties.

5

Causing the Majority Illusion

In this chapter, we will use a threshold model to simulate the spread of the misperception caused by the majority illusion. We introduce a novel approach to estimate a subset of vertices in a network to target in order to artificially promote the spread of an attribute in a network by exploiting the majority illusion. We compare the performance of our approach to the performance of the baseline heuristic to target high-degree vertices. We present the results of our baseline evaluation and draw conclusions on the performance of our approach.

5.1 Preliminaries and Model

Granovetter (1978) introduced what is called a threshold model. It describes how social behavior spreads within a group of people using a simple threshold rule. When a certain number, or proportion, of an individual's neighbors adopt a certain social behavior, the individual adopts the social behavior itself. This number is its threshold $T \in [0, 1]$. In a threshold model, the adoption of a behavior is a binary decision between two distinct and mutually exclusive alternatives. We will use a threshold model to simulate the spread of a binary attribute in a network $G = (V, E)$ of interacting individuals, such that each vertex $v_i \in V$ has a threshold T_i. Each vertex v_i has an attribute $a_i \in \{0, 1\}$, which indicates whether v_i does or does not exhibit attribute a. If v_i does not have attribute a, i.e. $a_i = 0$, but the proportion of its neighbors that do exhibit attribute a is equal to or larger than threshold T_i, v_i adopts the behavior of its neighbors. Likewise, if v_i has attribute $a_i = 1$, but the proportion of v_i's neighbors that have attribute $a_j = 0$ is equal to or larger than T_i, v_i will adopt the behavior of its neighbors and will discard the attribute, such that a_i becomes 0. Vertices are fully aware of the attributes of their neighbors and in each time step, a vertex observes all the attributes among its neighbors and either adopts it or not. Whenever a vertex v_i changes its attribute, i.e. it was persuaded by its neighbors, each neighbor observes the change and re-evaluates whether to change its attribute as well. Eventually, the network will converge and vertices will stop changing their attributes. Under certain network configurations, this might lead to influence cascades, in which individuals are persuaded, which persuades other individuals, which persuades others, and so on, such that the number of vertices that have adopted attribute a is significantly higher than the number of vertices that initially exhibited the attribute.

We use a threshold model with threshold $T_i = 0.5$ for every vertex $v_i \in V$. This is the model that Lerman et al. (2015) used to demonstrate the majority illusion. A threshold model with $T = 0.5$ accurately captures the notion of an illusion in that an individual affected by the illusion will observe a majority of its neighbors exhibit some attribute a, while in the global perspective still belonging to a minority.

Let us illustrate the threshold model and such an influence cascade with an example.

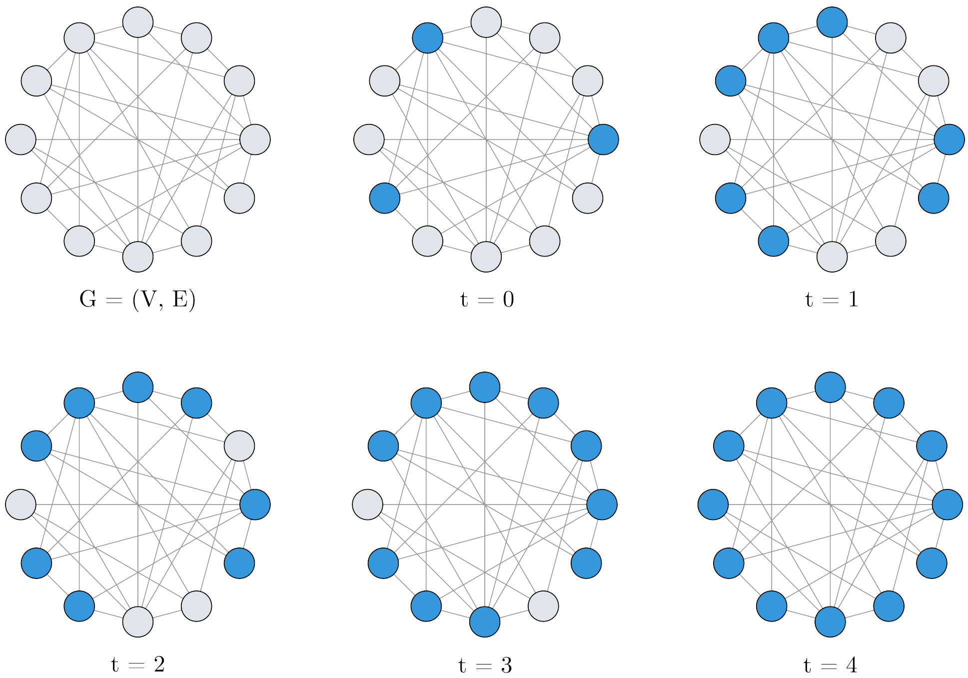


Figure 5.1: Illustration of threshold model and influence cascade using $T_i = 0.5$ (Granovetter, 1978)

Figure 5.1 depicts the same network in five different stages t of a diffusion process in a threshold setting, whereby t indicates the time step the network is in. The color of a vertex indicates whether it exhibits attribute a, i.e. whether it is active. Grey vertices are inactive and blue vertices are active. Each vertex v_i in the example network has a threshold of $T_i = 0.5$. That is, a vertex in this network adopts attribute a and changes its color from grey to blue if the proportion of active neighbors is equal to or larger than 0.5. The network configuration where $t = 0$ is the initial configuration and attribute distribution, from which the diffusion process starts. After the first diffusion step, the network is in time step $t = 1$. Before the transition between $t = 0$ and $t = 1$, four inactive vertices observed that the proportion of their neighbors that are active exceeds their threshold and thus adopted attribute a and changed their color to blue. Likewise, in the following time steps, more and more vertices observe a proportion of active neighbors that is larger than their threshold and also adopt the attribute. Note that the number of vertices that were initially active is significantly lower than the number of vertices that were subsequently activated. The majority of inactive vertices, that were convinced to adopt attribute a, were not convinced to do so solely by the vertices that were initially active. It required additional vertices to adopt attribute a beforehand, in order to supply those vertices with the amount of social evidence needed to persuade them. Nevertheless, the influence of the minority vertices that were initially active sufficed to persuade some fraction of their neighbors, which then sufficed to activate additional vertices, such that eventually, the entire network could be activated. That is, initially activating three influential vertices has triggered an influence cascade that spread across the entire network and managed to activate a significant number of additional vertices - in this case, the entire network.

Influence cascades are a good example to show the importance of the interplay of multiple individuals and their positions in a network for diffusion processes. An individual does not necessarily have to influence a high number of other individuals in order to be influential. An individual is most influential, if it is able to persuade many other *influential* individuals, which then further influence more (influential) individuals, and so on. Likewise, an individual is not necessarily influential, i.e. important for diffusion dynamics, if it is highly connected. Finding a good set of individuals to target in order to intentionally promote the spread of an attribute is a very complex task. The problem of finding such sets is known as influence maximization.

5.2 Causing the Majority Illusion

Domingos and Richardson (2001), Richardson and Domingos (2002) and Domingos (2005) have studied the problem of influence maximization in the context of viral marketing as a probabilistic model of interactions. A customer is said to have a network value, which is the expected profit from sales to people the customer might influence to buy the same product. That is, activating a specific customer might trigger an influence cascade in which the social behavior that is spread is the purchase of a product. Kempe et al. (2015) considered the problem of selecting a set of most influential individuals in a network as a problem in discrete optimization and have shown that the problem is NP-hard for threshold models. Kempe et al. (2015) introduced approximation algorithms to compute sets of most influential individuals in a network and showed that their algorithms significantly out-perform vertex selection heuristics based on degree centrality and distance centrality. Goyal et al. (2010) studied alternative optimization problems which address the same influence maximization problem. Particularly, Goyal et al. (2010) investigated what they call the *minimum target set selection*, which is the problem of finding the minimal set of vertices to activate, such that the influence cascade that is caused activates at least n vertices. Furthermore, Goyal et al. (2010) have shown that the alternative problems are also NP-hard for linear threshold models.

We introduce an approach to estimate a (minimum) target set selection, such that $n = |V|$, i.e. the entire network, for a network $G = (V, E)$ in a threshold model, such that threshold $T_i = 0.5$ for all vertices $v_i \in V$. The approach uses insights derived from the friendship paradox as discovered by Feld (1991). We take a similar approach as Christakis and Fowler (2010) and Garcia-Herranz et al. (2014) did to place sensors in social networks for an early detection of contagious outbreaks. Instead of searching for individuals that are influential on a global perspective, i.e. individuals that exhibit characteristics that appear influential in comparison to all other individuals in the same network, we build on the assumption that global influence can be estimated by finding locally influential individuals. That is, those individuals that are important for diffusion dynamics on a network level are presumably also important for diffusion dynamics on a vertex-centric level. Furthermore, vertices that appear non-influential on a global perspective, might in fact be influential in their local neighborhood, which might be very hard to detect from a global perspective.

In our approach, each vertex $v_i \in V$ has a counter c_i. Initially, $c_i = 0$ for all vertices in V. For each vertex v_i, search for the n most influential vertices v_j in the neighborhood N_i of v_i, such that n is equal to or larger than the number of neighbors required by v_i to be active, i.e. the proportion described by threshold T_i, such that v_i would adopt a. Influence can be defined in various ways and encompass several vertex characteristics. In our case, we measure the influence of a vertex using its degree. That is, the influence of a vertex is equal to its degree, whereby a high degree equals higher influence. Add neighboring vertices v_j to a list L and increment the counter for each vertex v_j that was found to belong to the most influential vertices among the neighborhood of v_i. The result of this operation is a list of vertices L, that contains all vertices that are considered locally influential.

L is not guaranteed to be minimal, but has an important property: Every vertex required to activate in order to persuade any arbitrary vertex $v_i \in V$ is contained in L. Therefore, initially activating each vertex $v_j \in L$ would trigger an influence cascade that would eventually activate the entire network. However, in order to adopt attribute a, some vertices in L only require vertices to be active that are also in L, or that are persuaded by vertices in L. Only a subset of vertices in L is required to cause an influence cascade that activates the entire network.[1] Therefore, we want to decrease the size of L and remove vertices that are not required to achieve the same influence cascade. In the next step, L is sorted by c_i in ascending order, i.e. less important vertices appear earlier in L. For each vertex $v_j \in L$, we check whether all vertices that v_j considers most influential are still in L or will be activated by vertices that are still in L or will be activated by vertices that will be activated by vertices that are still in L, and so on. If this condition is true, vertex v_j can be removed from L, as it is not required to be initially active for the influence cascade.

The result of this process is a list of vertices L^* which is not necessarily minimal, but sufficient to trigger an influence cascade that eventually activates the entire network.

The network shown in Figure 5.2 is the same as in Figure 5.1 and illustrates the

[1] Consider the following example: L contains every vertex that is required to be active in order to activate any arbitrary vertex $v_i \in V$. Thus, if we pick an arbitrary vertex $v_i \in V$, such that $v_i \in L$, we can remove v_i from L and still have a set of vertices that triggers an influence cascade that activates the entire network, because those vertices that are required to activate v_i are still L.

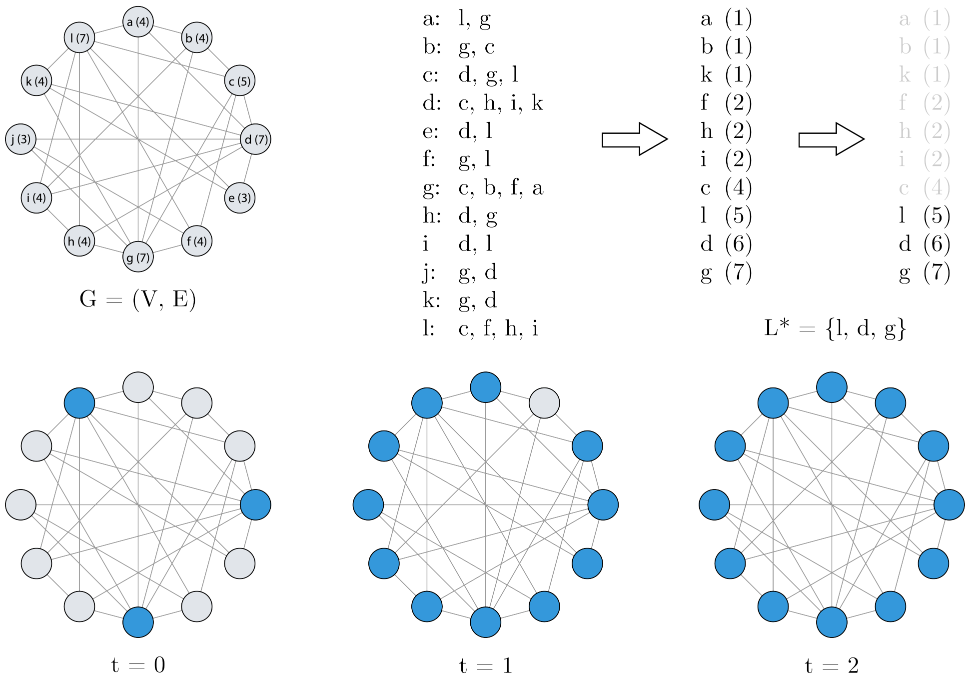


Figure 5.2: Illustration of finding target set selection using influential neighbors

approach. The uncolored network in the upper left shows each vertex with its degree. Next to the network, we can see the n most influential vertices for each vertex. Again, each vertex in the network has threshold 0.5. For instance, for vertex a with degree $k_a = 4$, the two most influential neighbors, i.e. the neighbors with highest degree, are l and g. Next to this, we find the list L, which holds all locally influential neighbors together with their count. For instance, vertex g is among the most influential neighbors of seven other vertices. This is the list L we retrieve, after checking all vertices and finding their most influential neighbors. After this, unnecessary vertices are removed from L. Every vertex in L that will be activated by vertices that are also in L, or vertices that will be activated by vertices that will be activated by vertices in L, and so on, can be removed from L. For instance, we will remove a from L, because a requires vertices l and g to be active, which are still in L. Next, we can remove b, because b requires g and c, which are also in L. Eventually, this gives the list $L^* = \{l, d, g\}$.

The network in the lower left shows the same network with vertices l, d and g initially activated, ready for the diffusion process to start. Before the transition from $t = 0$ to $t = 1$, almost the entire network observed a majority of their neighbors to be active and thus, were also activated. From $t = 1$ to $t = 2$, the remaining inactive vertex could also be persuaded, such that the entire network is now active.

5.3 Baseline Evaluation

The example in Figure 5.2 has shown that using our approach, we were able to find a good set of vertices to initially target, to cause an influence cascade, that eventually activates the entire network. Though, the set we retrieved with our approach is not minimal. Removing one of the three vertices l, d or g, and only target two of them, e.g. l and g, will cause a similar influence cascade with the same effect on the network. Nevertheless, searching for locally influential neighbors, as compared to traditional approaches, has given us a good result with comparably small effort. Therefore, we want to compare the performance of our method to the performance of the baseline heuristic to target high-degree vertices. That is, we want to compare the number of target vertices that we retrieve using our approach, to the number of target vertices we retrieve by searching for a set of highest-degree vertices that, if initially activated, cause a similar influence cascade.

5.3.1 Networks

Milgram (1967) and Travers and Milgram (1969) described the famous experiment in which packets that were passed from person to person were able to reach a designated target person in only around six steps. With this, Milgram (1967) and Travers and Milgram (1969) were the first to demonstrate the small-world effect, the fact that individuals in most networks appear to be only a few steps apart. Various networks were since shown to display the property that the average distance between two vertices is around the famously known six degrees (Nunes Amaral, Scala, Barthelemy, & Stanley, 2000; Newman, 2003). Kumar et al. (2006) have shown that the Flickr and Yahoo! 360 online social networks display small-world properties. Ugander et al. (2011) have shown that the Facebook social graph not only also displays small-world properties, but with an average distance between pairs of users of 4.7, their world is even smaller.

Watts and Strogatz (1998) proposed a model to generate networks with small-world properties, i.e. small average distances between pairs of vertices. In this model, vertices are initially arranged in a ring and connected to their k closest neighbors. Then, depending on a probability β, the edges of a vertex are randomly rewired to another vertex chosen uniformly at random. Both parameters, k and β are user-specified and can be set independently. Degrees in networks generated with this model are normally distributed.

As de Solla Price (1965) discovered, the degree distribution of a network of citations that he studied follows a power-law. Price (1976) introduced a theory that explains the power-law distribution that he discovered, which he called cumulative advantage - the effect that "the rich get richer". The idea is that authors who published many papers are more likely to publish additional papers, compared to authors that have published only few papers. Also, authors that are often cited, and thus popular in their field of research, are likely to be cited again, whereas authors that are unknown are unlikely to be cited. Barabási and Albert (1999) re-introduced the idea of cumulative advantage and reported a collaboration network, a computer network and a power grid network to display the same property. Reka et al. (1999), Faloutsos et al. (1999), Adamic and Huberman (2000) and Broder et al. (2000) confirm in their experiments that the degree distribution of computer networks appears to follow a power-law. Kumar et al. (2006) have shown that the degree distributions of the online social networks Flickr and Yahoo! 360 both follow a power-law.

Barabási et al. (2000) introduced a model to generate networks that have a degree distribution that follows a power-law. In their model, whenever a vertex is added to the network, the probability that the vertex will be connected to some other vertex depends on the degree of that vertex, which leads to a random, scale-free network. Dorogovtsev and Mendes (2002) proposed an alternative model to generate networks that have a degree distribution that follows a power-law. In their model, a newly added vertex connects to both ends of an edge, chosen uniformly at random. As high degree vertices share many edges in the network, this mechanism leads to a scale-free network.

We used the three models we described to generate networks to test the performance of our approach, as these three models provide networks that exhibit specific characteristics that we would expect in real-world networks. We will refer to a network generated according to the model of Barabási et al. (2000) as *BA-network*, likewise, we will refer to a network generated as proposed by Dorogovtsev and Mendes (2002) as *DM-network* and to a network generated with the model of Watts and Strogatz (1998) as *WS-network*. We generated numerous BA- and DM-networks, ranging from networks with a thousand vertices to networks with a million vertices, as well as numerous WS-networks for different values of k and β, varying in their size. That is, we created WS-networks, ranging from ten vertices to hundreds of thousands of vertices, each for both, a fixed value for $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and a fixed value for k, defined as a fraction of the number of vertices, i.e. $k = \frac{n}{10}$, with $k \geq 1$.

5.3.2 Experimental Setup

For every network that we created, we estimated a target set selection using two approaches. First, we calculated a set of target vertices using our neighbor approach as described in Section 5.2. Then, we calculated a set of target vertices using a highest-degree approach. That is, we searched for the minimal number of vertices, ordered by their degree, that have to be initially activated, in order to trigger a similar influence cascade. We searched for sets of highest-degree vertices by activating a number of highest-degree vertices and checking whether the influence cascade that is triggered activates the entire network. We repeated this process until we found a minimal set of highest-degree vertices, such that, only if all of these vertices are activated, the influence cascade that is triggered activates the entire network. After this process we analysed the structural properties of the networks we have created. We calculated various metrics such as clustering coefficients, degree assortativity coefficients, and average degrees, which we used to further analyse the relationship between the network structure and the performance of our neighbor approach.

Finding target sets of highest-degree vertices, as well as analysing the networks for their structural properties, are computationally intensive processes. Thus, we could not evaluate our approach on all networks that we generated. We used 109 BA-networks, 105 DM-networks and 1602 WS-networks.

5.3.3 Results

We evaluated the efficiency of the two approaches by comparing the sizes of the target sets we retrieve, whereby smaller vertex sets are desired. We present the sizes of the sets retrieved from applying both approaches as fractions of the total number of vertices in the networks. Additionally, we compare the performance of our neighbor approach as a fraction of the sizes of the two target vertex sets, such that values below 1 indicate that the neighbor approach finds a smaller vertex set. We will refer to this metric as *performance measure*. We present our results separately for each of the three network types.

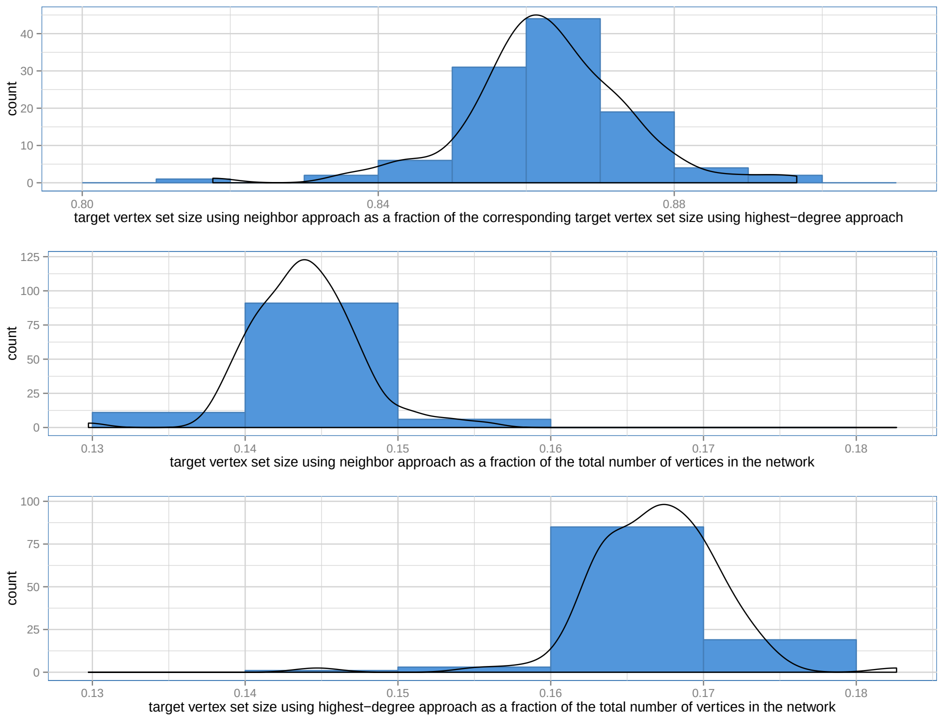


Figure 5.3: Distribution of relative target vertex set sizes for BA-networks (Barabási et al., 2000)

BA-Networks Figure 5.3 shows the results of the two approaches applied to the BA-networks. In the first diagram, we can see the distribution of the sizes of the target vertex sets retrieved with our neighbor approach as a fraction of the size of the corresponding target vertex set retrieved with the highest-degree approach. It shows that the neighbor approach consistently out-performs the highest-degree approach in all BA-networks.

The mean target vertex set size retrieved with our neighbor approach is 0.8627 of the size of the corresponding target vertex set size retrieved with the highest-degree approach with variance 0.00012. The middle diagram shows the distribution of the relative target vertex set sizes from the neighbor approach, expressed as a fraction of the total number of vertices in the network. The mean relative target vertex set size using the neighbor approach in BA-networks measures 0.1439. The minimum and maximum relative target vertex set sizes are 0.1297 and 0.1557 respectively. The bottom diagram shows the distribution of the relative sizes of the target vertex sets using the highest-degree approach. The mean relative target vertex set size using the highest-degree approach in BA-networks is 0.1668. The minimum and maximum relative target vertex set sizes are 0.1447 and 0.1826 respectively.

Further analysis of the structural parameters and the results of BA-networks reveals that the structural properties, and hence the results we retrieve, remain fairly stable. Networks that are generated using the Barabási et al. (2000) model, result in networks with almost identical structural parameters regardless of their size.

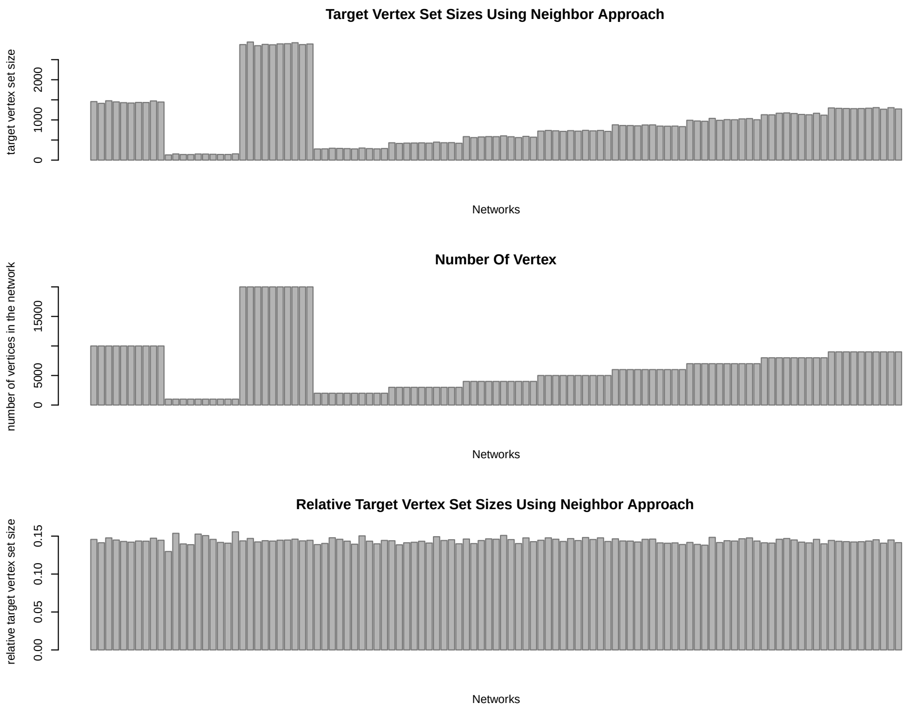


Figure 5.4: Actual and relative target vertex set sizes for neighbor approach in BA-networks (Barabási et al., 2000)

Figure 5.4 shows the actual target vertex set sizes using the neighbor approach in the top diagram, the sizes of the networks, in the middle diagram, and the relative target vertex set sizes as a fraction of the total number of vertices in the bottom diagram. Each grey bar represents a single network. The illustration demonstrates that the results we retrieve are very similar. We think this is because the structural parameters of all BA-networks that we generated are almost identical.

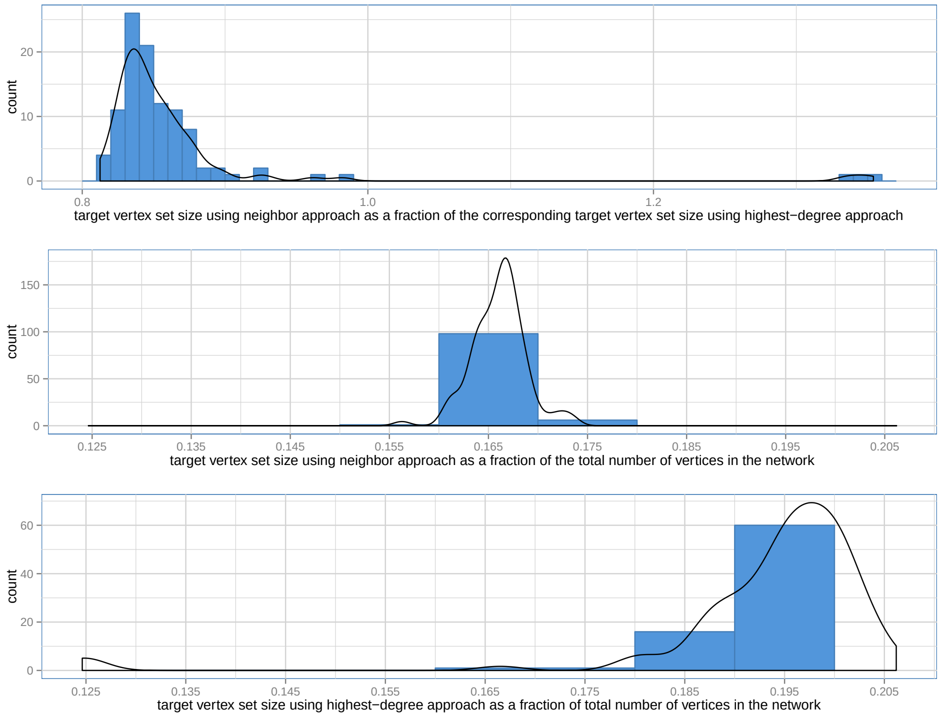


Figure 5.5: Distribution of relative target vertex set sizes for DM-networks (Dorogovtsev & Mendes, 2002)

DM-Networks Figure 5.5 shows the results of the two approaches applied to the DM-networks. The upper diagram shows the distribution of the sizes of the target vertex sets retrieved with our neighbor approach expressed as fractions of the size of the corresponding target vertex set retrieved with the highest-degree approach. It shows that the neighbor approach predominantly out-performs the highest degree approach. The mean target vertex set size from the neighbor approach measures 0.8444 of the size of the corresponding highest-degree target vertex set with variance 0.00756. The minimum and maximum values are 0.8125 and 1.3540 respectively. The middle diagram shows the distribution of the target vertex set sizes from the neighbor approach as fractions of the total number of vertices in the network. On average, the target vertex sets retrieved with the neighbor approach contain 0.1661 of the total number of vertices in the network with variance 7.229e−6. The minimum and maximum relative target vertex set sizes are 0.1563 and 0.1735. The bottom diagram shows the distribution of the relative target vertex set sizes using the highest-degree approach. The mean relative target vertex set size is 0.1933 with variance 0.00018. The minimum and maximum relative target vertex set sizes are 0.1246 and 0.2062 respectively.

Further analysis of the structural parameters and the results of DM-networks shows a similar consistency in network structure and target vertex set sizes as in BA-networks.

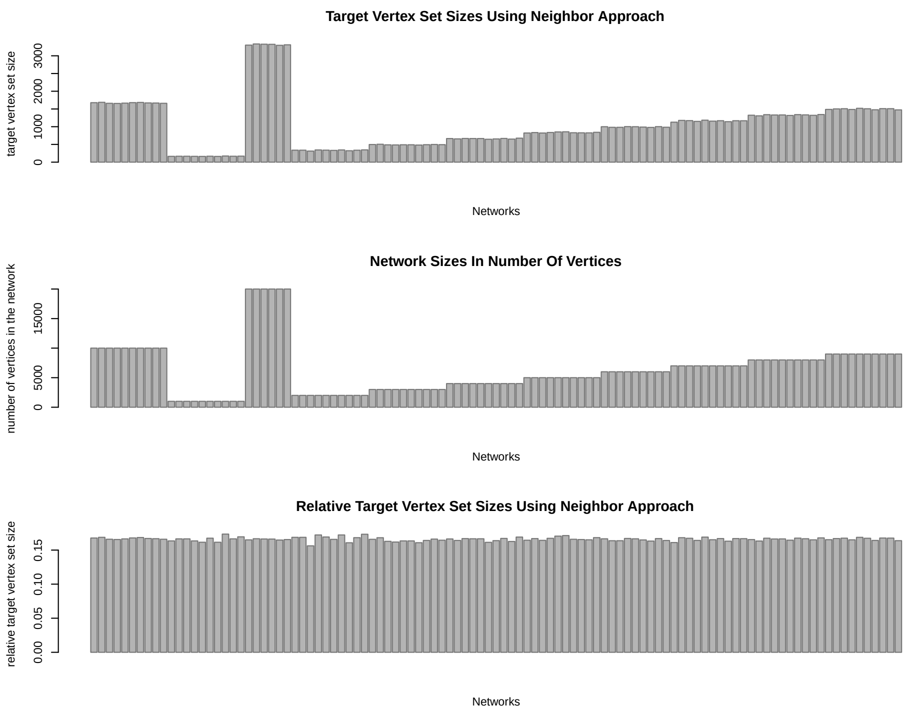


Figure 5.6: Actual and relative target vertex set sizes for neighbor approach in DM-networks (Dorogovtsev & Mendes, 2002)

Figure 5.6 shows the actual target vertex set sizes using the neighbor approach in the top diagram, the sizes of the networks in the middle diagram, and the target vertex set sizes as a fraction of the total number of vertices in the bottom diagram. Each grey bar represents a single network. The diagrams show that, similar to what could be observed in Figure 5.4, increasing the network size does not lead to structural changes and hence, the target vertex set sizes we retrieve with the neighbor approach remain fairly stable. Although there were a few single cases in which the highest-degree approach yielded a better target vertex set, the vast majority of the result sets retrieved with the neighbor approach were smaller.

WS-Networks Figure 5.7 shows the results of the two approaches applied to the WS-networks for all values of k and β. The top diagram shows the distribution of the sizes of the target vertex sets retrieved with the neighbor approach expressed as a fraction of the size of the corresponding target vertex set from the highest-degree approach. The distribution shows that the highest-degree approach significantly outperforms the neighbor approach with the neighbor approach returning target vertex sets of up to 24 times the size of the corresponding highest-degree target vertex set. The mean

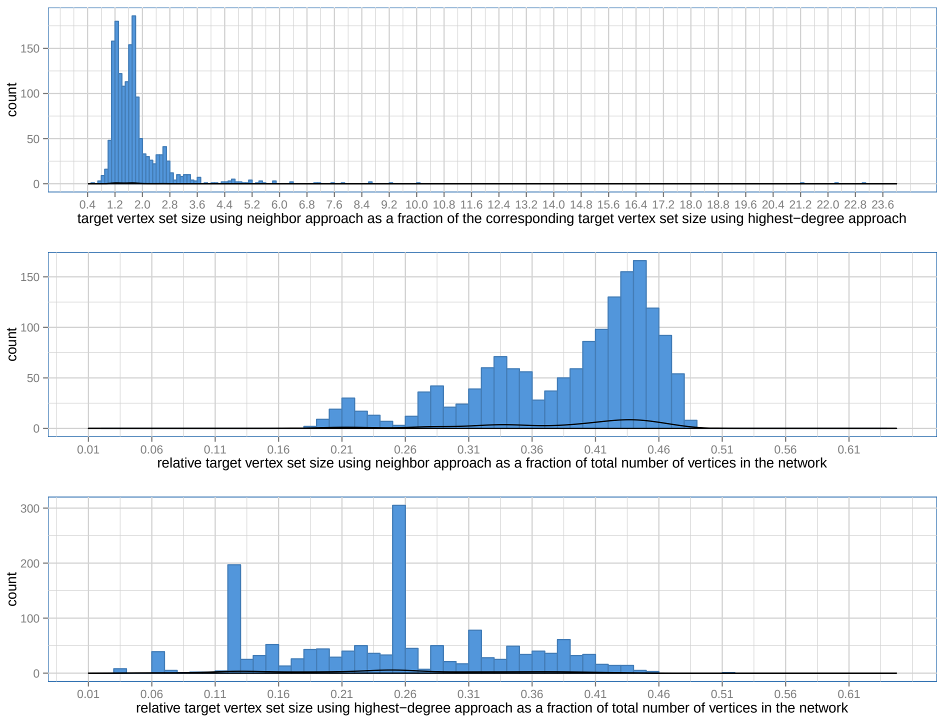


Figure 5.7: Distribution of relative target vertex set sizes for WS-networks (Watts & Strogatz, 1998)

performance measure is 1.8610, with variance 2.11853. The minimum and maximum values are 0.4247 and 24.0 respectively. The middle diagram shows the distribution of the relative target vertex set sizes expressed as fractions of the total number of vertices in the network retrieved with the neighbor approach. The mean target vertex set size is 0.3863 with variance 0.00496. That is, the highest-degree approach, on average, returns results that are approximately half the size of the target vertex sets retrieved with the neighbor approach. The minimum and maximum values are 0.1800 and 0.4817 respectively. The bottom distribution shows the relative target vertex set sizes retrieved with the highest-degree approach expressed as fractions of the size of the network. The mean relative target vertex set size is 0.2475 with variance 0.00905. The minimum and maximum values are 0.0100 and 0.6475.

Further analysis of the structural parameters and results of WS-networks reveals some very interesting details. The highest performance measures of 7, 10 or 20 and higher, which can be seen in the top diagram in Figure 5.7, come from small networks in which vertices have a very low average degree. The upper diagram in Figure 5.8 shows the average degree on the y-axis along with the performance measure on the x-axis. The lower diagram in Figure 5.8 shows the number of edges in a network on the y-axis and the performance measure on the x-axis. Opposed to BA- and DM-networks, WS-networks are not scale-free. Vertex degrees in WS-networks are normally distributed and the majority of the vertices in a WS-network have degree equal to the average degree. The number of vertices that are required for an arbitrary vertex v_i in a network to be initially active in order to activate vertex v_i equals half its degree. Hence, the average number of vertices required to activate an arbitrary vertex v_i in a WS-network is half the average degree. In WS-networks, the majority of vertices have degree equal to the average degree and thus, the average neighbor degree in WS-networks is approximately equal to the average degree. In very small WS-networks with very low average degree, minimal target vertex sets only require a very small subset of all vertices, which the neighbor approach is not able to capture. This explains why in such network configurations the highest-degree approach is able to find target vertex sets that are significantly smaller than those retrieved with the neighbor approach.

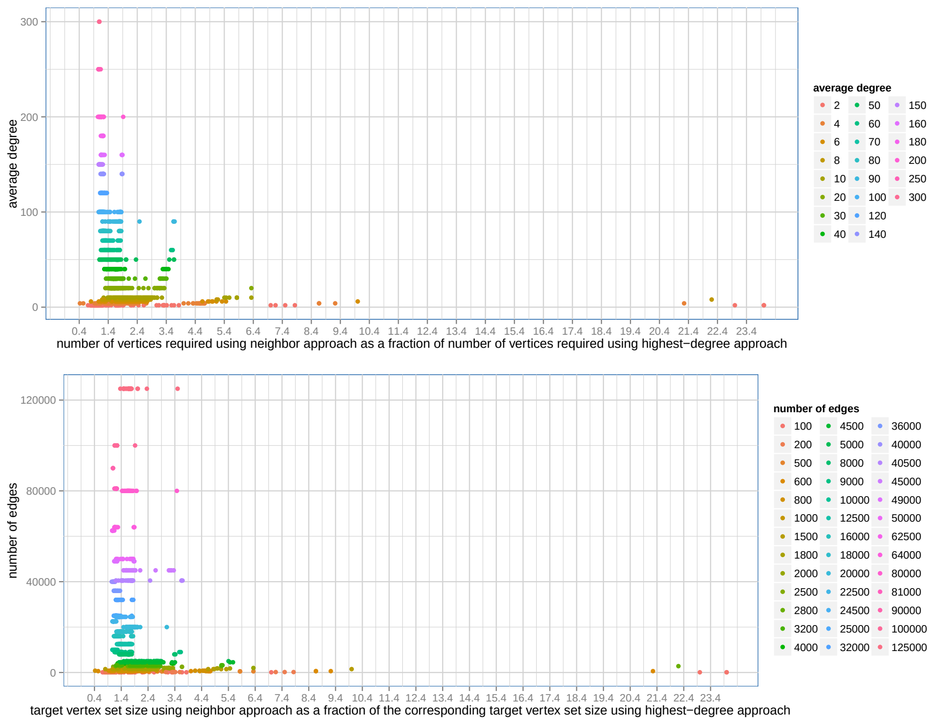


Figure 5.8: Comparison between performance of the neighbor approach and different average degrees and network sizes for WS-networks (Watts & Strogatz, 1998)

Akin to what could be observed in BA- and DM-networks, WS-networks generated with fixed parameters k and β tend to have similar structures. Figure 5.9 shows the results of our experiments separately for different combinations of k and β. Target vertex set sizes using the neighbor approach for fixed values for k and β exhibit a similar stability as we could observe in BA- and DM-networks, whereas target vertex set sizes retrieved with the highest-degree approach display an unaccountable variation. That is, the highest-degree approach finds different target vertex sets for similar networks. We think the variation that is observed can be explained with how the highest-degree approach compares vertices and with the degree distribution of WS-networks. Vertex degrees in WS-networks are normally distributed and the majority of vertices have degree and average neighbor degree equal to the average vertex degree in the network. The highest-degree approach does not compare vertices other than with their degree, due to which there is a high number of vertices that could potentially be selected, which leads to many different possible target vertex sets with substantially different results.

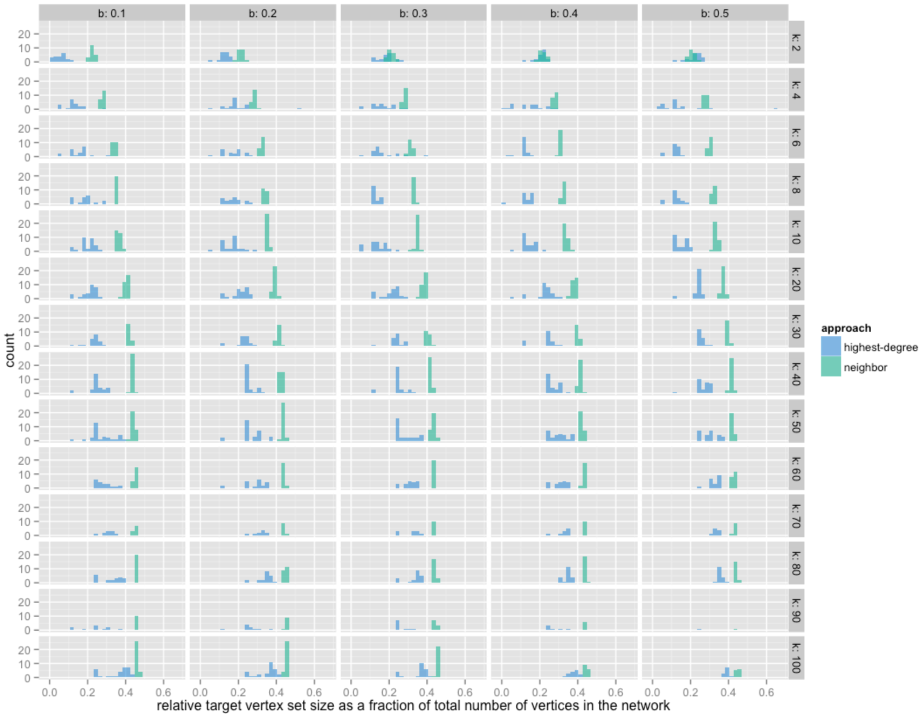


Figure 5.9: Relative target vertex set sizes for both approaches for different values of k and β in WS-networks (Watts & Strogatz, 1998)

5.3.4 Conclusions

We have introduced an approach to estimate target vertex sets for networks in a threshold model with $T_i = 0.5$ for all vertices $v_i \in V$, such that the influence cascade that is triggered by initially activating all vertices in the target vertex set eventually activates the entire network. The approach was shown to consistently out-perform a highest-degree approach in scale-free networks that exhibit structures as described by Barabási et al. (2000) and Dorogovtsev and Mendes (2002). In small-world networks as described by Watts and Strogatz (1998), the neighbor approach returns target vertex sets that, on average, contain twice as many vertices as target vertex sets retrieved with the highest-degree approach. We think this is because in WS-networks, vertices with high degrees are naturally more important than in BA- and DM-networks. Additionally, as the majority of vertices in WS-networks have the same degree, targeting high-degree vertices is reasonably a suitable approach, which the neighbor approach is unable to capture. The neighbor approach is therefore not generally a better approach to find target vertex sets for networks with any arbitrary network structure.

Nevertheless, the neighbor approach was shown to return target vertex sets with consistent sizes in similar network configurations for all three network models. While the target vertex sets retrieved with the neighbor approach are larger in WS-networks, the neighbor approach may be computationally more efficient and it is effective in causing an influence cascade that eventually activates the entire network. We conclude that the neighbor approach has proven itself to be a good candidate to estimate target vertex sets that cause influence cascades that activate the entire network.

6

Dynamic and Complex Diffusion Model

In this chapter, we introduce a new dynamic diffusion model that incorporates assumptions that we make on how individuals in the real world influence each other. We apply our diffusion model on three different network models and present the results of the experiment.

6.1 Preliminaries and Definitions

Granovetter (1978) introduced the threshold model to describe how social behavior spreads within a group of individuals. In his model, individuals observe a social behavior among their neighbors and depending on a simple threshold rule, individuals might adopt the social behavior that they observe. In the same context, Goldenberg et al. (2001b, 2001a) studied the effects of word-of-mouth communications in networks of interacting individuals in a cellular automata model. In a cellular automata model, individuals that exhibit the social behavior persuade their neighbors according to a predefined probability. If an individual was persuaded by one of its neighbors, it too tries to persuade its neighbors with a predefined probability. If an individual fails to persuade its neighbors, it does not retry. Eventually, either all individuals will be persuaded or there will be no individuals left that try to persuade their neighbors. These are the two models Kempe et al. (2015) and Goyal et al. (2010) used in their studies of approximation algorithms for influence maximization.

We introduce a new dynamic diffusion model that incorporates assumptions that we make on how social behavior spreads in the real world within a network $G = (V, E)$ of interacting individuals represented by vertices $v_i \in V$. Our model treats the decision between two distinct and mutually exclusive alternatives as the formation of a preference for either of the two. That is, in our model, individuals do not choose to exhibit a social behavior or attribute a in a binary fashion, but express the extent to which they agree or disagree with it, using an implicit and continuous attribute a_i. That is, each individual v_i has an attribute $a_i \in [0, 1]$ which indicates whether the individual agrees or disagrees with attribute a, such that $a_i = 0$ implies that the individual is entirely against what is represented by a and $a_i = 1$ implies that the individual is entirely in favor of or exhibits a in its fullest form. All values for a_i between 0 and 1 indicate to what extent an individual agrees or disagrees with a. Individuals that are undecided or neutral and do not have a preference for or against a exhibit attribute $a_i = 0.5$. Individuals do not influence each other following a threshold rule or probability, but rather examine a selected set of properties for each of their neighbors and according to these properties, decide to what extent they let themselves be influenced by the social behavior they observe. That is, an individual v_i examines a selected set of properties and estimates a weight w_{ij} for each of its neighbors $v_j \in N_i$, which indicates to what extent v_i is influenced by v_j. Also, individuals may be resistant to external influence and only allow a certain proportion of external influence to reach them. This *influence resistance* is expressed with a parameter $r_i \in [0, 1]$ for each vertex $v_i \in V$, such that $r_i = 1$ indicates that v_i is entirely resistant against all external influences and will never change its attribute. Similar to the threshold model, an individual is influenced by all its neighbors in our model. However, unless an individual v_i is entirely resistant to external influence, i.e. $r_i = 1$, it always changes its attribute if it is influenced by its neighbors. But rather than switching between binary values, an individual incorporates the influence it receives into its own attribute, conditioned upon its resistance and the properties it examined. Furthermore, the influence that individuals adopt from their neighbors decreases linearly with the number of diffusion steps t_n. That is, we make the assumption that the strength of influence deteriorates over time, as individuals get used to hear about a certain topic and give it less importance.

6.2 Parameters and Diffusion Process

Let us introduce and formally define the selected set of properties an individual examines for each of its neighbors, in order to estimate a weight w_{ij} that indicates, to what extent individual v_i is influenced by a neighbor v_j.

McPherson et al. (2001) suggest that individuals who are connected tend to be similar in various characteristics. Furthermore, McPherson et al. (2001) propose that this is because individuals tend to interact with others that are similar. We use this to make the assumption that the influence from interactions between individuals that are similar tends to be higher than from interactions with individuals that do not share any similarities. We characterize the similarity between two neighboring vertices v_i and v_j and use it to adjust the weight of the edge between v_i and v_j, i.e. the degree of influence the vertices have on each other. We say that two vertices v_i and v_j are similar to each other, if they have similar neighbors. That is, the more mutual neighbors two vertices have, the closer these two vertices are, i.e. the higher is the influence they have on each other.

Let N_i be the neighborhood of vertex v_i, such that v_i shares an edge with all vertices $v_j \in N_i$. Then $N_i \cap N_j$, the intersect of the neighborhood of vertices v_i and v_j, is the set of vertices that share an edge with both vertices, v_i and v_j. We define $H(v_i, v_j) \in [0, 1]$ as the similarity between two vertices v_i and v_j, such that $H(v_i, v_j)$ is the proportion of neighbors that v_i and v_j share, including the vertices v_i and v_j themselves.

$$H(v_i, v_j) = \frac{2*(|N_i \cap N_j|+1)}{|N_i|+|N_j|} \tag{6.1}$$

Additional to $H(v_i, v_j)$, we say that two vertices v_i and v_j are similar, if the attributes a_i and a_j that they exhibit are also similar. Furthermore, we make the assumption that individuals that have opposing opinions are less engaging in their interactions and that the influence an individual experiences from an individual that promotes an opposing opinion is less likely to affect the individual. We define $O(v_i, v_j) \in [0, 1]$ as the compliance of the attribute a_i to the attribute a_j of neighboring vertices v_i and v_j, such that $O(v_i, v_j)$ indicates to what extent vertices v_i and v_j find agreement in their opinions.

$$O(v_i, v_j) = 1 - |a_i - a_j| \tag{6.2}$$

For our third parameter, we make the assumption that individuals are more receptive to influence from popular individuals. We define $P(v_i, v_j) \in [0, 1]$ as the popularity bonus a vertex v_j earns on the influence on vertex v_i due to its popularity, whereby the bonus is dampened for highly popular individuals.

$$P(v_i, v_j) = 1 - e^{-log(k_j)} \tag{6.3}$$

Note that $P(v_i, v_j) = 0$ for vertices v_j with degree $k_j = 1$. We think that this accurately captures the behavior in configurations in which a single charismatic individual leads the opinion formation of a community, i.e. in network configurations known as hubs or stars.

The weights w_{ij} that indicate the extent to which vertex v_i is influenced by vertex v_j is the sum of the three parameters.

$$w_{ij} = D(v_i, v_j) + O(v_i, v_j) + P(v_i, v_j) \tag{6.4}$$

The individual weights are then normalized with the weights that v_i appoints to its other neighbors, such that the sum of all weights w_{ij} for all vertices $v_j \in N_i$ equals one.

$$w_{ij}^{'} = \frac{w_{ij}}{\sum_{j=1}^{k_i} w_{ij}} \tag{6.5}$$

Initially and whenever a vertex v_i changes its attribute, all of its neighbors v_j observe the change and re-evaluate the weight w_{ij}. A vertex v_i calculates the *weighted average attribute* $\overline{a_{i,w}}$ among its neighbors, such that $\overline{a_{i,w}}$ denotes the average attribute v_i experiences when v_i applies the individual weights w_{ij} on the attributes a_j it observes from its neighbors v_j.

$$\overline{a_{i,w}} = 0.5 + \sum_{j=1}^{k_i} (a_j - 0.5) * w_{ij}^{'} \tag{6.6}$$

Vertex v_i then estimates the difference between the weighted average attribute $\overline{a_{i,w}}$ and its own attribute a_i and decides to adopt a proportion of the difference depending on its resistance r_i and the diffusion step t_n.

$$a_i^{'} = a_i - \frac{((a_i - \overline{a_{i,w}}) * (1 - r_i))}{t_n} \tag{6.7}$$

Note that as the number of time steps t_n increases, the proportion of the weighted average attribute $\overline{a_{i,w}}$ that is adopted decreases. Furthermore, as vertices update their attribute whenever at least one of their neighbors changes its attribute and as the proportion of the weighted average attribute that is adopted decreases over time, the diffusion process is not guaranteed to automatically halt after a finite number of steps.

6.3 Experiments

We used the BA-, DM- and WS-networks that we generated to simulate the spread of an attribute a within our dynamic diffusion model. In order to have comparable starting conditions for all three network models, we initially activated and set attributes $a_i = 1$ for one percent of the top most influential neighbors of the top most influential vertices in the network. Influence was measured by vertex degree, such that vertices with a high degree were considered highly influential. All other vertices $v_i \in V$ were initially neutral and had attribute $a_i = 0.5$. Furthermore, we set $r_i = 0$ for all vertices $v_i \in V$, except for those vertices that were initially activated, these vertices had $r_i = 1$. We configured the diffusion process to halt for a single vertex after $10,000$ steps. That is, the diffusion process on the entire network stopped after all vertices have participated in exactly $10,000$ steps. For each network, we calculate S_{a_0}, the sum of all attributes a_i of all vertices $v_i \in V$ before the diffusion process starts and S_{a_1}, the sum of all attributes a_i of all vertices $v_i \in V$ after the diffusion process has stopped. The difference between S_{a_0} and S_{a_1} shows how far the attribute could spread within the network. We call the fraction $\frac{S_{a_1}}{S_{a_0}}$ the *diffusion coverage*.

Figure 6.1 shows the distribution of the diffusion coverage for BA-networks in the top diagram, for DM-networks in the middle diagram and for WS-networks in the bottom diagram. It shows that the diffusion coverage is highest in DM-networks and, on average, lowest in WS-networks. It also shows that WS-networks display the highest variation in diffusion coverage. We think that this is because of the degree distribution in WS-networks. In WS-networks, vertex degrees are normally distributed and the majority of vertices have the same degree, i.e. the average degree, and hence, similar to what could be observed in Chapter 5, randomly selecting a subset of highest-degree vertices can lead to a high number of different possible target sets with substantially different results for the diffusion process in networks with similar structures.

Centola (2010) and Centola and Baronchelli (2015) have studied the effects of network structure on the spread of behavioral adoption in networks of interacting individuals. In their experiments, they could show that in highly clustered networks social behaviors spread more quickly. Initially, our hypothesis was that we might experience a similarly quicker attribute diffusion in highly clustered BA-, DM- and WS-networks. The clustering coefficients for all BA-networks we generated are consistently equal to 0. The mean clustering coefficients for all DM-networks we generated is 0.7387 with variance 3.95607e−6, i.e. very stable. The mean clustering coefficients for WS-networks for which we also had enough computational resources to both, calculate clustering coefficients and simulate our model, is 0.13720 with variance 0.03233. Calculating the correlation between the diffusion coverage and the clustering coefficient for the three network models separately shows only a very weak correlation between the two metrics. Nevertheless, we

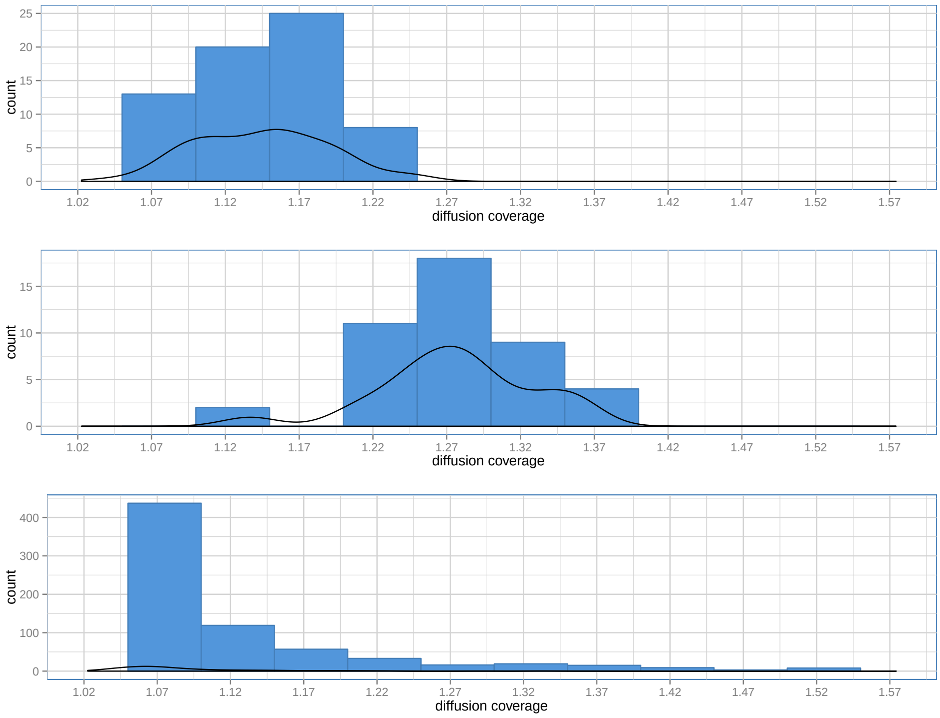


Figure 6.1: Distribution of diffusion coverage for BA-, DM- and WS-networks (Barabási et al., 2000; Dorogovtsev & Mendes, 2002; Watts & Strogatz, 1998)

found that in DM-networks, which on average have the highest clustering coefficient, the diffusion coverage was on average the highest and in BA- and WS-networks, which on average have very low clustering coefficients, the diffusion coverage was also signficantly lower. Although from this perspective, it appears that we find a similar correlation between attribute diffusion and clustering coefficients, the exact nature of the relationship could not be revealed and the hypothesis could neither be entirely invalidated nor proven.

7

Limitations

The experiments we conducted to evaluate the efficiency of our approach to estimate target vertex sets that we discussed in Chapter 5 and the experiments we conducted on the diffusion process within the diffusion model that we introduced in Chapter 6 both use networks that were generated using models to create scale-free networks as described by Barabási et al. (2000) and Dorogovtsev and Mendes (2002) and small-world networks as described by Watts and Strogatz (1998). Various real-world network were shown to display structures similar to those explained by the three models (Newman, 2003; Nunes Amaral et al., 2000; Barabâsi et al., 2002; Kumar et al., 2006; Ugander et al., 2011). Hence, we expect the networks we generated according to these models to exhibit network structures that can also be found in real-world networks. Nevertheless, using only artificially generated networks does not allow to make strong claims about the performance in real-world networks. All claims that are made in this thesis concerning the efficiency of the approach discussed in Chapter 5 and the quality of the diffusion model discussed in Chapter 6 refer to networks that exhibit network structures exactly as in Barabási et al. (2000), Dorogovtsev and Mendes (2002) or Watts and Strogatz (1998). The diffusion model we proposed in Chapter 6 is based on assumptions we consider reasonable but could not verify in the scope of this work. Additionally, to provide comparable starting conditions for the diffusion processes in the three different network models, we initially activated the top one percent most influential neighbors of the most influential vertices in each network. This may not accurately reflect realistic starting conditions for a diffusion process. Furthermore, the threshold model according to Granovetter (1978) and the restrictions we set on the threshold for each vertex we used for our approach in Chapter 5 to simulate the spread of an attribute in a network might not be the best approximation of real-world networks.

8

Future Work

In this chapter we share our thoughts on future work regarding our approach to estimate target vertex sets as well as our dynamic diffusion model. We suggest a set of experiments that could be conducted to study the efficiency of our approach and suggest how the efficiency of our algorithm might be improved. Furthermore, we propose a set of experiments that could be conducted to investigate the validity of our dynamic diffusion model and suggest ways to enhance and extend our model.

8.1 Estimating Target Vertex Sets Using Influential Neighbors

In Chapter 5 we introduced a novel approach to estimate target vertex sets that, if initially activated, cause an influence cascade that eventually activates the entire network for networks in a threshold model with threshold $T_i = 0$ for each vertex $v_i \in V$. Our approach has proven to be a good candidate for estimating such target vertex sets. We have shown that the approach consistently returns good results in scale-free networks and fails to provide good results in small-world networks in which vertex degrees are normally distributed. The exact nature of the relationship between the efficiency of our approach and underlying network structures could not be revealed. We suggest that in future work further experiments regarding this relationship could be conducted, such that it becomes clearer why the approach does not perform well in small-world networks, or generally, if and in what network configurations the approach that we propose is able to find valuable estimates for target vertex sets. Particularly, the efficiency of our neighbor approach in real-world networks could be evaluated. Moreover, we suggest to compare the efficiency of our neighbor approach to approximation algorithms as introduced by Kempe et al. (2015) and Goyal et al. (2010). Furthermore, we suggest that the neighbor approach can be enhanced and extended by using additional and or different metrics to select locally influential neighbors. Particularly, experiments with a degree discount heuristic as introduced by Chen et al. (2009) might significantly improve our approach.

8.2 Dynamic Diffusion Model

In Chapter 6 we introduced a new dynamic diffusion model to simulate the spread of a social behavior in a network of interacting individuals. With this model, we tried to map real-world social behavior into a diffusion model. The results of the experiments we conducted neither support nor invalidate a hypothesis as proposed by Centola (2010) and Centola and Baronchelli (2015) that diffusion processes profit from highly clustered network structures. We suggest that in future work further experiments regarding the relationship between the efficiency of a diffusion process and the clustering of a network in our dynamic model could be conducted. Moreover, we suggest to conduct experiments regarding the validity of the model we propose. Particularly, future work could investigate whether a model like ours is able to accurately capture a real-world diffusion process using real-world networks with real-world diffusion information. Furthermore, we suggest that our dynamic diffusion model can be enhanced and extended using additional parameters and or fine-tuning existing parameters.

9

Conclusions

We investigated the majority illusion in social networks, a phenomenon discovered and introduced by Lerman et al. (2015) that tricks individuals in a network into thinking that something is common, when in reality, it is not. In Chapter 3, we discussed the importance of disproportionate experiences caused by the illusion in the context of social contagion. In Chapter 4, we have shown how the majority illusion can be quantified on a network level as well as a vertex-centric level. In Chapter 5 we investigated how a misperception such as the majority illusion can be exploited in order to artificially promote the spread of a social behavior in a network of interacting individuals. We introduced an approach to estimate target vertex sets that uses ideas from the friendship paradox, Feld (1991), Christakis and Fowler (2010) and Garcia-Herranz et al. (2014). We have shown that the approach we propose is able to find good estimates for target vertex sets with comparably small effort. In scale-free networks that exhibit structures as described by Barabási et al. (2000) and Dorogovtsev and Mendes (2002), the approach consistently out-performs a highest-degree approach, in which high-degree vertices are selected as target vertices. In small-world networks that exhibit structures as described by Watts and Strogatz (1998) and in which vertex degrees are normally distributed, the approach that we propose returns target vertex sets which are, on average, twice the size of the target vertex sets retrieved with a highest-degree approach. A relationship between the vertex degree distribution and the efficiency of our approach for a given network is suspected, but could not be shown. We think that the proposal we gave to estimate target vertex sets has introduced an approach that has proven itself a good candidate to estimate valuable target vertex sets with comparably small effort. Nevertheless, the approach requires further investigation with real-world networks as well as other networks that exhibit various other structures, to be able to finally conclude whether it is capable of performing with the same quality as approximation algorithms that were introduced by Kempe et al. (2015), Goyal et al. (2010) and Chen et al. (2009). In Chapter 6 we introduced an alternative dynamic diffusion model to simulate the spread of a social behavior in a network of interacting individuals. Our model incorporates assumptions about human behavior in the real world and considers the time dimension. In our model we were unable to confirm a relationship between the extent and speed at which a social behavior spreads and the degree of clustering as suggested by Centola (2010) and Centola and Baronchelli (2015). But we were also unable to disprove this claim in our diffusion model. With our diffusion model we were able to capture some of the behavior that we could expect from real-world human interaction in the context of social contagion. Nevertheless, the parameters we introduce and the model were not thoroughly evaluated with regards to the quality of the diffusion process with real-world data and thus, the accuracy of the model is unknown. We think that this thesis provides valuable scientific contributions in the field of influence maximization in the context of network phenomena such as the majority illusion.

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science*, *287*(5461), 2115–2115.

Baer, J. S., Stacy, A., & Larimer, M. (1991). Biases in the perception of drinking norms among college students. *Journal of studies on alcohol*, *52*(6), 580–586.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Barabási, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, *281*(1), 69–77.

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

Bearak, J. M. (2014). Casual contraception in casual sex: Life-cycle change in undergraduates' sexual behavior in hookups. *Social Forces*, sou091.

Berkowitz, A. D. (2005). An overview of the social norms approach. *Changing the culture of college drinking: A socially situated health communication campaign*, 193–214.

Bettencourt, L. M., Cintrón-Arias, A., Kaiser, D. I., & Castillo-Chávez, C. (2006). The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, *364*, 513–536.

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

Centola, D. (2010). The spread of behavior in an online social network experiment. *science*, *329*(5996), 1194–1197.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *arXiv preprint arXiv:1502.06910*.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 199–208).

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PloS one*, *5*(9), e12948.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515.

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

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining* (pp. 57–66).

Dorogovtsev, S. N., & Mendes, J. F. (2002). Evolution of networks. *Advances in physics*, *51*(4), 1079–1187.

Eom, Y.-H., & Jo, H.-H. (2014). Generalized friendship paradox in complex networks: The case of scientific collaboration. *arXiv preprint arXiv:1401.1458*.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Acm sigcomm computer communication review* (Vol. 29, pp. 251–262).

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 1464–1477.

Feld, S. L., & Grofman, B. (1977). Variation in class size, the class size paradox, and some consequences for students. *Research in Higher Education*, *6*(3), 215–222.

Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PloS one*, *9*(4), e92413.

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

Goldenberg, J., Libai, B., & Muller, E. (2001b). Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, *2001*, 1.

Goyal, A., Bonchi, F., Lakshmanan, L. V., & Venkatasubramanian, S. (2010). Approximation analysis of influence spread in social networks. *arXiv preprint arXiv:1008.2005*.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 1420–1443.

Granovetter, M. (1984). Small is bountiful: Labor markets and establishment size. *American Sociological Review*, 323–334.

Hemenway, D. (1982). Why your classes are larger than 'average'. *Mathematics Magazine*, *55*(3), 162–164.

Hodas, N. O., Kooti, F., & Lerman, K. (2013). Friendship paradox redux: Your friends are more interesting than you. *arXiv preprint arXiv:1304.3480*.

Kempe, D., Kleinberg, J., & Tardos, É. (2015). Maximizing the spread of influence through a social network. *Theory of Computing*, *11*(4), 105–147.

Kooti, F., Hodas, N. O., & Lerman, K. (2014). Network weirdness: Exploring the origins of network paradoxes. *arXiv preprint arXiv:1403.7242*.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 611–617).

Lerman, K., Yan, X., & Wu, X.-Z. (2015). The majority illusion in social networks. *arXiv preprint arXiv:1506.03022*.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415–444.

Milgram, S. (1967). The small world problem. *Psychology today*, *2*(1), 60–67.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Nunes Amaral, L. A., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of behavior of small-world networks. *arXiv preprint cond-mat/0001458*.

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, *27*(5), 292–306.

Reka, A., Jeong, H., & Barabasi, A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*(6749), 130–131.

Richardson, M., & Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 61–70).

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *science*, *311*(5762), 854–856.

Schelling, T. C. (1973). Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *The Journal of Conflict Resolution*, *17*(3), 381–428.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 425–443.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, *393*(6684), 440–442.

Young, H. P. (2011). The dynamics of social innovation. *Proceedings of the National Academy of Sciences*, *108*(Supplement 4), 21285–21291.

A

Appendix

A.1 Information About Relative Target Vertex Set Sizes

Tables A.1, A.2 and A.3 show the metrics we mentioned in Section 5.3.3 for the relative target vertex set sizes we calculated from applying the neighbor approach and highest-degree approach to the networks we generated.

	Neighbor Approach	Highest Degree Approach	Performance Measure
Minimum	0.1297	0.1447	0.8176
1st Quantile	0.1415	0.1640	0.8572
Median	0.1437	0.1669	0.8626
Mean	0.1439	0.1668	0.8627
3rd Quantile	0.1459	0.1696	0.8688
Maximum	0.1557	0.1826	0.8966

Table A.1: Relative target vertex set size metrics and performance measure metrics for neighbor approach and for highest-degree approach in BA-networks (Barabási et al., 2000)

	Neighbor Approach	Highest Degree Approach	Performance Measure
Minimum	0.1563	0.1246	0.8125
1st Quantile	0.1643	0.1911	0.8333
Median	0.1664	0.1959	0.8444
Mean	0.1661	0.1933	0.8649
3rd Quantile	0.1675	0.1995	0.8617
Maximum	0.1735	0.2062	1.3540

Table A.2: Relative target vertex set size metrics and performance measure for neighbor approach and for highest-degree approach in DM-networks (Dorogovtsev & Mendes, 2002)

	Neighbor Approach	Highest Degree Approach	Performance Measure
Minimum	0.1800	0.0100	0.4247
1st Quantile	0.3389	0.1729	1.2860
Median	0.4100	0.2500	1.6310
Mean	0.3863	0.2475	1.8610
3rd Quantile	0.4400	0.3150	1.9090
Maximum	0.4817	0.6475	24.0000

Table A.3: Relative target vertex set size metrics and performance measure for neighbor approach and for highest-degree approach in WS-networks (Watts & Strogatz, 1998)

A.2 Information About Diffusion Coverages

Table A.4 shows metrics for the diffusion coverages we retrieved with simulating the spread of an attribute in a network in the dynamic diffusion model we introduced in Chapter 6. These are the metrics we presented in Section 6.3.

	BA-Networks	DM-Networks	WS-Networks
Minimum	1.042	1.133	1.023
1st Quantile	1.108	1.248	1.057
Median	1.148	1.275	1.075
Mean	1.145	1.276	1.111
3rd Quantile	1.175	1.307	1.126
Maximum	1.247	1.365	1.574

Table A.4: Diffusion coverage metrics for BA-, DM- and WS-networks (Barabási et al., 2000; Dorogovtsev & Mendes, 2002; Watts & Strogatz, 1998)

List of Figures

List of Tables