Information Diffusion and Social Influence in Online Networks

by

Eytan Bakshy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Lada A. Adamic, Chair
Professor Michael D. Cohen
Professor Mark E. J. Newman
Assistant Professor Eytan Adar

For Jerry Uhl

ACKNOWLEDGEMENTS

I would like to thank my advisor, Lada Adamic, whose creativity and enthusiasm has made my research perpetually rewarding. I am greatly indebted to my committee members, Michael D. Cohen whose guidance and probing questions I can only describe as incredibly wise and valuable, and Eytan Adar, whose open door and healthy dose of cynicism has benefited me greatly this past year.

This work would also not be possible without the help of many others: my mentors outside of Michigan, Spiro Maroulis and Itamar Rosenn; my good friends and colleagues, Brian Karrer and Sean Munson, who have provided me with endless insightful conversations that have shaped my work; Judy Bakshy, Miron Bakshy, Aric Bakshy, and Catherine Le for their love and unwavering support.

I would also like to recognize the lasting impact of my other collaborators over the past four years: Mark Ackerman, Lars Backstrom, Jake Hofman, Jon Kleinberg, Tom Lento, Cameron Marlow, Winter Mason, Matt Simmons, Edwin Teng, Duncan Watts, and Jun Zhang. I also appreciate the helpful and encouraging feedback on my work from Sinan Aral, Dean Eckles, Emily Falk, James Fowler, Matthew Salganik, and Cosma Shalizi.

This thesis is dedicated to the memory of Jerry Uhl, who taught me to fight.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Information Diffusion and Social Influence in Online Networks
by
Eytan Bakshy

Chair: Lada A. Adamic

The explosive growth of online social systems has changed how individuals consume and disseminate information. In this thesis, we conduct large-scale observational and experimental studies that allow us to determine the role that social networks play in information diffusion online, and the factors that mediate this influence. We first examine the adoption of user-created content in a virtual world, and find that social transmission appears to play a prominent role in the adoption of content. Ultimately, we are faced with a critical problem that underlies all contemporary empirical research on social influence: how do we measure whether individuals in a network influence one another, when the basis for their interaction rests upon commonalities that are predictive of their future behavior? We use two coupled experiments to address this question. In our first experiment, we randomize exposure to social signals about friends' information sharing behavior to determine the causal effect of networks on diffusion among 253 million subjects in situ. Our second experiment further tests how social information affects individual sharing decisions when viewing content. Finally, this thesis concludes with a study that examines how individuals allocate attention across their network of contacts, which has implications for influence and information diversity in networks.

CHAPTER I

Introduction

Quantifying social influence is crucial to understanding a range of behavioral phenomena, from the dissemination of information, to the adoption of political opinions, products, and health-related behaviors (*Granovetter*, 1978; *Watts and Dodds*, 2007; *Christakis and Fowler*, 2007). As social interactions move online, exposure to information is increasingly mediated through online social networks. Nearly 75% of Americans who read news online are directed to news articles through email or online social networking sites, and over half also share links to news with their online contacts (*Purcell et al.*, 2010). Moreover, such networks also represent real world social connections. A recent survey found that over 55% of U.S. adults over the age of 18 use Facebook, and among these individuals, 48% of their real-world contacts were also their contacts on Facebook (*Hampton and Rainie*, 2011). Therefore, the systematic study of online diffusion not only sheds light on how Internet technologies alter the face of human communication, but also has the potential to inform basic social science research beyond the digital domain.

Each day hundreds of millions of Internet users consume and propagate information shared by friends. Given the right set of resources, one can capture these users' networks, their sharing behavior, and their contacts' sharing behavior. Assuming all information exchange occurred within these networks, such data would enable us to study the dynamics of social contagion with physics-like precision. Unfortunately, no dataset will ever encompass all sources of information. To make matters worse, one of the most robust findings in the social networks literature is that of homophily: individuals tend to associate with others that are similar to themselves (*McPherson et al.*, 2001). As a result, individuals have common interests and activities, which make them more likely to be tuned into the same, potentially unobservable, information sources.

This situation reflects a grave problem underlying many observational studies of social contagion: individuals' attributes are predictive of their past, present, and future behavior, as well as of the friends they maintain. Therefore, an individual's behavior may be predicted by their friend's behavior, without that friend causally influencing that individual in any way (*Shalizi and Thomas*, 2011). Long lines of research are based on explicitly modeling social contagion as a function of the number of "infected" friends. Other research suggests that strong ties are more influential than weaker ones. When confronted with data that shows that individuals are more likely to engage in the same behavior as their friends, it is tempting to assume that the data corroborates with these theories. However, since homophily implies that those who interact more often, or have overlapping friendship ties, tend to be more similar to one another, such evidence for influence may simply be mere social correlation.

Not all contagion studies have to infer whether one individual influenced another. Many studies (e.g. *Katz and Lazarsfeld* (1955); *Greenberg* (1964); *Brown and Reingen* (1987)) use interviews and surveys to directly ask respondents to identify sources of contagion. Still, even if the source is a friend, the data does not tell us about the relative importance of social ties in the spread of information. For example, though 50% of respondents in Greenberg's 1964 study of news diffusion learned about the Kennedy assassination from interpersonal ties, many of the respondents may have gotten the news at a slightly later point in time from the very same media outlets as their friends. Therefore, a complete understanding of how social networks affect information diffusion not only requires us to identify interpersonal contagion, but also requires a counterfactual understanding of what would happen if certain interactions did not take place.

In this thesis, we take a twofold approach to understanding social influence and information diffusion in online networks. First, we conduct broad data-intensive exploratory studies to examine how individuals interact with one another and share information online. In doing so, we identify similarities and differences between online and offline interactions that can guide further research. While observational studies tend to yield many insights into how individuals behave, they struggle to identify mechanisms that generate this behavior. For that reason, we combine the exploratory approach with more systematic experimental methods which isolate specific causal effects and explain diffusion phenomena.

We begin in Chapter 2 with an exploratory study that analyzes the diffusion of content on Second Life, a virtual world made up entirely of user-created content. We discover regularities that are consistent with research on social contagion and

suggestive of interpersonal influence. Using detailed information about when users adopt content and their social networks over time, we develop a model of adoption to explain the diffusion of content within time-evolving social networks. Our results provide intuition for the mass spread of content in online social systems.

While our model of diffusion based on friends' behavior appears to provide a reasonable description for how individuals behave, the homophily confound makes it impossible to determine how well such a model reflects the underlying mechanisms responsible for content spread. To understand the causal effect of social information on content diffusion in real-world settings, we conduct two very large randomized field experiments. Chapter 3 presents an experiment that randomizes exposure to Web content shared by friends among 253 million subjects on Facebook. We show that individuals exposed to content from their strong ties are more likely to propagate that information, but that weak ties expose individuals to information that they would not have otherwise spread. Using the underlying distribution of known tie strengths, and causal estimates of how much influence is exerted by strong and weak ties, we are able to conclude that weak ties are collectively more influential.

Our experiment shows that networks surface information that would not have otherwise been shared; but to what extent do social signals affect an individual's decision to share content? Chapter 4 presents a second field experiment that isolates the effect of social information on sharing decisions for individuals visiting pages on the Web. We show that while the number of sharing friends and a subject's strength of ties with those friends is predictive of whether or not content they choose to share, much of this correlation exists even when friends' sharing behavior is not shown to subjects.

One friend's ability to influence another is undoubtedly mediated by the amount of attention given to the influencer. In Chapter 5, we examine how individuals allocate their attention across friends. This new measurement, which we call the balance of attention, captures the extent to which individuals focus their attention upon few or many contacts, and has important implications for how information flows in networks.

Chapter 2 was done in collaboration with Brian Karrer and Lada Adamic, and is published as *Bakshy et al.* (2009). Chapter 5 was done in collaboration with Lars Backstrom, Jon Kleinberg, Tom Lento, and Itamar Rosenn, and is published as *Backstrom et al.* (2011). Chapters 3 and 4 are unpublished work done in collaboration with Itamar Rosenn, Cameron Marlow, and Lada Adamic.

CHAPTER II

The Spread of User-Created Content

Social influence determines to a large extent what we adopt and when we adopt it. This is just as true in the digital domain as it is in real life, especially with the proliferation of user generated content that one must first become aware of and then select from. We present an empirical study of user-to-user content transfer occurring in the context of a time-evolving social network in Second Life, an immersive massively multiplayer virtual world. We model social influence based on the change in adoption rate following the actions of one's friends and find that the social network plays a significant role in the adoption of content. Adoption rates quicken as the number of friends adopting increases and this effect varies with the connectivity of a particular user. We further find that sharing among friends occurs more rapidly than sharing among strangers, but that content that diffuses primarily through social influence tends to have a more limited audience. Finally, we examine the role of individuals, finding that some play a more active role in distributing content than others, but that these influencers are distinct from the early adopters.

2.1 Introduction

In the digital age, the creation and distribution of digital goods has been democratized. On YouTube, users view millions of videos created by millions of users, on Flickr users upload their own photos and view others', and news are reported on, consumed, and commented on by a distributed network of bloggers and media sources. Perhaps the purest example of a market for user-generated content is that of the virtual world Second Life. The vast majority of the content, in fact pretty much all of

This chapter is published as *Social Influence and the Diffusion of User-Created Content* in the EC 2009 Proceedings of the 10th ACM Conference on Electronic Commerce (*Bakshy et al.*, 2009).

the virtual world itself, from buildings to objects to fashion, is created, distributed, and consumed by the users themselves.

The unique property of studying social contagion in Second Life is that one can observe not just adoption in the context of an explicit social network, but also trace direct transfers of user-contributed content owned by users, which we will refer to as *assets*. In Second Life, you can search for interesting places to visit on your own, or a friend or business can give you a landmark – a bookmark that allows you to teleport directly to a location. If upon arriving, you would like your avatar to dance, wave, or make a certain sound, you need to retrieve that *gesture* from your inventory of assets. That gesture may have been given to you by a friend, or you may have purchased it from a store. Such transfer of assets and information presents a unique opportunity to compare diffusion via word-of-mouth to adoption resulting from broadcasts. Depending on the intellectual property rules attached to each object, some assets can be freely copied and shared; one Second Life user can pass on a gesture, hairstyle, or article of clothing to another.

The chapter proceeds as follows. After reviewing related work and motivating our approach in Section 2.1.1, in Section 2.1.2 we describe the Second Life data set and the characteristics of information diffusion among Second Life users. In Section 2.1.3 we quantify the properties of asset transfer cascades and their relationship to the social network. We find that assets that are passed from friend to friend tend to produce deeper cascades, but the overall popularity of the asset is lower. We demonstrate that this insight can be used to predict how many additional individuals will adopt an asset over a period of time. Section 2.2 models the rate of adoption which we find to strongly depend upon the number of adopting friends a user has at any given time. As might be expected, when users have no previously adopting friends, their rate of adoption is related to the popularity of the asset in the population overall. However, once a friend has adopted, the adoption rate increases significantly, especially for less popular, niche assets. In Section 2.3 we identify two kinds of individuals, influencers who directly influence many of their friends to adopt, and early adopters. We find that early adopters are more likely to adopt without having to first observe their friends, but that they are not necessarily influential in subsequent adoptions. Section 2.4 concludes and discusses future directions.

2.1.1 Background & motivation

The context in which our study occurs, the virtual world Second Life (*Ondrejka*, 2004b) has been studied for aspects of its economy (*Ondrejka*, 2004a; *Castronova*,

2008) and social conventions (*Yee et al.*, 2007; *Friedman et al.*, 2007). Our study provides a complementary perspective on how individuals influence one another, while contributing to a larger body of work in the measurement of large-scale social phenomena relating to the dynamics of content consumption in online communities.

In the marketing science literature there is a wealth of macro-scale studies of new product diffusion (*Mahajan et al.*, 1990). For example, the Bass model is a differential equation model that predicts adoption based on relative populations of "innovators" that are not influenced by the decisions of others and "imitators" whose adoption depends of the total number of adoptions in the system (*Bass*, 1969). Extensions to these models have traditionally not taken into account social structure, nor the individual decision making processes of the adopters. On the other hand, micro-level studies, such as (*Chatterjee and Eliashberg*, 1990), do model factors that influence the adoption of a product, but have only been studied in the context of small laboratory experiments.

Although the theory of information diffusion in social networks was developed decades ago (*Rogers*, 1995), social contagion has only recently been measurable on a large scale through the digital traces that modern communication leaves behind. Social contagion can be distinguished from viral, unintentional sharing of e.g. human (*Pastor-Satorras and Vespignani*, 2001; *Newman*, 2002) or electronic (*Newman et al.*, 2002) malaises over networks. One feature of social contagion is that there may be thresholds to infection, with many individuals waiting for several of their friends to adopt before taking the plunge themselves (*Centola and Macy*, 2007). Unlike disease spread, this diffusion typically has the property that an individual decides whether to accept the contagious object.

The availability of large scale social network data has lead to a number of studies quantifying various aspects of social contagion. Of interest in all these studies is how one might maximize the spread of influence through a social network by selecting a subset of influential individuals to initially infect with an idea or product (*Kempe et al.*, 2003). On the other hand, one may simply wish to find out early what assets are "hot" by monitoring a subset of individuals that are likely early adopters of popular assets (*Leskovec et al.*, 2007). Although some have modeled adoption simply as a function of observing strangers' actions (*Salganik et al.*, 2006; *Wu and Huberman*, 2007), principally, these studies measured the likelihood that an individual takes an action as a result of their friends' choosing the same action.

Social network information has successfully been used, for example, to predict whether a customer will sign up for a new calling plan once one of their phone contacts

does the same (*Hill et al.*, 2006). The photos we view and the stories we "Digg" are often the ones we observed our friends consuming (*Lerman*, 2007; *Lerman and Jones*, 2007). LiveJournal bloggers are more likely to join a group if many of their friends joined, and if those friends belong to the same clique (*Backstrom et al.*, 2006). Blogs are likely to link to content that other blogs have linked to (*Song et al.*, 2007). The insight that individuals tend to like (or like to have) the same things that their friends like can be used to improve collaborative filtering algorithms (*Zheng et al.*, 2007).

On the other hand there are relatively few studies that have included direct transfers between users. A study of person-to-person book and video recommendations found conditions under which such recommendations are successful (*Leskovec et al.*, 2006a,b). A study of online chain letters discovered that as messages diffuse through individuals' email contact networks, they form cascades that are far deeper than one would expect at random (*Liben-Nowell and Kleinberg*, 2008). However, information cascades spreading through email were not studied in the context of an explicit social network that would allow one to measure both direct or indirect influence simultaneously.

In contrast to prior work, we are able to analyze social influence not just indirectly through separate information about the social network and user adoption, but also by accounting for direct transfer of assets between individuals. The direct transfers allow us to more precisely identify influencers who are responsible for a disproportionate fraction of the asset adoptions. Furthermore, we develop a simple model of adoption rates, as opposed to probabilities, that can incorporate information about the evolving social network without needing to make arbitrary decisions about how to subdivide time intervals. This model allows us to clearly illustrate the importance of network effects in the adoption of content.

2.1.2 Description of data

Our data set includes time-stamped content ownership data and weekly snapshots of the complete social network over a 130 day period between September 1, 2008 and January 16, 2009, with the exception of the weeks of September 19th and November 14th. We do not have the exact time stamps of when the friendship ties were formed or dissolved, but by using weekly snapshots, we can approximate the coarse evolution of the social graph. At the user level, we have information on when the user first joined Second Life and how many hours they have played. The data also includes the social network of users. The data were provided directly by Linden Lab, the maker of Second Life and no personally identifying information of Second Life's users was

shared with the authors.

The observed social network we observe is made explicit by the users themselves, who add one another as "friends". By default, friends are aware of when their other friends are in Second Life, and if they grant additional permissions, those friends can see where in the virtual world they are located. Some users will even grant each other permission to modify each others' objects. This tends to occur among a small group of users for the purpose of collaboratively creating content. In other cases, users may not grant one another any permissions. Friend permissions do not necessarily need to be reciprocal.

As in many online social networks, the meaning of a friendship tie is somewhat ambiguous and can denote anything from casual acquaintanceship to a close relationship. One user may add another as a friend because they met in Second Life and wished to continue interacting. Or a Second Life friendship may reflect a "first life" relationship that has been carried into the virtual realm. While privacy preferences can vary from user-to-user, we consider the user's "social network" to consist of all friendship linkages that have, at a minimum, reciprocated permissions to see one another's online status. Throughout the chapter, we will refer to two users connected in this fashion as "friends" or "neighbors" in the social graph.

Since the subject of the work is on the diffusion of user-created content, we focus on studying content that is freely available, non-trivial to produce, and widely distributed amongst users. Content that can be carried around by a user is called an asset and is stored in the users' inventories. We chose to study *gestures*: transferable animations that allow a user's avatar to carry out programmed physical movements or make sounds. The choice of this type of asset was made because gestures are discrete and simple to trace. In our analysis, we use Linden Lab's definition of an active user: those we have logged in in the 60 days prior to the last observation date (Jan. 2009) and have used Second Life for more than six hours. In addition we focus on the users who have exchanged at least one object with another user between September 2008 and January 2009. We chose gesture assets that had at least sixteen unique owners and were never directly distributed to users by Linden Lab. The former exclusion rule omitted gestures that had not diffused, and the latter excludes gestures the users may have received without opting to. With these restrictions, our sample population contains 100,229 users and 106,499 assets. Because of the long-tail of asset popularity, this represents only a small fraction of the unique 5,327,671 gestures.

Most assets in our data set are owned by a relatively small number of users, and very large assets of size 1,000 or greater make up less than 10% of all assets. This

is the familiar long tail, shown in Figure 2.1, of content popularity; a few gestures are widely adopted by users, while the majority remain of little or niche interest. Interestingly, none of the content has saturated the user population, with the largest assets owned by roughly 10% of the population.

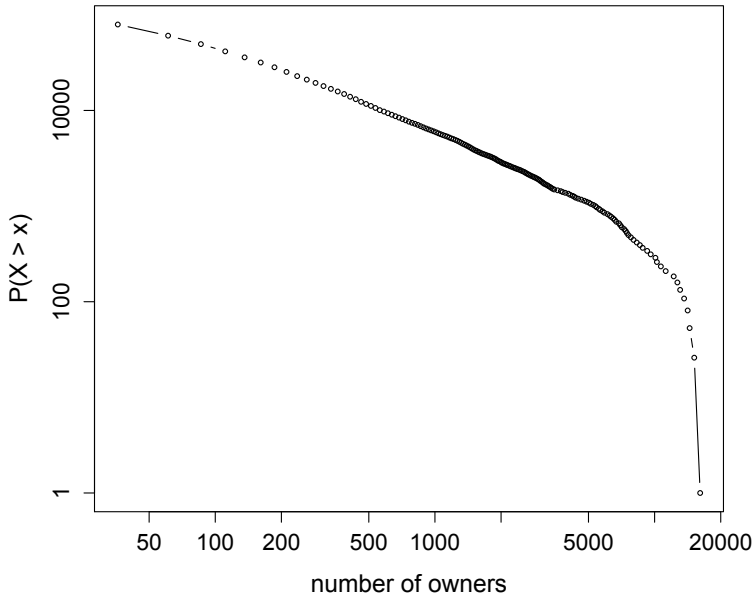

Figure 2.1: Cumulative distribution of the number of unique owners per asset in our sample population.

The content ownership data comes in the form of asset transfers, that contain the asset, previous owner, next owner, and time-stamp. It indicates that the previous user had given a copy of the asset to the next user. There are a total of 12,585,298 asset transfers over the observation period, 3,409,630 (23%) of which have accurate information about the previous owner. On average, approximately 43% of the observations in each asset have previous owner information. The average is higher than the total percentage because for larger assets there are more observations without previous owner information. Information can be lost, for example, when a user copies or moves assets in their inventory. The extent to which individual assets are missing previous ownership information does not appear to vary systematically with the owner's experience level, their connectedness to others, or how many gestures they own.

The transfers of each asset can be visualized as a cascade forest, with edges drawn between each owner and the previous owner, showing an "infection" path that represents the direct flow of content between users. Where previous owner information

is missing, we start a new tree in the forest. Figure 2.2 shows a cascade forest for one particular gesture. We note a fanning pattern, with some users transferring the gesture to many others.

Of the assets transfers for which we have accurate previous ownership information, 1,754,852 (approx. 48%) of the transfers occurred between friends. This suggests that direct social influence over the social network plays a considerable role in the distribution of content. In addition to direct influence, we find that indirect influence along the social network also plays a large role in adoption. Of those transfers that did not occur between friends, 678,908 (approx. 38%) of the users who had acquired a new asset did so after at least one of their friends had also adopted.

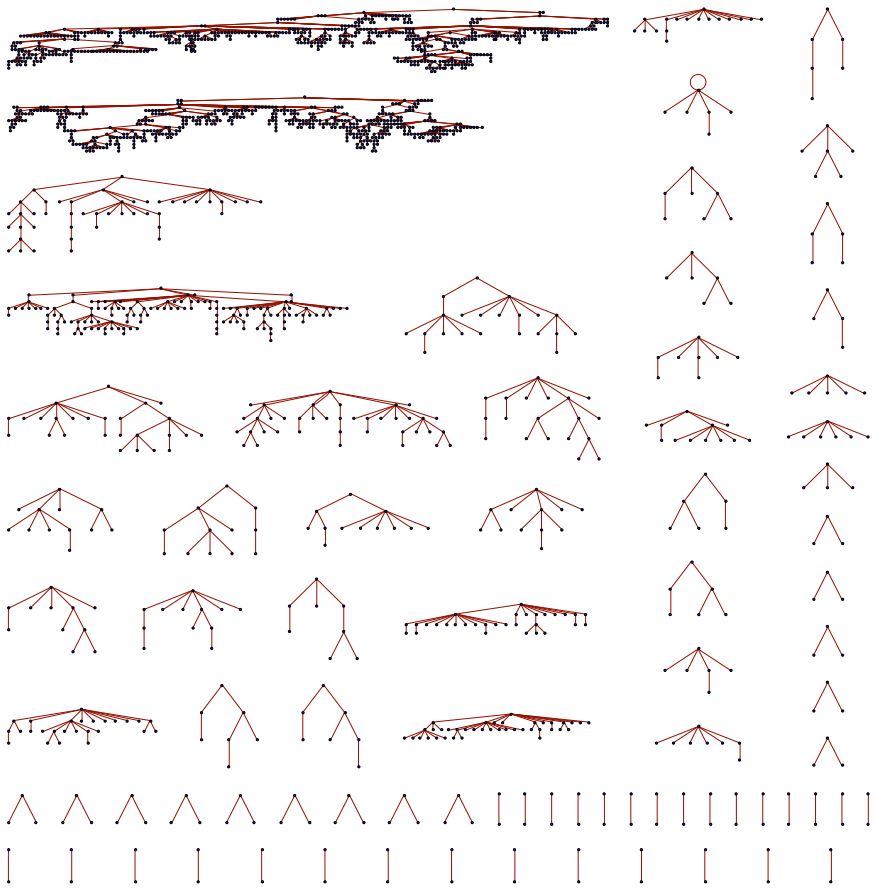

Figure 2.2: Example of a cascade forest for the Aerosmith(916) gesture. Edges denote transfers of the gesture between users.

2.1.3 Friend-to-friend vs. one-to-many

Given the above observations on the role of the social graph in the transfer and adoption of content, an important question a viral marketer may wish to answer is how much of a boost one can expect from having customers themselves advertise to one another and distribute the assets (*Domingos and Richardson*, 2001). Previous work on book and DVD recommendations found that viral marketing is more effective for niche products as opposed to widely popular ones (*Leskovec et al.*, 2006a). We find a similar trend here.

In order to quantify between-user transfers, we look at the following variables for each asset: the total number of adopters for the asset (the asset size or popularity), the percentage of the transfers that were between friends (% direct), and the percentage of transfers that resulted in subsequent transfers by the adopting user (% non-leaf). We find the percentage of non-leaf nodes, which can be thought of as a measure of cascade depth, to be correlated with the percentage of the adoptions that can be accounted for by the social graph ($\rho = 0.42$), indicating that the diffusion along the social network produces deeper cascades for which users actively participate in the transfer of the asset. But while these cascades tend to be deeper, they are not wider. The average popularity of the asset falls as the proportion of non-leaf nodes and social influence increases. As Figure 2.3 shows, having more adopters actively transferring assets is actually indicative of the asset not being broadly popular.

One can use the above observation of asset size and the role of social influence to predict the growth in the number of adoptions for a particular asset. We differentiate social influence (having a friend adopt before you do), and direct influence (obtaining an asset from a friend). Not all assets can be obtained from a friend, even if the friend has said asset, because of copy permissions. We therefore separate the assets where no transfers occur between friends (these likely cannot be copied), and ones that do.

We observe the number of adoptions in the first 30 days since the asset is created. We then run a regression to model the number of adoptions in the following 60 days. Besides the initial number of adoptions, we also included the following statistics from the first 30 days: whether the adoption occurred after at least one other friend adopted (% social), the percent of adoptions that are direct transfers along the social network, and the percent of adoptions occurring directly through the social network that resulted in further adoptions. Just two variables yielded the greatest explanatory power: the number of initial adoptions, and the percentage of initial adoptions that can be explained by the social network. We further find that using those same two variables, assets that are transfered from friend-to-friend at least sometimes are more

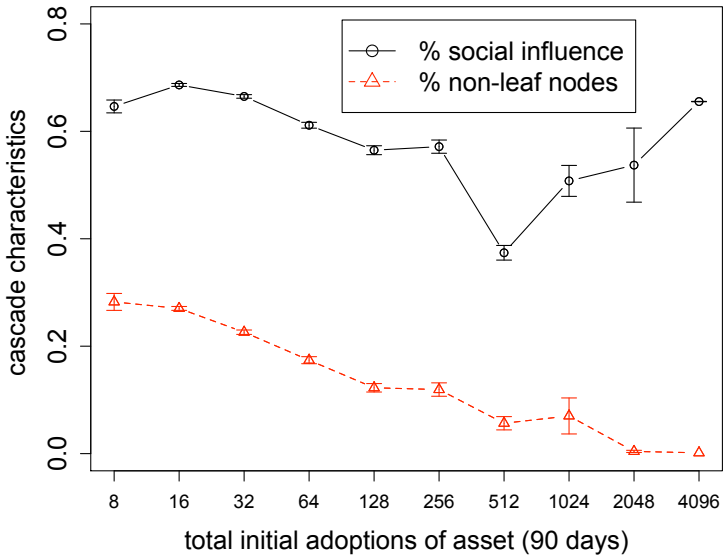

Figure 2.3: Percentage of non-leaf nodes vs. asset size for assets over the first 90 days of their spread.

predictable than those that are never passed between friends. A possible reason is that if friends are unable to share assets due to copying restrictions, then the distribution falls on a limited set of individuals, making the sharing of the assets more variable. Although information diffusing through a social network may lead to unpredictable cascades (*Watts*, 2002), in this case being able to observe such diffusion actually makes the cascade more predictable.

As Table 2.1 shows, unsurprisingly, a higher initial rate of spread translates to a higher number of subsequent adoptions. What is interesting is that the percentage of social adoptions (those that can be easily attributed to friends' adoptions) is *negatively* correlated with the the number of additional users who adopt. This suggests that assets that are diffusing through the social network may be of interest to a smaller subset of individuals. Because of homophily, the tendency of like to associate with like, these individuals are more likely to be friends with one another. So while a niche product may be shared more readily through the social network because the social network reflects niche tastes, the product does not have a wide susceptible audience, and therefore will not be adopted as widely.

While the regression suggests that the overall rate of spread through the social network is slower than through alternate paths, we find individual transfers to be more rapid between friends. Figure 2.4 shows the distribution of lags between when

	all assets	all assets	d	d
log(initial size)	0.362	0.388	0.508	0.476
% social		-0.808		-0.897
R^2	0.112	0.161	0.164	0.196

Table 2.1: Regressing the subsequent number of adoptions on the initial adoptions and percentage that can be explained by social influence. d is the restricted set of assets that were observed to have been transfered on the social network.

an individual becomes infected and when they infect either a friend, vs. when they infect a non-friend. First, we note that individuals are most likely to share a gesture within a short time of receiving it, while the context and novelty of the asset are still fresh.

Furthermore, we find that friends will more rapidly share with one another than with strangers: the average time lag between when a user acquires an asset and when they give it to a friend is 53.1 days, compared to the 75.6 it takes them to transfer it to a non-friend. The average time lag between one friend adopting after another (without sharing the asset with one another directly) is 105.2 days, compared to 228.3 days for adopters who are not friends. Although there is a mild cohort effect (with friends being more likely to join Second Life around the same time), it alone would not explain why friends are adopting so closely in time. That there is variation in speed depending on the relationship type is of interest because the speed of a interpersonal link can dramatically effect the fastest route information will take as it spreads, to the point where some slower links play little role at all (*Kossinets et al.*, 2008). It is therefore of interest to model the *rate* of adoption following a friend's adoption, and this is what we undertake in the next section.

2.2 Modeling adoption

As a Second Life user observes other users' avatars adopting particular assets, she may not only be more likely to adopt the asset herself, but the rate at which she does so may quicken as she observes more and more of her friends adopting. In order to characterize this social influence effect that occurs through Second Life's social network, we utilize a simple model of users' adoption rates. We show how with slightly different assumptions the same model can be applied to adoption rates both at the asset and at the user levels. Our results are compared with a Cox proportional hazards model with time-varying covariates that incorporates other possible influences such

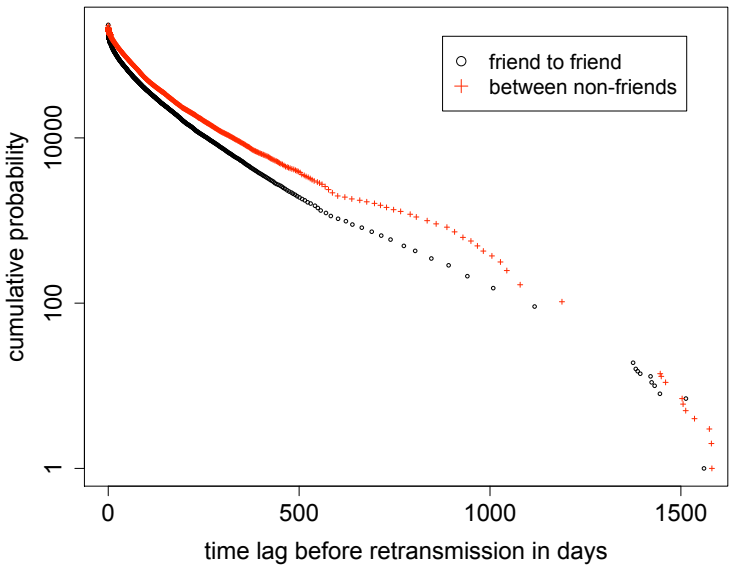

Figure 2.4: Lags between a users' adoption and retransmission times, for assets with 100-200 adopters.

as the total number of adopters in the user population. We show that the estimates produced by the Cox model are consistent with our simple model.

2.2.1 Formulation

One way in which this neighbor influence has been measured before is by computing the probability of adoption as a function of the number of neighbors who have already adopted in some time interval (*Backstrom et al.*, 2006). To be more precise, one counts the number of individuals who have not adopted that have k neighbors who have adopted at the beginning of the time interval and then compute the fraction of these individuals who have adopted at the end of the time interval.

An improved and related approach, used by (*Anagnostopoulos et al.*, 2008), considers the probability of adoption within many identical discrete time intervals, rather than just one. Our approach presents a further refinement by utilizing a continuous time model of adoption where we have stochastic rates of adoption rather than probabilities of adoption. We consider rates of adoption from two perspectives: at the level of adopting a particular asset and at the level of the user. In the former case, we assume that the rates of adoption are characteristic of a particular asset, are fixed in time, and the same for all users. These assumptions are analogous to the assumptions used in (*Anagnostopoulos et al.*, 2008) and (*Backstrom et al.*, 2006). At the level of

the individual user, we assume that a particular user's rates are fixed in time and equivalent for all assets, but that they differ from user to user.

We first explain the model formulation from the perspective of a particular asset computed over the entire population of users. A user enters into state k at the moment that their k_{th} friend adopts the particular asset. The model assumes that once an individual is in state k, the time until they adopt, T_k, is exponentially distributed, i.e. they draw an exponentially distributed random variable T_k with mean $1/\lambda_k$ where λ_k will be referred to as the adoption rate for state k. If an avatar's state changes before they reach their adoption time, they discard that time and draw a new time from the next exponential distribution corresponding to their new state. There are three ways in which a user can exit state k. If one of their existing neighbor adopts or they become friends with someone who has already adopted (adding an edge in the social network), they advance to state $k+1$. If they end a friendship with an adopter (deleting an edge in the social network), they return to state $k-1$.

We use maximum-likelihood to estimate λ_k from the available data for each asset. To do this, we have to compute the probability of observing the data given the model. Let t_k^i be the total amount of time the ith user spent in the k state and θ_i be one if the user adopted by the end of our observation period or zero if the avatar did not adopt. For the users that did adopt an asset, let a_i be the state from which that avatar adopted. Then the probability (density) of the data given the model is

$$\prod_i \lambda_{a_i}^{\theta_i} exp(-\sum_k \lambda_k t_k^i). \tag{2.1}$$

We can further simplify the probability of the data given the model by defining A_k to be the number of individuals who adopted from state k and $M_k = \sum_i t_k^i$ to be the total amount of time spent in state k over all individuals. Then

$$\prod_k \lambda_k^{A_k} exp(-\lambda_k M_k). \tag{2.2}$$

Maximizing with respect to the model parameters yields

$$\lambda_k = A_k/M_k, \tag{2.3}$$

as the maximum-likelihood estimate of the rates, assuming a uniform prior over the model.

We make a further distinction based on the population of measurements used to

calculate the characteristic rates in our model. For a particular asset, it's unclear whether the entire population of users should be included in the calculation. The reason for not including all users is that some individuals may never want to acquire the asset regardless of the number of their neighbors that adopt. Including all users for each asset is what has been done previously, which carries the assumption that all individuals considered will adopt if one waits a sufficiently long time. However, a user may never want to adopt, no matter how long they have been exposed to it. For example, Aerosmith gestures may be a taste that a particular user will never acquire. Rather, individuals are selective in their adoptions, and will resist both advertising and social influence if an asset does not match their tastes or interests. Therefore, our alternative approach is to estimate the rates only using measurements from the observed user population that has adopted the asset. We can be sure that this population wants the asset, but of course, there may be other individuals who want the asset but have just not acquired it yet.

Since there are advantages and disadvantages in including the non-adopting population in our measurements, we report our results for both specifications, referring to the respective calculations as utilizing the entire population and the adopting population of users. We note that our population of all users is still restricted to users who have adopted at least one asset during the time period, which means that all users were susceptible to adopting in general. To specify to the adopting population only, we follow the above derivation only including users that were observed to adopt the asset. This adjustment again leads to Eq. 2.3, where now M_k is the total amount of time spent in state k over individuals that adopted the asset.

As we mentioned above, one can model many users adopting the same asset, or one can model a particular user as they adopt different assets. Calculating adoption rates for a particular user over the entire population of assets is also simple. We again use maximum-likelihood to estimate λ_k for each individual using every asset. Let t_k^i be the total amount of time a user spent in the k state for the ith asset, θ_i be one or zero if the avatar adopted or did not adopt the ith asset respectively, and a_i be the state from which that user adopted the ith asset. Then the probability (density) of the data for that individual given the model is again Eq. 2.1. Defining A_k to be the number of assets adopted from state k and $M_k = \sum_i t_k^i$ to be the total amount of time the individual spent in state k over all assets, and then maximizing with respect to model parameters leads to Eq. 2.3. As in the analysis for particular assets, we also can decide to only include assets that the user was observed to acquire. This specification results in M_k being the amount of time that an individual has spent in

state k over all assets that they were observed to adopt. Again, we report our results for both cases for each user, which we refer to as either utilizing the entire population and the adopted population of assets.

In all cases, because our social network data begins on September 1st, 2008, we only consider times t_k^i calculated after the beginning of September. The state of adoption on September 1st, 2008 is treated as the initial condition to the model and no rate estimation is done using pre-September timing information for which the social network is uncertain.

2.2.2 Analysis

We first report on the differences in adoption rates as a function of the number of adopting neighbors for small and large assets separately. Asset size denotes simply the total number of adopting users for the asset. We also consider the trends across all assets, and "new" assets that appeared after Sept. 1, 2008. Examining new assets helps us avoid confounds such as large assets being in the later stages of their adoption curve. Figure 2.2.2 shows that adoption rates increases with the number of previously adopting neighbors a user has, whether one considers all users or just the adopters, and whether one includes all assets or just newer ones. When one considers all users, the rate increase is initially convex, suggesting that having two, rather than just one adopting increases the likelihood that a user will adopt at all. This is in agreement with previous analyses (*Leskovec et al.*, 2006a; *Backstrom et al.*, 2006; *Anagnostopoulos et al.*, 2008), which found that the probability initially increases steeply with k but then shows diminishing returns as k increases further.

Once we consider the population of just the adopting users, the rates do not show as steep of an initial gain as they did for all users. This is because now the rates do not reflect a binary outcome of whether or not the user adopts at all, but rather how much more quickly a susceptible user adopts following the adoption of multiple neighbors. For smaller assets that have between 50 and 500 adopters, the rate doubles between having no adopting neighbor to having one, with the increase more pronounced for new assets. It then increases roughly another 60% when a second neighbor adopts.

What is most striking, however, is that this rate of adoption as a function of the number of neighbors increases more rapidly for smaller assets. These plots confirm our intuition from Section 2.1.3 concerning the relationship between relative popularity and channels of influence. The increase in rate appears most strong for more niche items, whereas neighborhood effects appear to play less of the role for more popular assets. This suggests that what is driving the adoption of more popular assets must

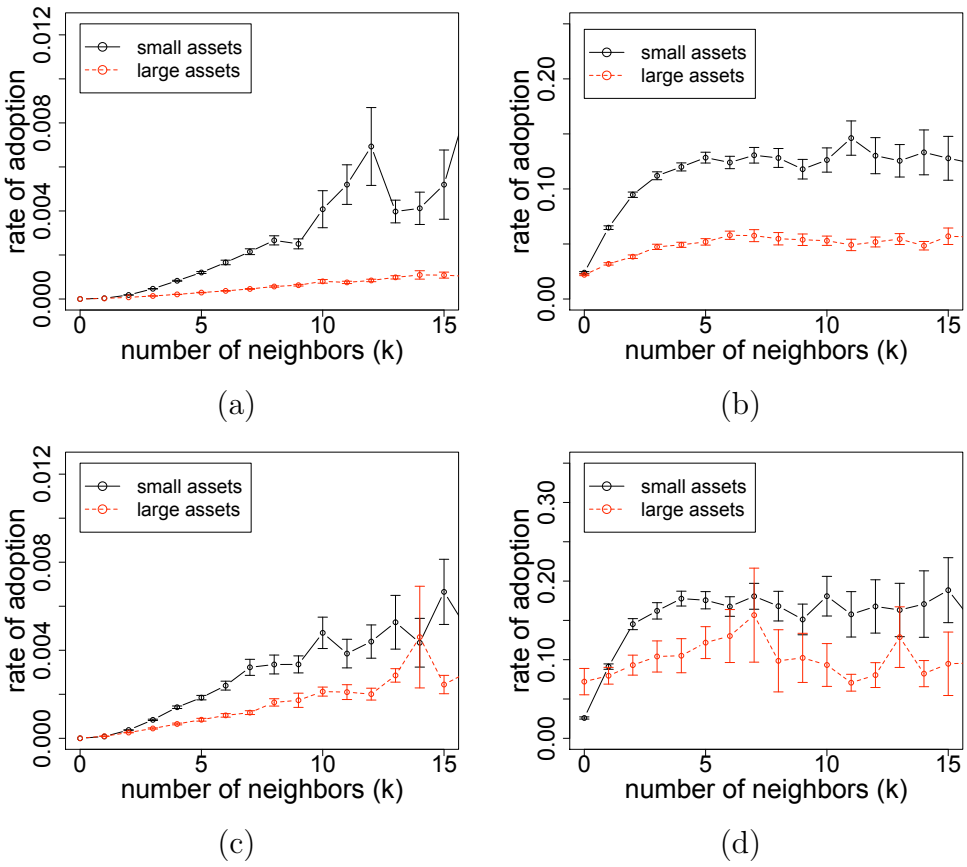

Figure 2.5: The average rate of adoption of assets as a function of adopting neighbors, k. The black curve corresponds to assets that are owned by 50-500 users, and the red curve corresponds to assets owned by 500 or more users. (a) entire population, all assets (b) adopting population, all assets (c) entire population, new assets (d) adopting population, new assets. The rates are in units of inverse days.

lie at least partly outside of the social network. For large assets, those with ≥ 500 adopters, λ_0 is 4.73 times higher than for smaller assets. For newer assets, this ratio is 7.43. Because collectively users spend much more time in the $k = 0$ state (having no adopting neighbors) than in the $k > 0$ states, a small difference in λ_0 can lead to significant differences in asset size. For example, across assets with between 50 and 500 adopters, the total length of time spent by all users in the $k = 0$ state is a factor of 190 times greater than the total length of time spent with at least one adopting neighbor.

We next turn to an analysis of user-characteristic adoption rates. We average the data over all individuals, where we divide the data into high and low degree, as shown in Figure 2.2.2. The first column and second column use the entire population and

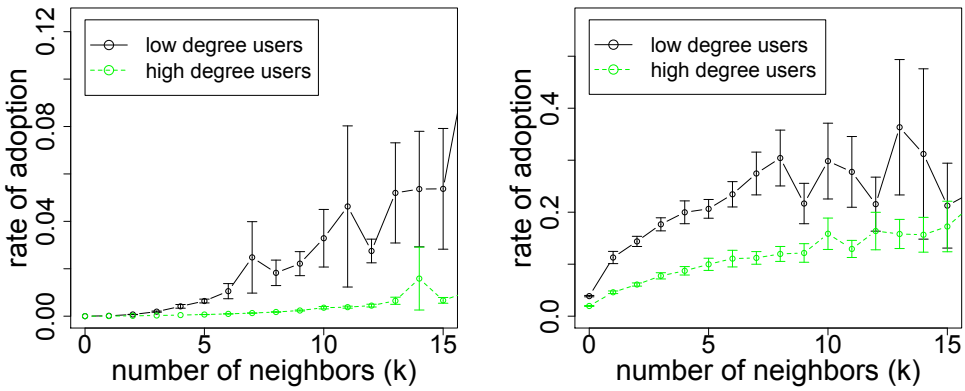

Figure 2.6: The average rate of adoption for users as function of adopting neighbors, k. The black curve corresponds to users of low degree that have 15-100 friends, and the green curve corresponds to users with 100-1000 friends. Left: entire population of assets. Right: adopted population of assets. The rates are in units of inverse days.

adopted population of assets, respectively. Interestingly, the users with high degree tend to adopt at comparably lower rates than their lower degree counterparts. This suggests that individuals may accumulate many friends, especially in online contexts, but consequently any individual friend holds less influence.

2.2.3 Comparison with the hazard model

A more general approach to adoption rates that can also be applied to our data is the Cox proportional hazards model with time-varying covariates. For the regression we included a fixed average degree observed over the time period, the number of assets owned, the user's cohort (0-5, 5 being most recent), and usage (in days). We include the number of adopting neighbors and number of adopting users as time-varying covariates. The number of assets, usage, and number of adopting users were log-transformed. Results from the regression are shown in 2.2.

As in the previous model, we find that the number of adopting neighbors has a significant and positive effect. We also see that high average degree does indeed have a negative effect on the adoption rate. By itself, the overall popularity of an asset does increase the rate of adoption, as suggested in 2.2.2(d). In combination with the other factors, however, overall popularity has a weakly negative effect in the rate of adoption. Finally, we see that users that have signed up recently tend to adopt friends content more rapidly, and that this effect decreases with experience.

parameter	estimate	error
mean degree	-0.00134	0.00009
assets owned	0.03391	0.00601
cohort	0.62933	0.01176
usage	-0.18349	0.00470
adopting neighbors	0.32795	0.00902
adopting users	-0.04634	0.00754

Table 2.2: Cox proportional hazards model with time-varying covariates. All estimates have $p < 0.001$.

The results indicate substantial heterogeneity in user behavior, which we further investigate in the next section where we look for influential users and early adopters.

The above results indicate substantial heterogeneity in user behavior and we further investigate this in the next section, where we look for influential users and early adopters.

2.3 Influencers and early adopters

Thus far we have observed social influence from the point of view of the adopter – finding that the rate of adoption increases as one observes more and more friends adopting. This suggests that each friend holds some influence, and that having more adopters among one's friends increases the "hazard" that one will catch the bug and adopt as well. But one may also pose the question of whether all adopters are equally contagious to their friends. More specifically, using data on user-to-user asset transfers among friends, we can examine whether a few individuals are responsible for distributing assets.

2.3.1 Concentration of influence

First, we look at the distributions of transfers per individual, shown in Figure 2.7. The distributions are heavy tailed, indicating that a majority of individuals play a negligible to small role in distributing assets, while a handful of users disproportionately contribute to the dissemination of content. Some of the heavy-tailedness may be explained by primary content providers (i.e. store owners) whose role includes marketing assets to individuals. While approximately 52% of the transfers occur between non-neighboring users, many transfers occur at similar scales between users that are affiliated with one another, or more strongly, have at least three other friends

in common.

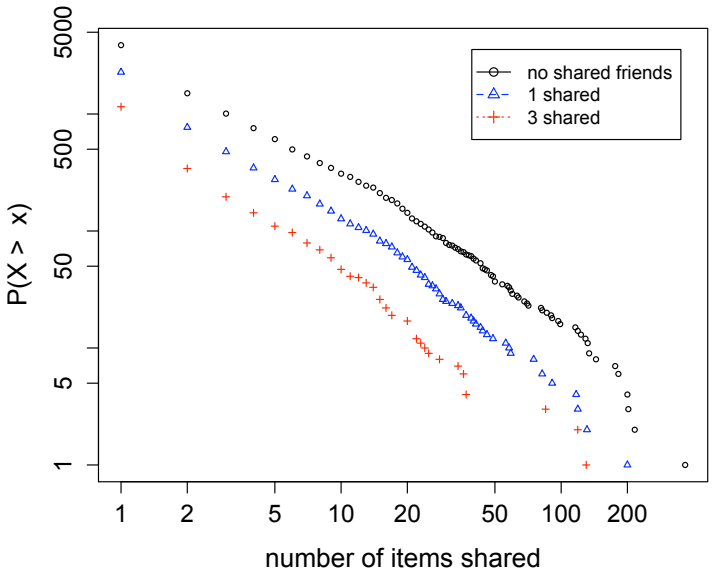

Figure 2.7: Distributions of the number of assets shared by users with other users with whom they share a specified number of friends in common.

We can also measure the entropy of users who are responsible for transfers and compare it against a null model where each subsequent adopter receives the asset from a randomly chosen previous adopter. The entropy is simply computed using the proportion of transfers that can be attributed to each user in the cascade who shared at least one asset. The null model has two parameters, the total number of owners of the asset n, and the proportion p of missing edges in the cascade. At each time step, the null model adds a new owner, who with probability p starts a new tree, and with probability $(1-p)$ picks one of the previous owners uniformly at random as its parent node. The null model was computed for each asset using the corresponding (n, p). We find that the distribution of entropies from the data, measured in bits, have a mean of 2.72, which is significantly lower than that of the null model (3.48), (t = 97.08, p = 0). This indicates that the actual distribution of assets is more concentrated than one would expect if every previous adopter participated with equal probability.

An obvious distinction between the null model and the actual cascades is the tendency of the observed cascades to be concentrated on the social graph, with many users adopting after their friends do. As we mentioned before, 48% of the direct transfers occur on the social graph. A null model that takes just any previous adopter

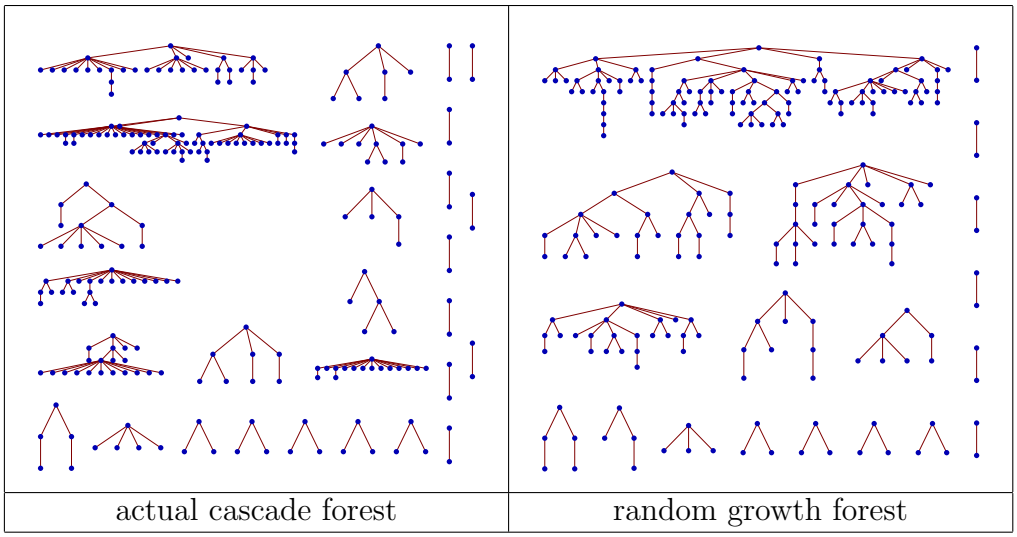

Figure 2.8: Comparison between actual growth of cascades and a null model where each previous adopter is equally likely to be sharing assets.

as the source of an asset would pick a friend 6.6% percent of the time. This is calculated by computing the fraction of previous adopters who are friends for each transfer with accurate previous owner information and dividing by the total number of such transfers. Unsurprisingly, direct sharing is much more a feature of friendship ties than simply a desire to share with others of similar tastes.

2.3.2 Strength of influence

The number of times a user transfers assets is an unambiguous influence measure. However, it doesn't capture how successful a user would be in a competition where one's friends could obtain assets from others. We propose a simple measure, γ, that compares the number of times a user A infected one of its friends B, against the expected number given the odds that B was not infected by one of their other adopter friends. For example, if B had 2 other friends besides A who had previously adopted, and B obtained the asset through a friend, then the probability that A was the infector is 1/3. This adds 1/3 to the expected number of transfers for A.

We measure γ =*(transfers - expected)/expected* for all users who had at least 20 instances where one of their friends acquired an asset through a social tie after they did. If odds were even that the adopter receives the asset from any one of their friends, the user's γ scores would be narrowly distributed around 0 – they would be doing no worse or better than odds. Figure 2.9 shows what the distribution of gamma scores would be if all the observed transfers occurred from a randomly chosen previously

adopting friend. In contrast, the distribution of observed γ scores is highly skewed – approximately 74%, fall below 0 and while the remaining 26% are more influential than odds. The actual gammas have a mean of -0.286.

A further question one might have is whether a user can be influential in distributing many assets or just a few. The overall correlation between the number of transfers a user made and the number of assets they were sharing was highly positive ($\rho = 0.63$), but still displayed a wide range of behaviors. For example, one user influenced 73 transfers to friends (when just 4 were expected) involving just 2 different assets. In another case, a user who was expected to have made 13.3 transfers, but made 104, had similarly high γ but these transfers involved 16 different assets. In yet another case, 47 transfers involved 46 assets, implying that one user is repeatedly transferring items to the same user.

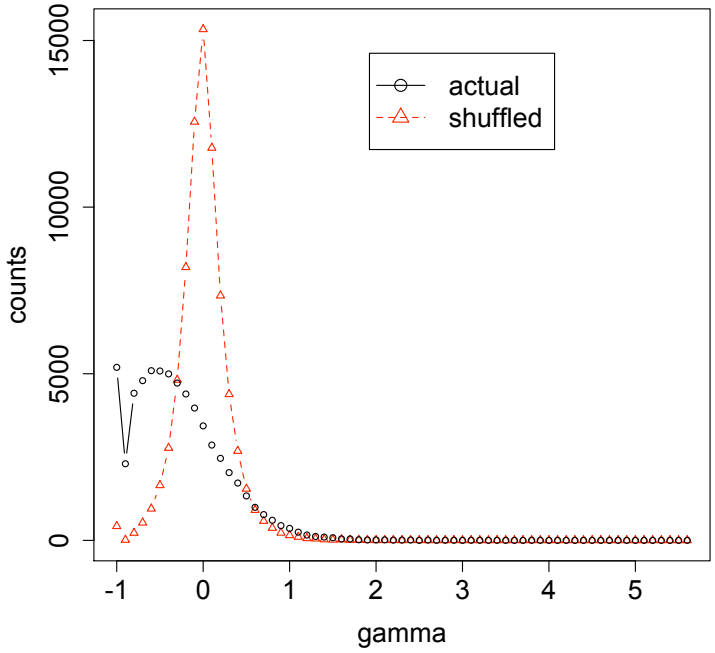

Figure 2.9: Users' influence using the γ measure, for actual and randomized adoptions.

Who are these influencers and what are their characteristics? Interestingly, even though users with a higher number of friends tend to have been around longer ($\rho = 0.13$), have more assets($\rho = 0.16$), and have made more transfers in total ($\rho = 0.14$), a

user's γ score is negatively correlated with their number of friends ($\rho = -0.17$[1]). This is likely because maintaining strong ties with many individuals is more difficult, hence influencing any single one is less likely. We observe, for example, that the number of assets shared by two friends is correlated with the strength of their tie ($\rho = 0.10$), as given by the number of friends the two have in common.

A higher γ is slightly negatively correlated with the number of assets ($\rho = -0.05$), but highly positively correlated with the number of transfers to friends per asset owned ($\rho = 0.35$). This means that influencers don't necessarily have more assets than others, but the ones that they do have, they like to share with their friends. Overall, we find that users who are sharing a higher number of assets and making more transfers tend to be sharing less popular ones $\rho = -0.15$, again suggesting, as in Section 2.1.2 that assets shared tend to be niche products.

We also examine whether those who are directly responsible for their friends' adoptions tend to acquire assets earlier. While users who have more transfers per asset tend to be "earlier" in their adoption ($\rho = -0.06$), both in terms of absolute rank (they were the r^{th} person to adopt) and relative rank (they were among the first p% of users to adopt), a user's γ score and relative adoption rank are uncorrelated. Altogether, combining the age, number of friends, number of assets, average adoption rank, and average number of transfers in a linear regression model yields an R^2 of 0.17 for a user's γ score.

2.3.3 Early adopters

This still leaves the question of whether the very earliest adopters might be different as a group from other users. We select 779 users who have 20 or more gestures and have an average relative rank of 0.05 (meaning that they are on average among the first 5% of adopters for all the assets that they own). This corresponds to being the 15th adopter on average across the assets one owns. For the analysis below we obtained qualitatively similar results when we selected an early adopter group of approximately the same size, but slightly different criteria: adopting 40 or more gestures, and being among the first 10% of users to acquire them.

We compare the early adopter group against the group of 50,000 users who have also acquired 20 or more gestures, but are on average in the latter half of adopters for those gestures. We can immediately rule out some factors relating to whether a user becomes an early adopter. The early adopters were on average born just 68 days earlier, meaning that joining Second Life earlier yields only a slightly higher

[1] number of friends and assets were log-transformed before their correlation was measured

advantage in being one of the first adopters of an asset. Early adopters have actually had a bit less playtime than the later adopters (40 hours), and have an average of 8 fewer friends (for an average of 61 and median of 33). Clearly the early adopters are neither especially early, active, nor gregarious.

The very earliest adopters distinguish themselves in other ways. For the assets that they eventually adopt, the rate of adoption before any of their friends adopt, $\lambda_{k=0}$, is twice as high as that of the laggard group ($t = 4.2, p < 0.0001$), as is their rate of adoption under initial social influence, $\lambda_{k=1}$, though this difference was not as significant ($t = 2.3, p < 0.05$). This indicates that they are more susceptible to adopting assets early (when none or one of their friends have adopted), although on average they own 20 fewer assets than late adopting users (t = 10.3, p = 0). Possibly it is not so much that they are early adopters, but, being trendsetters, they resist acquiring assets that have become too common.

Finally, we examine the direct influence that these early adopters wield, and find that their γ scores, though closer to odds (-0.08) than that of the later adopters (-0.22) are not particularly impressive. The number of transfers they make is not significantly higher than the laggard group, even though the assets they adopt eventually grow to be more popular than those owned by laggard group ($t = 5.5, p < 10^{-7}$). Previously simulated models of social influence over social networks have established a negative link between being an early adopter (easily succumbing to a new trend) and therefore been less influential (*Watts and Dodds*, 2007). This is not the case for the most extreme early adopters in Second Life. But the overall trend for all users is a very slight but statistically significant negative correlation between the probability that one adopts before one's friends do, and both γ ($\rho = -0.015, p < 0.001$) and number of transfers the user makes ($\rho = -0.02, p < 10^{-7}$).

In summary, we identified some users as influential, and others as early adopters. They don't appear to be one and the same, with the early adopters being more easily susceptible early on, but not being more likely to share their finds. We were able to identify some characteristics of both early adopters and influencers, however, these characteristics alone cannot be used reliably to identify such users. The size of a users' social network is just one of the variables that was of little help in identifying influencers, although the social network itself is responsible for many of the transfers.

2.4 Conclusion

We examined the interplay of social networks and social influence in the adoption of online content. Roughly 48% of transfers occur along the social graph, the remainder occurring between users who are not friends. We find that assets whose transfers typically occur through the social graph tend to have deeper transfer cascades measured as a higher proportion of non-leaf nodes, but tend to grow more slowly. This suggests that social networks are an important medium for diffusion of *niche* information in Second Life.

We applied models of social contagion that capture the rate at which users adopt following the adoption by one of more of their friends. We find that the rate of adoption increases as more of one's friends adopt, and that this is more significant for smaller, niche assets. We also find that someone who has many friends is less likely to be influenced by any particular one. A user with many ties would have difficulty maintaining all of them, increasing the probability that many of the ties are weak and therefore hold less influence. Indeed, we found a slight correlation between the strength of a tie and number of assets that are transferred between two friends.

We further find that some individuals play a more active role in the transfer of assets than others. A random cascade model, where any node is equally likely to produce another leaf node, yields a higher entropy than the empirically observed cascades. But the variability in influence cannot be attributed to the social network alone: when we measure the direct influence an individual has on a particular friend, this influence is negatively correlated with the number of friends. Finally, the early adopters, while being more susceptible to adopting content without waiting for many of their friends to so, do not wield greater influence over others.

CHAPTER III

The Effect of Social Networks on Information Diffusion

Social networks are thought to play an important role in the dissemination of information. Quantifying this effect not only requires that one identify who influences whom, but also how individuals would share information in the absence of interpersonal influence. We determine the role of networks in information diffusion with a large-scale experiment that randomizes exposure to information about friends' sharing behavior among 253 million subjects in situ. We find that those who are exposed are significantly more likely to spread information, and share sooner than those who are not. Although stronger ties are individually more influential, we show that the majority of content is disseminated via the more abundant weak ties who expose their friends to novel information that would not have otherwise been spread.

3.1 Introduction

The structure of social interactions can have a substantial effect on social contagion and other spreading processes (*Granovetter*, 1978; *Watts and Strogatz*, 1998; *Newman*, 2002). This insight has generated academic and commercial interest in quantifying influence in face-to-face (*Christakis and Fowler*, 2007) and online networks (*Aral et al.*, 2009; *Goldenberg et al.*, 2009; *Cha et al.*, 2010; *Bakshy et al.*, 2011). But to what extent are individuals actually influencing one another, and how critical are these interactions to the overall spread of information within a network? Regardless of scale, empirical studies face a persistent challenge of distinguishing influence from mere correlation (*Manski*, 1993; *Aral et al.*, 2009). Homophily, the tendency of individuals with similar characteristics to associate with one another (*McPherson et al.*, 2001; *Adamic and Adar*, 2001; *Kossinets and Watts*, 2009), provides an alternative

explanation of why connected individuals may share the same content: they regularly frequent similar information sources, such as web sites (*Adar et al.*, 2009). Statistical methods that attempt to disambiguate homophily from influence (*Christakis and Fowler*, 2007; *Anagnostopoulos et al.*, 2008; *Aral et al.*, 2009) cannot entirely account for unobserved factors that fundamentally confound the two (*Shalizi and Thomas*, 2011).

Moreover, since individuals are more similar to those with whom they interact often (*Granovetter*, 1973; *McPherson et al.*, 2001), the relative role of strong and weak ties in information diffusion becomes even less clear. On one hand, individuals who interact more often have greater opportunity to influence one another and have more aligned interests, increasing the chances of contagion (*Brown and Reingen*, 1987; *Hill et al.*, 2006). However, this commonality amplifies the potential for confounds: those who interact more often are more likely to have increasingly similar information sources. As a result, inferences made from observational data may overstate the role of strong ties in information spread. Conversely, individuals who interact infrequently have more diverse social networks that provide access to novel information (*Granovetter*, 1973; *Burt*, 1992). But because contact between such ties are intermittent, and the individuals tend to be dissimilar, any particular piece of information is less likely to flow across weak ties (*Centola and Macy*, 2007; *Centola*, 2010). While data on how often individuals communicate has historically been biased (*Marin*, 2004; *Bernard et al.*, 1984), the increasing ubiquity of networked communications technologies have allowed researchers to more precisely analyze the relationship between social structure and information diversity (*Aral and Van Alstyne*, 2011).

3.2 Experimental design

We use a randomized field experiment on Facebook to determine the causal effect of networks on information diffusion, which allows us examine how tie strength is reflective of common information sources and interpersonal influence. A recent survey of US Facebook users shows that the average user maintains 48% of their social network on Facebook, and that 40% of users are connected to all core discussion confidants on Facebook (*Hampton and Rainie*, 2011). Given the strong connection to real-world contacts and scale of usage, Facebook is an ideal platform to test theories of diffusion in situ. Facebook users primarily interact with information through an aggregated history of their friends' recent activity, called the news feed, or simply feed for short (Figure 3.1). A user can share a link (URL) to content she finds on the Web, which

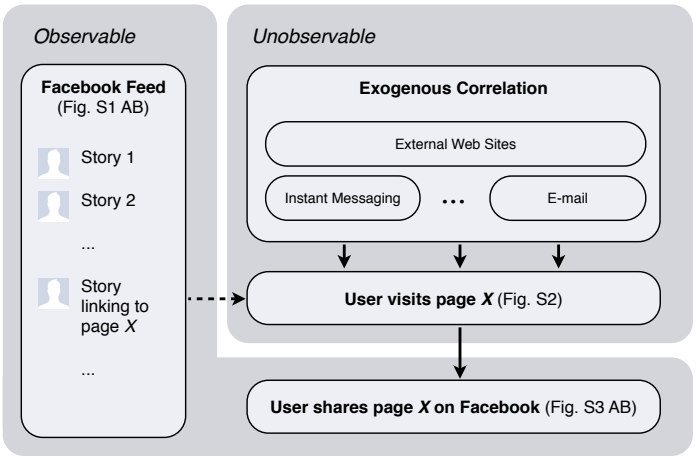

Figure 3.1: Causal relationships resolved by our experimental design. Information presented in users' news feeds and other sharing behavior on `facebook.com` are observed. Exogenous events that cause users to be exposed to information outside of Facebook cannot be observed and may explain their sharing behavior. Our experiment blocks the causal relationship (dashed arrow) between Facebook and user visitation by randomly removing stories about friends' sharing behavior in subjects' feeds. Thus, our experiment allows us to compare situations where both endogenous social influence and exogenous correlations exist (the *feed* condition), to situations in which only exogenous correlations exist (the *no feed condition*).

is then broadcast to her friends via the feed; a user can also re-share a URL that her friend previously shared, which will subsequently be broadcast to her other friends. By randomly selecting individuals and removing their exposure to social information about randomly selected content, we are able to directly compare the overall probability with which subjects share links that they were or were not exposed to on the feed. We then combine this information with data on users' interaction frequencies, which allows us to contrast the amount of correlation and influence between strong and weak ties.

URLs shared by subjects' friends are assigned to two experimental conditions: the *feed* and *no feed* conditions. Subject-URL pairs assigned to the *feed* condition are presented to subjects, whereas those in the *no feed* condition are not shown. Pairs are deterministically assigned to a condition at the time of display, so any subsequent share by any of a subject's friends is assigned to the same condition. Because removal occurs on a subject-URL basis, and we include only a small fraction of subject-URL

pairs in the *no feed* condition, a URL is on average delivered to over 99% of its potential targets. All viewing and sharing activity related to the subject and URL are logged. Our experiment took place over the span of seven weeks in 2010 and includes 253,238,367 subjects, 75,888,466 URLs, and 1,168,633,941 unique subject-URL pairs.

3.3 Results

We find that subjects who are exposed to content on the feed are many times more likely to share, and share sooner than those who are not exposed. To measure the relative increase in sharing due to exposure, we compute the risk ratio: the likelihood of sharing in the *feed* condition (0.260%) divided by the likelihood of sharing in the *no feed condition* (0.044%), and find that individuals in the *feed* condition are 5.59 (95% $CI = [5.53, 5.65]$) times more likely share. Subjects who share the same URL as their friends typically do so within a time that is proximate to their friends' sharing time. For those URLs that were assigned to both experimental conditions, the median sharing latency in the *feed* condition is 6 hours, compared to 20 hours when assigned to the *no feed* condition (Wilcoxon rank-sum test, $p < 10^{-16}$; see Appendix A).

Models of biological and social contagion posit that the likelihood of contagion increases as a function of activated contacts *Granovetter* (1978); *Newman* (2002); *Centola and Macy* (2007); *Centola* (2010). Our experiment shows that the probability of sharing a link on Facebook increases with the number of sharing friends in both experimental conditions (Figure 3.2A). The presence of this relationship in the *no feed* condition provides evidence of correlation among individuals due to factors outside of social information in the feed. The effect of the feed relative to these other exogenous factors can be measured as either the difference or ratio between the probability of sharing in the *feed* and *no feed* conditions (Figure 3.2BC). While the difference in sharing likelihood grows with the number of activated friends, the relative risk ratio falls. This contrast suggests that social information in the feed is most likely to influence a user to share a link that many of her friends have shared, but the relative impact of that influence is highest for content that few friends are sharing.

In addition to claims about the number of contacts, social contagion research suggests that tie strength contributes to the amount of influence between two individuals, and that weak ties provide access to more novel information *Granovetter* (1973). Our experiment enables us to evaluate these hypotheses using measures of tie strength that are informed by the complete set of interactions between users on

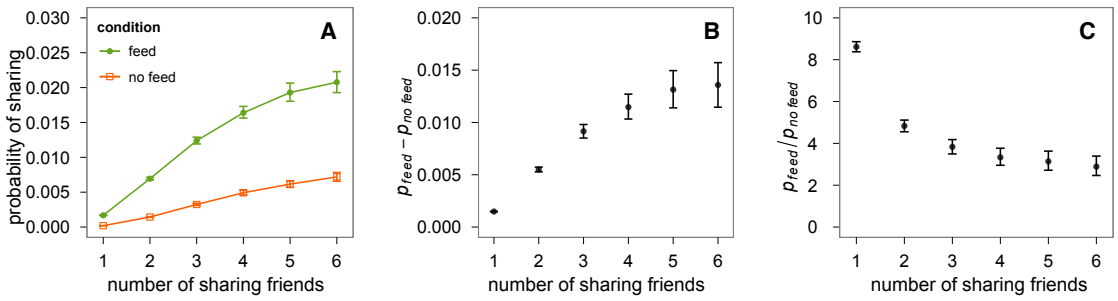

Figure 3.2: Users with more friends sharing a Web link are themselves more likely to share. (A) The probability of sharing for subjects that were (*feed*) and were not (*no feed*) exposed to content increases as a function of the number sharing friends. (B) The absolute effect of the feed is greater when subjects have more sharing friends, but additional friends have decreasing marginal influence. (C) The multiplicative impact of the feed, measured as the ratio of probabilities in the two conditions, is greatest when few friends are sharing. Error bars represent the 95% bootstrapped confidence intervals clustered on the URL (see Appendix A for details).

Facebook. We construct two measures of tie strength using data from a time period directly prior to our experiment: (i) the frequency of interaction on Facebook in terms of comments the subject receives from a friend on her posts, during the past three months, and (ii) number of photo coincidences between the two individuals, during the past year. The measurements are chosen to reflect the strength of online and offline interactions; two other measures are further analyzed in Appendix A.

We then examine how the likelihood of sharing a URL in the *no feed* and *feed* conditions varies according to the strength of tie between a subject and her friend, for subjects with exactly one sharing friend. In both conditions, a subject is more likely to share a link when her previously sharing friend is a strong tie (Figure 3.3AB). For example, subjects who were exposed to a link shared by a friend from whom the subject received three comments are 2.83 times more likely to share than subjects exposed to a link shared by a friend from whom they received no comments. For those who were not exposed, the same comparison shows that subjects are 3.84 times more likely to share a link that was previously shared by the stronger tie. We find that the risk ratio of sharing between the *feed* and *no feed* conditions is highest for content shared by weak ties (Figure 3.3CD). This suggests that weak ties carry information that one is unlikely to be exposed to otherwise, thereby increasing the diversity of information propagated within a portion of a social network. Using the

experimental data, we can compute the average amount of contagion on the feed due to strong and weak ties. Even under a generous classification of *strong ties* as friends with whom the subject had at least one interaction, we find that the vast majority of information shared originates from *weak ties*, with whom the subject had no interactions[1] (Figure 3.4).

By separating exogenous correlation from endogenous social influence, our experiment sheds light on the role of social ties in propagating information within networked communication technologies. We are able to estimate the causal effect of ties on information diffusion through randomization, and show that although strong ties are individually more influential, the majority of information travels between individuals who interact infrequently. Unlike the spread of items that are subject to positive externalities *Aral and Van Alstyne* (2011) or require effort to adopt *Centola and Macy* (2007); *Centola* (2010), mere exposure to information is often sufficient to induce large increases in spread. This suggests that in highly connected networked environments, the capacity for weak ties to disseminate information may differ significantly from situations examined by previous work. As online social networks play an increasingly important role in the dissemination of information, these findings carry important implications for the diversity of information individuals are exposed to. In addition, we also expect that other non-network features play an important role in diffusion processes. Future work may investigate how properties of the individual, such as age, gender, and nationality, or features of the content, such as popularity and breadth of appeal, relate to influence and its confounds.

[1]Since only 4.2% of shares occur when more than one friend has shared, consideration of cases in which subjects have more than one sharing friends would not effect the significance of our findings.

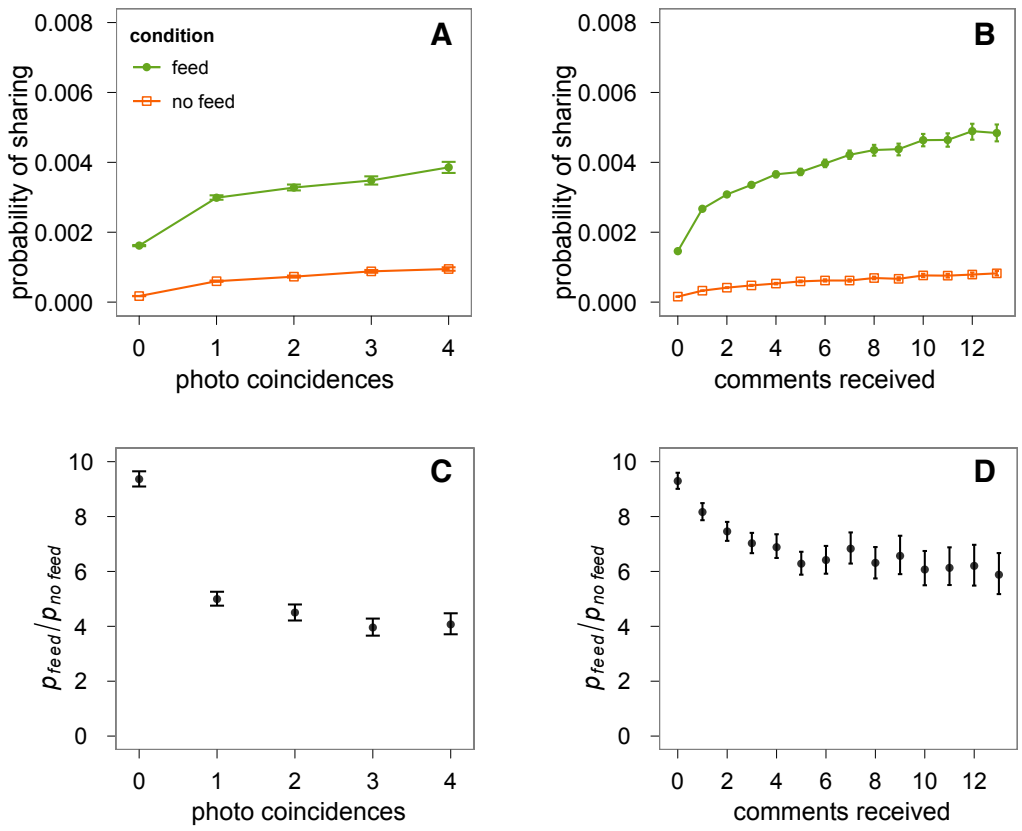

Figure 3.3: Strong ties are more influential, and weak ties expose friends to information they would not have otherwise shared. Figures show the effect of two measures of tie strength on sharing: the number of photo coincidences and the number of comments received from a subject's friend. (A) and (B) show the increasing relationship between tie strength and the probability of sharing a link that a friend shared in the *feed* and *no feed* conditions. (C) and (D) show that the multiplicative effect of feed diminishes with tie strength, suggesting that exposure through strong ties may be redundant with exogenous exposure, while weak ties carry information one might otherwise not have been exposed to.

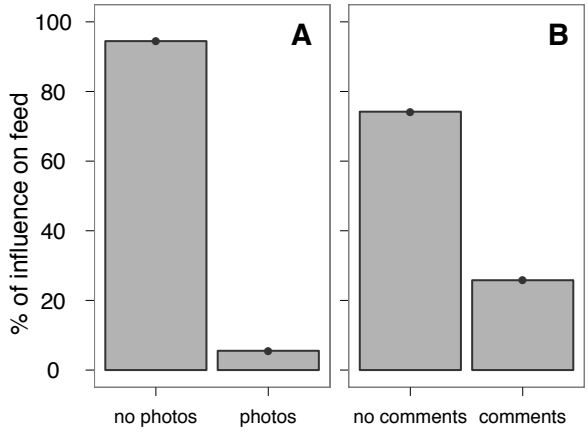

Figure 3.4: Weak ties are collectively more influential than strong ties. (A) and (B) show the percentage of information spread by weak and strong ties. We classify a tie as *strong* and *weak* if two users did or did not appear in the same photograph (A), or if the influencer had or had not previously commented on their friend's post (B). The percentage of influence on feed is computed by taking the number of links appearing in the feed from each tie type, and weighting them by the expected probability of sharing due to influence on the feed, given the tie type. Although the probability of influence is significantly higher for those with any trace of interaction, most information travels along weak ties, which are more abundant (Appendix A).

CHAPTER IV

The Effect of Social Information on Sharing Decisions

Social networks act as pathways for information diffusion, but to what extent does social information affect individual sharing decisions? We present an experiment that randomizes the amount of information about friends' sharing behavior given to 1.1 million Facebook users visiting 470 thousand web pages. Our design evaluates how social information increases content sharing beyond its ability to expose others to novel content, while simultaneously controlling for confounding factors related to homophily. We find that the number of sharing friends an individual has, and the strength of ties with those friends, is a significant predictor of whether an individual shares a page, even when no social information is displayed. In addition, the number of friends shown has a significant but relatively modest increase in the likelihood of sharing compared to the baseline condition in which no friends are displayed. Lastly, we find that stronger ties are more predictive of an individual's sharing behavior, but that strong ties do not appear to be any more influential than weak ties. Our results quantify the effect of social information on individual sharing decisions in online settings, and highlight the importance of experimentation in the evaluation of such systems.

4.1 Introduction

Many models of social influence in networks exist, but to what extent do they capture the mechanisms responsible for the spread of information online? In the spread of a disease, every exposure to a contagion has some independent probability of infecting the exposed individual (*Anderson and May*, 1992; *Newman*, 2002). In contrast, the spread of *social* contagions is thought to depend critically on properties and behaviors

of a given individual's network of contacts, which may affect the individual's adoption beyond simple exposure (*Schelling*, 1973; *Granovetter*, 1978). Whether this contagion be an idea, political stance, rumor, or news bulletin, its spread occurs in two stages: an individual must first be exposed to information, then make a purposeful decision to propagate that information. Measures of social influence can pertain to the first stage, second stage, or the entire process. In this study, we focus on the second stage of the diffusion process using a large-scale randomized field experiment that controls for correlations between the sharing behavior of subjects and their friends.

There are many reasons that may explain why one's network of contacts can affect decision making processes. For example, the utility of adopting a technology may depend on the current number of adopters (*Schelling*, 1973). The behavior of one's close friends increases the perceived legitimacy of social movements (*Finkel et al.*, 1989) or clothing styles (*Crane*, 1999). The credibility of one's peers plays a role in the spread of innovations (*Coleman et al.*, 1966; *Markus*, 1987) and folk knowledge (*Granovetter*, 1978). These examples form the basis of threshold models of contagion, which assert that one's willingness to adopt a behavior or idea depends critically on the total number of "infected" contacts that one observes. Because the spread of these behaviors requires affirmation from multiple sources, this process is sometimes referred to as complex contagion (*Centola and Macy*, 2007; *Centola*, 2010), and stands in contrast to simple contagion, where the effect of every exposure on an individual's behavior is independent of signals received from other contacts. Online information diffusion can be thought of as a composite of both models: exposure is a prerequisite for being able to make a decision about sharing content, and it seems reasonable to expect that knowledge of ones' network would affect the ultimate decision to share.

Identifying the process of influence is further complicated by external factors. Connected individuals in networks are similar to one another (*McPherson et al.*, 2001; *Adamic and Adar*, 2001; *Kossinets and Watts*, 2009), and in Chapter 3, we show that they are more likely to share the same information as one another, and therefore have similar information sources. We capture the effect of social information along both stages of the contagion process by censoring exposure to content in subjects' Facebook feeds: subjects in the *no feed* condition are neither exposed to content their friends share via the feed, nor can this social information come to bear on subjects' decision to share that content if they find it independently. To separate the effect of social signals on the decision making process from the exposure process, one must manipulate signals given to subjects after they have discovered content independent

of their friends.

In this study, we examine how social information specifically affects sharing decisions by randomizing the amount of information subjects see about their friends' sharing behavior on web pages the subjects visit independent of the Facebook feed. Our experimental design builds upon the experiment discussed in the previous chapter: by focusing solely on subjects who were not shown their friends' sharing behavior in feed (those assigned to the *no feed* condition), we are able to control for exposure and isolate the effect of social information on sharing decisions. In Section 4.2, we describe the experimental design and population. Section 4.3 examines how the likelihood of sharing varies as a function of the number of friends shown, controlling for the actual number of sharing friends. We show that subjects are much more likely to share when they have at least one sharing friend, even when no social information is presented, suggesting that much of the variation may be attributed to factors unrelated to so-called influence response functions. In Section 4.4, we analyze the relationship between tie strength and influence by comparing the likelihood of sharing for subjects who did or did not see a friend with whom they had interacted at a certain frequency. We show that while strong ties are most predictive of an individual's sharing behavior, they are at most marginally more influential than weak ties.

4.2 Experimental design

4.2.1 Setup

We conduct our experiment on the population of Facebook users visiting pages with *like widgets*. Like widgets are interface elements that can be embedded on any web page and allow users to share the page on Facebook with a single click to a button, labeled "Like" or "Recommend". When an individual visits a page with a like widget while logged into Facebook, the like widget renders the names and faces of the individual's friends who have shared that page. Our experiment randomizes the amount of social information displayed to a small percentage of visitors by randomly removing friends' names and faces from the widget (Figure 4.1). For example, if a visitor has two friends that have shared a page, she may see zero, one, or two friends, depending on her assignment to an experimental condition. Using this design, we can compute the increase in the likelihood of sharing as a function of the number of friends displayed.

Our experimental population consists only of subject-URL pairs that were, or would have been, assigned to the *no feed* condition of the experiment discussed in the

(a) o = 0, k = 2 (b) o = 1, k = 2 (c) o = 2, k = 2

Figure 4.1: An example of the like widget interface for a user with two sharing friends in the three possible treatment conditions where (A) no friends are shown (B) one friend is shown (C) all friends are shown.

previous chapter. Therefore, our population consists solely of cases where a subject was never shown the URL on the Facebook feed, regardless of whether her friends ever shared that content. Subject-URL pairs in our population are then deterministically assigned to one of five experimental conditions, $0, 1, 2, 3, 4$, denoted by ω, which specifies the maximum number of friends shown to that subject. Therefore, the amount of social information given to a subject within the experiment is $o = min(\omega, k)$, where k is the actual number of sharing friends. Note that this implies that for subject-URL pairs with a given value of k that is less than 4, more pairs will be assigned to the $o = k$ condition than other conditions. Subjects who arrive at a page via Facebook through other means of sharing on Facebook (e.g. direct messaging) are also removed from the analysis.

4.2.2 Population statistics

We randomize the number of observed friends for subjects with a particular value of k sharing friends, so the population for any particular k may differ. Table 4.2.2 compares basic self-reported demographic information for subjects with varying values of o and k for up to three sharing friends[1]. Our experimental population only includes subjects where k is greater than zero; for comparison, the table also presents demographic information for $k = 0$, meaning subjects who visit web pages that none of their friends had previously shared. Within any particular k, the demographics across o remain relatively stable. Those with no sharing friends tend to be somewhat older and are slightly less likely to identify as male, and those with more than one sharing friend tend to all have similar demographic features. In total, there were

[1]We leave out the $k = 4$ case from the majority of the analysis, since there are too few observations to make comparisons with the necessary power.

1,156,608 unique subjects visiting 470,089 distinct URLs[2].

	$k=0$	$k=1$		$k=2$			$k=3$			
Treatment (o)	**0**	**0**	**1**	**0**	**1**	**2**	**0**	**1**	**2**	**3**
Num. Trials	166,970,891	300,923	1,220,197	51,808	55,167	163,066	18,877	19,971	20,308	41,451
Num. Subjects	63,247,399	216,941	811,954	37,261	39,985	112,265	13,664	14,972	14,633	29,883
Age										
13-17	9.7%	19.4%	20.1%	20.8%	21.9%	21.9%	23.1%	25.1%	26.1%	24.2%
18-24	37.2%	37.6%	38.5%	37.8%	37.8%	37.6%	37.6%	37.7%	37.4%	36.5%
25-34	28.9%	23.2%	23.0%	23.0%	22.4%	21.8%	22.0%	21.2%	19.8%	21.3%
35-44	13.0%	9.9%	9.9%	9.6%	9.8%	9.9%	8.9%	8.8%	8.3%	9.5%
45-54	5.9%	5.3%	4.7%	4.6%	4.5%	4.7%	4.6%	3.9%	4.7%	4.5%
55-64	3.0%	2.7%	2.2%	2.2%	2.0%	2.4%	2.1%	2.1%	2.0%	2.3%
65+	2.3%	1.9%	1.6%	1.8%	1.4%	1.7%	1.7%	1.1%	1.5%	1.7%
Gender										
Female	50.5%	47.2%	45.7%	48.9%	46.1%	48.3%	48.3%	46.6%	49.1%	48.6%
Male	47.7%	51.7%	53.1%	50.0%	52.5%	50.5%	50.5%	52.1%	49.9%	50.1%
Unknown	1.8%	1.1%	1.2%	1.2%	1.4%	1.1%	1.1%	1.3%	1.0%	1.3%

Table 4.1: Summary of demographics for subjects with 0, 1, 2, or 3 sharing friends (k), who were randomly assigned to a treatment condition (o) between 0 and k. Those with $o < k$ did not see all of their sharing friends, while those with $o = k$ were not affected by the experiment. More subjects are randomly assigned to $o = k$ for the reasons described in Section 4.2.1.

4.3 The marginal effect of social information

In empirical studies of diffusion, contagion effects are commonly inferred by computing the likelihood that an individual engages in a certain behavior given the current number of friends with that behavior. Such relationships have been examined in the context of numerous online behaviors, including participation in groups (*Backstrom et al.*, 2006; *Cha et al.*, 2009), games (*Wei et al.*, 2010), product purchases (*Leskovec et al.*, 2006a), tagging (*Anagnostopoulos et al.*, 2008), and the adoption of user-created content (*Bakshy et al.*, 2009). In all of these settings, the relationship between the number of contacts with a behavior and a user's likelihood of adopting that behavior is concave increasing, and widely believed to be suggestive of influence.

In our experiment, subjects visiting pages with like widgets were not exposed via the Facebook feed, but still exhibit similar correlated sharing behavior (Figure 4.2). Empirically, a subject's likelihood of sharing, given one sharing friend, is over five times greater compared to situations in which the subject has no sharing friends. While a subject's friends may be predictive of whether or not the subject chooses to share a page, this fact does not on its own demonstrate that the display of social information affects the decision making process. There are two possible hypotheses

[2]URLs that were classified as malicious or "spam" content by Facebook's security team were removed from the analysis.

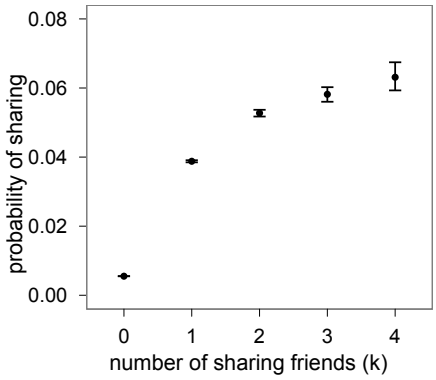

Figure 4.2: The probability of sharing a link as a function of the number of sharing friends for subjects who visit pages with like widgets that display all sharing friends ($o = k$). Error bars represent 95% bootstrapped confidence intervals.

that easily explain the phenomenon: (1) visitors are motivated to share a page when they see that friends have also shared the page, or (2) because of homophily, friends' sharing behaviors reflect the habits, tastes, and preferences of the visitor, and are therefore predictive, but not causal. Another more complex explanation is the possibility that one of the subject's Facebook friends had shared the page with the subject using a communication medium other than Facebook, which motivates her to share that page on Facebook.

We measure the effect of social information on sharing decisions by randomizing the number of friends shown to subjects. Figure 4.3 shows the marginal effect of social information (o), for subjects with some fixed number of sharing friends (k). In the figure, each panel represents some fixed value of k, the x-axis represents the number of friends displayed, and each point is the probability that a subject decides to share the page she has visited. For subject-URL pairs where $k = 1$, we find that the increase in the likelihood of sharing is only about 12% (relative risk ratio = 1.12, 95% $CI = [1.10, 1.15]$). Similarly, we find that for subjects with two and three sharing friends, there is a significant difference between being shown zero and one friend: for $k = 2$ the increase in likelihood is estimated at about 12% (relative risk ratio = 1.12, 95% $CI = [1.03, 1.21]$), and for $k = 3$ the increase is about 9% (relative risk ratio = 1.09, 95% $CI = [1.03, 1.15]$). In both cases, the estimated difference is significant at a $p < 0.005$ significance level. Despite the presence of these effects, a subject's behavior is mostly predicted by the number of sharing friends regardless of whether those friends are visible. Furthermore, for cases with more than one sharing

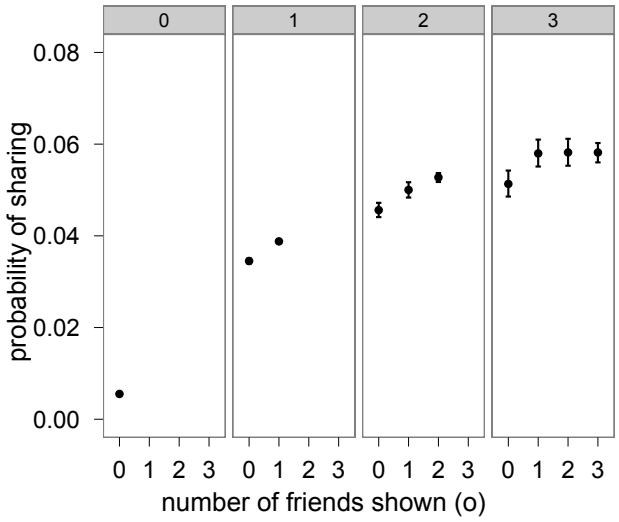

Figure 4.3: The probability of sharing a link as a function of the number of observed sharing friends (o), given the number of actual sharing friends (k). Error bars represent 95% bootstrapped confidence intervals.

friend, increasing amounts of social information beyond the first displayed friend do not have a statistically significant effect on individual decision making. For example, in the case of $k = 2$, we fail to detect a significant difference between cases where one friend is shown and cases where two friends are shown. These results suggest that, in situations when individuals have already been exposed to content, the presentation of a single friend's sharing behavior provides a significant yet modest increase in the likelihood of sharing, and that there is no significant marginal effect of presenting additional peers beyond the first.

4.4 The effect of tie strength

The strength of a tie is also relevant to influence. Strong ties are more likely to be perceived as influential, and have a higher likelihood of spreading information (*Weimann*, 1982; *Brown and Reingen*, 1987; *Hill et al.*, 2006). However, given that an individual has already been exposed to content, and is aware of their friends' willingness to share that content, to what extent does the strength of her relationship with that contact impact the decision to share? Because of homophily, we expect that strong ties better reflect individuals' preferences, and are thus more predictive of subjects' behavior. To separate these effects, we first examine the extent to which strong ties are predictive, then examine how the display of strong versus weak ties

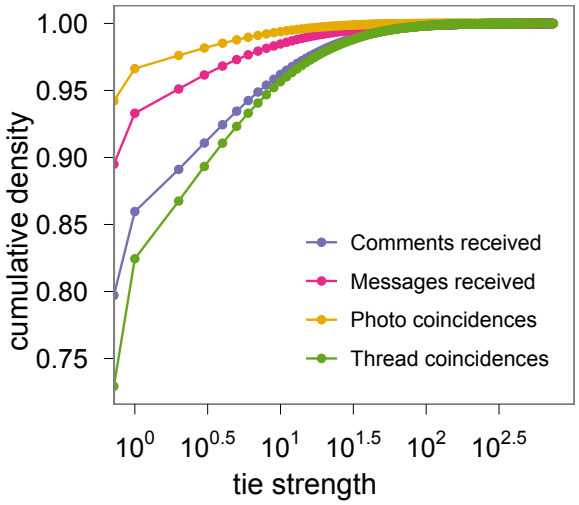

Figure 4.4: The distribution of the four tie strength measures between subjects and their alters in the $k = 1$ conditions.

affects subjects' sharing decisions.

We examine four measurements of tie strength using data from three months directly prior to the experiment: (1) the number of comments the subject received from their alter, which reflect public online interaction; (2) the number of messages the subject received from their alter, which reflects private interpersonal communication; (3) the number of photos the subject and alter were tagged in together, which reflects real-world coincidences; and (4) the number of thread coincidences, which indicates the number of times the subject and alter had participated in the same public discussion thread. The distribution of these measurements for subjects in the $k = 1$ group is shown in Figure 4.4.

4.4.1 Predictive power of tie strength

To test whether the sharing behaviors among strong ties more closely reflect one another compared to the behaviors among weak ties, we consider the likelihood of sharing for subjects with one sharing friend ($k = 1$) who is a weak or strong tie. Figure 4.5 shows the probability of sharing for subjects with alters with whom they had or had not interacted in the past three months (tie strength > 0). From this figure, we can see that individuals with strong tie alters are more likely to share, and that this effect exists even when no social information is shown to the subject.

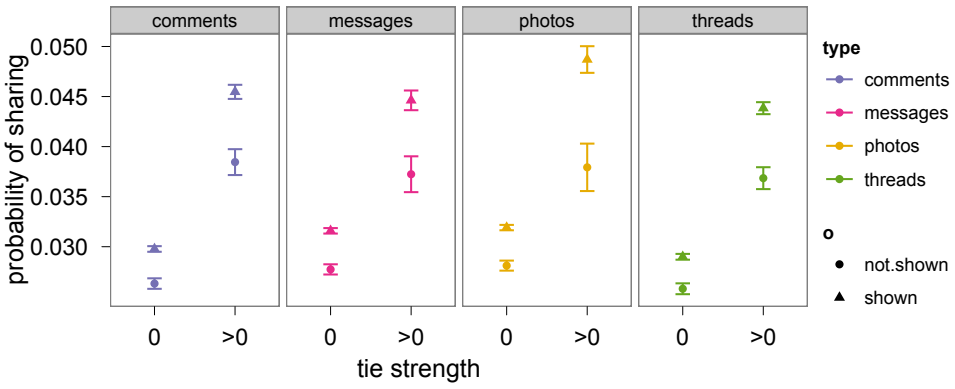

Figure 4.5: The relationship between tie strength and likelihood of sharing. Subjects are more likely to share when their alter is a friend with whom they have interacted at least once, and are more likely to share when that friend is shown. Error bars represent 95% bootstrapped confidence intervals.

If subjects who communicate often with friends that share pages on Facebook are themselves more likely to share on Facebook, then this correlation may partially explain the relationship we observe in Figure 4.5. Using logistic regression, we simultaneously estimate the effect of tie strength as well as viewing a friend, while controlling for subjects' propensity to share. Equation 4.1 gives a logistic regression model of the effect of a friend with whom the subject had interacted at least once (*interacted*), whether the alter was shown to the subject (*shown*), and the number of times the subject had used a like widget over the three months directly prior to the experiment. We fit the model in Eq. 4.1 separately for each of the four measurements of tie strength. The regression coefficients are summarized graphically in Figure 4.6A.

$$shared \sim \beta_0 + \beta_t interacted + \beta_s shown + \beta_p log(1 + prev.likes) \quad (4.1)$$

$$shared \sim \beta_0 + \beta_t interacted + \beta_s shown + \beta_{t:s} interacted * shown + \beta_p log(1 + prev.likes) \quad (4.2)$$

Our results support the hypothesis that individuals are significantly more likely to share the same content as their stronger ties, even when no friends are displayed, and that this effect is not simply a result of correlations between subjects' propensity to share and tie strength. We use a similar regression model with an interaction term (Eq. 4.2) to test if the presence of social information affects pages shared by stronger ties compared to weaker ties. Figure 4.6B shows that there is no significant effect for comments or messages, and that there is a weak, but significant positive effect for subjects with alters whom they had appeared in at least one photo or thread discussion. This positive effect implies that for

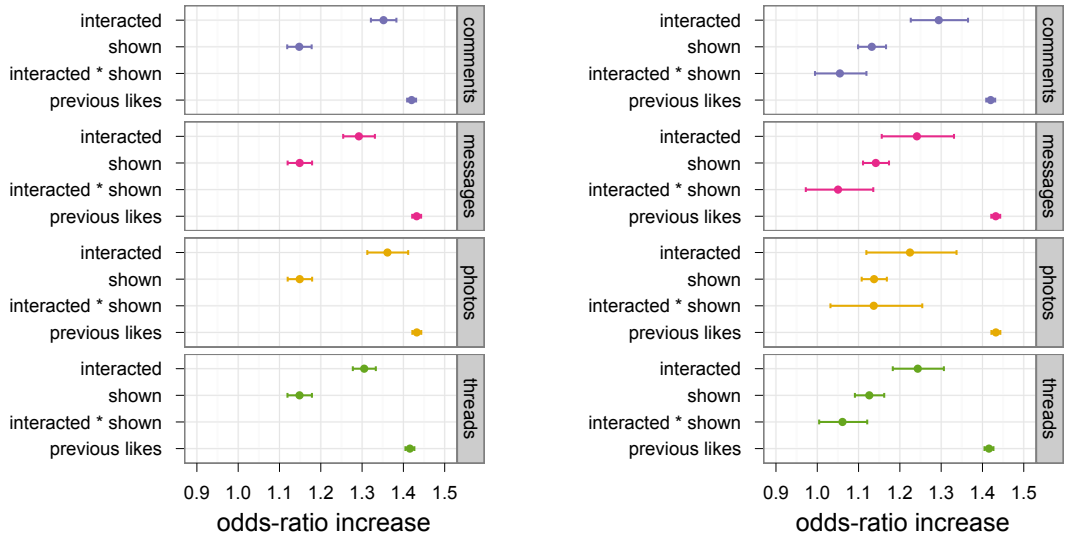

Figure 4.6: The effect of social information and tie strength on the probability of liking a page, for subjects with one sharing friend, using the logistic regression models in Eq. 4.1 (left) and Eq. 4.2 (right). *Previous likes* is one plus the logarithm of the number of times the subject had used like widgets over the three months directly prior to the experiment. For each regression coefficient β, we plot the point estimate, $exp(\beta)$, and its corresponding 99% confidence interval.

these two measures of tie strength, the presence of social information has a stronger impact on the subject's propensity to share when that information pertains to a strong tie rather than a weak tie. This may suggest that strong ties exert more influence on a subject's decision to share, or it may simply mean that the presence of social information has greater impact for pages that are shared by strong ties.

4.4.2 Influence of strong ties

To answer the question of how much more or less influence strong ties exert, we consider subjects who visit a page that has been shared by exactly one weak tie and one strong tie, and are only shown one of those two friends on the like widget (that is, a subset of those in the $k = 2, o = 1$ condition). In these cases, whether the strong or weak tie was shown is randomly assigned, so we can directly evaluate the effect of tie strength by comparing the difference in likelihood of sharing among the two groups.

Table 4.2 summarizes the propensity to share for subjects that were randomly assigned to see either one weak or one strong tie. We do not observe a statistically significant difference in likelihood of sharing when comparing whether the displayed friend is a strong

or weak tie, along any of the four measures of tie strength. To show the robustness of this result, we report the minimum detectable difference our data affords, at 90% power and a 95% significance level. The likelihood of sharing given the display of a weak tie varies between 6.05% and 7.23% while the minimum detectable difference if the displayed friend is a strong tie is roughly an order of magnitude smaller, varying between 0.44% and 0.90%. This suggests that if our test fails to detect a true difference between the influence exerted by strong versus weak ties, that difference is modest at most.

	comments		messages		photos		threads	
Tie Strength	**0**	**> 0**	**0**	**> 0**	**0**	**> 0**	**0**	**> 0**
Num. trials	9800	9680	5217	5414	3487	3496	11376	11401
Num. clicks	614	591	331	322	252	265	688	702
% shared	6.27%	6.11%	6.34%	5.95%	7.23%	7.58%	6.05%	6.16%
$p_{>0} - p_0$	−0.15%		−0.39%		+0.35%		+0.10%	
Min. det. diff.	±0.49%		±0.68%		±0.90%		±0.44%	
p-value	0.66		0.42		0.60		0.75	

Table 4.2: The difference in the propensity to share for subjects who were shown their strong or weak tie, among subjects with one weak or one strong tie. All differences are not significant. We also give the minimum detectable difference computed at the 90% power level to show that even with more randomized trials, the difference between the the two treatment conditions would be small.

Admittedly, our formulation of strong ties is generous, since it only requires that pairs of users interact once within the past three months. Since there is no a priori obvious cutoff for what should be considered a strong tie, we also perform comparisons between those who have not interacted and those with some minimum number of interactions (τ). As we increase this threshold, the effect size tends to increase slightly, but the number of subjects in our population decreases rapidly, thereby widening the minimum detectable difference (Table 4.3). Across all four measures and a wide range of cutoffs, we are unable to detect a significant difference between the amount of influence due to exposure to information about strong and weak ties.

4.5 Conclusion

Unlike many theories of social contagion, the behavior of multiple friends or strong ties hold little influence in subjects' decision to follow the behavior of their friends. In combination with our experimental results from the previous chapter, it appears that the primary role of social networks in information diffusion is to expose individuals to relevant content. Relational aspects, such as the number of sharing friends or strength of tie with a

	Comments		*Messages*		*Photos*		*Threads*	
τ	$p_{>\tau} - p_0$	N	$p_{>\tau} - p_0$	N	$p_{>\tau} - p_0$	N	$p_{>\tau} - p_0$	N
1	+0.04%	6277	-0.50%	3183	+0.50%	1920	+0.31%	6433
2	-0.27%	4735	-0.27%	2220	+0.78%	1309	+0.26%	4459
3	+0.30%	3750	+0.57%	1753	+0.90%	1011	0.00%	3407
4	+0.42%	3075	+0.53%	1420	+2.01%	781	-0.04%	2778
5	+0.16%	2593	+0.51%	1213	+1.59%	638	+0.11%	2296
6	+0.63%	2286	+0.92%	1007	+1.96%	539	+0.08%	1977
7	+0.25%	1998	+1.15%	845	+2.32%	471	+0.35%	1712
8	+0.46%	1782	+1.65%	751	+2.28%	392	+0.70%	1486
9	+0.66%	1607	+1.54%	686	+3.14%	351	+0.85%	1312
10	+0.92%	1442	+1.42%	627	+1.75%	293	+1.41%	1189
11	+0.84%	1327	+1.72%	579	+2.04%	266	+1.79%	1090
12	+0.56%	1207	+2.05%	532	+1.68%	245	+1.18%	977
13	-0.02%	1138	+2.18%	489	+3.26%	219	+0.75%	861
14	-0.15%	1054	+1.96%	452	+1.87%	209	+0.20%	771
15	+0.39%	1000	+1.30%	424	+1.47%	188	+0.38%	674
16	+0.23%	927	+1.68%	387	+0.51%	174	+0.30%	637

Table 4.3: Even under more strict definitions of tie strength, for which pairs of individuals must have at least some minimum threshold of interaction (τ) to be considered strong ties, we are unable to detect a statistically significant difference in the propensity to share for subjects who are shown a weak versus a strong tie. N is the number of comparisons available for our test of proportions, and falls off rapidly with τ due to skew in the distribution of tie strengths. All differences have $p > 0.1$.

sharing alter are predictive of whether or not an individual will take interest in and share content. Consequently, these features may be critical to how information spreads, but have little causal impact on an individual's ultimate decision to share once she has arrived at a particular piece of content.

As practitioners move toward enhancing technologies with social data, we expect that the use of experimentation will become increasingly important for how information systems are evaluated. In our study, a naive estimate might suggest that the display of a single friend increases use by over 800%, but more careful analysis through experimentation shows that the increase is only about 12%. Such differences do not just represent a problem with measuring diffusion, but reflect a fundamental problem caused by homophily that can affect the evaluation of systems that draw upon friends' behavior.

CHAPTER V

Allocating Attention

An individual's *personal network* — their set of social contacts — is a basic object of study in sociology. Studies of personal networks have focused on their size (the number of contacts) and their composition (in terms of categories such as kin and co-workers). Here we propose a new measure for the analysis of personal networks, based on the way in which an individual divides his or her attention across contacts. This allows us to contrast people who focus a large fraction of their interactions on a small set of close friends with people who disperse their attention more widely.

Using data from Facebook, we find that this balance of attention is a relatively stable property of an individual over time, and that it displays interesting variation across both different groups of people and different modes of interaction. In particular, activities based on communication involve a much higher focus of attention than activities based simply on observation, and these two types of modalities also exhibit different forms of variation in interaction patterns both within and across groups. Finally, we contrast the amount of attention paid by individuals to their most frequent contacts with the rate of change in the identities of these contacts, providing a measure of *churn* for this set.

5.1 Introduction

People maintain a broad range of personal relationships. In the language of social networks, these relationships can be thought of as the links connecting an individual to her network neighbors, a set of people we will refer to as her *contacts*. A significant body of research in sociology has focused on an individual's contacts — her *personal network* — as an important attribute in settings that range from professional opportunities *Granovetter* (1973); *Burt* (1992) to social support and advice on important matters. *Fischer* (1982); *McPherson et al.* (2006); *Wellman and Wortley* (1990).

This chapter is published as *Center of Attention: How Facebook Users Allocate Attention across Friends* in the ICWSM 2011 Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (*Backstrom et al.*, 2011).

This line of work has considered variations in both the *size* and the *composition* of personal networks. Size is most naturally defined simply as the number of contacts (*Killworth et al.*, 1990). Composition has generally been studied in terms of discrete variables that include the number of kin and non-kin contacts, and the distinction between close friends and more distant acquaintances. Earlier research has considered how the composition of personal networks differs across attributes including age, race/ethnicity, gender, and educational level (*Marsden*, 1987; *McPherson and Smith-Lovin*, 1993; *Moore*, 1990) while more recent work has examined personal network composition within the context of social media (*Chang et al.*, 2010; *Gilbert et al.*, 2008).

We propose a new measure for analyzing personal networks that addresses a dimension distinct from network size and composition. This measure expresses the way in which an individual divides his or her attention across contacts. Everyday experience suggests that some people focus most of their attention on a small circle of close friends, while others disperse their attention more broadly over a large set. As a specific property of an individual, this contrast between focus and dispersion — the individual's *balance of social attention* — is distinct from the properties discussed above: two people with personal networks of similar size and composition can differ greatly in the extent to which their attention is focused on a small or large subset of their personal network. Furthermore, the balance of social attention is not a purely structural measure, since it takes into account both the links in the underlying social network and the amount of time that an individual allocates to these links.

We believe this type of measure can play a useful role in illuminating the fine structure of an individual's personal network in both on-line and off-line settings. An understanding of how the balance of attention varies across individuals can also help to inform the design of social media applications, many of which must manage a tradeoff between diversity and relevance. These applications attempt to avoid stale content, while at the same time ensuring that everything that appears is personally relevant. Designers of such social products can use an individual's balance of social attention to help customize this tradeoff on a per user basis. For example, stratification of users by a measure capturing balance of social attention recently led to increased interaction with the Facebook News Feed.

Although a metric for balance of social attention is potentially useful and theoretically interesting it has been difficult to study empirically. Even the size and composition of friendship networks are notoriously difficult to measure, and generally have been captured through self-reports aided by elicitation mechanisms (*Campbell and Lee*, 1991). Measuring the balance of attention requires an even higher resolution, as it depends on a careful estimation of the volume of interaction between an individual and each member of her personal network. In order to overcome these measurement difficulties we use data from Facebook to analyze the interaction volume. After reviewing further related work, we turn in the

next section to a precise formulation for the balance of social attention. Subsequent sections present analysis that shows how this measure exhibits interesting patterns of variation across groups of people and across different modalities of interaction.

Further related work. Recent work in on-line social networks has articulated the contrast between the links in a network and the activity that takes place on these links. This is also the distinction that motivates our work, although our focus differs from earlier papers to address this issue: *Kossinets et al.* (2008) study how link activity can lead to different pathways for information flow over multi-step paths, and *Wilson et al.* (2009) focus on aggregate measures for how activity is distributed, and the network structures that result from thresholding the links by activity level. In contrast, we are interested in the distribution of attention levels as an attribute operating at the individual level — in understanding how this attribute varies across people and groups, and how it relates to other individual attributes.

From a theoretical perspective the balance of social attention is related to the distinction between strong and weak ties (*Granovetter*, 1973), but this is not simply a different measure of tie strength. Although tie strength is ultimately a synthesis of several factors, including volume of interaction and affective closeness (*Marsden and Campbell*, 1984), our measure begins from the aspects of tie strength related to volume and synthesizes them into a node-level measure in the network that takes into account an individual's full set of ties. Furthermore, our approach also relates to arguments by *Milgram* (1970) and *Mayhew and Levinger* (1976) that settings such as dense urban areas, which produce many interactions ought to result in less time spent on any one of these interactions. Our measure enriches these considerations by formulating multiple ways in which an individual can manage a large personal network: either by slicing her attention relatively evenly over all contacts, or by focusing on a few at the expense of the others. Finally, our measure is related to other quantitative trade-offs between focus and dispersion in an individual's personal network, such as the geographic spread of one's friends and the searchability of social networks (*Kleinberg*, 2006; *Backstrom et al.*, 2010). The focus of our work is to quantify this trade-off in terms of the volume of interaction, rather than embedding the analysis in external frames of reference such as geography or social categories.

5.2 The balance of social attention

Consider a population of n individuals, and a person i in this population who sends messages to her contacts. (Later we will consider a range of different interaction modalities, but for purposes of exposition it is useful to think about messages.) Suppose m_j is the number of messages sent by person i to person j in her set of contacts. If the total number

of messages sent by i (over all contacts) is m, we say that the fraction of i's attention that she devotes to j is $a_j = m_j/m$.

As a function of k, what fraction of i's attention does she devote in total to her k most frequent contacts? We sort all of i's contacts j in order of decreasing a_j, and we say that i's *top k contacts* are the people corresponding to the first k positions in this sorted list. The fraction of i's attention devoted to her top k contacts, denoted f_k, is the sum of a_j over all individuals j in this set of top k contacts. If i has n contacts, then the vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ are each complete descriptions of how i divides her volume of interaction across her contacts, and these vectors serve as our starting point for measuring the balance of social attention.

The full vectors turn out to be a highly redundant representation. For much of our analysis, we find that individual coordinates of the vector $\mathbf{f}$ can serve as relatively stable summaries of aggregate properties computed from the full vector. Specifically, if we compare individuals simply by the single number f_k, we get extremely similar aggregate comparison results for all k in a broad middle range where most of the volume of interaction takes place, i.e., in the interval between $k = 5$ and $k = 25$. There is a natural reason for this: in general, if user A has lower f_k value than user B, then A will also typically have a lower f_ℓ value compared to B, when k and ℓ are in this middle range from 5 to 25. As a result, any coordinate from this range produces roughly similar results, which allows us to collapse a collection of measurements to something that is effectively a single-dimensional question.

5.3 Balance of attention across modalities

5.3.1 Data

To examine how an individual balances her attention across her friends, we compute metrics for a number of different modalities of attention. These modalities can be divided into two distinct groups: communication and viewing. The communication modalities encompass directed interaction, such as sending a private message or posting a public comment on a photo, while viewing behavior is derived from users visiting pages on Facebook. Thus, in the communication modalities the target is aware of the user's actions (since they receive the communication), but in the viewing modalities they are not; only the user is aware of the viewing activity. (See also *Jiang et al.* (2010) for further discussion of this contrast in on-line social networks.)

- *Messages.* Individuals can send each other private messages similar to email.
- *Comments.* When a user shares a piece of content, such as posting a photo or a link, other users can typically leave a public comment on the item.

- *Wall Posts.* A user's Facebook profile includes a publicly viewable 'wall', on which other users can post content.
- *Profile Views.* This measures how many times one user views another's profile page.
- *Photo Views.* This measures how many times a user views photos posted by another user.

An individual might focus her attention toward differing subsets of her contacts through each of the modalities mentioned above. Therefore, we compute attention measurements independently for each modality by collecting the sum of all actions for each user-target pair within each modality from January 2010 to December 2010. All data has been anonymized and aggregated prior to analysis. For comments and messages, each individual post counts as a single action directed at a given target. For wall posts we consider only the subset of items posted outside of the target user's birthday window, defined as the time span from two days prior through one day after the target's birthday.[1]. We exclude birthday wall posts from our measurements because they are typically triggered by a birthday reminder on Facebook rather than some user-specific mechanism, and are therefore not representative of the directed attention the communication modalities generally capture. The viewing modalities require the user to make a direct navigation to view a target user's profile page or a photo owned by the target user. Simply encountering a target user or a target user's photo in the News Feed or on another user's profile does not constitute a view. Henceforth, when we talk about a user's level of *activity* in a given modality, we refer to the number of discrete actions the user performed in this modality, as measured according to the definitions above.

In order to minimize the impact of behavioral trends related to the overall growth of Facebook, we restrict our analysis to users who were already members as of January 1, 2009. In addition, we are interested in measuring the behavior of users for whom Facebook represents a non-trivial medium of communication and social attention, so that we can see the balance of attention among people in a context where this is a relevant quantity. Therefore, we select only those Facebook users who have visited the site on at least 80% of the days in 2009 and 2010. This user sample represents a population of 16 million heavily active Facebook users.

Figure 5.1 shows the distribution of activity for each modality, with the percentile rank within the modality along the x-axis and the total number of actions within each modality on the y-axis. A given user's volume of viewing actions is likely to be an order of magnitude higher than her volume of communication actions, while non-birthday wall posts are least common. As even active users may not use some of the features (for instance, 27% of these

[1]This time span represents the average time window in which the number of wall posts received is significantly higher than normal.

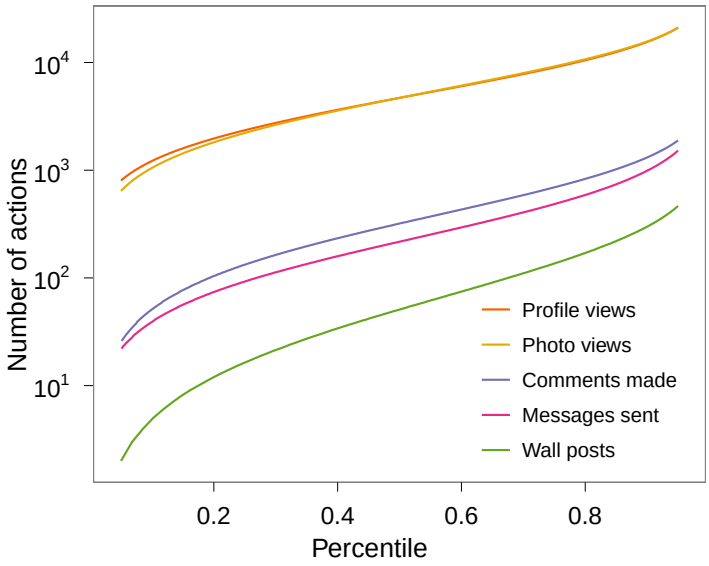

Figure 5.1: Distribution of volume of activity per user for each modality between January 2010 to December 2010.

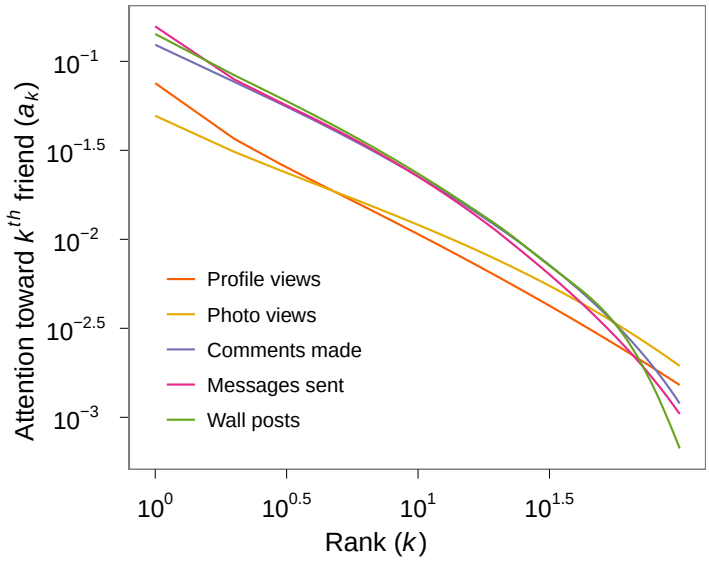

Figure 5.2: Fraction of attention devoted to a given contact, based on the contact's rank in terms of overall volume of attention received.

active users users sent less than 100 messages in a year), some of our subsequent analyses will further restrict the user set studied.

5.3.2 The average balance of attention

Figure 5.2 shows a_k, the fraction of attention given to the k^{th} friend, as a function of k for the five modalities. We only consider the users in the 70^{th} to 95^{th} percentiles of activity level for each modality. In doing so, we filter out individuals who do not significantly use each modality as well as the extreme outliers at the top end.

The communication curves begin somewhat higher than the viewing curves, but tail off more quickly at the higher ranks. This happens because many users, even within this relatively active set, have not communicated with more than 50 unique targets, causing us to average in zeros. All of the communication modalities and profile viewing have very similar slopes for low k in a log-log plot, each fitting Cx^{α} for α between 0.75 and 0.78. Thus, while the viewing modalities account for an order of magnitude more activity than the communication modalities, and the quantities $a_1, a_2, a_3, \ldots$ are smaller in absolute terms, they fall off at a very similar rate (proportional to about $k^{-3/4}$) for both viewing and communication. The one modality that behaves differently is photo views and one possible explanation is that the viewing target is less clear: user A might look at a photo created by user B not because of interest in B, but because of interest in one of the photo's subjects.

Although the curves in Figure 5.2 are restricted to users in the upper percentiles of activity level, they still aggregate over users with varying activity levels, which may hide the importance of a user's overall activity level in impacting the shape of her attention curve. Indeed, the impact of overall activity on these curves is not immediately clear. It may be that people who communicate more are doing so because they communicate more with their lower ranked contacts. But it could also be the case that most people are only capable of maintaining a small number of direct contacts, and increased activity occurs mostly within this fixed set.

To understand the impact of activity on attentional balance, we consider the fraction of activity f_{15} as a function of activity level. (Results for f_k are very similar for all k in the range between 5 and 25, and somewhat beyond this as well.) In order to enable comparisons between modalities like messaging (which a typical user performs a few hundred times a year) and activity types like profile viewing (which occur in the thousands), we examine f_{15} as a function of a user's percentile rank of activity level within each modality. Figure 5.3 shows that within each modality, there is a sharper initial decrease for users at low activity levels, but then a long section that is relatively more gradual. Indeed, while users with low activity necessarily have high f_{15} (those with 15 or fewer contacts have $f_{15} = 1$), the middle region of activity levels decreases quite slowly, and in the case of messaging and especially profile views is approximately flat. One can also look at the distribution of f_k

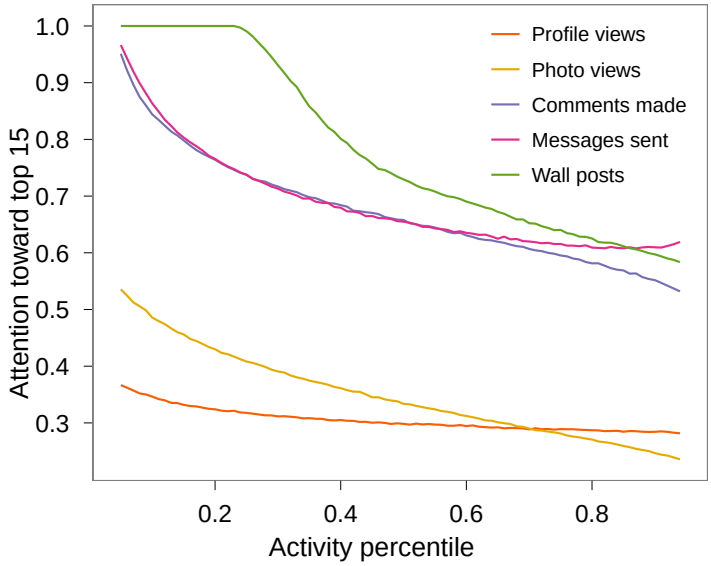

Figure 5.3: Average fraction of attention devoted to top 15 contacts within a given modality, against level of activity.

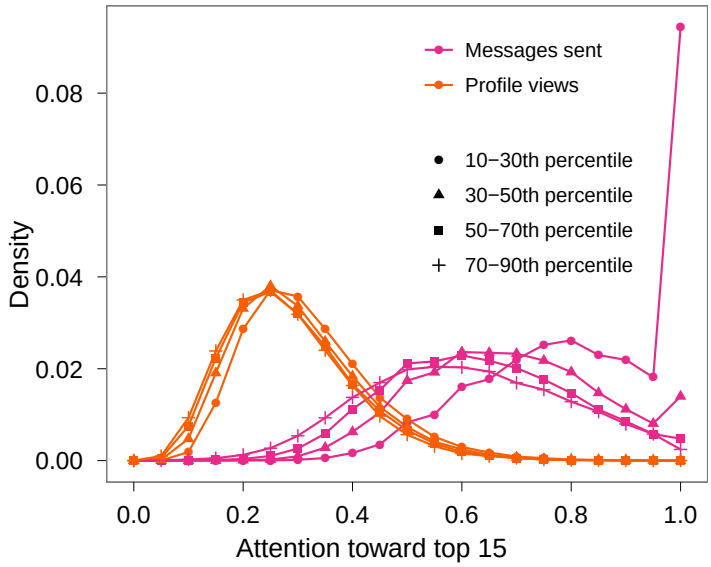

Figure 5.4: Distribution of fraction of attention devoted to top 15 contacts for messaging and profile views, broken down by activity level.

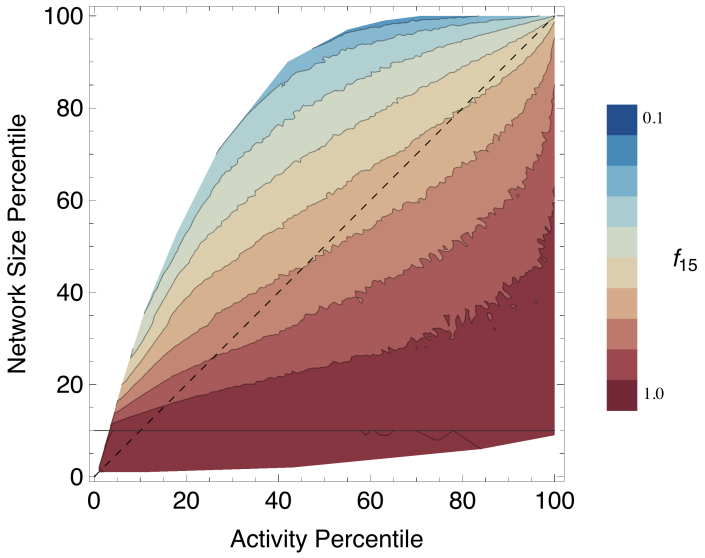

Figure 5.5: Average fraction of messages sent to top 15 contacts as a function of network size and activity. Horizontal line indicates the percentile at which the number of contacts exceeds 15.

values at a given activity level rather than their mean: Figure 5.4 shows this for varying levels of activity within messaging and profile viewing. While higher activity groups are slightly left-shifted, the distributions are qualitatively similar once we move beyond the 30^{th} percentile (many below the 30^{th} percentile have messages to 15 or fewer people), and are nearly identical for profile views.

These graphical comparisons show evidence of a broad distinction between viewing and communicating modalities. In general, communication is much more focused, with a high fraction going towards top contacts, while viewing is significantly more dispersed across contacts.

In addition to a user's activity level within a given modality, the size of her personal network within that modality will also affect the value of f_{15}. Naturally, as the network size increases for a fixed activity level, f_{15} tends to decrease, since the individuals added to the network must receive some share of this fixed activity. On the other hand, Figure 5.5 shows that among users with comparable personal network size, those with higher activity level are more focused. Thus, larger networks tend to lead to smaller f_{15}, while more activity tends to lead to larger f_{15}. Due to the high correlation (0.83 for messaging, 0.91 for profile viewing) between network size and activity level, this effect is lost when looking at f_k only as a function of activity level.

5.4 Variation by individual characteristics

5.4.1 Variation across individuals

The distributions in Figure 5.4 show that, even for a fixed activity level, some individuals seem significantly more focused than others in their attention. Although it's possible that this variation arises primarily from the inherent randomness in all of our interactions over time, the more intriguing possibility is that some individuals are genuinely more focused or dispersed than others, and that these differences persist over time.

In order to examine whether users who are active in two distinct time periods have consistent attention patterns across both observation windows, we compare data from early 2010 and late 2010. A simple regression that attempts to predict a user's f_5 value in Oct-Dec 2010 from just her f_5 in Jan-Mar 2010 yields R^2 values of 0.45 and 0.23 for viewing and messaging that — while relatively modest in absolute terms — show a non-trivial level of stability in this quantity over time. This is all the more notable given that this computation has access only to this single f_5 number for predicting the corresponding value close to a year later. Moreover, using only the user's activity level and personal network size in Jan-Mar 2010 performs worse at predicting f_5 in Oct-Dec 2010 than simply using the Jan-Mar f_5 by itself for this prediction (0.23 vs. 0.19 and 0.45 vs. 0.31).

5.4.2 Age and gender

Figure 5.6 shows the average value of f_{15} for users between the ages of 13 and 60. We restrict to users in 70^{th}-95^{th} percentiles of activity in each modality, for the reasons discussed above. Each modality exhibits a roughly monotonic relationship with age, but the relationship for viewing moves in the opposite direction of communication: we find that older users are more focused in their viewing behavior, but more dispersed in their communication behavior. Moreover, these two directions of change appear at different rates as we consider older users: the decreasing focus in communication is rapid over ages ranging roughly from 13 to 30, with slower changes beyond this point, while the increasing focus in viewing is much steadier over the full range of ages considered.

Compared to males, females tend to focus more of their attention toward their top k friends in all modalities (Figure 5.7). This difference may be partly explained by differences in the underlying distribution of activity and network size for each gender. For example, female Facebook users tend to maintain larger active networks than their male counterparts.[2] To adjust for this, we perform a regression analysis to explain the relationship between f_5, number of contacts, activity level, age, and gender (Table 5.1), using data from all users that self-report their gender and fall within the $5^{th}-95^{th}$ percentile in terms of total activity and number of contacts. The regression indicates that the fraction of attention allocated

[2]See http://www.facebook.com/note.php?note_id=55257228858 for a comparison.

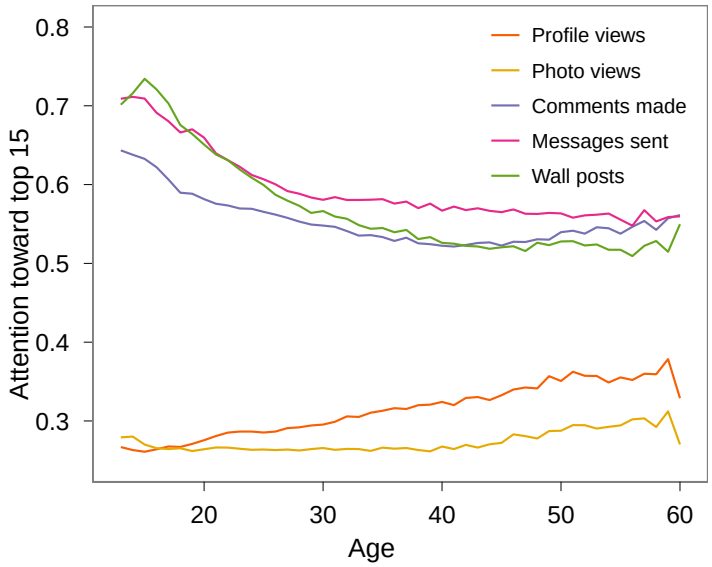

Figure 5.6: f_{15} as a function of age, for users in the 70th to 95th percentile of activity in the given modality.

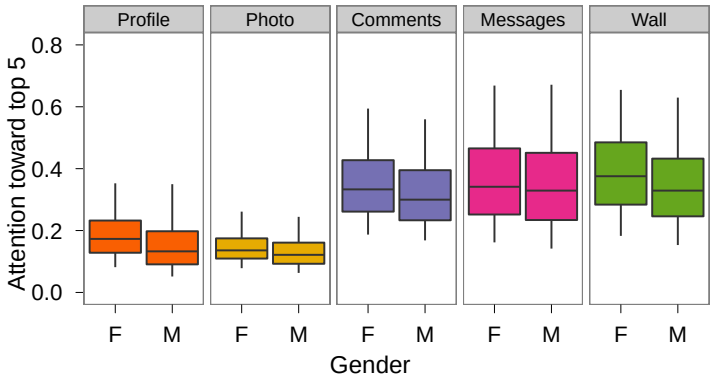

Figure 5.7: Distribution of attention given to top 5 friends for females and males, for users in the 70^{th}-95^{th} percentile of activity in the given modality.

Modality	*inter*	*con*	*act*	*age*	*male*	R^2
Profile	0.18	-0.53	0.44	0.03	0.02	0.38
Photo	0.20	-0.47	0.21	-0.01	0.01	0.53
Comment	0.43	-0.81	0.41	-0.03	-0.01	0.67
Message	0.44	-0.87	0.48	0.03	0.00	0.59
Wall	0.51	-1.48	0.92	-0.02	0.00	0.62

Table 5.1: Regressions explaining the variation in the fraction of f_5 as a function of individual characteristics ($N = 1,037,885$) for different modalities. Activity (*act*) and number of contacts (*con*) are centered percentiles within each modality, and range between -0.45 and 0.45. Age is given in terms of centered percentiles, with -0.5, -0.25, 0, -0.25, and 0.5 corresponding to 13, 21, 25, 33, 65 years, respectively. The intercept (*inter*) shows the expected f_5 score for a 25 year old female with a median number of contacts and activity level. All coefficients are significant at the $p < 10^{-16}$ level and have standard errors that are at least two orders of magnitude less than the coefficient.

to the top five friends depends to a large extent on the number of contacts and activity level of an individual: more contacts are associated with lower values of f_k, and this effect is balanced by higher levels of activity (as seen in Figure 5.5). For a given level of activity and number of contacts, gender and age both have a small but significant effect. For example, a male user with the same activity level and number of contacts as a female would be expected to have a f_5 score that is 0.02 higher than that of a female. Thus, while there are significant differences in f_5 that depend solely on age and gender, the primary effect on f_5 seems to stem from the fact that total activity and number of contacts vary significantly with age and gender. Note that in some cases, the coefficients on age and gender appear to contradict Figure 5.6 and Figure 5.7. However, this is not a contradition and shows that, for example, while older users tend to be less focused in their messaging overall, comparing users with the same activity level and number of contacts, the older ones will be slightly more focused.

5.4.3 Interactions within and between genders

In general, the structure of social ties among people of the same gender is quite different than the structure of social ties across genders (*McPherson et al.*, 2001). Thus, we further refine the gender analysis to separately consider the interaction of users within their own gender and across genders. We find that females send 68% of their messages to females, while males send only 53% of their messages to females. This distinction is consistent with *gender homophily* — in which each gender has a bias toward within-gender communication — modulated by the overall distribution of Facebook messages. On the other hand, we see much smaller differences in viewing: for typical activity levels, both females and males

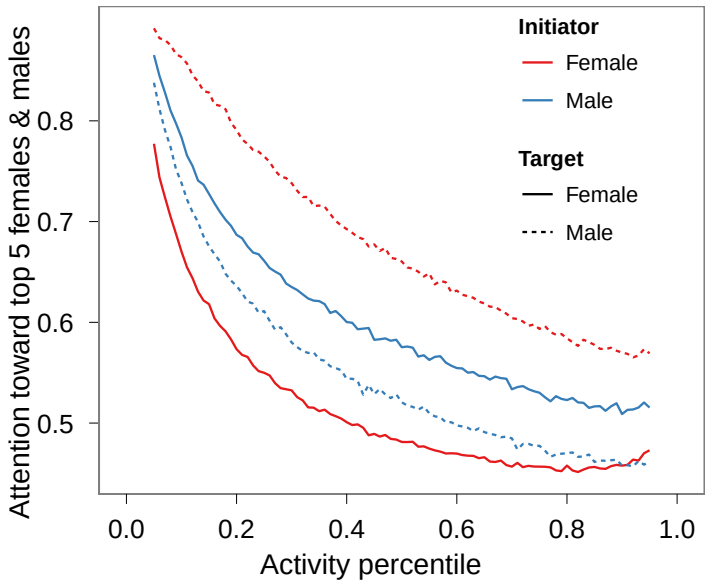

Figure 5.8: f_5 for messages, fixing the gender of both the initiator and the target of the action.

direct roughly 60% of their profile viewing activity to female users.

We then examine how a person balances their social attention separately to their contacts of each gender. That is, we partition each user's set of actions into two subsets — one for the actions directed at females, and one for the actions directed at males — and then compute the quantities a_k and f_k separately for these subsets. Figure 5.8 shows the results for messaging: the average f_5 value for actions by users of gender X toward users of gender Y, for each choice of X and Y. We see that there is greater concentration in across-gender communication than within-gender communication. Furthermore, females are more concentrated than men with respect to across-gender communication, and more dispersed than males with respect to within-gender communication. Viewing behaviors provide (in Figure 5.9) an interesting contrast with messaging: females and males have roughly equivalent levels of focus in viewing profiles of female users, but markedly differing levels of focus in viewing profiles of male users, where female viewers are much more focused.

5.4.4 Relationship status

The effect of gender on interaction patterns is further influenced by factors such as marital status — unmarried people display different network structures than married ones (*McPherson and Smith-Lovin*, 1993). To understand the effect of these factors, we consider the subset of active users in our population whose listed *relationship status* on Facebook remained unchanged throughout 2010 and was set to one of the following three values:

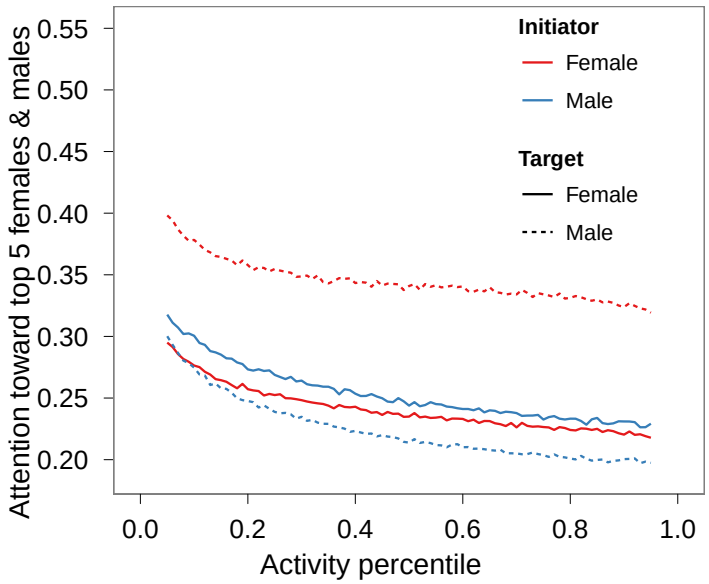

Figure 5.9: f_5 for profile views, fixing the gender of both the initiator and the target of the action.

single, in a relationship, or married.[3] Therefore, we consider 12 categories of behavior for each modality: for users of gender X and relationship status S we look at the balance of attention in their interactions with users of gender Y. Table 5.2 shows the average f_5 values for these 12 categories for messaging, and Table 5.3 shows them for profile viewing. For messaging, we see a refinement of the gender homophily effects in Figure 5.8. For viewing, a striking "non-monotonic" effect shows up clearly in interactions across the genders: for both females viewing males' profiles and males viewing females' profiles, the level of focus for single and married users is roughly the same, while the focus for users in a relationship is significantly higher than either of these.

5.5 Attention over time

We have shown that the fraction of attention to users' top k contacts, f_k, decreases as a function of activity when averaged over all individuals. One might expect that as a result the top k contacts will tend to change more rapidly over time for those with higher activity. However, we find that increased levels of activity are actually associated with higher levels of stability over time. We examine the overlap between a user's top k contacts in two consecutive time periods; Jan-Feb 2010 constitute the first time period, and Mar-Apr 2010 form the second period. We find the relationship between activity level, number of

[3]Note that since most relationships in a broad population are heterosexual (*Black et al.*, 2000), such relationships will be the bulk of our computed averages.

Initiator-Target	Rel. Status	f_5 (Msg.)
$F \to F$	married	0.476
$F \to F$	single	0.523
$F \to F$	relationship	0.524
$M \to M$	married	0.570
$M \to M$	relationship	0.578
$M \to M$	single	0.584
$M \to F$	single	0.631
$M \to F$	married	0.637
$F \to M$	single	0.663
$M \to F$	relationship	0.678
$F \to M$	relationship	0.700
$F \to M$	married	0.715

Table 5.2: Focus in messaging, grouped by gender and relationship status.

Initiator-Target	Rel. Status	f_5 (Profile)
$F \to F$	married	0.225
$F \to F$	relationship	0.225
$M \to M$	married	0.225
$M \to M$	relationship	0.227
$M \to M$	single	0.228
$M \to F$	single	0.232
$M \to F$	married	0.242
$F \to F$	single	0.244
$M \to F$	relationship	0.274
$F \to M$	single	0.311
$F \to M$	married	0.329
$F \to M$	relationship	0.364

Table 5.3: Focus in profile viewing, grouped by gender and relationship status.

contacts, and overlap to be qualitatively similar for ranges of k between 1 and 20; we report on $k = 10$ for concreteness.

5.5.1 Stability and activity

Figure 5.10 shows the overlap between the two time periods as a function of activity. Although we found in Section 5.3 that users' attention to their top k contacts was lowest for profile and photo views, we find that they are among the highest in terms of overlap. We also find differences between modalities of similar total volume and aggregate f_k values; for example, commenting exhibits significantly higher overlap than messaging, and in fact its overlap is quite similar to that of profile viewing. It is an interesting question to consider the possible bases for these contrasts and similarities; one possibility is that a large fraction of profile views are initiated from the News Feed. While the news feed may in part be responsible for the stability of top contacts, the act of sending messages or leaving wall posts are not directly affected by stability introduced by the Facebook news feed. We find that the messages and wall posts tend to have the greatest churn, although higher levels of activity lead to relatively large increases in stability.

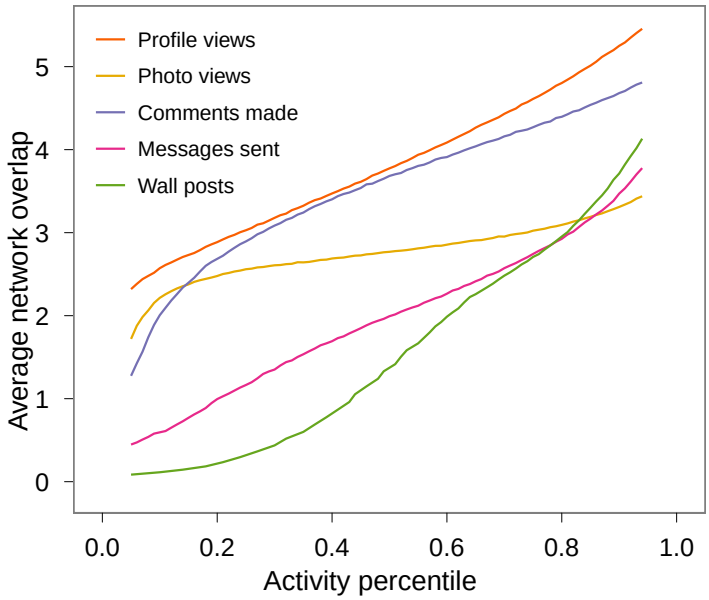

Figure 5.10: Overlap between top 10 users from January & February to March & April as a function of activity. Higher activity levels are associated with greater stability in users' networks over time.

5.5.2 Regression analysis of model stability

To further explore the predictability of network churn and the relationship between network size and activity, we perform a regression analysis for each modality shown in Figure 5.10. We consider modalities independently, and attempt to predict the fraction of each user's top 10 neighbors that persist between two time periods. We use two basic properties of users' networks that factor into attention: activity level and network size.

The regressions are summarized in Table 5.4. One can see that in all regressions, the *con* coefficient is negative, meaning an increase in the number of contacts decreases the expected amount of overlap for a user with median activity level. Conversely, an increase in activity generally increases the expected degree of overlap, as *act* is positive.

We find that the stability of users' networks can depend significantly on the tradeoff between network size and activity level. For example, wall posts exhibit a relatively large and positive interaction term, meaning that high levels of activity mitigate the negative effect of additional contacts. In other cases, such as comments or profile views, highly active users are even less likely to retain top contacts as they interact with more contacts. The models explain between 7% and 62% of the variance, which suggests that activity and network size are useful but not sufficient for predicting shifts in attention over time.

Modality	*intercept*	*act*	*con*	*act* × *con*	R^2
Profile	0.39	0.83	-0.56	-0.11	0.31
Photo	0.24	0.38	-0.31	-0.02	0.07
Comment	0.38	0.65	-0.40	-0.26	0.30
Message	0.20	0.53	-0.24	-0.09	0.26
Wall post	0.15	1.07	-0.64	0.20	0.62

Table 5.4: Regressions explaining the variation in the fraction of top 10 contacts that persist over time. Independent variables are given as percentiles, centered at zero, so that the intercept captures the expected level of overlap between the two time periods for users with a median level of activity (*act*) and median number of contacts (*con*). N = 103, 058 All coefficients are significant at the $p < 10^{-16}$ level and have standard errors that are at least an order of magnitude smaller than the coefficient itself.

5.6 Conclusion

We have provided a way of analyzing individuals' personal networks in terms of the way they balance their attention across social contacts. This measure exposes properties that are distinct from traditional analyses of personal networks based on size and composition, and it enables a comparison of different interaction modalities and different patterns within and between groups. In addition, the measure has important practical implications: by modeling an individual's balance of social attention, product designers can properly tailor that individual's experience to match her preferences for keeping in touch mostly with her top contacts, or with a more diverse set of people.

While our analysis here is based on Facebook data, the framework is very general, and can be applied to any context where detailed interaction data is available, including other social media sites as well as communication modalities such as phone and e-mail. It is an interesting open question to see how the balance of social attention varies across different domains, and in principle these measures can provide a way of categorizing such domains as more focused or more dispersed. It also becomes promising to consider using the balance of attention as a potential feature of individuals in user-based classification and learning tasks, since we have seen that it captures sources of variation among individuals in ways that other measures may miss.

CHAPTER VI

Conclusions

Online social networks increasingly pervade people's daily lives, and as we show in this thesis, can be used to study human behavior at an unimaginable scale. Online networks create new ways for information to flow, and necessitates the use of data-driven exploratory studies. In doing so, one can discover phenomena unique to online networks, and draw parallels between online and offline behaviors. More significantly, online systems do not just give us the ability to observe human behavior at a mass scale, but also enable us to conduct controlled experiments in situ, which can identify causal effects that would have been impossible to rigorously establish in offline situations.

Our first study examines the adoption of user-created content in a virtual world, and finds evidence for social influence. In the subsequent two chapters, we evaluate causal claims about the role of networks in the daily spread of information using two very large field experiments on Facebook. The first experiment shows that while strong ties are more likely to influence others to share information, weak ties are collectively responsible for the majority of diffusion that takes place. Our second experiment further evaluates the effect of friends' behavior on sharing decisions. Together, the studies suggest that the primary role of social networks in information diffusion is to expose individuals to relevant content, but that social signals only play a minor role in individuals' ultimate decision to share content. The experiments also demonstrate that observational evidence can greatly overrepresent the effect of social influence by giving the appearance that individuals are subject to previously studied theories of social contagion. This likeness can alternatively be be explained by homophily, and highlights the importance conducting experiments. Finally, using the complete set of viewing and communication activity between users on Facebook, we analyze how individuals allocate their attention across friends.

As individuals continue to share more online, the success of future online technologies will depend critically upon their ability to predict and shape how information will flow in networks. Our experiments focus on average effects, where we primarily consider whether or not subjects had been exposed to information, and the strength of tie with their friends.

Other features of the individual and her alters, such as age, gender, or location may modulate the effect of social influence. Further study is needed to understand how these factors, as well as differences in spread according to topic or content type (e.g. advertisements, music, software) relate to influence and its confounds. Work along these lines can be used to develop models that predict whether individuals will engage with or re-share content, which can be used in social news ranking algorithms, recommender systems, and ad targeting. For those who wish to maximize the spread of information, more empirical work must be done to understand how the timing of activity and algorithmic ranking systems affect information flow. For example, whether an individual is exposed to information shared by a source might depend on when the information is shared, when the individual logs in, and what other content might take priority over the information. Finally, since there are few technical limits on the number of contacts an individuals has online, and how much information those contacts share, the ability to route information effectively will be of great practical significance. I hope to resolve many of these outstanding problems and help build better online social networking systems in my future work.

APPENDIX

APPENDIX A

Supporting Material for The Effect of Social Networks on Information Diffusion

A.1 Experimental design

All users visiting `facebook.com` are presented with a list of stories about their friends' activity on the site, in an interface called the feed. Some of these stories contain links to content on the World Wide Web, uniquely identified by URLs (Uniform Resource Locators). Our experiment evaluates how much exposure to these URLs on the feed increases sharing beyond correlations that one might expect among Facebook friends. For example, friends with whom a user interacts more often may be more likely to visit sites that the user also visits. As a result, those friends may be more likely to share the same URL as the user before she has the opportunity to share that content herself. Other unobserved correlations may arise due to external influence via e-mail, instant messaging, etc.

Subject-URL pairs are randomly assigned to control (*no feed*) and treatment (*feed*) conditions at the time of display. Stories that contain links to a URL assigned to the *no feed* condition for the subject are never displayed in the subject's feed. Those assigned to the *feed* condition are displayed. All activity relating to subject-URL pairs assigned to either experimental condition is logged, including feed exposures, censored exposures, and clicks to the URL (from the feed or other sources, like messaging). Because we expected the probability of sharing to be much lower for subjects assigned to the *no feed* condition, we assign twice as many subject-URL pairs to the *no feed* condition to improve the statistical significance of our results. Directed shares, such as a link that is included in a private Facebook message or explicitly posted on a friend's wall, are not affected by the assignment procedure. If a subject-URL pair is assigned to an experimental condition, and the subject clicks on content containing that URL in any interface other than feed, that subject-URL pair is removed from the experiment.

A.2 Subject experience

Stories that contain URLs assigned to the *no feed* condition are removed from subjects' feeds, but the experience of subjects assigned to each condition are otherwise indistinguishable from one another. Figure A.1A and A.1B show an example of a Facebook feed for a hypothetical user assigned to the *feed* and *no feed* conditions. In the latter condition, a share present in the *feed* condition is not displayed.

A subject in the *feed* condition may share a link in three ways. First, the subject can click on the "share" link located directly below a feed story containing that link (Figure A.1A, red arrow). Second, the subject may visit the page being linked to (Figure A.2), and copy and paste the URL of the site into her status update box (Figure A.3B). Status updates are the primary way in which Facebook users broadcast content without directly responding to content that others have shared. Finally, the subject may click on "Share", "Like", or "Recommend" action links on the external page that she is visiting (an example of such links can be seen in Figure A.2, red arrow). These action links are deployed on millions of pages throughout the web, and the exact interface and wording of these links may vary from site to site. Subjects in the *no feed* condition may only share links using the latter two methods.

A.3 Population

Subject-URL pairs are randomly assigned to experimental conditions at the time of display, and therefore our experimental population consists of a random sample of all Facebook users who visited the site during the experiment (August 14th to October 4th 2010) and had at least one friend sharing a link. At the time of the experiment, there were approximately 500 million Facebook users logging in at least once a month. Our sample consists of approximately 283M of these users. All Facebook users report their age and gender, and a user's country of residence can be inferred from the IP address with which she accesses the site. In our sample, the median and average age of subjects is 26 and 29.3, respectively. Subjects originate from 236 countries and territories, 44 of which have one million or more subjects. Additional details on user demographics are summarized in Table A.1.

A.4 Ensuring data quality

Threats to data quality include using content that was or may have been previously seen by subjects on Facebook prior to the experiment, content that subjects may have seen through mediums on Facebook other than feed, and malicious content. We address these issues in several ways. First, we only consider content that was shared by any of the subjects' friends after the start of the experiment. This enables our experiment to

Figure A.1: An example of the Facebook News Feed interface for a hypothetical subject who has an NPR link assigned to the *feed* (A) or *no feed* condition (B). The red arrow points to a link that triggers the re-share action depicted in Figure A.3A.

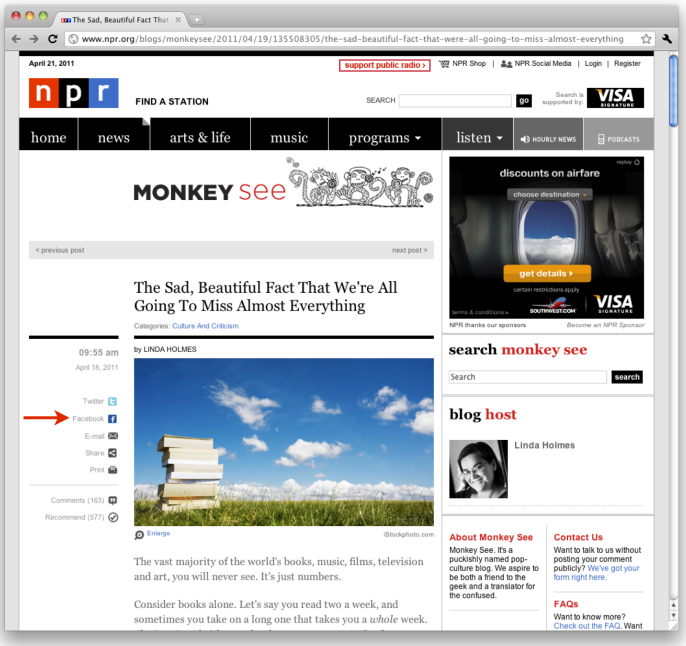

Figure A.2: An example of a web page that can be shared on Facebook. The address bar at the top of the window gives the URL of the page, which may be copied and pasted into the status update interface on `facebook.com` (e.g. Figure A.3B), or shared by clicking the "Facebook" share link (red arrow) embedded on the lefthand side of the web page.

Figure A.3: Links may be shared on `facebook.com` either by (A) re-sharing a link on the feed or (B) copying and pasting a URL directly into the status post update interface.

Demographic Feature	**feed** (% of subjects) ($N = 160,688,092$)	**no feed** ($N = 218,743,932$)
Gender		
Female	51.6%	51.4%
Male	46.7%	47.0%
Unspecified	1.5%	1.5%
Age		
17 or younger	12.8%	13.1%
18-25	36.4%	36.1%
26-35	27.2%	26.9%
36-45	13.0%	12.9%
46 or older	10.6%	10.9%
Country (top 10 & other)		
United States	28.9%	29.1%
Turkey	6.1%	5.8%
Great Britain	5.1%	5.2%
Italy	4.2%	4.1%
France	3.8%	3.9%
Canada	3.7%	3.8%
Indonesia	3.7%	3.5%
Philippines	2.1%	2.3%
Germany	2.3%	2.3%
Mexico	2.0%	2.1%
226 Others	37.5%	37.7%

Table A.1: Summary of demographic features of subjects assigned to the *feed* and *no feed* condition (total $N = 253,238,367$)

accurately capture the first time a subject is exposed to a link in the feed, and ensures that URLs in our experiment more accurately reflect content that is primarily being shared contemporaneously with the timing of the experiment. We also exclude potential subject-URL pairs where the subject had previously clicked on the URL via (i) any interface on the site at any time up to two months prior to exposure, or (ii) any interface other than the feed for content assigned to the *no feed* condition. Finally, we use Facebook's internal classification of "spam" and malicious sites to remove URLs that may not reflect ordinary users' purposeful intentions of distributing content to their friends.

A.5 Data analysis

Our assignment procedure allows us to directly compare the overall probability that subjects share links they were or were not exposed to on the feed. Although the assignment is completely random, subjects and URLs may differ in ways that impact our measurements. For example, certain users may be highly active on Facebook, so that they are assigned to experimental conditions more often than other users. If these users were to vary significantly in terms of their information sharing propensities, such as sharing or re-sharing greater or fewer links than others, the disproportionate inclusion of these users may bias our measurements and threaten the population validity of our findings. Similarly, very popular URLs may also introduce biases; they may be more or less likely to be re-shared because of their inherent appeal or more likely to be discovered independently of Facebook because of their relative popularity amongst friends. To provide control for these biases, we use bootstrapped averages clustered by the subject or URL. We find that in all of our analyses, clustering by the URL rather than the subject yields nearly identical probability estimates that have marginally wider confidence intervals, so we chose to present our results using means and 95% confidence intervals clustered by URL.

Risk ratios are obtained by computing the 95% bootstrapped confidence intervals of likelihood of sharing in the *feed* and *no feed* conditions. To obtain the lower bound, we divide the lower bound of the probability of sharing in the *feed* condition by the upper bound for the *no feed* condition. The upper bound is obtained by dividing the upper bound in the *feed* condition by the lower bound of the *no feed* condition. The additive analog of the same procedure is used to obtain confidence intervals for the absolute differences.

The presence of strong temporal clustering A.4 in both conditions highlights the difficulty with attributing temporal proximity in sharing times to influence: sharing may be due to a host of other factors, such as regular revisitation to sites that link to the same content.

We examined four measures of tie strength between a subject and her sharing friend: (i) the frequency of private online communication between the two users in the form of

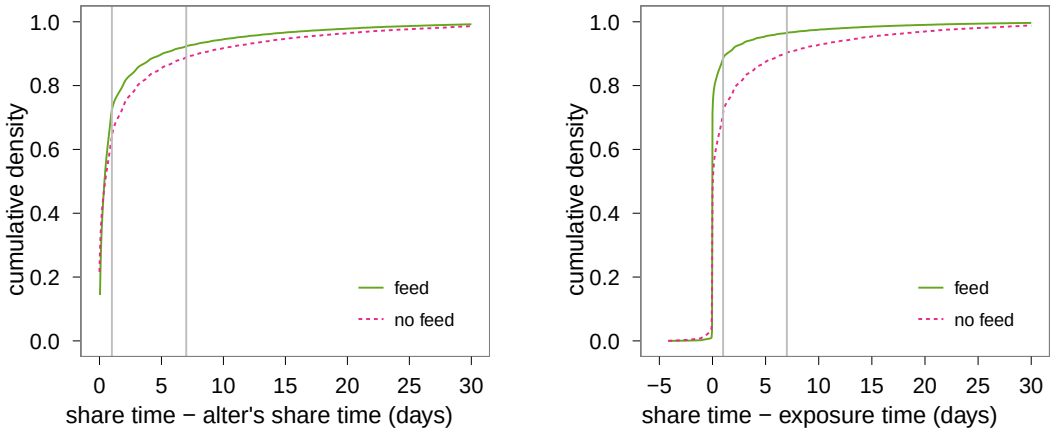

Figure A.4: Temporal clustering in sharing the same link as a friend in the *feed* and *no feed* conditions. (A) The difference in sharing time between a subject and their first sharing friend. (B) The difference between the time at which a subject was first to exposed (or was to be exposed) to the link and the time at which they shared. The difference is negative when a subject shares content after their friend, but did not log into Facebook until a later time. Vertical lines indicate one day and one week.

Facebook messages; (ii) the frequency of public online interaction in the form of comments left by one user on another user's post; (iii) the number of real-world coincidences captured on Facebook in terms of both users being labeled as appearing in the same photograph; and (iv) the number of online coincidences in terms of both users responding to the same Facebook post with a comment. All four measurements yield qualitatively similar results (e.g. Figure A.5, and Figure 2 in the main text.) In our main text, we report on comments because they are more widely used than Facebook messages, and report on photos because they are the best available proxy for offline interaction. For plots showing probabilities or risk ratios as a function of tie strength, the horizontal axes range from zero to the 99^{th} percentile of tie strength (from all impressions).

Comments and messages are directed actions in the sense that they have a sender and recipient, while photo and thread coincidences are undirected. Thus, for our measurements of comment and message frequency between a subject and her sharing friend, we had the option of using the number of communications that the subject received, sent, or a combination of the two, such as the geometric mean. Figure A.6 shows the probability measurements (A, B) and risk ratios (C) for the number of comments received, sent, and the geometric mean of the two (results for messages are qualitatively similar). We chose to present our main results in terms of comments received because it is easiest to interpret, and because this measure produces a more generous classification of weak ties that provides a conservative estimate of the differences in sharing probability and risk ratios for strong

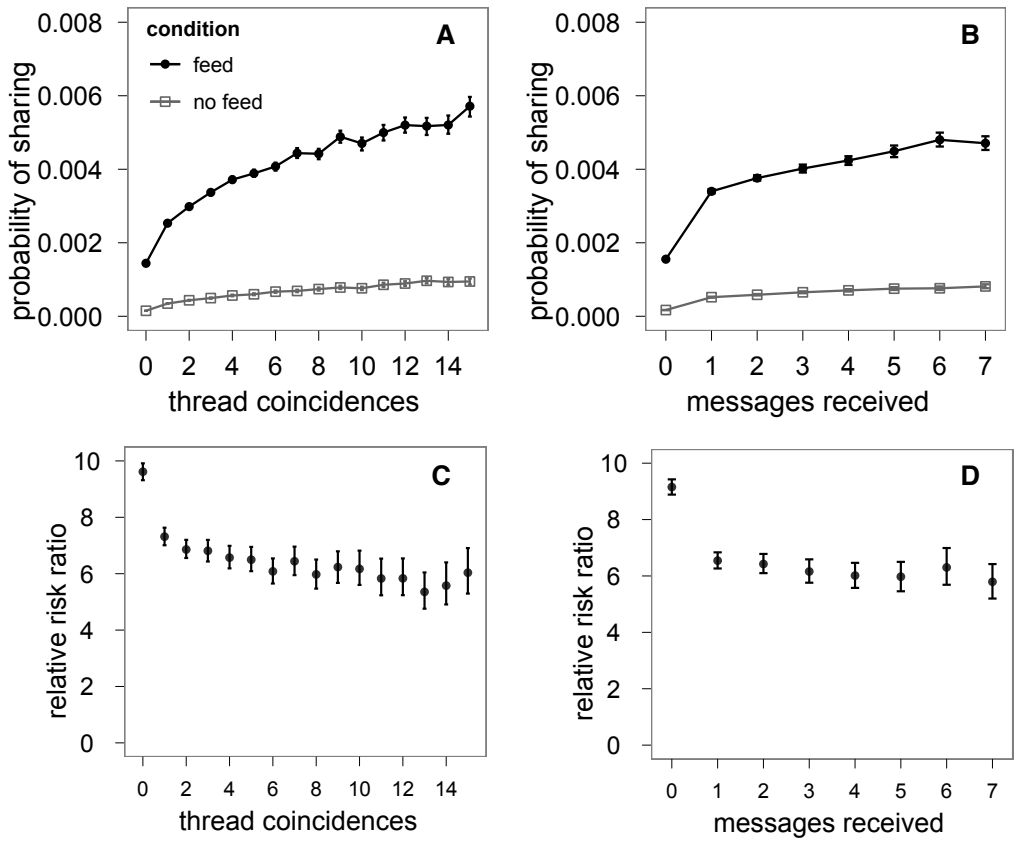

Figure A.5: The relationship between likelihood of sharing and two other measures of tie strength, thread coincidences and personal messages received. (A) and (B) show the relationship between tie strength and the probability of sharing a link that a friend in the *feed* and *no feed* condition. (C) and (D) show the multiplicative effect of feed ($p_{feed}/p_{no\ feed}$).

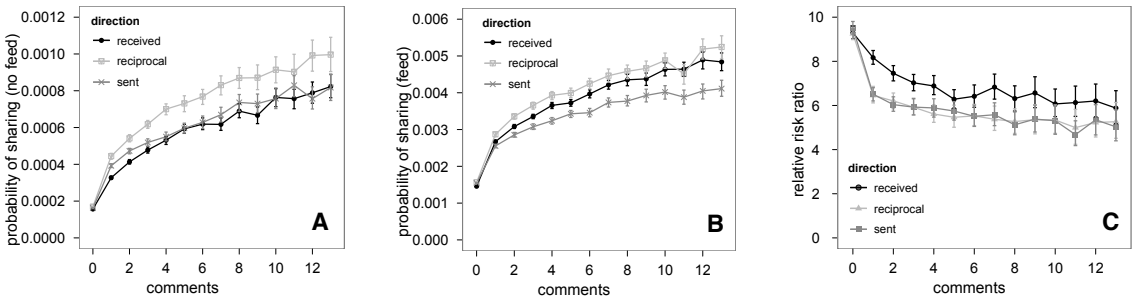

Figure A.6: Sensitivity of probability estimates to choice of directed tie strength measurement for (A) *no feed* and (B) *feed* conditions. The main text expresses tie strength in terms of comments received (•). Comments sent (×) and the geometric mean of the two directions (□), $\lfloor\sqrt{sent * received}\rfloor$, yield similar results. (C) Shows the risk ratio between the *feed* and *no feed* conditions, illustrating that slightly different formulations would lead to more substantial differences in the multiplicative effect of feed with respect to tie strength.

versus weak ties.

While the causal effect of feed is larger for stronger ties, the overall amount of influence generated by weak and strong ties depends on the empirical distribution of tie strength. Figure A.7 shows this distribution, and illustrates that the majority of ties are *weak* in the sense that subjects had no trace of previous interaction with those ties.

To estimate the fraction of influence generated by ties of strength k, we first compute the average treatment effect for subjects with a tie of strength k: $ATE(k) = p_{k,\ feed} - p_{k,\ no\ feed}$ (e.g. the difference in probabilities in Figure A.5AB). We then multiply the average treatment effect at k by n_k, the fraction of links displayed in users' feeds that are of strength k. To compare the impact of *weak* and *strong* ties, we must set a cutoff value for the minimum amount of interaction required between two individuals in order to consider that tie *strong*. Setting the cutoff at $k = 1$ provides the most generous classification of strong ties while preserving some meaningful distinction between strong and weak ties, thereby giving the most influence credit to *strong* ties. Therefore, in our main text, we consider the comparison of $ATE(0) * n_0$ and $\sum_{k=1}^{N} ATE(k) * n_k$, and find that the majority of influence is generated by *weak* ties. In Figure 4 of the main text, these quantities are expressed in terms of overall percentage of influence on feed, which can be obtained by dividing by the estimated total fraction of shares due to exposure on Facebook, $\sum_{k=0}^{N} ATE(k) * n_k$.

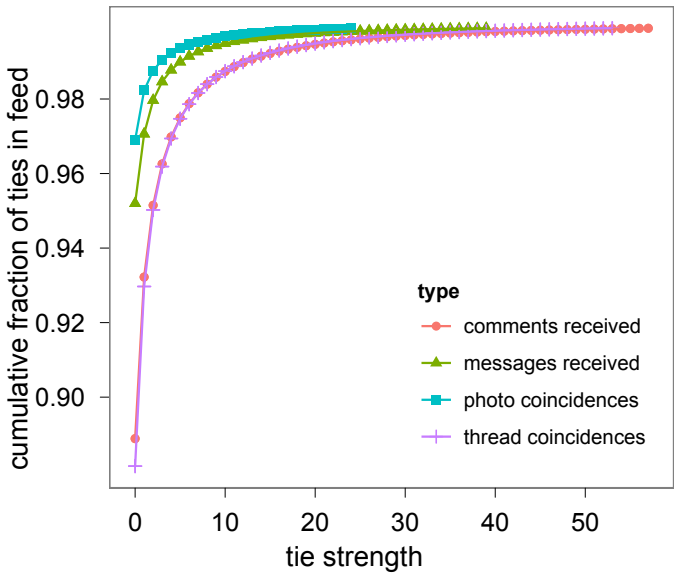

Figure A.7: Tie strength distribution amongst friends displayed in subjects' feeds using the four measures. Points are plotted up to the 99.9^{th} percentile. Note that the vertical axis is collapsed.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adamic, L. A., and E. Adar (2001), Friends and neighbors on the web, *Social Networks*, *25*, 211–230.

Adar, E., J. Teevan, and S. T. Dumais (2009), Resonance on the web: web dynamics and revisitation patterns, in *Proceedings of the 27th International Conference on Human factors in Computing Systems*, CHI '09, pp. 1381–1390, ACM Press, New York, NY, USA.

Anagnostopoulos, A., R. Kumar, and M. Mahdian (2008), Influence and correlation in social networks, in *Proceedings of the 14th Internal Conference on Knowledge Discover & Data Mining*, pp. 7–15, ACM Press, New York, NY, USA.

Anderson, R., and R. M. May (1992), *Infectious diseases of humans: Dynamics and control*, Oxford University Press, Oxford.

Aral, S., and M. W. Van Alstyne (2011), Networks, Information Brokerage: The Diversity-Bandwidth Tradeoff, *Am. J. Sociol.*

Aral, S., L. Muchnik, and A. Sundararajan (2009), Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, *Proc. Natl. Acad. Sci.*, *106*(51), 21,544–21,549.

Backstrom, L., D. Huttenlocher, J. Kleinberg, and X. Lan (2006), Group formation in large social networks: membership, growth, and evolution, in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44–54, ACM, New York, NY, USA.

Backstrom, L., E. Sun, and C. Marlow (2010), Find me if you can: Improving geographical prediction with social and spatial proximity, in *Proc. 19th International World Wide Web Conference*.

Backstrom, L., E. Bakshy, J. Kleinberg, T. Lento, and I. Rosenn (2011), ePluribus: Ethnicity on social networks, in *Proc. 5th International Conference on Weblogs and Social Media*.

Bakshy, E., B. Karrer, and L. Adamic (2009), Social influence and the diffusion of user-created content, in *Proceedings of the Tenth ACM Conference on Electronic Commerce*, pp. 325–334, ACM.

Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts (2011), Everyone's an influencer: Quantifying influence on twitter, in *3rd ACM Conference on Web Search and Data Mining*, ACM Press, Hong Kong.

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

Bernard, H. R., P. Killworth, D. Kronenfeld, and L. Sailer (1984), The problem of informant accuracy: The validity of retrospective data, *Annu. Rev. Anthropol.*, *13*, 495–517.

Black, D., G. Gates, S. Sanders, and L. Taylor (2000), Demographics of the gay and lesbian population in the United States: Evidence from available systematic data sources, *Demography*, *37*(2), 139–154.

Brown, J. J., and P. H. Reingen (1987), Social ties and word-of-mouth referral behavior, *J. Consumer Research*, *14*(3), pp. 350–362.

Burt, R. S. (1992), *Structural holes: The social structure of competition*, Harvard University Press, Cambridge, MA.

Campbell, K. E., and B. A. Lee (1991), Name generators in surveys of personal networks, *Social Networks*, *13*(3), 203–221.

Castronova, E. (2008), A Test of the Law of Demand in a Virtual World: Exploring the Petri Dish Approach to Social Science, *SSRN eLibrary*.

Centola, D. (2010), The Spread of Behavior in an Online Social Network Experiment, *Science*, *329*(5996), 1194–1197.

Centola, D., and M. Macy (2007), Complex Contagions and the Weakness of Long Ties, *American Journal of Sociology*, *113*(3), 702–734.

Cha, M., A. Mislove, and K. P. Gummadi (2009), A measurement-driven analysis of information propagation in the flickr social network, in *Proceedings of the 18th international conference on World wide web*, WWW '09, pp. 721–730, ACM, New York, NY, USA.

Cha, M., H. Haddadi, F. Benevenuto, and K. P. Gummad (2010), Measuring user influence on twitter: The million follower fallacy, in *Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media*, Washington, DC.

Chang, J., I. Rosenn, L. Backstrom, and C. Marlow (2010), ePluribus: Ethnicity on social networks, in *Proc. 4th International Conference on Weblogs and Social Media*.

Chatterjee, R., and J. Eliashberg (1990), The innovation diffusion process in a heterogeneous population: A micromodeling approach, *Management Science*, *36*(9), 1057–1079.

Christakis, N. A. A., and J. H. H. Fowler (2007), The spread of obesity in a large social network over 32 years., *N. Engl. J. Med.*, *357*(4), 370–379.

Coleman, J. S., E. Katz, and H. Menzel (1966), *Medical innovation: A diffusion study*, Bobbs-Merrill, New York.

Crane, D. (1999), Diffusion models and fashion: A reassessment, *Annals of the American Academy of Political and Social Science*, *566*, pp. 13–24.

Domingos, P., and M. Richardson (2001), Mining the network value of customers, in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 57–66, ACM, New York, NY, USA.

Finkel, S. E., E. N. Muller, and K.-D. Opp (1989), Personal influence, collective rationality, and mass political action, *The American Political Science Review*, *83*(3), pp. 885–903.

Fischer, C. S. (1982), *To Dwell Among Friends*, University of Chicago Press.

Friedman, D., A. Steed, and M. Slater (2007), Spatial Social Behavior in Second Life, *Lecture Notes in Computer Science*, *4722*, 252.

Gilbert, E., K. Karahalios, and C. Sandvig (2008), The network in the garden: An empirical analysis of social media in rural life, in *Proc. 26th ACM Conference on Human Factors in Computing Systems*, pp. 1603–1612.

Goldenberg, J., S. Han, D. R. Lehmann, and J. W. Hong (2009), The role of hubs in the adoption process, *J. Marketing*, *73*(2), 1–13.

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

Granovetter, M. S. (1978), Threshold models of collective behavior, *Am. J. Sociol.*, *83*(6), 1420–1443.

Greenberg, B. S. (1964), Person to person communication in the diffusion of news events, *Journalism Quarterly*, *41*, 489–494.

Hampton, K., and L. Rainie (2011), The power of social networking sites, *Tech. rep.*, Pew Internet & American Life Project.

Hill, S., F. Provost, and C. Volinsky (2006), Network-Based Marketing: Identifying Likely Adopters via Consumer Networks, *Statistical Science*, *21*(2), 256.

Jiang, J., C. Wilson, X. Wang, P. H. an d Yafei Dai, and B. Y. Zhao (2010), Understanding latent interactions in online social networks, in *Proc. 10th ACM SIGCOMM Internet Measurement Conference*.

Katz, E., and P. F. Lazarsfeld (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communications*, Free Press, Glencoe, Ill.,.

Kempe, D., J. Kleinberg, and E. Tardos (2003), Maximizing the spread of influence through a social network, in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, New York, NY, USA.

Killworth, P. D., E. C. Johnsen, H. R. Bernard, G. A. Shelley, and C. McCarty (1990), Estimating the size of personal networks, *Social Networks*, *12*(4), 289–312.

Kleinberg, J. (2006), Complex networks and decentralized search algorithms, in *Proc. International Congress of Mathematicians*.

Kossinets, G., and D. J. Watts (2009), Origins of homophily in an evolving social network, *Am. J. Sociol.*, *115*(2), 405–450.

Kossinets, G., J. Kleinberg, and D. Watts (2008), The structure of information pathways in a social communication network, in *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 435–443, ACM, New York, NY, USA.

Lerman, K. (2007), Social information processing in news aggregation, *IEEE Internet Computing*, *11*(6), 16–28.

Lerman, K., and L. A. Jones (2007), Social browsing on flickr, in *ICWSM*.

Leskovec, J., L. A. Adamic, and B. A. Huberman (2006a), The dynamics of viral marketing, in *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pp. 228–237, ACM, New York, NY, USA.

Leskovec, J., A. Singh, and J. Kleinberg (2006b), Patterns of influence in a recommendation network, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Springer.

Leskovec, J., A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance (2007), Cost-effective outbreak detection in networks, in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 420–429, ACM, New York, NY, USA.

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

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

Manski, C. F. (1993), Identification of endogenous social effects: The reflection problem, *Rev. Econ. Stud.*, *60*(3), 531–42.

Marin, A. (2004), Are respondents more likely to list alters with certain characteristics? implications for name generator data, *Social Networks*, *26*(4), 289–307.

Markus, M. L. (1987), Toward a critical mass theory of interactive media, *Communication Research*, *14*(5), 491–511.

Marsden, P. V. (1987), Core discussion networks of americans, *American Sociological Review*, *52*(1), 122–131.

Marsden, P. V., and K. E. Campbell (1984), Measuring tie stength, *Social Forces*, *63*(2), 482–501.

Mayhew, B. H., and R. L. Levinger (1976), Size and the density of interaction in human aggregates, *American Journal of Sociology*, *82*(1), 86–110.

McPherson, M., and L. Smith-Lovin (1993), You are who you know: A network approach to gender, in *Handbook of Statistical Genetics*, edited by P. England, pp. 223–251, Aldine.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001), Birds of a feather: Homophily in social networks, *Annual Review of Sociology*, *27*, 415–444.

McPherson, M., L. Smith-Lovin, and M. E. Brashears (2006), Social isolation in America: Changes in core discussion networks over two decades, *American Sociological Review*, *71*(3), 353–375.

Milgram, S. (1970), The experience of living in cities, *Science*, *167*(3924), 1461–1468.

Moore, G. (1990), Structural determinants of men's and women's personal networks, *American Sociological Review*, *55*(5), 726–735.

Newman, M. (2002), Spread of epidemic disease on networks, *Physical Review E*, *66*(1), 16,128.

Newman, M., S. Forrest, and J. Balthrop (2002), Email networks and the spread of computer viruses, *Physical Review E*, *66*(3), 35,101.

Ondrejka, C. (2004a), Aviators, moguls, fashionistas and barons: Economics and ownership in second life, Available at SSRN: http://ssrn.com/abstract=614663.

Ondrejka, C. (2004b), A piece of place: Modeling the digital on the real in second life, *Social Science Research Network Working Paper Series*.

Pastor-Satorras, R., and A. Vespignani (2001), Epidemic Spreading in Scale-Free Networks, *Physical Review Letters*, *86*(14), 3200–3203.

Purcell, K., L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead (2010), Understanding the participatory news consumer, *Pew Internet and American Life Project*, *1*.

Rogers, E. M. (1995), *Diffusion of Innovations*, fourth ed., Free Press, New York.

Salganik, M. J., P. S. Dodds, and D. J. Watts (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, Daylight Saving: A Study of Binary Choices with Externalities, *J. Conflict Resolution*, *17*(3), 381–428.

Shalizi, C. R., and A. C. Thomas (2011), Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, *Sociological Methods and Research*, *27*, 211–239.

Song, X., Y. Chi, K. Hino, and B. Tseng (2007), Information flow modeling based on diffusion rate for prediction and ranking, in *Proceedings of the 16th international conference on World Wide Web*, pp. 191–200, ACM Press New York, NY, USA.

Watts, D. (2002), A simple model of global cascades on random networks, *Proceedings of the National Academy of Sciences*, *99*(9), 5766.

Watts, D., and P. Dodds (2007), Influentials, Networks, and Public Opinion Formation, *J. Consumer Research*, *34*(4), 441.

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

Wei, X., J. Yang, L. A. Adamic, R. M. de Araújo, and M. Rekhi (2010), Diffusion dynamics of games on online social networks, in *Proceedings of the 3rd conference on Online social networks*, WOSN'10, pp. 2–2, USENIX Association, Berkeley, CA, USA.

Weimann, G. (1982), On the importance of marginality: One more step in the two-step flow of communication, *American Sociological Review*, *47*(6), 764–773.

Wellman, B., and S. Wortley (1990), Different strokes from different folks: Community ties and social support, *American Journal of Sociology*, *96*(3), 558–588.

Wilson, C., B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao (2009), User interactions in social networks and their implications, in *Proc. 2009EuroSys Conference*, pp. 205–218.

Wu, F., and B. Huberman (2007), Novelty and collective attention, *Proceedings of the National Academy of Sciences*, *104*(45), 17,599.

Yee, N., J. Bailenson, M. Urbanek, F. Chang, and D. Merget (2007), The Unbearable Likeness of Being Digital: The Persistence of Nonverbal Social Norms in Online Virtual Environments, *CyberPsychology & Behavior*, *10*(1), 115–121.

Zheng, R., F. Provost, and A. Ghose (2007), Social Network Collaborative Filtering, working paper CeDER-8-08. Center for Digital Economy Research, Stern School of Business, New York University.