Characterizing and Leveraging Social Phenomena in Online Networks

Zeinab Abbassi

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Characterizing and Leveraging Social Phenomena in Online Networks

Zeinab Abbassi

Social phenomena have been studied extensively in small scales by social scientists. With the increasing popularity of Web 2.0 and online social networks/media, a large amount of data on social phenomena have become available. In this dissertation we study online social phenomena such as social influence in social networks in various contexts. This dissertation has two major components: 1. Identifying and characterizing online social phenomena 2. Leveraging online social phenomena for economic and commercial purposes.

We begin the dissertation by developing multi-level revenue sharing schemes for viral marketing on social networks. Viral marketing leverages social influence among users of the social network. For our proposed models, we develop results on the computational complexity, individual rationality, and potential reach of employing the Shapley value as a revenue sharing scheme. Our results indicate that under the multi-level tree-based propagation model, the Shapley value is a promising scheme for revenue sharing, whereas under other models there are computational or incentive compatibility issues that remain open. We continue with another application of social influence: social advertising. Social advertising is a new paradigm that is utilized by online social networks. Social advertising is based in the premise that social influence can be leveraged to place ads more efficiently. The goal of our work is to understand how social ads can affect click-through rates in social networks. We propose a formal model for social ads in the context of display advertising. In our model, ads are shown to users one after the other. The probability of a user clicking an ad depends on the users who have clicked this ad so far. This information is presented to users

as a social cue, thus the click probability is a function of this cue. We introduce the social display optimization problem: suppose an advertiser has a contract with a publisher for showing some number (say B) impressions of an ad. What strategy should the publisher use to show these ads so as to maximize the expected number of clicks? We show hardness results for this problem and in light of the general hardness results, we develop heuristic algorithms and compare them to natural baseline ones.

We then study distributed content curation on the Web. In recent years readers have turned to the social web to consume content. In other words, they rely on their social network to curate content for them as opposed to the more traditional way of relying on news editors for this purpose – this is an implicit consequence of social influence as well. We study how efficient this is for users with limited budgets of attention. We model distributed content curation as a reader-publisher game and show various results. Our results imply that in the complete information setting, when publishers maximize their utility selfishly, distributed content curation reaches an equilibrium which is efficient, that is, the social welfare is a constant factor of that under an optimal centralized curation.

Next, we initiate the study of an exchange market problem without money that is a natural generalization of the well-studied kidney exchange problem. From the practical point of view, the problem is motivated by barter websites on the Internet, e.g., swap.com, and u-exchange.com. In this problem, the users of the social network wish to exchange items with each other. A mechanism specifies for each user a set of items that she gives away, and a set of items that she receives. Consider a set of agents where each agent has some items to offer, and wishes to receive some items from other agents. Each agent would like to receive as many items as possible from the items that she wishes, that is, her utility is equal to the number of items that she receives and wishes. However, she will have a large dis-utility if she gives away more items than what she receives, because she considers such a trade to be unfair. To ensure voluntary participation (also known as individual rationality), we require the mechanism to avoid this. We consider different variants of this problem: with and without a constraint on the length of the exchange cycles and show different results including their truthfulness and individual rationality.

In the other main component of this thesis, we study and characterize two other social phenomena: 1. friends vs. the crowd and 2. altruism vs. reciprocity in social networks. More specifically, we study how a social network user's actions are influenced by her friends vs. the crowd's opinion. For example, in social rating websites where both ratings from friends and average ratings from everyone is available, we study how similar one's ratings are to each other. In the next part, we aim to analyze the motivations behind users' actions on online social media over an extended period of time. We look specifically at users' likes, comments and favorite markings on their friends' posts and photos. Most theories of why people exhibit prosocial behavior isolate two distinct motivations: Altruism and reciprocity. In our work, we focus on identifying the underlying motivations behind users' prosocial giving on social media. In particular, our goal is to identify if the motivation is altruism or reciprocity. For that purpose, we study two datasets of sequence of users' actions on social media: a dataset of wall posts by users of Facebook.com, and another dataset of favorite markings by users of Flickr.com. We study the sequence of users' actions in these datasets and provide several observations on patterns related to their prosocial giving behavior.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Vishal Misra, for his great help and guidance through this dissertation. I am indebted to him for his insightful comments on my research and his support during my PhD studies to pursue my broad interest in studying various aspects of online social phenomena. I am also grateful to my committee members Augustin Chaintreau, Xi Chen, Muthu Muthukrishnan and Dan Rubenstein, for reviewing my dissertation and for their valuable comments.

I have benefited greatly from collaborations with my coauthors and for that I am thankful to them. The research in this thesis is based on joint work with Christina Aperjis, Aditya Bhaskara, Nima Haghpanah, Nidhi Hegde, Bernardo Huberman, Laurent Massoulie, Vahab Mirrokni, Vishal Misra, and Eric Schwartz.

I would also like to thank the staff of the Computer Science department at Columbia University especially Sophie Majewski, and my fellow graduate students for making my years at Columbia truly memorable.

Last, but not least, I would like to express my sincere gratitude to my family: my parents, Laya and Abbas; my brother and his wife, Hossein and Sara; my husband Vahab and my beautiful daughter Sama, whose endless love and support has always been a bliss for me. I could not have completed this work without your limitless support. I dedicate this thesis to my family. Thanks for everything.

Chapter 1

Introduction

Phenomenon is defined as a fact or situation that is observed to exist or happen, especially one whose cause or explanation is in question. Social phenomena include all behavior which influences or is influenced by organisms sufficiently alive to respond to one another. Social sciences have been studying social phenomena for a very long time – in small scales. With the increasing popularity of Web 2.0, online social networks and social media, a large amount of data has become available. By employing the principles of computer science to deal with large-scale data, we can look at social science problems such as marketing, social influence, and model the ways that very large groups of people interact on the web. Our ability to quantify this behavior sheds light on interesting phenomena and ideally makes further developments possible for online platforms. In addition to identification and characterization of these phenomena, different technics can be employed in order to leverage these phenomena in different ways. As an example, for improving user experience with better targeted advertising.

This thesis has two major components: 1. Leveraging social phenomena for economic and commercial purposes 2. Identifying and characterizing social phenomena that have not been studied well so far.

We begin in Chapter 2, where we develop multi-level revenue sharing schemes for viral marketing on social networks. Viral marketing leverages social influence among users of the social

network. For the proposed models, we develop results on the computational complexity, individual rationality, and potential reach of employing the Shapley value as a revenue sharing scheme. Our results indicate that under the multi-level tree-based propagation model, the Shapley value is a promising scheme for revenue sharing, whereas under other models there are computational or incentive compatibility issues that remain open. This chapter is published as [14].

We continue in Chapter 3 with another application of social influence: social advertising. Social advertising is a new paradigm that is utilized by online social networks. Social advertising is based in the premise that social influence can be leveraged to place ads more efficiently. The impact of social influence among users has been confirmed in sociological studies, statistical models, and in online randomized experiments (see [24] and references therein). The goal of our work is to understand how social ads can affect click-through rates in social networks. We propose a formal model for social ads in the context of display advertising. In our model, ads are shown to users one after the other. The probability of a user u clicking an ad depends on the users who have clicked (or taken a certain action on) this ad so far. This information is presented to u as a social cue (which could be of different kinds, as we will see in Chapter 3), thus the click probability is a function of this cue. We introduce the social display optimization problem: suppose an advertiser has a contract with a publisher for showing some number (say B) impressions of an ad. What strategy should the publisher use to show these ads so as to maximize the expected number of clicks? We show hardness results for this problem and in light of the general hardness results, we develop heuristic algorithms and compare them to natural baseline ones. This chapter is published as [10].

In Chapter 4 we study distributed content curation on the Web. In recent years readers have turned to the social web to consume content. In other words, they rely on their social network to curate content for them as opposed to the more traditional way of relying on news editors for this purpose – this is an implicit consequence of social influence as well. In this Chapter we study this phenomenon. In particular, we study how efficient this is for users with limited budgets of attention. We show that the centralized optimization, while being NP-complete, can be reduced to a separable assignment problem, thus admitting a $(1 - 1/e)$-approximation algorithm. We model distributed content curation as a reader-publisher game and show that the price of anarchy is at most 2. When in addition the readers are selective in the items they choose, we show that the price of anarchy is bounded by 2. Our results imply that in the complete information setting,

when publishers maximize their utility selfishly, distributed content curation reaches an equilibrium which is efficient, that is, the social welfare is a constant factor of that under an optimal centralized curation. This chapter is published as [12].

In Chapter 5, we initiate the study of an exchange market problem without money that is a natural generalization of the well-studied kidney exchange problem []. From the practical point of view, the problem is motivated by barter websites on the Internet, e.g., swap.com, and u-exchange.com. In this problem, the users of the social network wish to exchange items with each other. A mechanism specifies for each user a set of items that she gives away, and a set of items that she receives. Consider a set of agents where each agent has some items to offer, and wishes to receive some items from other agents. Each agent would like to receive as many items as possible from the items that she wishes, that is, her utility is equal to the number of items that she receives and wishes. However, she will have a large disutility if she gives away more items than what she receives, because she considers such a trade to be unfair. To ensure voluntary participation (also known as individual rationality), we require the mechanism to avoid this. We consider different variants of this problem: with and without a constraint on the length of the exchange cycles and show different results including their truthfulness and individual rationality. This chapter is published as [11].

The above chapters: Chapters 2, 3, 4, and 5 belong to the first component of this thesis: leveraging social phenomena.

In the other main component of this thesis (Chapter 6), we study and characterize two other social phenomena: 1. friends vs. the crowd and 2. altruism vs. reciprocity in social networks. More specifically, in Section 6.1 we study how a social network user's actions are influenced by her friends vs. the crowd's opinion. For example, in social rating websites where both ratings from friends and average ratings from everyone is available, we study how similar one's ratings are to each

In Chapter 6.3, we aim to analyze the motivations behind users' actions on online social media over an extended period of time. We look specifically at users' likes, comments and favorite markings on their friends' posts and photos. These actions are analogous to what social scientists and economists call prosocial giving or in general prosocial behavior. Prosocial behavior is defined as "voluntary behavior intended to benefit another person" and has been a puzzle to researchers of different fields because while this type of behavior benefits others, it is often costly for the person

performing it. Most theories of why people exhibit prosocial behavior isolate two distinct motivations: Altruism and reciprocity. In our work, we focus on identifying the underlying motivations behind users' prosocial giving on social media. In particular, our goal is to identify if the motivation is altruism or reciprocity. For that purpose, we study two datasets of sequence of users' actions on social media: a dataset of wall posts by users of Facebook.com, and another dataset of favorite markings by users of Flickr.com. The main difference between these two datasets is that in the Flickr dataset, a user can mark any other user's photo as favorite but in the Facebook dataset, interactions are limited to friends only. We study the sequence of users' actions in these datasets and provide several observations on patterns related to their prosocial giving behavior.

In this dissertation we study social phenomena in particular social influence and social exchange on the Web with different approaches and from different perspectives. Users are spending more and more time on the Web and therefore generating large-scale data. These data can be leveraged in order to improve user experience.

Chapter 2 was done in collaboration with Vishal Misra and is published as [14]. Chapter 3 was done in collaboration with Vishal Misra and Aditya Bhaskara and is published as [10]. Chapter 4 was done in collaboration with Nidhi Hegde and Laurent Massoulie and is published as [12]. Chapter 5 is done in collaboration with Nima Haghpanah and Vahab Mirrokni and is published as [11]. Chapter 6.1 was done in collaboration with Christina Aperjis and Bernardo Huberman and is published as [8]. Chapter 6.3 is unpublished work and was done in collaboration with Vishal Misra and Eric Schwartz.

Chapter 2

Multi-level Revenue Sharing for Viral Marketing

2.1 Introduction

In recent years, the Web has been, among other things, leveraged to harness the power of users to carry out tasks that require collective efforts. Wikipedia, crowdsourcing platforms such as Amazon's Mechanical Turk, and Yahoo! Answers are only a few examples. Users participate with different incentives such as monetary compensation, social recognition, altruism or a combination of these. Some of these tasks require networked structures to succeed. The recent tremendous growth in the popularity of online social networks suggests leveraging these networks for those types of tasks. The recent DARPA network challenge is an example of such a task. The challenge was to locate ten red balloons spread all over the Unites States on a given date as quickly as possible for the prize of $40000 [1]. A task that would be considered impossible using conventional information gathering methods [2]. The winning team implemented a recursive incentive scheme over a social network of referrals with the objective to locate the balloons in minimum time [90].

Refer-a-friend marketing (which is a type of viral marketing) is another example of such tasks because the network of trust between the referrer and the potential adopters plays an important role in the adoption of a product in a viral marketing campaign. In a refer-a-friend advertising campaign, typically, a current user gets some form of discount for referring a product to her friend that ends up in the adoption of the product. Referring a friend might trigger a cascade if the new

adopter recommends the product to her friends as well (hence the term viral marketing).

Another motivating scenario inspired by the above ideas can be for advertising over online social networks. The increasing need for monetizing social networks more effectively is causing social network platforms to look for alternatives to online behavioral targeting. A specific example for this model is as follows (illustrated in Figure 2.1): A social network platform may build the following system to target ads and coupons more effectively. This system allows users to opt in and help the platform in exchange for a share in the revenue. To be more precise, a user who opts in this system is presented with a number of ads/coupons to assign to a limited number of her friends according to their interests. In this example a social network user, named Alice, is presented with two ads: one about audiobooks and the other about places to travel with a baby. Alice knows that Bob has a baby and might be interested in going on a trip, so she would select Bob for the ad: "Best places to travel with a baby". On the other hand, knowing that Carol enjoys listening to audio books, Alice would assign the ad for free audio books to her. This will help the social network route ads to the right users. Bob sees the ad about traveling with a baby as he logs into his account. He also thinks that another fellow parent might be interested in this ad, therefore, he suggests it to be shown to her. The source of the ad can be transparent to the recipient of the ad. In other words, this system can be incorporated in the ad network that the social network platform implements.

A significant challenge for this model to work is to give the users the proper motivation to earnestly contribute to the system in finding the most relevant match. To incentivize users, the social network platform shares the added revenue with the users that opt in and have impact. In this chapter, we introduce various revenue sharing models and discuss fairness and individual rationality of such incentive schemes, and design efficient algorithms to compute and implement such schemes. In particular, we propose models in which referrals (both for products and ads) are rewarded either for one level or multiple levels, and discuss Shapley value revenue sharing. In other words, we discuss single-level and multi-level propagation models, and identify tree-based multi-level propagation as a special case of the graph-based propagation model. We compare these models in terms of the polynomial-time computability of Shapley revenue shares, individual rationality of the revenue shares, and potential reach or expected effectiveness of these models. First, we prove that finding Shapley value revenue shares is $\#P$-hard for general graph-based multi-level

models, but it is polynomial-time solvable for single-level and tree-based multi-level propagation models. Further, by showing supermodularity of revenue function for tree-based propagation models, we show that for the tree-based propagation model, (i) Shapley value revenue shares lie in the core and thus satisfy certain individual rationality conditions, (ii) the nucleolus of these games is polynomial-time computable, and (iii) one can implement budget-balanced group-strategyproof mechanisms, extracting people's willingness to participate in the process. Finally, via simulations on real-world datasets, we conclude that with a fixed amount of revenue share, multi-level tree-based propagation will result in larger expected reach. Overall, we conclude that the tree-based multi-level propagation model for revenue sharing is more effective and more efficient than single-level or graph-based multi-level propagation models.

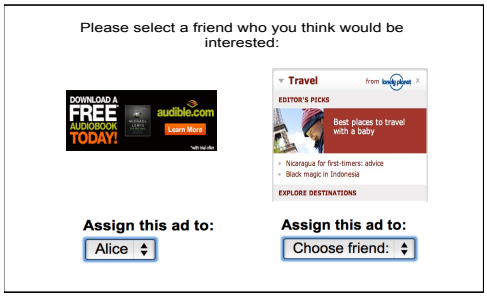


Figure 2.1: Users select most relevant friends.

2.2 Related Work

Viral marketing over social networks has been studied for the purpose of influence maximization [37, 70], or revenue maximization [56, 16]. In the influence maximization models [37, 70], a person's decision to buy a product is influenced by the set of other people who own the product. In the revenue maximization model[56, 16], people don't simply adopt products, but rather must pay money to buy them. A person's decision to buy a product is influenced by the set of other people that own the product as well as the price at which the item is offered. On the subject of propagation of information and influence on social networks, various work has been done. In a recent paper [25], the authors argue that viral marketing would be more effective if a large number of ordinary users are picked as influencers. None of the above work however studies the effect of revenue

sharing in incentivizing users of the social networks in maximizing the reach and effectiveness of online ads.

Recently, a social ad model considering user influence, called AdHeat, has been explored [26]. In this model, the advertising platform may diffuse hint words of influential users to others and then matches ads for each user with aggregated hints. They perform experiments on a real-world data set, and show that AdHeat outperforms the traditional relevance models by a large factor. Although this study shows the effectiveness of using social network information in online advertising, however they do not consider active propagation of ads by users of the social network.

The applications of the Shapley value to network and in particular Internet economics is also been of interest recently: from applications to peer-assisted services [86] to the settlement issue between Content, Eyeball, and Transit ISPs [78, 79, 33]. In a less recent work [71], authors propose that certain private information should only be disclosed by users if they get compensated fairly. The paper determines the value of private information in the context of online surveys, collaborative filtering and in general recommender systems by the Shapley value.

A related but different problem is the cost sharing of Steiner tree or multicast network design problem in cooperative game theory[65, 41]. It is not hard to see that cost function of a multicast tree for a subset of nodes is a submodular function, and this implies the existence of budget-balance cross-monotone cost sharing methods for this problem. This might seem related to our proof of supermodularity of the revenue function in the tree-based propagation model. However, in the multicast cost sharing context, the submodularity of the cost function holds for any general cost function on the edges while, it is not hard to show that supermodularity of the revenue function does not hold in the general case and we show that it holds in the case that the revenue share of each node in the graph is the same (Section 2.5).

2.3 Model

Let U be the set of users of a social network platform (SNP) $\mathcal{P}$. The SNP $\mathcal{P}$ implements a viral marketing program helping an ad campaign or an online retailer, by giving incentives to its users to participate in marketing for the advertiser. To be more specific, the SNP suggests coupons, products or ads to a subset of its users, and asks them to refer a *limited* number of their friends for

each. If their friends buy the coupon or product or click on the ad (or simply do anything through which the SNP gains revenue), the revenue gained would be shared with the referrers. If the friend also refers some of her friends a cascade of referrals would be initiated. In this chapter, if a user adopts a product, buys a coupon or clicks on an ad or takes any action through which revenue is earned by the SNP (directly or through a cascade), we say that the node has become *active*. The nodes that had a role in activating a node are called the *activators* of that node.

Looking at this model as a cooperative game, the set of players are denoted by U, where $N = |U|$. We call any nonempty subset $S \subseteq U$ a coalition of players. For each coalition S, we denote by $f(S)$ the worth function ($f : 2^{U \cup \{\mathcal{P}\}} \rightarrow R$), which measures the total revenue from an advertiser produced by the system when all players of this coalition S are active. Clearly the revenue from a subset S of users without $\mathcal{P}$ is zero as any coalition needs the SNP to implement the marketing/advertising strategy. Let $f_u(S)$ denote the revenue of player u in the coalition S, we then have: $f(S) = \sum_{u \in S} f_u(S)$.

We suggest two models for revenue sharing if a node becomes *active*. In the first model, which is very similar to what happens commonly in referral-based marketing, the SNP shares the added revenue only with direct *activators* of that node. The other model suggests that the revenue be shared with the whole set of users who contributed in activating a node.

Before we get into specifics of our proposed models we define the Shapley value [101]:

Definition 1 (Shapley Value) *Shapley value of player u in coalition S is denoted as $\phi_u(S, f)$ and is computed as: $\forall u \in U, \phi_u(S, f) = \frac{1}{|S|!} \sum_{\pi \in \Pi} \Delta_u(f, S(\pi, u))$*

where Π is the set of all $|S|!$ orderings of S and $S(\pi, u)$ is the set of players preceding u in the ordering π. The Shapley value of player u can thus be interpreted as the expected marginal contribution $\Delta_u(f, S')$, where S' is the set of players in S preceding u in a uniformly distributed random ordering of S.

Shapley value of each player u in coalition S satisfies the three following axioms:

Axiom 1: Efficiency $\sum_{u \in S} \phi_u(S, f) = f(S)$ This axiom means that the total revenue share of each user should be equal to the total revenue gained in the coalition.

Axiom 2: Symmetry If for all $S' \subseteq S \backslash \{u, v\}$, $f(S' \cup \{u\}) = f(S' \cup \{v\})$ then $\phi_u(S, f) = \phi_v(S, f)$ This axiom states that players contributing equal amounts to a coalition should receive same amount of the revenue.

Axiom 3: Fairness For any $u, v \in S$, v's contribution to u equals u's contribution to v, or in other words $\phi_u(S, f) - \phi_u(S\backslash\{v\}, f) = \phi_v(S, f) - \phi_v(S\backslash\{u\}, f)$

We formulate the network of referrals by constructing a directed graph representing users influencing their friends in becoming *active*. More specifically, nodes in this graph are users of the SNP and there is an edge from user w to user u if w activates u. We also add a root node for the SNP and connect this root to all original users who become active. Given this graph, users who are on paths from the root to each node u share the revenue from the activity of this node. In fact, since there might be several users contributing in activation of a specific node, there are different ways to construct this graph. We will describe our models below.

Note that similar to other advertising and revenue sharing systems, such revenue sharing schemes should also be accompanied with a reasonable fraud detection and reputation system that can control for malicious user behavior. Also, It is important to note that in this model, a user is only allowed to refer an ad/product to a *limited* number of her friends, therefore, increasing her incentive to pick the most relevant ones.

2.3.1 Single-level Propagation Model.

In the single-level propagation model the amount paid by the advertiser for a node becoming active is shared among the SNP and the set of its direct activators. Considering this model as a cooperative game, the players of this game are the SNP and the users who influence other nodes to become active. If k users refer a product/ad to user u and she becomes active resulting in revenue gain, the coalition would consist of these k users and the social network platform. The worth function is described in the following.

To be more precise, assume that the advertiser is willing to pay Q_u for each action, and the probability of an action by user u is p_u, i.e., the expected revenue of referring a product to u is $p_u Q_u$. In this setting, the revenue $f_u(S)$ from a user u given a subset $S \subseteq U \cup \{\mathcal{P}\}$ of users is computed as follows: if $\mathcal{P} \in S$ and there exists a user $u' \in S$ who is connected to user u and activates u, then $f_u(S) = p_u Q_u$, otherwise $f_u(S) = 0$ We should note that it is possible that k of u's friends play role in activating u, therefore fair ways to share the revenue among the contributing users should be explored. We will discuss such fair methods in Section 2.4.

2.3.2 Multi-level Propagation Model.

The multi-level model generalizes the single-level model by sharing the revenue with all the users collaborating in a cascade of referrals. In other words, the multi-level model keeps track of the path of users activating their friends in the network and shares revenue with all of them. This way, users get credit from their friends' friends activation. This propagation can grow for multiple levels, which gives more incentive to users to make referrals because they will earn money not only from direct activations but also from all the chains they are part of. Note that in this model, the assumption is that users do not have an incentive to refer every item to all their friends, since that will result in losing their credibility and ruining their reputation.

Modeling this as a cooperative game, the players are the SNP and all the users who participate The worth function would be as follows. In such a model, referrals are propagated through paths, and the expected revenue for a referral from a user u is $p_u Q_u$ if there is a path of users propagating the referral iteratively to user u, i.e., $f_u(S) = p_u Q_u$ if $\mathcal{P} \in S$ and there is a path of users $v_1, v_2, \ldots, v_l \in S$ where v_1 originally has become active, and then each user v_i activates user v_{i+1} after receiving it from v_{i-1}, and finally v_l activates u, otherwise $f_u(S) = 0$.

There are various ways for the SNP to keep track of activation paths. These various ways would, in turn, impose different revenue sharing schemes. For example, the SNP may only keep track of the first user who makes a referral to each user. Alternatively, the SNP may keep track of all users who made referrals to another user. Different methods for keeping track of users making referrals can be divided into two main categories. Consider a set $w_1, w_2, \ldots, w_k$ activating user u.

Graph-based Model: The SNP may keep track of all users who activated u, i.e., we may put edges from all nodes $w_1, w_2, \ldots, w_k$ to node u. **Tree-based Model**: We may put an edge only from one node w_i to user u, e.g. we may put an edge only from w_1 who is the first user who activated u encourage users to make referrals as early as possible.

For each of the above multi-level models, we can define a *k-level propagation model* in which the revenue sharing for a node u happens among at most k users on the path to u, i.e, the revenue from node u is shared only among the last k nodes on a path to node u. For example, in the *k-level tree-based model*, the revenue share for a user u is shared among k top parents of node u in the corresponding propagation tree. Throughout this chapter, we mainly study the general multi-level model both for the graph-based and tree-based models, but all of our results hold for the k-level

variant of these propagation models. In parts of this chapter that we need to distinguish between different k-level models, we specify the k-level propagation model.

2.4 Computing the Shapley Value

The challenge in the above mentioned models is to design a fair mechanism to share the added revenue with users that participated in the process. Shapley value as described in details in Section 2.3 not only ensures fairness, but also has other desirable properties. In what follows, we discuss computation of the Shapley value for the proposed models. We either show that Shapley value can be computed polynomially, or prove a hardness result and provide approximate solutions.

2.4.1 Single-level propagation model.

Assume that k users have activated user u. Here, we discuss simple fair ways to share the revenue gained among the contributing users and the SNP. One way is to just consider the first user who starts the propagation, in which case the revenue should only be divided between that first user and the SNP. Alternatively, we may share the revenue with all these k users[1]. Consider user u, let K_u be the set of k_u users who activated user u, and let the revenue of user u becoming active be $p_u Q_u$. Then, the Shapley value revenue share of SNP for each user u is $\frac{k_u}{k_u+1} p_u Q_u$ and the Shapley value revenue share of each of these k_u users is $\frac{1}{k_u(k_u+1)} p_u Q_u$. Letting k_u be the set of users who have activated user u, and summing up the Shapley value revenue share for all users, the revenue share of the SNP in the single-level model is $\sum_{u \in U} \frac{k_u}{k_u+1} p_u Q_u$. Also letting A_i be the set of users who have been referred by user i, Shapley value for user i would be $\sum_{u \in A_i} \frac{1}{k_u(k_u+1)} p_u Q_u$.

2.4.2 Graph-based Multi-level propagation model

In this section, we show that computing the Shapley value in the graph-based multi-level model is computationally hard, and in fact is $\#P$-hard. The proof is by reduction from a node variant of the NETWORKRELIABILITY problem, called NODERELIABILITY. Both problems are defined below:

[1]Considering the influence model known as the threshold model, it is reasonable to assume that all referrers should receive some credit.

NETWORKRELIABILITY**Problem**
Instance: Graph $G = (V, E)$, a rational failure probability $p(e)$ for each $e \in E$, nodes s and t.
Question: If edge failures are independent from each other, what is the probability that there exists a path from s to t in this graph?

NODERELIABILITY**Problem**
Instance: Graph $G = (V, E)$, a rational failure probability $p(v)$ for each $v \in V$, nodes s and t.
Question: If node failures are independent from each other, what is the probability that there exists a path from s to t in this graph?

The NETWORKRELIABILITY is known to be $\#P$-complete, even for a fixed probability $p(e) = \frac{1}{2}$ for each edge e [35, 69]. Using this, it is not hard to show that NODERELIABILITY is also $\#P$-hard by giving a reduction from from the edge variant: Given an instance of the NETWORKRELIABILITY problem, construct an instance (s, t, G') of the NODERELIABILITY problem as follows: Let $V(G') = E(G) \cup \{s, t\}$, i.e., for each edge e in the graph G, put a node v_e in graph G' with $p(v_e) = p(e)$, and also put two nodes corresponding to s and t. Two nodes in G' are adjacent if their corresponding edges or nodes in G are adjacent. One can easily verify that the probability of having a path in a random sample of G in the NETWORKRELIABILITY problem is the same as the probability of having a path from s to t in the random sample of G' in the new instance of the NODERELIABILITY problem.

Theorem 1 *Computing Shapley value in the multi-level graph-based propagation model is $\#P$-complete.*

Proof. Consider an instance of NODERELIABILITY problem as follows: Given a graph $G = (V, E)$ and two nodes s and t, and probability $p(v) = \frac{1}{2}$ on each node, compute the probability

of having a path from s to t in a random graph constructed by including each node of G with probability $\frac{1}{2}$. From this instance, we construct an instance of the Shapley value computation in the multi-level graph-based model. The propagation graph G' is the same as graph G with an additional node v and two edges (s, v) and (v, t), i.e., $V(G') = V(G) \cup \{v\}$ and $E(G') = E(G) \cup \{(s, v), (v, t)\}$. Now consider the revenue share of nodes in G' toward node t, that is the revenue shares toward $p_t Q_t$. We claim that the total revenue share of nodes other than v in G' is $p_t Q_t$ times the probability of having a path from s to t in a random graph where each node of G is present with probability $\frac{1}{2}$. To see this, note that in computing the Shapley value revenue shares of nodes in $V(G')\backslash\{v\}$, the probability that each node $u \in V(G)$ appears before v in a permutation is $\frac{1}{2}$. Thus the probability that each node is before v is $\frac{1}{2}$ and is independent of any other node appearing before v. Also, if a path from s to t appears completely before v in a permutation, the marginal value of node t which is $p_t Q_t$ goes to one of the nodes in $V(G') - \{v\}$. Therefore, the total revenue share of nodes in $V(G') - \{v\}$ in the multi-level graph-based propagation model over G' is equal to the the probability of having a path from s to t in a random subgraph of G where each node is present with probability $\frac{1}{2}$. Thus if we can compute the Shapley value revenue shares, we can solve theNETWORKRELIABILITY problem which is $\#P$-hard. ■

2.4.2.1 Approximating Shapley Value

In light of the above hardness result, we design an algorithm based on sampling to approximate Shapley value for the graph-based multi-level model. It can be observed that by simply using polynomial number of samples we can compute Shapley values approximately in this general model.

Theorem 2 *If $\phi_u(S, f_v) > \frac{P_v Q_v}{n^3}$, then we can compute $\phi_u(S, f_v)$ within factor $(1 \pm \epsilon)$ with high probability (i.e., probability $1 - o(1)$), in time polynomial in $\frac{1}{\epsilon}$ and n. Otherwise, if $\phi_u(S, f_v) \leq \frac{P_v Q_v}{n^3}$, one can approximate it within multiplicative factor and $\frac{P_v Q_v}{n^3}$ additive error, in time polynomial in $\frac{1}{\epsilon}$ and n.*

2.4.3 Tree-based propagation model.

Here, we show that the Shapley value revenue share of each user in the tree-based model can be computed easily.

Lemma 3 *In the multi-level tree-based propagation model, the Shapley value revenue share of each user can be computed in time* $O(n^2)$.

2.5 Supermodularity

In this section, we observe a main advantage of the tree-based propagation model compared to the graph-based propagation model. In particular, we show that the revenue function for the multi-level tree-based model described above is supermodular, and this implies various nice properties of the Shapley value revenue shares for the tree-based propagation model. For example, this shows individual rationality of these revenue shares for the corresponding cooperative game. We first prove the supermodularity and then switch to summarizing the corollaries.

As we have explained at the end of Section 2.2, it is important to note that although this result might seem related to the multi-cast cost sharing problem, the results are different. Before stating the proof of supermodularity for the tree-based model, we observe that this property does not hold for the graph-based propagation model, and even for the single-level propagation model with uniform valuations for the revenue shares. To see this, consider the following example:

Example 1 *Consider a single-level model with 4 users* A, B, C, D *in which all 3 users* A, B, C *make referrals to user* D *and* D *becomes active. Let the revenue of each user be* 1. *In this case, the value of each subset of size 2 including* s *and one of* A, B *or* C *is 2 (since using* A, B, *or* C, *the path from* D *is formed). Also* $f(s) = 0$, $f(s, A, B) = 3$, *since there are three nodes reachable from* s *each with revenue* 1. *This example violates supermodularity as follows:* $f(s) = 0$, $f(s, A) = 2$, $f(s, B) = 2$ *and* $f(s, A, B) = 3$ *and* $f(s, A, B) - f(s, B) = 3 - 2 = 1 < 2 = 2 - 0 = f(s, A) - f(s)$.

Theorem 4 *The revenue function in the tree-based multi-level propagation model with uniform valuation for all users* $p_u Q_u = P$ *is a supermodular set function.*

Proof. A potential way to show that function f is supermodular is by showing that the set function f_u for any user u is supermodular. However, it is not the true in this case, i.e., for some users u, the set function f_u might not be supermodular. Nevertheless, we can show that the summation $f(S) = \sum_{u \in U} f_u(S)$ is supermodular. In the tree-based propagation model, for a subset $S \subset U \cup \{s\}$, $f(S) = 0$ if $s \notin S$, and otherwise, $f(S)$ is equal to $|T(S)|P$ where $T(S)$ is the maximal connected subtree rooted at s with all internal nodes in S. In other words, $f(S)$ is proportional to the number of nodes that are connected to s using a path whose all internal nodesare in S. Knowing $f(S) = P|T(S)|$ in the uniform valuation model, it is sufficient to prove that $|T(S)|$ is supermodular. We do so by verifying the supermodularity property of f by proving that for any two subsets $A \subset B$ and any element $i \notin B$,

$$|T(B \cup \{i\})| - |T(B)| \geq |T(A \cup \{i\})| - |T(A)|.$$

Letting $\Delta_i(B) = T(B \cup \{i\}) - T(B)$ and $\Delta_i(A) = T(A \cup \{i\}) - T(A)$, it is sufficient to prove that $\Delta_i(A) \subseteq \Delta_i(B)$. To show this, we consider two cases:

Case 1: i's parent is not in $T(A) \cap A$. In this case, adding i to A does not change $T(A)$, and thus $\Delta_i(A) = \emptyset \subseteq \Delta_i(B)$.

Case 2: i's parent is in $T(A) \cap A$. In this case $\Delta_i(A)$ contains all non-root nodes of the maximal connected subtree rooted at i with all internal nodes in A. In this case, i's parent is also in $T(B) \cup B$, and $\Delta_i(B)$ contains all non-root nodes of the maximal connected subtree rooted at i with all internal nodes in B. Now since $A \subseteq B$, the corresponding connected maximal subtree in the induced graph of B is larger than the maximal subtree rooted at i in A. The result follows from the above case analysis, as it shows that $\Delta_i(A) \subseteq \Delta_i(B)$, and thus $|\Delta_i(A)| \leq |\Delta_i(B)|$ which implies supermodularity of f. ■ The supermodularity of revenue shares, in turn, implies individual rationality, computability, and incentive compatibility results for the revenue shares based on the tree-based propagation model with uniform valuations.

2.6 Single-level vs. Multi-level

In the previous sections, we argued for an advantage of using tree-based propagation model compared to a general graph-based propagation model. In this section, we compare single-level and

multi-level propagation models, and show which revenue sharing strategy is more effective in maximizing the spread of viral marketing or advertising campaign.

Let M be the total revenue share for each new user who gets the referral, i.e., $M = pR$ where p is the probability that a new user becomes active, and R is the amount that the online retailer or the advertiser pays as the total revenue share with each user. We observe that there exists a direct relation between the amount of revenue shared with users and the amount of spread of the advertising campaign. Intuitively, the more revenue share there is, the more probable the spread is. Here, we observe that other than the parameter M, the revenue sharing scheme also plays an important role in the expected reach of the viral marketing method. Intuitively, for multi-level revenue sharing schemes, the potential gain of each user for making referrals is more than the potential gain for single-level sharing schemes. The reason is that the user not only gains a revenue share from her immediate neighbors, but also from her neighbors of neighbors and so on. This in turn increases the probability of making referrals by users, and thus it may result in higher expected reach of the marketing for multi-level revenue sharing schemes.

We model the potential reach of a revenue sharing strategy by simulating a random propagation process on real-world networks, and reporting the simulation results. In order to simulate this process, we consider the following model over a network: Each each user u has a random threshold t_u, where t_u is chosen uniformly at random from $[0, 1]$. Before making a decision to make a referral or not, each user computes her potential gain, Potential(u), and makes referrals if $Potential(u) \geq t_u$. This propagation model is inspired by the probabilistic threshold model that is widely studied as a model for viral marketing [70]. An important feature of our model is the way we compute the potential gain of users. In fact, the main difference in various revenue sharing policies is the way users compute their potential gain. If we use a k-level tree-based propagation model for a large k, there is a larger potential revenue from referrals (at the beginning of the propagation process). In such a setting along with a k-level tree-based revenue sharing, a user u may get a revenue share $\frac{M}{k}$ (or $\frac{M}{t}$ for some $t \leq k$) for each new user who gets a referral from u, and also u may get a revenue share of $\frac{M}{k}$ from each new user who becomes active through u, and so on. As a result, the total potential gain of user u for making referrals is proportional to the potential number of new users within a distance k of user u who became active through u. At the beginning of the propagation process, this potential is larger for each node u, but as time goes on, more people hear about it

through other friends, and the chance that a new user is informed by user u becomes lower. The more k is, the more potential gain users have at the beginning, and thus there is a higher chance that they start propagating referrals. On the other hand, for a larger k, after a fixed number of steps, more people have already heard about the product/ad, and the potential gain of users to propagate it to new users is lower. Because of this tradeoff, using different propagation models will result in different expected number of users who get the referral.

We simulate the above process for different k-level tree-based propagation models for several networks, and report the total expected number of users who end up getting the referral using different propagation models. We denote this expected number of users who hear about a product using a k-level tree-based model by $E(k)$. As we will explain in details, in all simulations, we observe that this expected number $E(k)$ increases as k increases.

Data. To evaluate the performance of our models we tested them on five large real networks. To avoid repetition, we report one of them in this chapter of the thesis as the behavior for all the networks were similar. The network presented here is a who-trusts-whom network of Epinions.com which is a consumer review website. Users of the site can explicitly indicate wether they trust another member or not. The graph contains a node for each user and there is an edge from each user to the other users that she trusts. The graph has 75879 nodes and 508,837 edges between the nodes [92].

Results. The performance of the k-level propagation model is depicted in Figure 2.2 for the Epinions network. The simulation is done for $k = 1, \ldots, 20$, for $M = 0.3$ and $M = 0.5$. The plots indicate the number of users that hear about the product for each k-level simulation. As the plots show the number increases as k increases up to 13 lvels for the Epinions network. At level $k = 13$, the network is saturated.

We observe that the saturation level is close to the diameter of the graph. This implies that implementing a k-level model beyond the diameter will not improve the number of informed users. The 90 percent effective diameter of the graph is 5.8. We also observe that the relative increase in the number of users getting the referral is at maximum between levels 4, 5 and 6. We also find that, especially in the case where the total revenue share M is smaller, having a k-level propagation model is even more effective to trigger users initiate propagation. This observation can be interpreted intuitively as follows: when M is small, the single-level model might not be incentivizing

Table 2.1: Epinions dataset properties.

Nodes	Edges	Avg. Degree	Diameter	largest CC
75879	508837	13.41	13	32223

enough for the users to opt in, however, by increasing the levels the potential increases which gives more incentive to users to take part in this process.

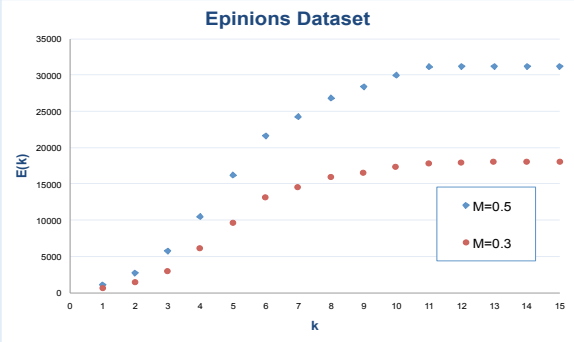


Figure 2.2: Epinions.com

2.7 Conclusion

In this chapter, we have developed multi-level revenue sharing schemes for viral marketing over social networks. For the proposed models, we develop results on the computational complexity, individual rationality, and potential reach of employing the Shapley value as a revenue sharing scheme. Our results indicate that under the multi-level tree-based propagation model, the Shapley value is a promising scheme for revenue sharing, whereas under other models there are computational or incentive compatibility issues that remain open. We also assess the effectiveness and potential reach of the single-level and k-level tree-based models through simulations, and our findings show that using a k-level tree-based model has higher potential for increasing the spread over the social network.

Chapter 3

Optimizing Display Advertising on Social Networks

3.1 Introduction

Advertising is one of the main sources – if not the main source – of revenue for most online social networks; e.g., online advertising comprised about 91% of Facebook's revenue in 2013 [3]. A significant portion of this revenue comes from display ads where certain business rules from traditional advertising apply. Conventional display ads (and more generally traditional online advertising methods) have not been as successful in generating revenue in social networks compared to other types of online advertising such as sponsored search advertising. To increase their revenue, online social networks have utilized a new paradigm called *social advertising*, which aims to leverage the social influence of a user on their friends. In this chapter, we develop a formal model to study social advertising in display ads, and formalize the algorithmic problem faced by a social network platform. We then present new theoretical and empirical results for this problem.

Let us start with a brief review of display advertising: An online social network has multiple pages where it displays ads in the form of images, video or text. Once a user visits a page, she views an ad. Her exposure to the ad is called an "impression." Advertisers buy blocks of impressions ahead of time via contracts, choosing blocks carefully to target a particular market segment. Once the contract is agreed upon, the advertiser expects a specified number of impressions to be delivered by the social network platform over an agreed-upon time period. In addition to deliver-

ing a predetermined number of impressions, the advertiser may choose to optimize other objectives like clicks or conversions [7, 6, 5].

It has been observed, however, that conventional methods of ad allocation are not very successful in the context of online social networks [24, 108], and the paradigm of *social advertising* was developed to address this shortcoming. The goal of social advertising is to leverage social influence among users. The impact of social influence among users has been confirmed in sociological studies, statistical models, and in online randomized experiments (see [24] and references therein). A social ad has been defined as "an ad that incorporates user interactions that the consumer has agreed to display and be shared. The resulting ad displays these interactions along with the user's persona (picture and/or name along with the ad content)" [64]. In other words, an ad being shown to a person can incorporate information about others who have clicked the ad in the past (assuming their consent).

The goal of our work is to understand how social ads can affect click-through rates in social networks. Recent field studies [108, 24] show that advertising using social cues are more effective than conventional demographic and behavioral targeting methods.

Our Contributions and Techniques. The main contributions of this Chapter are as follows:

- We propose a formal model for social ads in the context of display advertising. In our model, ads are shown to users one after the other. The probability of a user u clicking an ad depends on the users who have clicked (or taken a certain action on) this ad so far. This information is presented to u as a *social cue* (which could be of different kinds, as we see later), thus the click probability is a function of this cue.

- We introduce the *social display optimization* problem: suppose an advertiser has a contract with a publisher for showing some number (say B) impressions of an ad. What *strategy* should the publisher use to show these ads so as to maximize the expected number of clicks?[1]

- We show that this optimization problem is APX hard.[2] In fact, under a complexity assumption known as the *planted dense subgraph* conjecture (widely believed to be true, see [30]),

[1] We will discuss publishers' motivations for this in Section 3.2.1.

[2] Hard to approximate to some constant factor unless $P = NP$.

we prove that it is impossible to find a strategy that approximates the best one to a factor better than $n^{1/8-\epsilon}$, where n is the total number of users, and $\epsilon > 0$ is any constant.

- In light of the general hardness results, we develop heuristic algorithms and compare them to natural *baseline* ones. Inspired by influence-and-exploit strategies studied in [70, 56], we propose a two-stage algorithm: we first show the ad to αB users of highest influence, and then show it to users most likely to click (given the ones that clicked so far).

- Finally, we evaluate the performance of this as well as other heuristics on two real datasets with influence probabilities from Flixster.com and GoodReads.com.

Our model shares some common elements with previous work on influence maximization and viral marketing, but our aim is quite different: we care only about the users we show the ad to (which we are allowed to pick), and the *strategy* used for doing so (as opposed to creating a cascade over the *entire* graph). The strong inapproximability results mentioned above are also in contrast to constant-factor approximations known for influence maximization models.

Formally, a display *strategy* will consist of a set of users and a specificied order of showing ads, both of which are adaptive, in the sense they depend on which users have clicked the ad so far (thus it can be viewed as a decision tree). Indeed the number of such *strategy trees* can be doubly exponential in the size of the graph. Proving hardness results thus involves reasoning about arbitrary adaptive strategies. This is our main theoretical contribution that might find other applications.

This is also the reason proving worst approximation guarantees is difficult – because any approximation algorithm must (implicitly) certify that there is no strategy that could obtain a value much larger than that output by the algorithm. While this may suggest that the problem is hopeless, we will see in the experiments that our two-stage algorithm (outlined above) outperforms the baseline heuristics by a large margin – typically by about 11% to 100%.

3.1.1 Related Work

There is a rich and diverse body of work in the intersection of display advertising, social networks, marketing and influence maximization. For purposes of exposition, we will discuss these results in

three categories. The most relevant to us is work on social network advertising, which is what we intend to formalize in the context of display advertising.

Social Network Advertising: Works on social advertising have looked at the impact of displaying social signals (cues) to users. In other words, they measure the increase in the likelihood of a user clicking an ad, given she knows that her friends have already taken an action on that ad. In particular, in [24], the authors run an experiment on Facebook to measure this impact. They find that showing social cues increases the probability of clicks on fan pages. Tucker [108] studies the same problem on a different network and makes similar observations. In a recent paper [25], the authors argue that viral marketing would be more effective if a large number of ordinary users are picked as influencers. None of the above work, however, looks at how one could optimize the number of clicks, likes or conversions in display ads by leveraging these social cues.

Recently, a social ad model considering user influence, called AdHeat, has been explored [26]. In this model, the advertising platform diffuses hint words of influential users to others and then matches ads for each user with aggregated hints. They perform experiments on a real-world data set, and show that AdHeat outperforms the traditional relevance models by a large factor. Although this study shows the effectiveness of using social network information in online advertising, they do not consider active propagation of ads by the users of the social network.

Viral Marketing and Influence Maximization: The problem of influence maximization in social networks has received a lot of attention in the past decade or so, with applications to viral marketing, studying the spread of diseases, and a variety of other settings. Introduced in the seminal work of Kempe, Kleinberg and Tardos [70], the goal is to pick a small set of vertices to *influence*, with the goal of maximizing the expected number of nodes that this influence *cascades* to. We do not get into the formal definitions, but note that these works [37, 70, 56, 87] give formal ways to model the probability of a user buying a product based on her friends buying the product.

This is very similar to the way in which our work models the click probability, and our model is indeed inspired by this literature. However, as we stated earlier, our goal is to find good display strategies, which is quite different from finding good nodes from which to start a cascade. Thus it seems the algorithmic tools developed there do not apply to our setting (indeed our hardness results imply that we *cannot* obtain constant factor approximations, as in the case of influence maximization).

Another related work is the revenue maximization model [56, 85], in which a person's decision to buy a product is influenced by the set of other people who own the product, as well as the price at which the item is offered. The results in this line of work have a more economic focus, and thus have a very different flavor compared to ours.

Online display ad allocation: The problem of optimal allocation for display ads has been recently studied as an online optimization problem [43, 42, 109]. The display ad optimization model considered in these papers is similar to our model in that there is a goal of delivering the ad to a predetermined number of impressions while trying to maximize the expected number of clicks or conversions. Incorporating social influence into these settings is partly the motivation for our work.

3.2 Model

We will first give a brief background on display advertising, and thus motivate our main question of study. This will help set up the notation for describing the model formally.

3.2.1 Display Advertising

There are three major pricing models for online ads on the Internet: Cost-per-mille/impression (CPM), Cost-per-click (CPC), and Cost-per-action (CPA). In these models, the advertisers pay the platform (publisher) for the number of impressions, clicks, or actions[3] respectively. Even though a majority (in terms of revenue) of search advertising operates on the CPC and CPA models, a significant portion (roughly 33 percent, as of 2013 [4]) of display ads are sold based on the CPM model.

The CPM model is simple to describe: an advertiser enters in a contract with a web publisher for its ads to be shown to a fixed number of site's visitors. The advertiser may specify a segment of the market or some demographic criteria to target the ads to. The contract requires the publisher to show this number of impressions. Thus the publisher can choose which users to show the ad to (and on which pages). How should he make this choice?

Note that the more clicks an ad gets, the higher the chances that the advertiser would come

[3] Actions, or conversions correspond to a specific action by the online user, e.g., purchases of a product or signs up for newsletters on a website.

back for another advertising campaign. Thus it is in the publisher's interest to show the ads to users more likely to click. This is a well-recognized objective – in fact, most of the advanced display-ad platforms offer tools to optimize metrics like clicks, conversions, or even return-on-investment (ROI) while delivering a predetermined number of ads [7, 6, 5]. This motivates the question of optimizing the expected number of clicks (or conversions) subject to displaying a specified number of ads. We study this question formally in the setting of social networks.

3.2.2 Display ads in Social Networks

Let us consider the situation in which the publisher is a social network. The contracts now require the publisher to show an ad (from an advertiser) to a fixed number (say B) users of the social network. As we saw above, the publisher wishes to maximize the number of clicks or 'likes' the ad would receive. This, in turn, could be used for pricing the contract, or improving the customer-loyalty.

We will develop a way to model how users react to *social cues*, and use the model to optimize the number of *clicks* an ad is expected to get. Social cues can be of different kinds. In its most general form, the publisher could display to a user the entire set of friends who have clicked the ad so far. This has many problems – the first is the privacy of the users who have clicked the ad. A fix for this is (as in the experiments of [24]) is to only show users who have given consent. Even so, displaying a list is cumbersome; a realistic way is to show a small subset (say, ones with closest ties) of friends, or the fraction of friends, who clicked the ad (and consented to spreading the information). Additional information, such as the number or demographics of other people (non-friends) who clicked the ad may also be provided. We would like to have a model that is general enough to capture the probability of a user's click probability in all of these settings.

We now formally describe such a model. Let B denote the *budget*, i.e. the number of ads that are to be displayed. We show ads one by one to users. At some point of time, suppose S is the set of users who have clicked the ad. Then the probability that a user u clicks the ad is given by an *influence* function $p_u(S)$. We assume that $p_u(S)$ is increasing with S, i.e., the probability that a user clicks an ad only increases if more people click the ad (e.g. $p_u(S)$ may be a linear function or a submodular function). Given this setting, the objective of the publisher (the social network) is to find the optimal set of users and the optimal *display strategy* (described below) in order to

maximize the number of clicks the ad receives.

What kind of functions $p_u(S)$ are possible? From our discussion above, the user u does not see all of S, but only a *social cue*, which is something derived from S. Thus $p_u(S)$ is only a function of the social cue that u receives. In Section 3.2.2.1, we will see a list of reasonable candidates for $p_u(S)$ and the corresponding social cues.

Display strategy Formally a display strategy is a binary decision tree of depth B, the total number of impressions. The vertex at the root is the first user to be shown the ad. If the user clicks the ad, we follow the display strategy in the left subtree, else the right subtree, and so on. Note that even if the same user appears in the tree at the same depth, the probability of him/her clicking the ad will depend on the set of users who clicked so far (which is captured by the path from the root). Given a strategy, we can define the *expected clicks*, which is the expected number of users who click the ad if the publisher follows this strategy. This can be computed by a bottom-up computation in the tree.

The caveat here is that a strategy tree typically has exponential size (since it is a binary tree of depth B), thus computing the expected clicks is non-trivial. All the algorithms we consider will have a 'succinct description' of the tree (at each step, it will be a simple computation to pick the next root), however it is still not clear how to compute the expected clicks. We will compute this quantity using Monte-Carlo simulation. This can be done efficiently, because the variance is at most k^2 (in practice it is much smaller), and thus we get a good estimate of the expectation with only a few samples. The vast number of strategies is one reason it is difficult to reason about the optimal strategy (one with the largest expected clicks).

Adaptivity vs. non-adaptivity We have allowed our display strategy to be adaptive (the publisher can decide who to show the ad to based on which users clicked it so far). This is reasonable in most realistic cases. There are instances in which adaptivity gives a huge advantage (B vs. B^{ε}, for small ε). We do not get into them due to space constraints.

3.2.2.1 Influence Functions

In social advertising, the probability of a user u clicking or liking an ad could increase depending on the knowledge that certain other users have clicked the ad in the past, due to an inherent trust

in the taste or judgement of those users. This is what we capture using *influence* functions, as we defined earlier. Below we will list out some influence functions $p_u(S)$ we consider.

- *Linear influence*: Here for each pair (u, v) of users, we have a weight $w(u, v)$ (not necessarily symmetric), and $p_u(S) = c_u + \sum_{v \in S} w(v, u)$.[4] The constant term c_u could be zero for certain vertices. An interesting special case is one in which the weights $w(u, v)$ are all in $\{0, p\}$, i.e., given a graph $G(V, E)$ over users, $w(u, v) = p$ if $(u, v) \in E(G)$ and 0 otherwise. In this case, $p_u(S) = c_u + p \cdot |S \cap N_u|$. This special case is particularly interesting because it is very easy to communicate $p_u(S)$ via social cues – we can simply tell a user the number of friends who clicked the ad so far.

- *Independent Cascade Model*: This is discussed and motivated in [70]. Here as above, we have influences $p(u, v)$ for pairs of users, and $p_u(S) = 1 - \Pi_S(1 - p(u, v))$. In our context, we need to allow certain vertices u to have $p_u(S) = c_u$ for constants c_u (otherwise no one would click the ad to start with). We note that when $p(u, v)$ are all small, $\Pi_S(1 - p(u, v))$ is roughly $1 - \sum_S p(u, v)$, in which case this is a special case of linear influence.

- *Concave influence*: We have weights $w(u, v)$ as before, and have a concave function $g : R \to R$ such that $p_u(S) = g(\sum_{v \in S} w(v, u))$. Interesting examples of such functions are $g(x) = x^d$ for $d < 1$, and $g(x) = \log x$.

The linear and concave functions for influence are inspired by similar models considered in [56]. We could also have another *threshold based* functions $p_u(S)$, again inspired by [70].

- *Deterministic threshold function*: We have weights $w(u, v)$, and thresholds T_u. We have $p_u(S) = 1$ if $\sum_{v \in S} w(v, u) \geq T_u$, and 0 otherwise. We also need to have some vertices with $p_u(S) = c_u$ as explained earlier.

Allowing thresholds makes the problem extremely hard to approximate (and possibly unrealistic), thus we do not study algorithmic results for it.

Let us now formally define our problem.

[4] We cap probabilities at 1, though we do not explicitly write this.

Definition 1 ((Social Display Optimization)) *Given a tuple (U, B, p) of a set of users U, a bound B on the number of users to show an ad to, and influence functions $p_u(\cdot)$, the goal is to find a (possibly adaptive) strategy for showing the ad to B users so as to maximize the expected clicks. Sometimes we will simply refer to the problem as* display optimization.

3.3 Hardness Results

In this section, we will examine the complexity of the Display-Optimization problem (in terms of approximating the objective, which is the maximum expected clicks). We show that the problem is NP-hard to approximate up to a factor $(1+\varepsilon)$ for some small constant $\varepsilon > 0$.[5] Such a result is also called APX-hardness. Then, under a stronger hardness assumption, called the planted dense subgraph conjecture, we will show that we cannot approximate the optimal display strategy problem to a factor roughly $n^{1/8}$, where n is the number of users.

We first present the latter result — strong inapproximability under planted dense subgraph assumption — because it seems to highlight the crux of the problem, which is the following: if we wish to influence k users in a network, and we wish to take advantage of the graph structure, we should be able to find a set of k users who are well connected to each other, and this is hard in general. The reasoning below will make this rough intuition formal, and also illustrate how to argue about adaptive algorithms.

3.3.1 Strong inapproximability

We prove that for any $\varepsilon > 0$, the Display-Optimization problem cannot be approximated to a factor better than $n^{1/8-\varepsilon}$, unless we can approximate the random-planted version of the densest k-subgraph (DkS) problem to a factor better than $n^{1/4-\varepsilon}$ (conjectured to be hard [30]).

Let B be the budget, and suppose the probability that u clicks given S is the set of vertices that have clicked before, is given by

$$p_u(S) = \min\{1,\ p_0 + c \cdot |S \cap N(u)|\},$$

[5]Formally, it means that unless $P = NP$, it is impossible to tell if the optimal display strategy has expected clicks equal to M or expected clicks $\leq M/(1+\varepsilon)$ for some parameter M.

where p_0 and c will be picked appropriately. So a user has an 'independent' probability p_0 of clicking,[6] and there is an increase depending on the number of friends who clicked the ad. The aim, of course, is to maximize the expected number of clicks.

The planted DkS problem is the following: let $\varepsilon > 0$ be any constant; define two distributions over graphs as follows

$\mathcal{D}_1$: pick a graph from $G(n, n^{-1/2})$ (thus the expected degree is $n^{1/2}$).

$\mathcal{D}_2$: pick a graph from $G(n, n^{-1/2})$, and a random subset P of size $n^{1/2}$. Replace the induced subgraph on P by a graph from $G(n^{1/2}, n^{-(1/4+\varepsilon)})$.

To see that $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically different, we note that:

1. For a graph in $\mathcal{D}_1$, every induced subgraph on $n^{1/2}$ vertices has average degree $\leq O(\log n)$ with probability $1 - \exp(-n^{1/2})$. (Proof follows from Lemma 5.)

2. For a graph in $\mathcal{D}_2$, there exists an induced subgraph on $n^{1/2}$ vertices and average degree $\Omega(n^{1/4-\varepsilon})$ with probability $1 - \exp(-n^{1/2})$.

Conjecture 1 (Planted Dense Subgraph Conjecture) *Given a graph G, it is not possible in polynomial time to tell (with probability $> 2/3$) if $G \sim \mathcal{D}_1$ or $G \sim \mathcal{D}_2$. [30]*

Our theorem is now the following

Theorem 1 *Assuming Conjecture 1 (for some $\varepsilon \in (0, 1/16)$), it is not possible to approximate the Display-Optimization problem to a factor better than $n^{1/8-\varepsilon}$ in polynomial time.*

Proof. The reduction uses the same graph G (drawn from either $\mathcal{D}_1$ or $\mathcal{D}_2$), with parameters as follows:

$$B = n^{1/2} \;\; ; \;\; p_0 = \frac{1}{n^{1/8-\varepsilon/2}} \;\; ; \;\; c = \frac{1}{n^{1/8-\varepsilon/2}}.$$

We show that $\max \mathbf{E}[\#\text{clicks}]$ is (a) $O(\log^2 n) \cdot n^{3/8+\varepsilon/2}$ if $G \sim \mathcal{D}_1$, and is (b) $\Omega(n^{1/2})$ if $G \sim \mathcal{D}_2$, with high probability. These two claims easily imply the theorem.

[6] This is necessary for the clicking to *kick off*. We can simulate this by adding a new user who clicks with probability 1, and is connected to everyone with an edge of weight p_0.

It is easier to see (b). Suppose we are given a graph $G \sim \mathcal{D}_2$. Suppose P is the planted set of vertices, and suppose we show ads to the vertices in P (in a random order). Consider the situation after we show the ads to half the vertices in P. Of these vertices, $(B/2)p_0 = (1/2) \cdot n^{3/8+\varepsilon/2}$ vertices will have clicked the ad in expectation (and with very high probability, at least half this number). This means that for every remaining vertex in P, at least $\Omega(n^{1/8-\varepsilon/2})$ of its *neighbors* will have clicked the ad w.h.p. (Here we are using the fact that the planted subgraph is random and has degree $n^{1/4-\varepsilon}$). Thus by the choice of c, each subsequent vertex will click the ad with probability $\Omega(1)$, thus the expected number is $\Omega(n^{1/2})$ with high probability.

Now consider $G \sim \mathcal{D}_1$, and let $v_1, v_2, \ldots, v_B$ be any sequence of B users. Now suppose a display strategy shows the ad to these users in this order. Let S_i be the subset of $\{v_1, v_2, \ldots, v_{i-1}\}$ that clicked. We can upper bound $p_{v_i}(S_i)$ as:

$$p_{v_i}(S_i) = p_0 + c \cdot |S_i \cap N(v_i)| \leq p_0 + c \cdot |\{v_1, \ldots, v_{i-1}\} \cap N(v_i)|.$$

Now for $j \geq 1$, define vertex v_i to be in *level* j if $|\{v_1, \ldots, v_{i-1}\} \cap N(v_i)|$ lies in the interval $[2^{j-1}, 2^j)$.[7] Then, Lemma 5 (proved below) shows that with high probability ($\geq 1 - 1/n^2$, say), for *every* sequence $v_1, v_2, \ldots, v_B$, the number of v_i in level j is at most $\frac{O(\log n)n^{1/2}}{2^j}$.

Thus consider any (adaptive) display strategy. Suppose it shows the ad to users $v_1, v_2, \ldots, v_B$. Now divide these users into levels as above, and consider some level j. By the above, there are at most $O(\log n)n^{1/2}/2^j$ users in level j. Thus the expected number of these who click on the ad is

$$\frac{O(\log n)n^{1/2}}{2^j} \cdot \big(p_0 + \frac{2^j}{n^{1/8-\varepsilon/2}}\big) \leq O(\log n) \cdot n^{3/8+\varepsilon/2}.$$

By Chernoff bounds, the probability that the number who click is twice the expectation is $< 1/n^4$ in this case. Thus the total number of clicks is at most $O(\log^2 n)n^{3/8+\varepsilon/2}$ with probability at least $1 - 1/n^4$. This then implies that the expected number of clicks is at most $O(\log^2 n)n^{3/8+\varepsilon/2}$.

Note that the proof holds for every display strategy, thus concluding the proof. ■

It only remains to show Lemma 5.

Lemma 5 *Let $G = (V, E) \sim \mathcal{G}(n, n^{-1/2})$, and define $M = n^{1/2}$. Then with probability $(1 - 1/n^2)$, we have that for every $v_1, v_2, \ldots, v_M$, the number of edges in the induced subgraph is at*

[7] Include vertices with no edges to their predecessors into level 1.

most $(2\log n)M$. *Consequently, the number of* v_i *with* $> t$ *neighbors among* $\{v_1, \ldots, v_M\}$ *is at most* $(4\log n)M/t$.

Proof. Consider some $v_1, \ldots, v_M$. The probability that there are $\geq k$ edges is essentially

$$\binom{M^2/2}{k} p^k < \left(\frac{M^2 pe}{2k}\right)^k.$$

Now for $k = 2M\log n$, since $Mp = 1$, we have the probability above to be less than $e^{-2M\log n}$. Thus we can take a union bound over all choices of $v_1, \ldots, v_M$ (there are only $n^M = e^{M\log n}$ of them), completing the proof. ■

3.3.2 APX hardness

The theorem here is the following. Though the factor is much weaker, it is based on a much more standard assumption (NP hardness). This is the reason we include the result.

Theorem 2 *There is an absolute constant* $\epsilon > 0$ *such that it is NP-hard to approximate the Display Optimization problem to a factor* $(1+\epsilon)$.

We apply a result about the complexity of the k-uniform set cover problem (set cover in which all sets have size precisely k) and present a reduction from this. Formally, an instance of this problem consists of a family $\mathcal{S}$ of m subsets $S_1, \ldots, S_m$, each of size k, over a ground set of elements $[n] := \{1, 2, \ldots, n\}$. The goal is to find a subfamily of $\mathcal{S}$ of minimum size that covers all of $[n]$. The following hardness result is known:

Proposition 1 *[40] For every choice of constants* $s_0 > 0$ *and* $\varepsilon > 0$*, there exists a* k *(depending on* ε*) and instances of* $k-$*uniform regular set cover with* n *elements on which it is NP-hard to distinguish between the case in which all elements can be covered by* $t = \frac{n}{k}$ *disjoint sets (called YES instances), and the case in which every* $s \leq s_0 t$ *sets cover at most a fraction of* $1-(1-\frac{1}{t})^s+\varepsilon$ *of the elements (called NO instances).*

Reduction. We now give a reduction from k-uniform set cover to Display Optimization. Let V be the set of elements and $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ be a family of sets over V each of size k. The instance of Display-Optimization we construct is as follows: for each element $v \in V$, we have one user. For each $S_i \in \mathcal{S}$, we have a set U_i of $4\log n/p$ users, where $p = 1/n^{1/4}$. The edges are as

follows. We place an edge between $v \in V$ and $u \in U_i$ iff $u \in S_i$ (thus such a v has an edge to all the users in the 'group' U_i). Now the influence functions $p_u()$ are defined as follows. For users $u \in U_i$, $p_u(S) = p$ (i.e., these users click the ad with probability p independent of the rest). For users $v \in V$, we have

$$p_v(S) = \begin{cases} 1 \text{ if } |S \cap \Gamma(v)| \geq 1, \\ 0 \text{ otherwise.} \end{cases}$$

The number of users to show the ad to, i.e., the parameter B, is picked to be $(4t/p) + n$. (Recall $t = n/k$, defined above.)

We now show that if we started with a YES instance for set cover, there is a strategy which has expected clicks $\geq (1 - \delta)n + 4t$, and if we started with a NO instance, then *any* strategy has an expected clicks at most $(1 - \delta')n + 4t$, for some constants $\delta' > \delta > 0$. Recall that $t = n/k$ (k constant), so this will establish APX-hardness.

YES case. In this case there exists a sub-family of t sets – say $S_1, S_2, \ldots, S_t$ which cover all the elements V. Now consider the following display strategy:

1. Show the ad to precisely $4t/p$ users, according to the following algorithm: first show the ad to users in U_1 until either one of them clicks the ad, or we have exhausted all of U_1, then do the same with U_2, and so on, until the ad is either shown to $4t/p$ users, or we have one click in each each U_i for $i = 1, \ldots, t$. In the latter case, if we have not shown the ad to $4t/p$ users in total, show it to arbitrary (other) users in $\cup_j U_j$ so that the total is $4t/p$.

2. Then show the item to all the n users in V in any order.

Lemma 6 *With probability $1 - 1/n^2$, step 1 of the algorithm ends up with at least one click in each of $U_1, \ldots, U_t$.*

Proof. The full proof is technical, so we only give an outline in this version. The intuition is that in each of the U_i, by showing the ad to $1/p$ users, there is a probability roughly $1/2$ of some user clicking. Thus in roughly $t/2$ of the groups U_i, the algorithm will show the ad only to $1/p$ users in the group. Now a similar argument will show that in roughly $1/2$ of the rest, we require showing the ad to $2/p$ users, and in general, roughly $t/2^j$ of the groups will require showing the ad to j/p

users, for $j = 1, 2, \ldots, \log n$. Thus the total number of users to show the ad to, will be

$$\sum_{j=1}^{\log n} \frac{j}{p} \cdot \frac{t}{2^j} < 2t/p.$$

By allowing some slack in each bound, we can get high-concentration versions of these, which completes the proof. ■ Note that if we have one click in each U_i, $i \leq t$, then in step 2, we get n clicks. Further, the expected number of clicks in step 1 will be precisely $p \cdot (4t/p) = 4t$, and by Chernoff bounds, it will be $\geq 4t - \sqrt{40t \log n}$ with probability at least $(1 - 1/n^2)$. Thus with probability $\geq 1 - 2/n^2$, we have that in the YES case, the algorithm above gets $n + 4t - \sqrt{40t \log n}$ clicks. Thus the expectation is $\geq (1 - \delta)n + 4t$, for any constant $\delta > 0$ (for large enough n, since $t = n/k < n$). This completes the analysis.

NO case. Here we need to show that no display strategy can have expected clicks $> (1 - \delta')n + 4t$, for some absolute constant δ'. The key is to observe that an optimal strategy will (w.l.o.g.) first show the ad to users in U and then to users in V (this is because of the structure of our instance), and among the users in U, the order does not matter – the only thing that matters is the number of users in each U_i that are shown the ad.

Lemma 7 *In any strategy, the number of users in U who click the ad is at most $Bp + \sqrt{10Bp \log n}$ with probability $\geq 1 - 1/n^2$.*

Proof. Any strategy shows the ad to at most B users in U (B is in fact the total number of users it shows the ad to). For this lemma, it does not matter which users are shown the ad, because each user likes with probability p independent of all others. Thus the lemma follows by standard Chernoff bounds. ■ Note that the bound above is $Bp = 4t + np + O(\sqrt{t \log n})$, for our choice of p. We introduce a bit of notation: we will call a group U_i "good" if at least one of the users in U_i clicks the ad.

Lemma 8 *In any strategy, at most $40t$ of the groups are good with probability at least $1 - 1/n^2$.*

Proof. From the way n, t, p are related, we have $B < 5t/p$, thus any strategy shows the ad only to $5t/p$ users in U. This means that the number of groups in which the ad was shown to $> 1/(2p)$ users is at most $10t$ (else we would get a contradiction).

Thus if $40t$ groups are good in total, it means that in at least $30t$ groups, the strategy shows the ad to at most $1/(2p)$ users, and it manages to have one of the users like the ad. We show that this is very unlikely – can happen with probability at most $1/n^2$. Let us now perform a finer division. For each j, let n_j be the number of groups in which the ad is shown to $1/jp$ users, for $j = 2, \ldots, (1/p)$.[8] Define these groups to be in *level* j. By a calculation similar to the above, we have that

$$\sum_{j=2}^{1/p} n_j \cdot \frac{1}{jp} < \frac{5t}{p}.$$

For convenience, denote the quantity n_j/j by C_j. Then the above inequality becomes $\sum_j C_j < 5t$. We claim that for each j, the probability that there are $> 4(C_j + \log n)$ good groups in level j, is at most $1/n^2$.

The probability that some group in level j is good, is at most $1-(1-p)^{1/jp} \approx 1-\exp(-1/j) < 2/j$, for $j \geq 2$. Thus the expected number of good groups in level j is at most $2n_j/j = 2C_j$. Simple concentration bounds then give the claim above (because groups being good are independent events).

This then implies that with prob. $1 - 1/n^2$, the total number of good groups is at most $\sum_j 4(C_j + \log n) < 20t + t\log^2 n/p < 30t$. This completes the proof. ■

Once we have Lemmas 7, 8, it is easy to see that with probability $\geq 1 - 1/n^2$, the number of users who clicked the ad in total is at most (because we are in the NO instance of set cover)

$$4t + np + O(\sqrt{t \log n}) + \left\{\left(1 - 1/t\right)^{40t} + \varepsilon\right\} \cdot n < (1 - \delta')n + 4t,$$

for some absolute constant δ' (since ε can be picked small enough). This completes the proof of APX hardness.

Deterministic thresolds The proof above can be modified to show that if the influence function is allowed to be hard threshold, then there is no hope of approximating. More precisely we can show:

Theorem 3 *The Display-Optimization problem when some users are allowed to have a deterministic threshold function is NP-hard to approximate up to any polynomial* (n^c) *approximation factor.*

[8] Formally, we need to have the interval $(1/(j+1)p, 1/jp]$.

Proof. (Sketch) We modify the reduction above, noting that in the YES case, all n vertices of V would click the ad (in our algorithm, w.h.p.), while in the NO case, at most $(1 - \delta')$ fraction of V could click the ad w.h.p. (the same argument gives $1/n^{2c}$ instead of $1/n^2$ with minor changes). Now connect $M = n^{c+1}$ new users to all the users of V, and suppose these new users click the ad only if all n of V click the ad (threshold). Then in the YES case we get $M + O(n)$ clicks, while in the NO case we only get $O(n)$ clicks w.h.p. This shows inapproximability to an n^c factor. ■

3.4 Algorithms

In light of the above hardness results, we cannot hope to obtain algorithms with provably good approximation ratios. However, we will describe heuristics which perform much better than the natural baseline algorithms on real life data sets. The baseline algorithms are similar to those used in the context of influence maximization (and are known to give good algorithms for special cases).

3.4.1 Baseline algorithms

We present two natural heuristics for the problem, which we then use to compare the performance of our algorithms.

3.4.1.1 Largest probability greedy

One simple algorithm is to pick users that are most likely to click on the ad at each time step, i.e., given an input (U, B, p_u), pick users $\{a_1, a_2, \ldots, a_B\}$ as follows: at the i'th step, let $a_i \in U \backslash \{a_1, a_2, \ldots, a_{i-1}\}$ be the user maximizing $p_{a_i}(S_{i-1})$ where $S_{i-1} \subset \{a_1, a_2, \ldots, a_{i-1}\}$ is the set of users who have clicked on the ad so far. Ties are broken arbitrarily.

The problem with this heuristic is, intuitively, that it ignores the *future*. While picking users most likely to click, we may have picked ones that do not influence others (we will see examples).

3.4.1.2 Most influential greedy

The heuristic above ignores the influence of the chosen users on others. We can consider the other extreme: algorithms that pick the most influential users at each step (and ignore the click probability).

Most Influential Greedy Heuristic

Input: (U, B, p_u).

Output: A sequence $(a_1, a_2, \ldots, a_B)$, and set S.

Goal: Maximize $E[|S|]$.

1. **Initialize:** $S = \emptyset, A = \emptyset$
2. **For** $i := 1$ **to** B **do**
3. **Let** $a_i \in U \backslash A$ be the user maximizing top-infl(u)
4. **Let** $A := A \cup \{a_i\}$.
5. With probability $p_{a_i}(S)$, **let** $S = S \cup \{a_i\}$.

Figure 3.1: Most Influential Greedy Heuristic

Given an input (U, B, p_u) for an instance of social display optimization, we define a total influence infl(u) for each user u as $\sum_{v \in U} p_v(\{u\})$. The algorithm simply picks the top B users in the non-increasing order of infl(u). One issue with the above is that it considers the total influence of a vertex on all other nodes as opposed to the influence on at most B nodes. To deal with this, we define the *total top influence* top-infl(u) for each user u as $\sum_{v \in J(u)} p_v(\{u\})$, where $J(u) \subset U$ is a subset of users v with the B largest values of $p_v(\{u\})$. The greedy algorithm is to pick the top B users in the non-increasing order of top-infl(u).

This algorithm performs poorly in instances where the click probability and the influence are negatively correlated, but is reasonable on instances in which such correlations do not occur. We thus view the algorithm as a relevant baseline for evaluating better heuristics. The algorithm is formally stated in Figure 3.1.

3.4.1.3 Worst-case examples greedy heuristics

Let us describe concrete instances in which the baseline heuristics perform badly. These will inspire the more sophisticated algorithms.

For completeness, we also note that picking B vertices at random will perform very poorly. A simple example is a path with B vertices, in which the first vertex has a click probability 1

(always clicks), and each vertex has a click probability 1 if at least one neighbor has clicked, and 0 otherwise. It is easy to see that a random order does very badly, while the linear order obtains a value B. (The largest-probability greedy recovers this.)

Next we see that even in simple linear influence models with $B = n$, there are bad examples for the two greedy algorithms above. For the largest-probability greedy algorithm, consider a path as above, but with asymmetric weights: $w(i, i+1) = 1$ whereas $w(i+1, i) = 0$ (i influences the neighbor to the right, but not the one on the left). Suppose the probability $p_u(S) = \min\{1, \alpha + i\varepsilon + |S \cap N_i|\}$, where $\alpha = n^{\delta-1}$ for small δ, and $\varepsilon = 1/n^2$. The greedy algorithm picks vertices in the order $n, n-1, \ldots$ (because of the $i\varepsilon$ terms), and the expected number of clicks is only n^δ. However if shown in the order $1, \ldots, n$, then in roughly $1/\alpha$ steps, we see at least one click, and all vertices following that will definitely click, thus the expected value is $\Omega(n)$ for this strategy.

It is not hard to construct counter examples for most-influential greedy heuristic, with two sets of vertices, one with slightly higher influence but low click probabilities, which fool the greedy strategy.

3.4.2 Better heuristics

3.4.2.1 Adaptive Hybrid Heuristic

This heuristic is based on the simplest way to take into account both the influence and the click probability – the product of the two. More specifically, given an input (U, B, p_u) for an instance of the social display optimization, the algorithm greedily picks users $\{a_1, \ldots, a_B\}$ as follows: in the i'th step, let $a_i \in U \backslash \{a_1, \ldots, a_{i-1}\}$ be the user maximizing $p_{a_i}(S_{i-1}) \times \text{top-infl}(a_i)$ where S is the set of users who clicked so far, and $\text{top-infl}(a_i) = \sum_{v \in J(u)} p_v(\{u\})$, where $J(u)$ consists of users v who have not yet been shown the ad, and who have the $(B - i)$ highest values of $p_v(\{a_i\})$. The algorithm is shown in Fig. 3.2.

3.4.2.2 Two-stage heuristic

Inspired by the idea of the the influence-and-exploit strategies in viral marketing [70, 56], and the greedy algorithm for Densest k-subgraph [30], we propose the following two-stage heuristic: follow the most-influential greedy heuristics for the first stage of the algorithm, and then switch

Adaptive Hybrid Heuristic

Input: (U, B, p_u).

Output: A sequence $(a_1, a_2, \ldots, a_B)$, and set S.

Goal: Maximize $E[|S|]$.

1. **Initialize:** $S = \emptyset, A = \emptyset$
2. **For** $i := 1$ **to** B **do**
3. **Let** $a_i \in U \backslash A$ be the user maximizing $p_{a_i}(S) \times \text{top-infl}(a_i)$
4. **Let** $A := A \cup \{a_i\}$.
5. With probability $p_{a_i}(S)$, **let** $S = S \cup \{a_i\}$.

Figure 3.2: Adaptive Hybrid Heuristic.

to the largest-probability greedy algorithm in the second stage. More specifically, we can run the adaptive most-influential algorithm for the first αB steps, and then follow the naive greedy largest-probability heuristic in the last $(1 - \alpha)B$ steps. Our motivation for this greedy algorithm is to follow the intuition behind the greedy algorithm for influence maximization [70]. Although this algorithm does not provide a guaranteed approximation algorithm for this problem, we hope that this technique works well in practice, since the greedy heuristic has been very effective for influence maximization [70]. In fact, for the independent cascade model, the first part of the two-stage algorithm is the same as the maximum marginal greedy algorithm for influence maximization under cardinality constraint in this model. This algorithm achieves a $1 - 1/e$-approximation for the influence maximization problem with cardinality constraints. The optimal value of α certainly depends on the influence functions and the structure of the influence among users. Such an optimal value of α can be computed by trying a range of values for α and estimating the expected number of clicks via simulations. As part of our empirical study, we report a number of insights for the optimal choice of α for various settings.

Bad examples. Note that our bad example for largest-probability greedy (path with asymmetric influence) can be modified easily to give an $n^{1-\delta}$ gap for both the heuristics above. Influence of every vertex in that example is precisely 1 – so the adaptive hybrid works exactly like largest-

Two-stage Heuristic

Input: (U, B, p_u).

Output: A sequence $(a_1, a_2, \ldots, a_B)$, and set S.

Goal: Maximize $E[|S|]$.

1. **Initialize:** $S = \emptyset, A = \emptyset$
2. **For** $i := 1$ **to** αB **do**
3. **Let** $a_i \in U \backslash A$ be the user maximizing top-infl(u)
4. **Let** $A := A \cup \{a_i\}$.
5. With probability $p_{a_i}(S)$, **let** $S = S \cup \{a_i\}$.
6. **For** $i := \alpha B + 1$ **to** B do **do**
7. **Let** $a_i \in U \backslash A$ be the user maximizing $p_{a_i}(S)$
8. **Let** $A := A \cup \{a_i\}$.
9. With probability $p_{a_i}(S)$, **let** $S = S \cup \{a_i\}$.

Figure 3.3: The two-stage heuristic for specific α; the overall algorithm tries different α and picks the best

probability; if ties are broken badly (which is possible), this example is also bad for the two stage heuristic. However, the example is based on a *chain* of highly asymmetric influence, which is unlikely in real instances. We believe that this why our algorithms seem to perform quite well in real instances.

3.5 Empirical Evaluation

In this section, we evaluate variants of four heuristics discussed in the previous section on two families of instances taken from real-world datasets. After elaborating on our datasets, we report the improvement of the two-stage and adaptive hybrid algorithms over the two baseline algorithms.

3.5.1 Datasets

It is important to explain how we obtain the instances of Display-Optimization (in particular the influences $w(u, v)$) from real data. There is no a priori "correct" way.

Flixster[9]**:** Flixster is a social network for rating movies. We obtained the Flixster dataset from Goyal et al's work [52, 29]. This dataset contains 13,000 users with 192,400 directed edges between them. There are 1.84 million ratings done by these users. These statistics are presented in Table 3.1. The influence probabilities are learned by looking at the log of user ratings with time: $< u, i, t, r >$ (meaning user u rated item i at time t with rating r). We estimate the influence probability of user u on user v as the fraction of times user v rated an item after user u had rated that item. This fraction is then normalized over all neighbors of user v to make the sum of influence probabilities equal to 1.[10]

# Users	13,000
# Friendship links	192,400
# Ratings	1.84M

Table 3.1: Summary of Flixster Data Statistics

Goodreads[11]**:** Goodreads is a social book cataloging website where users can register books to create personal bookshelves and also form friendships with each other. The dataset contains 4,654 users with 445,947 edges between them [9] (statistics in Table 3.2). This time, we produce the influence probabilities according to the so-called *voter model*. This was introduced by [34] and [60] to model probabilistic influence. The model explains the diffusion of opinions in a social network as follows: in each step, each node changes her opinion by choosing one of her neighbors at random and adopting that neighbor's opinion. In [39], the authors show that degree is a good predictor of influence probabilities.

The voter model was introduced by [34] and [60] as a natural probabilistic influence model. The voter model explains the diffusion of opinions in a social network as follows: in each step, each

[9] www.flixster.com

[10] This typically overestimates causality; for the next dataset we consider a different way to estimate influences.

[11] www.GoodReads.com

# Users	592,081
# Friendship links	2,045,177
# Books	248,252

Table 3.2: Summary of Goodreads: owner-book information data

node changes her opinion by choosing one of her neighbors at random and adopting that neighbor's opinion. In [39], the authors show that degree is a good predictor of influence probabilities.

3.5.2 Experimental Setup

Here, we report the performance of our algorithms on the Flixster and GoodReads data sets. For each data set, we study the performance of these algorithms with $B = \beta n$ for four different values β, $0.02, 0.05, 0.1$ and 0.15. I.e., we set the the goal of showing the ad to 2%, 5%, 10% or 15% of the whole population, and report the results for each value of B. The way we compute the edge weights (probabilities) is described in Secton 3.5.1. As for the choice of the influence function, we examine the independent cascade model, and linear and concave influences. The concave functions we examined are $g(x) = \sqrt{x}$ and $g(x) = \log x$.

Finally, for each node u, the individual click probability $p_u(\emptyset)$ is drawn from a log-normal distribution with a large mean (between 0.1 and 0.45). We chose a large mean for these distributions to make sure that in the final click probability, the individual click probability is not dominated by the incremental probability due to influence.[12] The performance of the algorithms under the independent cascade model were almost the same as linear influence (as we noted, this is not surprising), thus we report the plots for the independent cascade model, and concave influences with $\sqrt{x}$ and $\log x$. The reason we report the results for both of these is to illustrate that our empirical observations are similar for seemingly very different influence functions.

As we discussed, we compare four algorithms (including two baseline heuristics). For the

[12]This probability could be much smaller for certain ad types. The goal of our empirical study is mainly to compare different heuristic methods. We observe that the magnitude of the click-through-rate numbers is not important for this comparison, and we expect to get similar relative performance if we scale all click-through-rates by the same factor. It is, however, important to choose the individual probability factors in such a way that their magnitude dominates that of the probabilities due to influence.

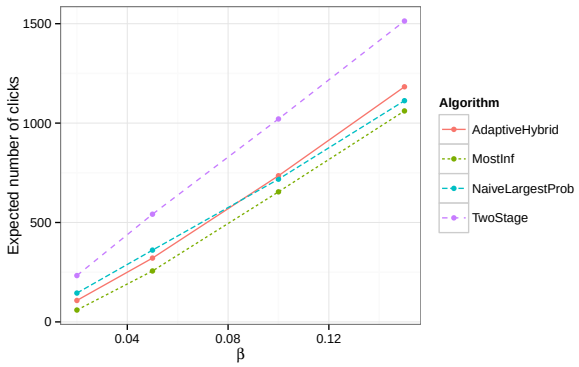


Figure 3.4: Performance of the heuristics on the Flixster dataset for the independent cascade model. The X axis shows β, where $B = \beta n$

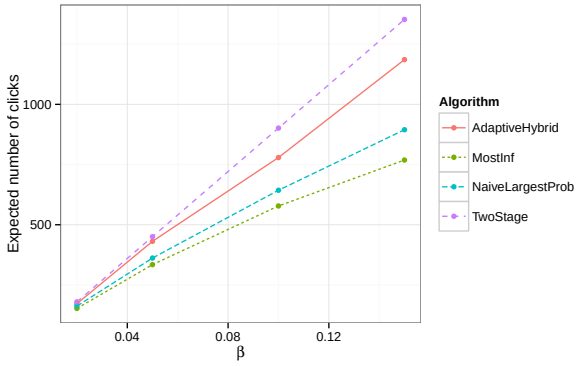


Figure 3.5: Performance of the heuristics on the Flixster dataset with a concave influence function, specifically $\log(x)$

two-stage algorithm, we try different values of α and choose the α with maximum expected value.

3.5.3 Observations

The empirical results for both data sets and for all propagation cases that we ran can be found in Figures 3.4, 3.5, 3.6, 3.7, 3.8, and 3.9. In these plots, the X axis changes β where $B = \beta n$. The Y axis is the expected number of clicks during the simulation. Here we summarize our main observations in these plots:

- Most notably, we observe that the two-stage heuristic algorithm consistently outperforms all the other heuristics. Across all instances, the gap between the performance of the two-

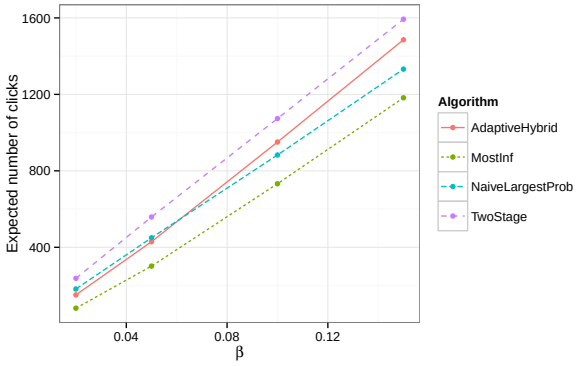


Figure 3.6: Performance of the heuristics on the Flixster dataset with a concave influence function, specifically $\sqrt{x}$

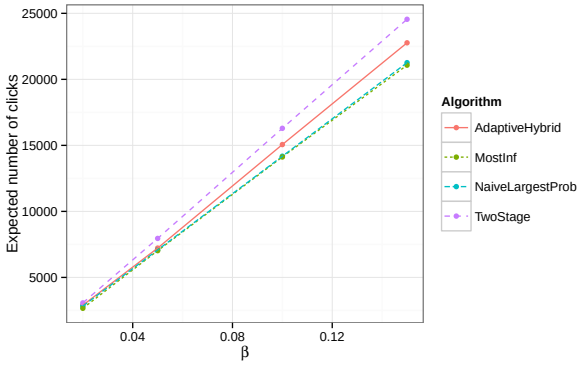


Figure 3.7: Performance of the heuristics on the GoodReads dataset for the independent cascade model

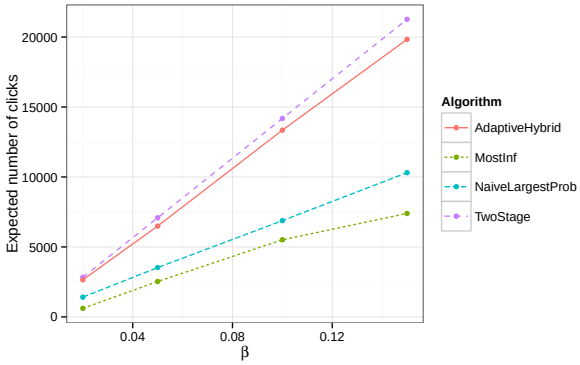


Figure 3.8: Performance of the heuristics on the GoodReads dataset with a concave influence function, specifically $\log(x)$

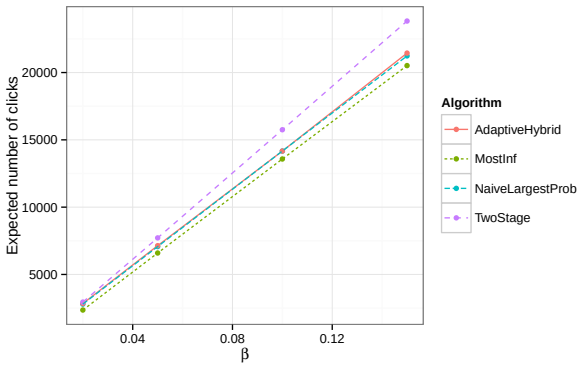


Figure 3.9: Performance of the heuristics on the GoodReads dataset with a concave influence function, specifically $\sqrt{x}$

stage heuristic and other algorithms increases as the the budget B increases. For example, for $\beta = 0.15$, for the Flixster dataset, the percentage increases for the two-stage heuristic from the best of other algorithms (i.e., adaptive hybrid) is around 25%, 14%, and 6% for the three different influence propagation models. For the GoodReads dataset, the percentage increases are around 7%, 5%, and 12%. The percentage increases from the output of the largest-probability heuristic (that ignores the influence function) to the output of the two-stage are 26%, 62%, and 23%. The same percentage increases for the GoodReads dataset are 11%, 100%, and 12%. The interesting parameter in two-stage algorithms is the α at which the best performance is achieved. See below for a discussion on how this behaves.

- Even our first heuristic, the adaptive hybrid greedy algorithm, outperforms the baselines for all values of β except $\beta = 0.02$ where the largest-probability heuristic is slightly better for two instances, and $\beta = 0.05$ where the largest-probability heuristic is slightly better for one instance. Again the performance increase from largest-probability heuristic to the adaptive hybrid heuristic increases as the budget increases.

- Finally, the most-influential greedy algorithm performs the worst. This was expected since it only focuses on picking the most influential users and not on their click probability.

Optimal α for the two-stage heuristic: As we discussed, in order to find the optimal α for the two-stage algorithm, we tried several values and chose the best one. The expected number of clicks for each value of α and for different budgets is plotted in Figure 3.11. The optimal choice of α

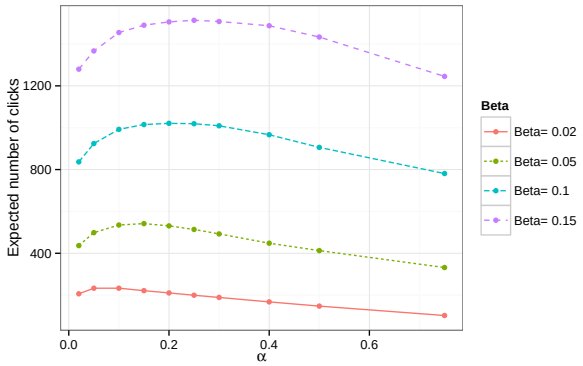


Figure 3.10: Two-stage heuristic for different values of α. Plot is for the independent cascade model on the Flixter dataset.

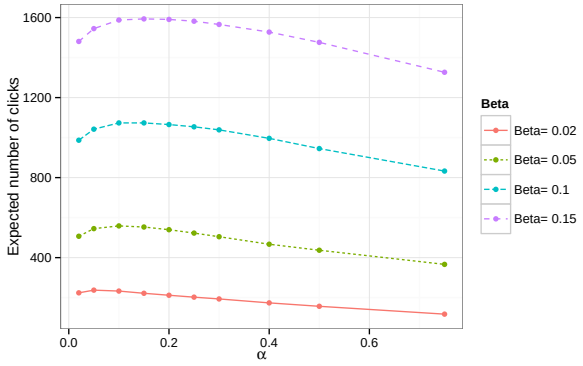


Figure 3.11: Two-stage heuristic for different values of α. Plot is for $\sqrt{x}$ influence function on the Flixter dataset.

for each instance may be related to the optimal way of influencing the network through ads. We observe that the optimal α for different instances vary from 0.05 to 0.2. It is worth noting the following points in these plots:

- The optimal choice of α increases as β increase from 2% to 15%. This suggests that with a higher budget, we can afford to spend a bit more fraction of time on 'exploration' (influencing) and attain higher expected clicks.
- The optimal choice of α decrease as the exponent of the concave influence function decreases. This is reasonable, because the lesser the influence among users, the lesser the importance of the influencing steps.
- Fixing an instance (i.e., fixing β and an influence propagation model), we observe that as α increases, the expected number of clicks from the two-stage heuristic is first monotonically increasing, and then it decreases monotonically (i.e., it is unimodal). This is a curious fact which would be nice to prove in restricted models. Of course, it also implies that the optimal α can be found by trying only a logarithmic number of values of α (by essentially binary search).

3.6 Conclusion

Social advertising has emerged as a promising alternative to conventional online advertising methods. In the setting of display advertising, relying on the notion that leveraging social cues can increase clicks in online social networks, we proposed a formal model for social display ad optimization, and initiated the study of the problem of optimally allocating display ads by modeling the impact of social influence on users' decisions. We showed that the social display ad optimization problem is APX-hard and is unlikely to be approximable within a factor much better than $n^{1/8}$. On the algorithms side, we proposed new algorithms which seem to perform significantly better than the baseline heuristics on datasets from real social networks. E.g., our two-stage algorithm achieved a 11% to 100% improvement over the output of baseline greedy algorithms. We also examined the question of the optimal "influence/exploit trade-off" for the two-stage heuristic under different choices of influence functions.

As a first step towards better display advertising in social networks, our work raises many interesting questions: can we develop algorithms with provable guarantees for synthetic graphs, such as the well-studied models for social networks? What other restricted influence models can we study for which we can prove provable approximation algorithms? What are good ways to learn influence weights from real data? Can we extend the model to incorporate multiple advertisers (each having a certain number of impressions)?

Our work suggests that the structure of the (induced) graph between the target users is a crucial parameter in this model of 'social display advertising'. It suggests that the graph structure, used well, could significantly increase the expected click rate. Running experiments to test this on real advertising platforms is an interesting direction for future research.

Chapter 4

Social Content Curation on the Web

4.1 Introduction

In recent years there has been an explosive growth of digital content in many forms, such as blogs, news feeds, as well as original and shared content on online social networks. Users are thus faced with a large collection on content that they receive through these various platforms. In particular, online social networks such as Google+, Twitter, and Facebook are rapidly turning into social reading or social content sharing networks, whereby each user shares content such as news, videos, etc., with her friends or followers. By re-sharing content, each user acts to filter items she receives or produces such that it is of interest to her followers. Indeed online aggregators also serve to perform a similar function of content curation where content is chosen such that it matches with subscribers' interest.

Content curation can be quite complex when users (following or subscribing to the same aggregator) have differing interests and a limited budget of attention. By the latter we mean that users have limited time to sift through contents in their stream. The curation function by the aggregator then must take this into account by choosing a relatively small set of contents that followers would find most interesting. We address the scenario of content curation where a set of aggregators, or publishers, aims to optimize the set of content to publish such that the followers receive contents of maximal interest to them.

We consider a very general setting with a set of aggregators or publishers, a set of followers or readers, and a set of contents or items of interest to the followers. Each reader has an intrinsic

interest in a given item, represented through a quantified value. The problem then is: how should the publishers choose a limited set of items to publish such that readers receive contents of maximal value? As we will show, solving the centralized optimization of this set of items is NP-complete, but admits approximation algorithms. We thus address the distributed content curation problem through a game-theoretic model where publishers are strategic agents with the aim of maximizing their utility, expressed in terms of feedback or incentives they receive from readers. In our analysis we aim to answer the following questions:

- Does distributed content curation based on selfish behavior of publishers converge to a pure equilibrium?
- How efficient is decentralized content curation? We define efficiency in terms of the price of anarchy, the ratio of social welfare under centralized curation to the sum of utilities achieved through distributed means.
- What is the rate of convergence to equilibria?
- How do the results change when readers are also strategic?

Our contributions. To the best of our knowledge our model is the first to address the problem of distributed content curation as a game. In the next section, we formulate the distributed content curation problem as a game where the readers receive utility in obtaining items of interest and the agents, the publishers, receive a reward from the readers for publishing items. The publishers objective is then to post items such that their rewards are maximized. We analyze this game in Section 4.3 and show that with a certain form of feedback, or reward, the distributed content curation game is efficient, in that the price of anarchy is bounded by 2. In particular, we show that the game has a pure Nash equilibrium(NE) and that best-response dynamics converge to a pure NE. We study the speed of convergence and show that in a polynomial number of best-response moves, players converge to a solution with an approximation factor of $2 + \epsilon$.

We study centralized curation in Section 4.4, where a central authority may optimize the set of items at each publisher so that the overall system utility is maximized. In this case, we show that the problem is NP-complete. We then show that there exists a $(1 - 1/e)$-approximation algorithm for this problem.

Lastly in Section 4.5 we consider the case the readers are strategic in a certain sense. Rather than obtaining all items posted at the publisher they are connected to, the readers are selective in choosing only a limited set of items of highest utility to them. We show that the price of anarchy in this case is also bounded by 2.

We conclude with a discussion on extensions to this model in Section 4.7.

4.2 Model

We consider a scenario where each user connects to, or follows content aggregators, and receives content such as videos, news updates, etc., posted by those aggregators or publishers. Our aim is to analyze content filtering behaviour whereby a set of users filter contents so as to optimize the content received by those connected to these users. As such, we consider a setting with a set of users that act as aggregators (or publishers), collecting and filtering contents from sources, and a set of users (or readers) that connect to these aggregators with the objective of receiving content of interest to them.We will assume a set of exogenous content sources, that is, sources that have no strategic actions. Examples of such sources are news sources such as newspaper and magazines, owners of online videos, etc.

We further consider the more realistic setting where readers have a limited budget of attention, that is, they spend a limited amount of time in the act of content consumption. In our model, this translates to a limited number of publishers that the readers consult, and a limited number of items obtained when a publisher is contacted. Further, each user has some intrinsic interest in the items, represented by a value that a user attaches to each item. We thus consider the dual problem of allocating links from each reader to a limited number of aggregators and selecting a limited number of items to host at publishers so as to maximize the total value received by users. We now formally specify our model.

4.2.1 System model

Our system consists of a set $\mathcal{P}$ of P publishers or aggregators, a set $\mathcal{R}$ of R readers or users, and a set $\mathcal{C}$ of C contents or items. This can be represented as a graph $G = ((\mathcal{C}, \mathcal{P}, \mathcal{R}), E)$ of three types

of nodes: content nodes, publisher nodes, and reader nodes[1], and of a set of directed edges. The edges are in the direction of content pull, in the following sense: a directed edge from a publisher node to a content node indicates that this publisher selects this content to host, and a directed edge from a reader node to a publisher node indicates that this reader follows this publisher. The term "following" in the social reading context has the meaning that a user receives contents posted by the publisher she follows. Define $\mathcal{F}_i$ to be the set of readers that follow publisher i, and $\mathcal{H}_j$ to be the set of publishers that reader j follows. We denote L to be the limit on a user's budget of attention, implying that a reader follows a limited number of publishers, or $|\mathcal{H}_j| \leq L$ for all readers $j \in \mathcal{R}$.

4.2.2 Publishing model

We assume that publishers post a set of items to their wall or stream and that this set is accessible to their followers. Examples are news aggregator streams or the Twitter stream of a user. We focus on the Twitter-like model where all the posts by publishers that a reader follows are presented to the reader in a combined stream format. Let $\mathcal{C}_p(i)$ denote the set of items posted by publisher i and $\mathcal{C}_r(j)$ the set of items received by reader j, that is, $\mathcal{C}_r(j) = \{k : k \in \cup_{i \in \mathcal{H}_j} \mathcal{C}_p(i)\}$. A limit on the budget of attention and filtering role of publishers implies that each publisher $i \in \mathcal{P}$ selects a set of items $\mathcal{C}_p(i)$ to post (or share), with $|\mathcal{C}_p(i)| \leq K$.

Readers have intrinsic interests in receiving certain items, and thus have valuations for these items. Denote $v(j,k)$ to be the value of item k to reader j. We define the utility received by a user j, u_j, as the total value of all items she receives: $u_j = \sum_{k \in \mathcal{C}_r(j)} v(j,k)$.

The publishers have no intrinsic interests in items, however receive a reward from their followers corresponding to the value of items posted. Specifically, publisher i receives a reward of $r_i(j,k)$ from reader j for posting item k. Note that these rewards can be in the form of feedback or incentives, through endogenous tools such as $+1$s, likes, retweets, or some exogenous form of monetary incentive. At this stage, we leave the type of this feedback quite general, and in the next section we formulate the reward structure that leads to desirable results.

We now define the content curation problem with strategic publishers as choosing a set of items to be posted at each publisher such that the global utility is maximized. We first consider the

[1]Note that we will use the terms *user* and *reader* interchangeably throughout this chapter, and similarly for the term pairs *content/item* and *aggregator/publisher*.

distributed case in Section 4.3, where publishers follow selfish dynamics to maximize their own payoffs. Next in Section 4.4 we study the centralized formulation. In Section 4.5 we will consider the problem with selective readers.

Note that in the present formulation we consider a set of discrete items to be shared. In reality, once an item is *consumed*, it is no longer of interest and a user would seek other items. However, our formulation includes the general case where content is classified by topics and users have intrinsic value for certain topics. Each *item* k in our formulation then represents a stream of content of *topic* k.

4.3 Content Curation Game

We first consider the case where the readers follow a fixed subset of publishers and collect all items shared by the publishers they follow and in return pay the publishers a reward. The strategic publishers then select a set of items to post that maximizes their payoff. Each reader j is assumed to follow a fixed subset of publishers $\mathcal{H}_j$ and receives a utility u_j that corresponds to the items she receives from those publishers.

Each publisher chooses a set of K items to post that maximizes her total reward. Denote by A_i the set of feasible actions available to publisher i, where a feasible action is a set of items, a_i, the publisher chooses, satisfying the limit of K items: $A_i = \{a_i \subseteq \mathcal{C} : |a_i| \leq K\}$. The reward, or payoff, that a publisher receives for action a_i is thus as follows:

$$W_i(a_i) = \sum_{k \in a_i} \sum_{j \in \mathcal{F}_i} r_i(j,k), \tag{4.1}$$

where $r_i(j,k)$ is reward she receives from follower j for posting item k. The form of these rewards will be made more precise below.

Each reader j receives a total utility $v(j,k)$ for each item k that she collects, whether she receives this from one or more publishers. The reader then sends a reward up to one or more publishers as a feedback representing her interest. In this work we consider bounded reward policies where the sum of all rewards sent by reader j for item k does not exceed $v(j,k)$, that is, $\sum_{i \in \mathcal{H}_j \cap \mathcal{P}_k} r_i(j,k) \leq v(j,k)$. In particular, we consider policies of the following form:

$$r_i(j,k) = v(j,k) b_i(j,k), \tag{4.2}$$

where $\sum_{i\in\mathcal{P}} b_i(j,k) = 1$ for all j, k. The tuple $\mathcal{G} = (\mathcal{P}, \{A_i\}, \{W_i(\prod_\ell A_\ell)\}$ now denotes our content curation game.

Remark. The reward paid by the reader might be of the form of likes, re-shares, favorite markings, etc., in the various online social networks accordingly. We do not go into detail on how the reward values are to be translated into such feedback, only noting that such a representation is possible.

Note that W_i represent the private utilities, or payoffs, of the agents in the game (here, these are the publishers). The social welfare function is defined as the sum of utilities received by all readers:

$$\mathcal{W}(\mathcal{A}) = \sum_{j\in\mathcal{R}} u_j = \sum_{j\in\mathcal{R}} \sum_{k\in \bigcup_{i\in\mathcal{H}_j} a_i} v(j,k),$$

where $\mathcal{A} = \{a_1, \ldots, a_P\}$ is the action profile of all users.

With the aim of characterizing the efficiency of distributed content curation, we consider the price of anarchy of the content curation game. The price of anarchy is defined as the ratio between the social welfare of an optimal allocation of items to publishers and that of the worst-case equilibrium. A high price of anarchy suggests that the distributed scheme is not efficient in terms of the achieved social welfare. A low price of anarchy that scales well with the system size, indicates an efficient distributed scheme, where even as the system grows the distributed scheme may achieve close to the optimal social welfare. We will now show that the price of anarchy is at most 2 in the content curation game. Note that a Nash equilibrium of this game may be one of mixed strategy, where an agent selects an action according to some probability distribution. We will show that the content curation game admits at least one pure Nash equilibrium.

Let Ω denote an optimal action profile in the content curation game, that is, one that maximizes $\mathcal{W}$, the social welfare.

Theorem 9 *Any Nash equilibrium of $\mathcal{G}$, the content curation game, results in social welfare at least half of the maximal social welfare:*

$$\mathcal{W}(\Omega) \leq 2\overline{\mathcal{W}}(A), A \in \mathcal{A},$$

where $\overline{\mathcal{W}}(A)$ is the expectation of $\mathcal{W}$ over the mixed strategy set A.

Proof. Vetta [110] has shown that valid utility games have a price of anarchy of at most 2. It

suffices to show that $\mathcal{G}$, the content curation game, is a valid utility game. A valid utility game has the three following properties:

1. non-decreasing submodularity: the social welfare function must be submodular and non-decreasing,

2. Vickrey condition: the utility of an agent is at least equal to the loss in the social welfare resulting from this agent declining to participate in the game,

3. cake condition: the sum of agent utilities under any set of strategies should be less than or equal to the social welfare.

We now show that the content curation game satisfies these three properties.

1. Since all item values $v(\cdot,\cdot)$ are non-negative, the social welfare function is non-decreasing. We denote by $n_A(j,k)$ and $n_{A'}(j,k)$ the number of publishers through whom reader j receives item k under strategy profile A and A' respectively, that is, $n_A(j,k) = |\mathcal{H}_j \cap \mathcal{P}_{A,k}|$, where $\mathcal{P}_{A,k} = \{i : k \in a_i\}$. We simplify the above notation to $n(j,k)$ and $n'(j,k)$ for strategy profiles A and A' respectively for the rest of the chapter and only include the subscript if the strategy profile under consideration is not clear. We consider profiles A and A' such that $a_i \subseteq a'_i$ for all i, and thus $n(j,k) \leq n'(j,k)$. We now show submodularity by studying the increase in social welfare due to the utility of any reader j when item k is added to A and A'. If both $n(j,k)$ and $n'(j,k)$ are non-zero, reader j's utility is not affected under either strategy and so the increase in social welfare is zero. If $n(j,k) = 0$ and $n'(j,k) > 0$, the social welfare has a non-zero increase when adding item k to A, but no increase when added to A' since reader k already receives the item under A'. Finally if $n(j,k) = n'(j,k) = 0$, the increase in adding item k to A and to A' is both $v(j,k)$. Summing over all readers, the increase in total social welfare due to adding any item k under A is not less than that under A'.

2. When publisher i declines to participate in the game (denoted by action set $\emptyset_i$), the loss in social welfare as compared to when she selects action a_i is $\mathcal{W}(\{a_1,\ldots,a_P\}) - \mathcal{W}(\emptyset_i, a_{-i}) = \sum_{k\in a_i} \sum_{j:j\in\mathcal{F}_i, n(j,k)=1} v(j,k)$; this corresponds to the loss of receiving an item provided only by this publisher to her followers. (Note that the readers who receive items from i also

from any other publisher do not lose any value.) The publisher's payoff, had she selected action a_i is $W_i(a_i) = \sum_{k \in a_i} \left(\sum_{j \in \mathcal{F}_i, n(j,k)=1} v(j,k) + \sum_{j \in \mathcal{F}_i, n(j,k)>1} v(j,k) b_i(j,k) \right) \geq \mathcal{W}(\{a_1, \ldots, a_P\}) - \mathcal{W}(\emptyset_i, a_{-i})$.

3. It is easily shown that $\sum_{i \in \mathcal{P}} W_i(A_i) \leq \mathcal{W}(\mathcal{A})$.

■

Bounded reward policies

We consider bounded reward policies, as described above, that allow us to show that the content curation game is a valid utility game and thus any Nash equilibrium of the game results in a social welfare at least half the optimal social welfare. This result however does not guarantee the existence of pure Nash equilibria. For this stronger result, we consider a more restrictive reward policy, that we call the Uniform policy. Under this policy, each reader divides her utility for an item equally among all publishers that she follows who post that item, that is, $b_i(j,k) = 1/n(j,k)$. In particular, the reward that a publisher i receives under action profile $\mathcal{A} = \{a_1, \ldots, a_P\}$ is as follows:

$$W_i(a_i) = \sum_{k \in a_i} \sum_{j \in \mathcal{F}_i} v(j,k)/n_{\mathcal{A}}(j,k). \tag{4.3}$$

Note that this is a special case of the set of bounded reward policies and thus the results presented thus far still apply. Such a policy is realistic in a scenario where a reader's stream of content is not presented in any particular order and it is equally likely that a given item is first read from any of the sources.

4.3.1 Convergence to Equilibria

We now show that the content curation game with the Uniform reward policy converges to a pure Nash equilibrium, by showing that it is a congestion game.

As defined by Rosenthal, a congestion game is a non-cooperative game $\mathcal{G}(N, \mathcal{M}, U)$ with a set of N players, a set of $\mathcal{M}$ resources, and utility functions U_n for each player. Each player has a set of feasible strategies, each of which specifies a subset of resources. A congestion game has two properties:

1. for any strategy selected by a player, this player receives a utility that can be written as the sum of the utilities from each resource in that strategy profile,

2. the utility that a player receives from each resource depends on the number of other players selecting the same resource.

Rosenthal [93] has shown that any congestion game has at least one pure Nash equilibrium by providing an exact potential function for such games. Further, this implies that best-response dynamics converge to an equilibrium.

If we let "items" be the set of "resources" of the congestion game, we cannot show that the content curation game is a congestion game. The reason is that the utility of a publisher i posting an item depends not only on the number of other users posting the same item, but also on the set of followers shared by those users and publisher i. However, in the following theorem we will show that the content curation game is a congestion game by letting the set $\mathcal{M}$ of resources be the set of pairs (j, k) of all followers $j \in \mathcal{R}$, and contents $k \in \mathcal{C}$.

Theorem 10 *The content curation game is a congestion game. In fact, it is a potential game and any best-response sequence of actions will converge to a Nash equilibrium of this game.*

Proof. We show that the content curation game is a congestion game by defining a resource as a reader-item pair. We thus have a set of resources $\mathcal{M} = \{(j, k) : j \in \mathcal{R}, k \in \mathcal{C}\}$. Let $r(j, k) = v(j, k)/n(j, k)$ denote the payoff received by a publisher that posts item k and has reader j as a follower. Since we have, from (4.3), for all $i \in \mathcal{P}$, $W_i(a_i) = \sum_{\substack{(j,k):\\ j\in\mathcal{F}_i, k\in a_i}} r(j, k)$, the first property of congestion games is verified. The second property is verified by the definition of publisher payoffs, W_i, under the Uniform policy. By considering the following potential function under strategy profile $\mathcal{A}$: $\Phi(\mathcal{A}) = \sum_{k\in\mathcal{C}, j\in\mathcal{R}} \sum_{t=1}^{n_{\mathcal{A}}(j,k)} v(j, k)/t$, we can easily show that the content curation game is a potential game. ■

4.3.2 Convergence to Approximate Solutions

In the previous subsection, we showed that any sequence of best-response dynamics will converge to a pure Nash equilibrium. However the rate of convergence (in terms of the number of best-response moves) is not guaranteed to be polynomial in the number of players. In this section,

we study the rate of convergence of approximate (α-Nash) dynamics to approximately optimal solutions. In particular, we show that the number of approximate best responses by players before they converge to a solution within a factor $2+\epsilon$ of the optimal solution is bounded by a polynomial. In our analysis, we use the concepts of α-Nash Dynamics and β-nice games, defined in Awerbuch et al. [22].

We first define approximate game dynamics and introduce some notation.

α-Nash Dynamics: An α-approximate best-response dynamics or α-Nash dynamics is a sequence of best responses by players in which each best response will increase the payoff of the player (who makes the change) by a factor of at least α. In an α-Nash dynamics with liveness property, each player gets a chance to play a best response after at most T steps.

β-nice games: Consider an exact potential game Λ, and let Ω be the optimal solution. Let $\mathcal{A} = (a_1, \ldots, a_P)$ be a strategy profile of the players and let $\mathcal{A}_i^{'}$ be a best response strategy for player i in strategy profile $\mathcal{A}$. The payoff of player i in strategy profile $\mathcal{A}$ is denoted by $W_i(\mathcal{A})$ and each player wants to maximize her payoff. In this setting, in a strategy profile $\mathcal{A}$, for each player i with the best response strategy $a_i' \in A_i$, we let $\Delta_i(\mathcal{A}) = W_i(\mathcal{A}_{-i}, a_i') - W_i(\mathcal{A})$, and $\Delta(\mathcal{A}) = \sum_{i \in \mathcal{P}} \Delta_i(\mathcal{A})$. We say that the game is a *β-nice game* if for any action profile $\mathcal{A}$,

$$\beta \cdot (\mathcal{W}(\mathcal{A}) + \sum_{i \in \mathcal{P}} \Delta_i(\mathcal{A})) \geq \mathcal{W}(\Omega).$$

1-bounded jump condition: We say that a game satisfies the *1-bounded jump condition* if for any action profile $\mathcal{A} = (a_1, a_2, \ldots, a_P)$, and any player i with best-response move a_i', and for every player i' the following two properties hold:

1. $W_{i'}(\mathcal{A}_{-i}, a_i') - W_{i'}(\mathcal{A}) \leq W_i(\mathcal{A}_{-i}, a_i')$.
2. for every improvement action $a_{i'}'$ of player i', it holds $W_{i'}(\mathcal{A}_{-i'}, a_{i'}') - W_{i'}(\mathcal{A}_{-\{i,i'\}}, a_i', a_{i'}') \leq W_i(\mathcal{A}_{-i}, a_i')$.

Awerbuch et al. [22] show that in order to prove the desirable result of this section, it is sufficient to prove that the content curation game satisfies the above two properties. Next, we show that the content curation game satisfies the above two properties.

First we state Theorem 5.6 of Awerbuch et al [22].

Theorem 11 *[22] Let $\frac{1}{8} > \delta \geq 4\alpha$. Consider an exact potential game Λ that satisfies the β-*nice *property and the 1-bounded jump condition. For any initial state $\mathcal{A}_{init}$ the unrestricted α-Nash best-response dynamics with liveness property generates a profile $\mathcal{A}$ with $\beta(1+O(\delta))\mathcal{W}(\mathcal{A}) \geq \mathcal{W}(\Omega)$ in at most $O\left(\frac{n}{\alpha\delta}\log(\frac{\phi^*}{\phi(\mathcal{A}_{init})})\cdot T\right)$ steps.*

Lemma 12 *The content curation game is a 2-nice game where Ω is the optimal solution, then for any action profile $\mathcal{A}$, we have $2\cdot(\mathcal{W}(\mathcal{A})+\sum_{i\in\mathcal{P}}\Delta_i(\mathcal{A}))\geq\mathcal{W}(\Omega)$.*

Proof. We show that $\mathcal{W}(\mathcal{A})+\sum_{i\in\mathcal{P}}W_i(\mathcal{A}_{-i},a'_i)\geq\mathcal{W}(\Omega)$ where a'_i is the best response of player i in strategy profile $\mathcal{A}$. Note that $W_i(\mathcal{A}_{-i},a'_i)\geq W_i(\mathcal{A}_{-i},\Omega_i)$. Let S be the set of pairs of posts and readers $(j,k)\in\mathcal{R}\times\mathcal{C}$ that are satisfied in the optimal solution , i.e., $\mathcal{W}(\Omega)=\sum_{(j,k)\in S}v(j,k)$. Let X be the set of pairs in S that are satisfied in $\mathcal{A}$ and X' be the rest of pairs in S. The value of all pairs in X appear in $\mathcal{W}(\mathcal{A})$, thus the sum of values of items in X is less than $\mathcal{W}(\mathcal{A})$. Moreover, for any pair (j,k) in X', if item k is covered by Ω_i and j follows i, then the utility $W_i(\mathcal{A}_{-i},\Omega(i))$ contains the whole value $v(j,k)$ for pair (j,k). This is because no other publisher that j follows posts k as otherwise the pair (j,k) would be in X and not X'. Therefore, $\sum_{j\in X'}v(j,k)\leq\sum_{i\in\mathcal{P}}W_i(\mathcal{A}_{-i},\Omega_i)\leq\sum_{i\in\mathcal{P}}W_i(\mathcal{A}_{-i},a'_i)$. The above inequalities imply the claim as follows:

$$\begin{aligned}\mathcal{W}(\Omega) &= \sum_{(j,k)\in T}v(j,k)=\sum_{(j,k)j\in T'}v(j,k)+\sum_{j\in T''}v(j,k)\\ &\leq \mathcal{W}(\mathcal{A})+\sum_{i\in\mathcal{P}}W_i(\mathcal{A}_{-i},a'_i)\leq 2(\mathcal{W}(\mathcal{A})+\Delta(\mathcal{A})).\end{aligned}$$

■

Lemma 13 *Content curation games satisfy the 1-bounded-jump condition.*

Proof. The proof lies in showing that the change in payoff for any publisher after a best-response move or improvement action by any other player is bounded by the new payoff of the publisher who makes the move. Consider two players i and i' in strategy profile $\mathcal{A}$. Consider each pair (j,k) of a reader j and content k with value $v(j,k)$. Recall that the number of publishers i that are followed by j and are posting content k in $\mathcal{A}$ is $n(j,k)$. Thus the payoff of publisher i in $\mathcal{A}$ is $\sum_{(j,k)\in\mathcal{F}_i\times a_i}\frac{v(j,k)}{n(j,k)}$. When a publisher i' changes her strategy to her best response $a'_{i'}$, for each

pair (j,k) of a reader and an item, the number of publishers followed by j and posting k changes from $n(j,k)$ to $n'(j,k)$ where $n(j,k)-1 \leq n'(j,k) \leq n(j,k)+1$. Therefore the increase in the payoff of publisher i is at most $\sum_{(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times[a_i\cap(a_{i'}\setminus a'_{i'})]}(\frac{v(j,k)}{n(j,k)-1}-\frac{v(j,k)}{n(j,k)})$. The payoff of player i' after changing her strategy from $a_{i'}$ to $a'_{i'}$ is at least $\sum_{(j,k)\in\mathcal{F}_{i'}\times a_{i'}}\frac{v(j,k)}{n(j,k)}$. For a pair $(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times[a_i\cap(a_{i'}\setminus a'_{i'})]$, at least two players i and i' followed by j are posting k in $\mathcal{A}$, thus $n(j,k)\geq 2$, and $(\frac{v(j,k)}{n(j,k)-1}-\frac{v(j,k)}{n(j,k)})\leq\frac{v(j,k)}{n(j,k)}$. Therefore,

$$\begin{aligned}\sum_{(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times[a_i\cap(a_{i'}\setminus a'_{i'})]}(\frac{v(j,k)}{n(j,k)-1}-\frac{v(j,k)}{n(j,k)})\leq&\\ \sum_{(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times a_i\cap(a_{i'}\setminus a'_{i'})}\frac{v(j,k)}{n(j,k)}&\\ \leq\sum_{(j,k)\in\mathcal{F}_i\times a_{i'}}\frac{v(j,k)}{n(j,k)}&\\ =W_{i'}(\mathcal{A}).&\end{aligned}$$

This implies the first condition of the bounded jump property, i.e, the increase in the payoff of player i is at most the payoff i'.

Consider a strategy profile $\mathcal{A}$ and two players i and i' with two best response strategies a'_i and $a'_{i'}$. When player i' changes her strategy to $a'_{i'}$, if she decreases the payoff for player i, then the decrease in payoff is at most $\sum_{(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times a'_i\cap(a'_{i'}\setminus a_{i'})}(\frac{v(j,k)}{n(j,k)}-\frac{v(j,k)}{n(j,k)+1})$. In this case, the payoff of i' from switching to her strategy is at least $\sum_{(j,k)\in\mathcal{F}_{i'}\times a'_{i'}}\frac{v(j,k)}{n(j,k)+1}$. Since for any pair $(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times[a'_i\cap(a'_{i'}\setminus a_{i'})]$, we have $n(j,k)\geq 1$, then $\frac{v(j,k)}{n(j,k)+1}\geq\frac{v(j,k)}{n(j,k)}-\frac{v(j,k)}{n(j,k)+1}$. These inequalities imply the second condition of the 1-bounded jump property as follows:

$$\begin{aligned}&W_i(\mathcal{A}_{-i},a'_i)-W_i(\mathcal{A}_{-\{i,i'\}},a'_i,a'_{i'})\\ &\leq\sum_{(j,k)\in\mathcal{F}_i\cap\mathcal{F}_{i'}\times[a'_i\cap(a'_{i'}\setminus a_{i'})]}(\frac{v(j,k)}{n(j,k)}-\frac{v(j,k)}{n(j,k)+1})\\ &\leq\sum_{(j,k)\in\mathcal{F}_{i'}\times a'_{i'}}\frac{v(j,k)}{n(j,k)+1}\leq W_{i'}(\mathcal{A}_{-i'},a'_{i'}).\end{aligned}$$

■

As stated earlier, Theorem 5.6 of [22] and Lemmas 12, 13 imply the following Theorem for the convergence of α-Nash dynamics to 2-approximate optimal solutions.

Theorem 14 *Let $\frac{1}{8} > \delta \geq 4\alpha$. Consider a content curation game Λ with any initial strategy profile $\mathcal{A}_{init}$. Any α–Nash best-response dynamics with liveness property generates a profile $\mathcal{A}$ with total welfare $\frac{1}{(2+O(\delta))}\mathcal{W}(\Omega)$ in at most $O\left(\frac{n}{\alpha\delta}\log(\frac{\phi^*}{\phi(\mathcal{A}_{init})}) \cdot T\right)$ steps.*

Numerical results on convergence

While we have shown in Section 4.3.1 that the content curation game converges to pure NE, we are not able to prove theoretical results on the convergence time. Since convergence to approximate equilibria in polynomial time has been shown above, we expect convergence time to pure NE to be similarly reasonable. We thus perform simulations to support this conjecture. We simulate the content curation game where each publisher chooses a best response strategy and each reader sends rewards to publishers. We consider a game with N nodes, including $R = 0.35N$ readers, $P = 0.3N$ publishers, and $C = 0.35N$ content sources, where each publisher may collect up to $K = 0.4C$ items. At each step of the simulation a publisher chosen uniformly at random selects his best strategy, the strategy that yields the optimal reward for that step. At each step readers also send up rewards to the publishers they follow according to the Uniform policy. Figure 4.1 plots the convergence time in number of steps against N, the system size. We observe that the increase in convergence time as the system grows appears to be reasonable, and polynomial in system size.

4.4 Centralized Content Curation

We now study the centralized curation problem, where a central authority with complete information optimizes the set of items each publisher must post so that social welfare is maximized.

We first show that the problem is NP-complete by giving a reduction to the Set Cover problem.

Theorem 15 *The centralized content curation problem is NP-complete.*

Proof. We prove the hardness by giving a reduction from the well-known Set Cover problem. In the Set Cover problem, a number l, and finite set S with a collection $\mathcal{T}$ of its subsets are given. A set cover for S is a subset $\mathcal{T}' \subseteq \mathcal{T}$ such that every element in S belongs to at least one member of $\mathcal{T}'$. The objective is to find a set cover with a cardinality of l. Given an instance $\mathcal{I}$ of Set Cover,

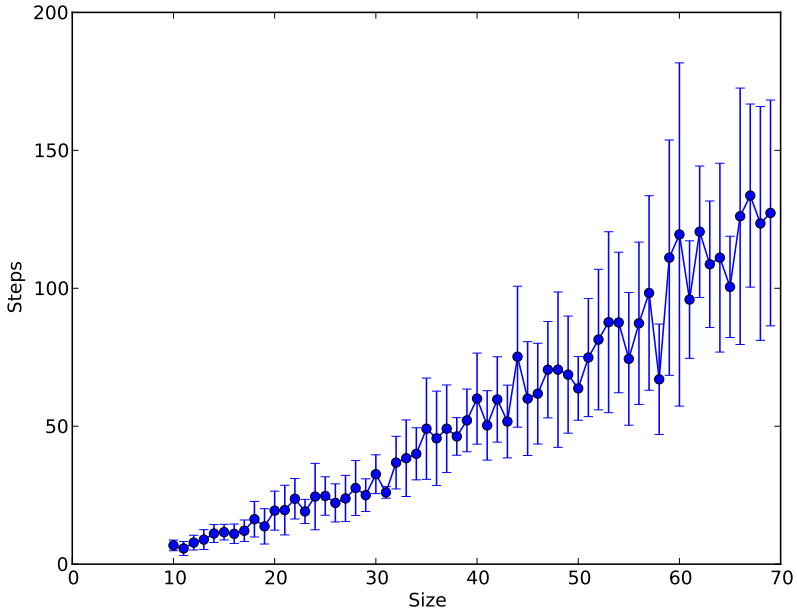


Figure 4.1: Simulation results on the convergence time of the content curation game.

we create an instance $\mathcal{J}$ of the curation problem. $\mathcal{I}$ consists of S and $\mathcal{T}$ as described above, where $S = \{j_1, \ldots, j_q\}$ and $\mathcal{T} = \{T_1, \ldots, T_n\}$.

Given an instance $\mathcal{I} = (S, \mathcal{T})$ of the set cover, we construct the following instance $\mathcal{J} = ((\mathcal{C}, \mathcal{P}, \mathcal{R}), \mathcal{E})$ of the content curation problem: Let $\mathcal{C} = \{k_0\} \cup \{k_1, k_2, k_3, \ldots, k_n\}$, $\mathcal{R} = \{j_1, j_2, \ldots, j_q, j'_1, j'_2, \ldots, j'_n\}$ and $\mathcal{P} = \{i_1, \ldots, i_n\}$. The structure of readers following publishers is as follows. Each publisher i_u, $u = 1, \ldots, n$, is followed by reader j'_u, and each publisher i_u is further followed by a number of readers from the set $\{j_1, \ldots, j_q\}$, such that each of this latter set of readers follows at least one publisher. Figure 4.2 provides an illustrative example where each of reader j_t, $t = 1 \ldots, q$, follows publisher i_t. Furthermore, for $1 \leq t \leq q$ and $1 \leq u \leq n$, we let $v(j_t, k_0) = M$,$v(j_t, k_u) = 0$, and $v(j'_u, k_u) = m$ and for $u' \neq u$, $v(j_u, k_{u'}) = 0$. We further assume that $M >> mn$. Finally, let $K = 1$ for this instance. The max curation problem then consists in choosing a subset of E, the set of edges between the publishers and content sources, such that each publisher connects to only one content source, and the total utility of the system is maximized. We claim that there exists a set cover of size l if and only if there exists a solution to the content curation problem with utility value of $Mq + (n - l)m$. To see this, consider a set cover solution of a family $\mathcal{T}'$ of l' subsets

in $\mathcal{I}$. We first show how to construct a solution of utility value $Mq + (n - l')m$ in $\mathcal{J}$. To do so, consider a solution to the curation problem where each publisher $i \in \mathcal{T}'$ posts the item k_0, and for each publisher $i_u \notin \mathcal{T}'$ publishes one item k_u. This is a feasible solution to the curation problem as each publisher posts one content item. In this solution, since all elements of S are covered by the set cover solution, each reader j_t, $t = 1, \ldots, q$, can read the item k_0 for a value of M each, summing to a total value of Mq from content item k_0. Moreover, for each of the $n - l'$ publishers $i_u \notin \mathcal{T}'$, reader j'_u receives item k_u since i_u posts item k_u, and therefore each such reader j'_u has value m, summing up to the value of $(n - l')m$. As a result, the total value of this solution is $Mq + (n - l')m$.

To complete the proof, we need to show that if we have a solution to the curation problem with total value $Mq + (n - l')m$, we can find a set cover solution of cardinality l' subsets. Let $\mathcal{T}'$ be a subset of publishers i_t that publish item k_0 and let $\mathcal{T}'' = \{i_t | 1 \leq t \leq n, i_t \text{ publishes } k_t\}$ (i.e., the set of publishers i_t that each post item k_t respectively). Since $M >> mn$, the value of the solution is at least Mq only when $\mathcal{T}'$ corresponds to a feasible set cover that results in having all readers $j_1, \cdots, j_q$ getting access to content item k_0. Given the above, the value of this solution is $Mq + m|\mathcal{T}''|$. Since the value of the solution to the content curation problem is $Mq + (n - l')m$, we conclude that $|\mathcal{T}''| = n - l'$. Moreover, since publishers in $\mathcal{T}'$ publish k_0, they may not publish any other content item, and thus $|\mathcal{T}''| \leq n - |\mathcal{T}'|$. Combining these two inequalities, we get:

$$|\mathcal{T}'| \leq n - |\mathcal{T}''| = n - (n - l') = l'.$$

We conclude that $\mathcal{T}'$ is a feasible set cover solution of cardinality at most l', and this completes the proof. ■

4.4.1 Approximation Algorithms

We now show that there exist $(1 - 1/e)$-approximation algorithms based on LP and greedy $(1/2)$-approximation algorithms by showing that the content curation problem is a special case of the separable assignment problem [44].

Lemma 16 *The maximum content curation problem is a special case of the separable assignment problem, thus admits a $(1 - 1/e)$-approximation factor through an LP-based algorithm and a $(1/2)$-approximation through by a greedy algorithm.*

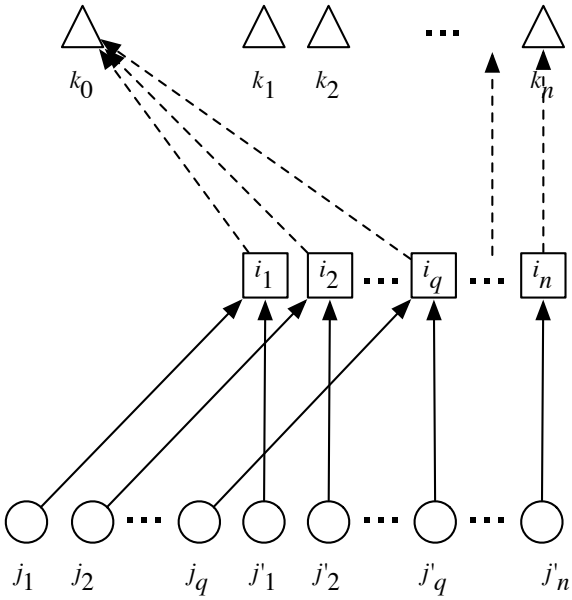


Figure 4.2: An example scenario for the reduction to the set cover problem. The solid arrows represent the follower structure while the dashed arrows represent one possible solution that results in a total utility of $Mq + (n - l')m$.

Proof. We show that the content curation problem is a special case of the separable assignment problem. In a separable assignment problem, we are given a set of items I to be assigned to a set of bins B and a packing maximization problem $\mathcal{P}_b$ for each bin $b \in B$. Each bin has a packing constraint and each item can be assigned to at most one bin. Any subset of items has a value for each bin, and the total value of the assignment is the sum of the value for the bins. The objective in the separable assignment problem is to find an assignment of items that maximizes the total value.

Assuming that there exists an α-approximation algorithm for a single-bin packing problem $\mathcal{P}_b$, it is shown that (i) one can get a $1 - \frac{1}{e^\alpha}$-approximation algorithm for the problem by solving and rounding an exponential-size configuration linear program, and (ii) one can get a $\frac{\alpha}{\alpha+1}$-approximation algorithm by applying a simple greedy algorithm [44].

We show that the centralized optimization problem of content curation is a separable assignment problem as follows: Each bin b corresponds to a publisher i, and an item in the separable assignment problem corresponds to a pair (j, k) of reader j and content (item) k. The packing problem $\mathcal{P}_i$ for the separable assignment problem of publisher i corresponds to choosing a set of

pairs (j, k) of content and readers maximizing the total utility $\sum_{j,k} v(j, k)$ such that j follows i and i posts k, i.e., a set Q of pairs (j, k) is feasible if the size of the set of posts that appear in the pairs of Q is not more than K, and for each reader j appearing in these pairs j follows i. We also observe that the optimization packing problem for each publisher (corresponding to the single-bin packing problem the case of the separable assignment problem) is polynomial-time solvable: in order to solve the packing optimization problem for a publisher i, it is sufficient to compute the utility of publisher i of posting content k, i.e. $\sum_{j \in \mathcal{F}_i} v(j, k)$, and choose the top K posts maximizing this value. Since the single-bin packing problem admits a polynomial time algorithm, $\alpha = 1$. Applying the theorems above with $\alpha = 1$, we get the aforementioned $1 - 1/e$-approximation (0.63) and greedy $1/2$-approximation algorithms. ■

4.5 Selective Readers

We have thus far assumed that readers are not strategic, that is they follow a fixed set of publishers reading all the items published by them. We now consider the readers' budget of attention not through a limit on the number of publishers they follow, but through a limit on the number of items they read. We thus assume that readers are selective and only read (or benefit) from the top Z items posted by publishers that they follow. The readers choose this set of Z items such that their utility is maximized. Now, we consider the same utility model as in the previous sections, but with selective readers, where user j chooses only Z_j items to read. The centralized optimization is still NP-complete.

From the game theoretic point of view, we can show that when followers are selective as described above, the price of anarchy is 2. We prove this by showing that the game is a valid utility game. The proof is similar to that of content curation games without selective readers.

Let Ω denote an optimal action profile in this game with selective readers, that is, one that maximizes $\mathcal{W}$, the social welfare.

Theorem 17 *Any Nash equilibrium of $\mathcal{G}$, the content curation game with selective readers, results in social welfare at least half of the maximal social welfare:*

$$\mathcal{W}(\Omega) \leq 2\overline{\mathcal{W}}(A), A \in \mathcal{A},$$

where $\overline{\mathcal{W}}(A)$ is the expectation of $\mathcal{W}$ over the mixed strategy set A.

Proof. We present a sketch of the proof by showing that the game is a valid-utility game, and apply the result of Vetta [110] who shows that valid utility games have a price of anarchy of at most 2. To do so, we show that the content curation game with selective readers satisfies the three properties for valid-utility games.

- *Non-decreasing submodularity.* Since all item values $v(\cdot, \cdot)$ are non-negative, the social welfare function is non-decreasing. To show its submodularity, consider two strategy profiles A and A' where $a_i \subseteq a'_i$ for each i. Now consider adding pair (j, k) of item k and reader j to the strategies of $a_{i'}$ and $a'_{i'}$ in A and A' respectively. There are two cases: if in strategy profile $(A'_{-i'}, a'_{i'} \cup \{(j, k)\})$ item k is not among the top Z_j items posted for j, then the increase in the social welfare by adding (j, k) to A' is zero, and thus less than or equal to that of adding this pair to A. For the case where item k is among the top Z_j items for j, we further consider three conditions. Let $n(j, k)$ and $n'(j, k)$ be the number of publishers through whom reader j receives item k under strategy profile A and A' respectively. Since $a_i \subseteq a'_i$ for all i, $n(j, k) \leq n'(j, k)$. In the first case, if both $n(j, k)$ and $n'(j, k)$ are non-zero, user j's utility is not affected under either strategy and so the increase in social welfare is zero. In the second case, if $n(j, k) = 0$ and $n'(j, k) > 0$, the social welfare has a non-zero increase when adding item k to A, but no increase when added to A' since reader k already receives the item under A'. Note that adding item k to A may replace another item ℓ from the top-Z_j list. Since the Z_j items are items of highest utility, this increase in social welfare, $v(j, k) - v(j, \ell)$ is non-negative. Finally if $n(j, k) = n'(j, k) = 0$, there is non-negative increase in adding item k to both A and to A'. Again, an item ℓ may be replaced in adding item k with value $v(j, \ell) \leq v(j, k)$. The increase in welfare in adding item k to A is at least that of adding to A'. Summing over all readers, the increase in total social welfare due to adding any item k under A is not less than that under A'.

- *Vickery condition.* When publisher i declines to participate in the game (denoted by action set $\emptyset_i$), the loss in social welfare as compared to when she selects action a_i corresponds to all items posted by publisher i that are only published by i and are among the top Z_j valued items for each i's followers, $j \in \mathcal{F}_i$. These values account for part of publisher i's payoff

from items she posts for readers who do not get the same items from anybody else. Moreover, publisher i receives payoffs from items she is posting that other publishers may be posting as well. Therefore her payoff is at least the reduction in social welfare due to dropping her strategy.

- *Cake condition.* Again by definition, it is not hard to show that $\sum_{i \in \mathcal{P}} W_i(A_i) = \mathcal{W}(\mathcal{A})$.

■

The above theorem shows that the price of anarchy of Nash equilibria in content curation games with strategic readers is at most 2. Since the content curation game with selective readers is not a congestion game, it is not necessarily a potential game, and the analysis of theconvergence of best-response (Nash) dynamics in these games remains open. However, one can argue that *regret-minimization* dynamics will converge to a $1/2$-approximately optimal solution. This result is implied in a result of Roughgarden [99] which shows that valid-utility games are $(1, 1)$-smooth. This implies that not only the price of anarchy for mixed Nash equilibria is $1/2$, the price of anarchy for correlated equilibria and no-regret dynamics is also $1/2$.

4.6 Related Work

To the best of our knowledge, our model is the first to address the problem of content curation by a set of aggregators for the optimization of utility of a set of readers. However, similar models of optimization and game theory have been considered in different contexts.

The content curation game is related to a previously studied game called the market sharing game [50]. In a market sharing game, there are a set of players (agents), each playing a subset of markets, and we are given a bipartite graph between the markets and the players indicating which markets are eligible to be played by each agent. A generalization of market sharing games is the distributed caching game where the strategy space of players is more general than playing a subset of markets [44]. Content curation games are different from both types of games in that in content curation games, there are three parties in the game: the players (or publishers), the items they post (corresponding to markets), and a third party that is the set readers who follow a subset of publishers. Neither of these three games is a special case of the other, and in particular content curation games are a generalization of ”uniform market sharing games” [50].

Our work is also related to another line of work that explores incentives in user-generated content systems where users are strategic. In particular, in [53], the authors study the problem of strategic news posting in online social networks. They model users as either greedy or courteous in their posting strategy. They analyze these two models on random graphs from a game theoretic point of view. They find that high quality information spreads in the network if users are greedy. Through simulations on Twitter data they show the same observation when users are modeled as courteous. In our work, rather than information spread, we are interested in maximizing the utility of readers where they have differing interests in content.

In another related work [81], the authors study the role of intermediaries in blogging and microblogging websites via a different model. They study the posting behavior of these intermediaries on real datasets and in a game theoretic setting.

Another line of work in the area of user-generated content models ranking mechanisms in a game theoretic setting, where utility is defined in terms of the attention (exposure) the contributor or their content receives. In this setting, generating higher quality content is assumed to be costlier. The objective of the system is to elicit high quality and high participation in equilibrium [48]. Our model is more general where we consider content aggregation, not modification of content quality.

4.7 Conclusion

We have considered the problem of content curation by a set of publishers aiming to maximize global utility of a set of readers. We show that the centralized optimization, while being NP-complete, can be reduced to a separable assignment problem, thus admitting a $(1 - 1/e)$ approximation algorithm. We model distributed content curation as a reader-publisher game and show that the price of anarchy is at most 2. When in addition the readers are selective in the items they choose, we show that the price of anarchy is bounded by 2. Our results imply that in the complete information setting, when publishers maximize their utility selfishly, distributed content curation reaches an equilibrium which is efficient, that is, the social welfare is a constant factor of that under an optimal centralized curation.

This result can be leveraged by online social network platform operators in the design of filtering mechanisms. For instance, when a user is about to share an item by posting it on her wall,

the social network platform can inform her about the number or fraction of her followers who have already seen that item. If a large fraction have already seen it, then it might be more efficient, or result in higher payoff, if that user shared another item that has not been seen by her followers as much.

We have presented an initial analysis into the phenomenon of social filtering of content. In this initial study we have made some simplifying assumptions to gain an insight into how these systems function, and what type of mechanisms lead to desirable behaviour. There are several avenues for the extension of our model that we intend to pursue. First, the bounded reward policies we consider are separable or decomposable, in the sense that the reward a user pays to a publisher for a given content is independent of the rewards to other publishers and of the rewards pertaining to other content. We can imagine a scenario where it would be beneficial to pay rewards to publishers offering the most contents of high value to a user. Further, the value derived by a user in receiving a content may also depend on the set of contents retrieved simultaneously through a given publisher. Under such conditions, the properties necessary for a valid utility game may no longer hold, in particular submodularity of the welfare function, but this does not imply that bounded price of anarchy and desirable equilibrium conditions may not hold. Other formulations for the analysis of such games will be considered. Second, we do not assume a cost structure for the publishers other than the limit of K items to post that implicitly puts an unbounded cost for more than K items. We may assume a soft budget of attention where the cost in posting each additional item increases with the number of items posted. This problem of maximizing the welfare of readers and minimizing the cost of the publishers is an interesting extension we leave for future work.

Chapter 5

Exchange Markets without Money

5.1 Introduction

Mechanism design without money has been a major subject of study in economics and mechanism design [94, 47, 95]. This line of research has been studied in the economics literature in the context of two-sided matching markets [47, 94], markets where monetary transactions are repugnant [96], and house allocation problems [105]. Recently this field has gained more attention in the computer science literature due to the fact that monetary compensations are not always easily applicable [91, 38]. In some cases, payments are hard to implement or to collect, e.g., implementing secure money transaction systems is costly in general and some people do not feel safe enough sharing sensitive information online fearing internet fraud [36, 100]. Moreover, in some repugnant markets, there may be legal or ethical issues with monetary transactions, e.g., in the case of kidney donation [97, 19]. In this chapter, we initiate the study of a fundamental exchange market problem without money that is a natural generalization of the well-studied kidney exchange problem. From the practical point of view, the problem is motivated by barter websites on the Internet, e.g., swap.com, and u-exchange.com.[1] We will elaborate on these applications after the problem description.

Consider a set of agents where each agent has some items to offer, and wishes to receive some items from other agents. A mechanism specifies for each agent a set of items that he gives away, and a set of items that he receives. Each agent would like to receive as many items as

[1] See `http://abcnews.go.com/International/buying-barter-economy-matures-niche-trend/story?id=18193023` for a recent news coverage.

possible from the items that he wishes, that is, his utility is equal to the number of items that he receives and wishes. However, he will have a large dis-utility if he gives away more items than what he receives, because he considers such a trade to be unfair. To ensure voluntary participation (also known as individual rationality), we require the mechanism to avoid this. We show that any individually rational exchange can be viewed as a collection of directed cycles, in which each agent receives an item from the agent before him, and gives an item to the agent after him. In addition to simplifying the statement of the problem, this suggests that we can implement an exchange by separately carrying out one-to-one trades among subsets of agents. In some settings, carrying out cycle-exchanges of large size is undesirable or infeasible. If there is a chance that each trade in a cycle fails, the chance that the whole cycle of exchanges is realized will exponentially decrease as the length of the cycle increases. Because of this and other problems with implementation, for example, most of the previous work on the exchange of kidneys focuses on short exchanges [97, 98]. Therefore, we distinguish the restricted problem in which the number of agents in each cycle is bounded above by some given constant $k \geq 2$. The most natural and commonly practiced cycles are of length 2, (i.e., swaps). Most of the results of this chapter are for the two extremes in which there is no limit on the length of the cycles, and $k = 2$.

As an example, consider a simple instance with 3 agents and 4 items, where agent a owns item 1, agent b owns item 2, and agent c owns items 3 and 4. Assume that agents a and c both wish to receive item 2, and agent b wishes to receive items $1, 3$, or 4. Therefore, agents a and c each would like to be the one who gets the chance to trade his item(s) for item 2. Now consider another instance in which agent c does not own item 4. Consider a mechanism that, given the first instance, picks agent a to trade with b, but given the second instance, picks agent c to trade with b. Then, if c truly owns both 3 and 4, he would prefer to claim that he only owns 3, and be the person who trades with b. The problem can be easily fixed here by making consistent decisions. The question is, then, can we design a mechanism that always finds the exchange that maximizes social welfare, and yet incentivizes truthfulness? Interestingly, we show that the answer differs in unconstrained and constrained problems. We elaborate upon this further after reviewing some real-world applications of this model.

Applications. Motivated by concerns about money transaction on the Internet, and simplicity

and convenience of swapping items in local economies, barter websites (also referred to as barter economy sites) have become more popular in the recent years.[2] Such barter websites help users exchange items with each other. Various types of items may be exchanged in these websites: from smaller used items like books, DVDs, cellphones, or children's clothing, to bigger items like boats, vehicles and vacation rentals. Some of these sites also support exchanging services like dental work and installing hardwood flooring. In most cases, users swap items with one another, i.e., only exchanges of size 2 are allowed. One can extend their setting to multiple exchanges over a cycle at the same time. We model such barter websites as networks amongst users where each user has two associated lists: an item list which consists of items the user is willing to give away to other users, and a wish list which consists of items the user is interested in receiving. A transaction involves a user giving an item to another user. Users are motivated to transact in expectation of realizing their wishes. Some examples of such marketplace applications are as follows:

- swap.com focuses on media like books and CDs. It claims about 1.2 million members and focuses on more local trades as they want to avoid expensive shipping fees.
- readitswapit.co.uk allows book lovers to exchange their already read books and receive new books in return. Almost all of the matching is done manually by the user herself, meaning that she has to go and find her desired book in a library and then mark it. The owner of the desired book will be informed by an email and will check the seeker's list of books and if willing to do the exchange, they will post the books for each other.

Other than these applications, the aforementioned exchange market problem without money is a natural generalization of the well-studied kidney exchange problem [97, 15, 19] where each agent wishes one item (a healthy kidney) and has only one item to offer.

Our Contributions. For the length-constrained variant of the problem, we rule out the existence of a $1 - o(1)$-approximate truthful mechanism for $k \geq 2$. We show that no truthful determinis-

[2]See examples at `http://mashable.com/2011/08/19/barter-sites/` and `http://gigaom.com/2012/07/07/summer-is-for-swapping-startups-boost-the-barter-economy/`. For a recent news coverage, see `http://abcnews.go.com/International/buying-barter-economy-matures-niche-trend/story?id=18193023`.

tic or randomized mechanism can achieve an approximation factor better than $\frac{3k+1}{3k+2}$ or $\frac{3k+1.89}{3k+2}$, respectively.

The above impossibility results are caused by incentive issues, and are not based on the computational complexity of the problem. We strengthen the hardness of the problem by proving that, even without the truthfulness requirement, the problem is APX-hard for any k. Finally, we present a $\frac{1}{8}$-approximately optimal truthful mechanism for the problem with $k = 2$. The mechanism visits pairs of agents in some fixed order, and considers adding a subset of exchanges when visiting a pair. The ordering of pairs is done such that, at any stage during the process, an agent can not affect the relevant future cycles by changing his strategy. We formalize this by defining an interaction set for each agent, which denotes the set of agents that are (possibly indirectly) affected by him at any stage, and making sure that an agent does not trade with any other agent who is currently in his interaction set.

For the unconstrained version, we present a class of polynomial-time algorithms solving the optimal exchange market problem, closely following algorithms for maximum flow and circulation problems. An algorithm maintains a set of feasible exchanges, and iteratively augments the current solution until the residual graph does not contain any more cycles.

5.1.1 Related Work

The Kidney Exchange Problem A related problem in exchange markets is the "national kidney exchange" problem. For many patients with kidney disease, the best option is to find a living donor – a healthy person willing to donate one of her two kidneys. The problem is that frequently, a potential donor and her intended recipient are blood or tissue-type incompatible. In the past, the incompatible donor's kidney was not used, and the patient had to wait for a deceased-donor kidney. However, now through regional kidney exchange programs in the United States, patients can swap their incompatible donors with each other, in order to each obtain a compatible donor [97, 98, 15, 19]. The kidney exchange problem is a special case of our problem where each user has only one item to offer and wishes one item. As a result, the kidney exchange problem is fundamentally simpler, and can be solved in polynomial time in the case of one-to-one exchanges (i.e., for $k = 2$), however, for the case of length-constrained exchanges even with $k = 2$, our problem is NP-hard (and also APX-hard as observed in this chapter). From the mechanism design point of view,

[19, 21] study the kidney exchange problem in the presence of strategic hospitals that may have an incentive not to list all their current available organ donors. By not listing donors, hospitals will still have the option of matching pairs internally. In our model, however, agents can have positive utility only by exchanging items.

Mechanism Design without Money Our work fits in a line of research that seeks to design strategy-proof mechanisms without monetary transfers. This line of research has been studied in the economics literature in the context of two-sided matching markets, repugnant markets, and house allocation problems:

- Two-sided matching markets have applications in college admissions and allocating interns to hospitals [95]. Incentive issues in such markets have been studied in several papers [62, 20]. A special class of the stable matching problem with dichotomous preferences, studied by [31], is remotely related to our problem as it employs the theory of bipartite matchings.

- Repugnant markets are markets that are considered by society to be outside of the range of market monetary transactions, due to moral issues. It applies to organ donation (like exchange of kidneys), and reproduction (e.g., child adoption and surrogate mothers). As discussed earlier, our problem can be thought of as a generalization of the kidney exchange problem.

- House allocation problems [105] are resource allocation problems where a set of items (houses) are to be allocated to a set of people each with a preference list over the items. These problems have applications in organ allocation (e.g., deceased donor waiting list), university dormitory room, parking space, and office space allocation.

Although about mechanism design without money, none of the above papers discuss approximately optimal truthful mechanisms. Recently, approximate mechanism design without money has become more popular in the computer science community. Initiated by [91], various assignment problems [38] as well as network design problems [76, 77] have been studied in this context. Our results fit in a similar framework, but our exchange market problem and the techniques we employ are different from all the above problems.

Algorithmic Results The length-constrained variant of the problem from an algorithmic perspective was studied in [13], where it is shown that the problem for length constraint of $k = 2$ is NP-hard, and a $5/3$-approximation is deriven using a reduction to the k-set packing problem. The bounded cycle cover problem, which constrains cycles to be of bounded size as well as simple and node-disjoint, was introduced by [59]. They present a heuristic for the problem along with empirical analysis. For the unconstrained exchange market problem, we design a polynomial-time algorithm which is a variant of the well-studied minimum cost circulation problem [51].

5.2 Preliminaries

Consider a set of n agents A and a set of m items I. In an instance $(A, I, \{(I_a, W_a)|a \in A\})$ of the *exchange market* problem, each agent a has an *item list* $I_a \subseteq I$ (items that he owns) and a *wish list* $W_a \subseteq I$ (items that he needs) such that $I_a \cap W_a = \emptyset$. An *exchange* $C : A \to I^2$ assigns to each agent a a set $C_1(a)$ of items that he receives in exchange for a set of items $C_2(a)$ that he gives away. An exchange is feasible if for each item i, $|\{a|i \in C_2(a)\}| \geq |\{a|i \in C_1(a)\}|$. The utility of agent a for exchange C is specified by a function u as follows:

$$u(a, C) = \begin{cases} |C_1(a) \cap W_a| & \text{if } C_2(a) \subseteq I_a \text{ and } |C_1(a) \cap W_a| \leq |C_2(a)|, \\ -\infty & \text{otherwise.} \end{cases}$$

In other words, in a feasible exchange, the utility of an agent is $-\infty$ if he must provide an item he does not own, or has to give away more items than he receives, and otherwise, his utility is the number of items that he receives from his wish list.

Our goal is to find a feasible exchange maximizing the *social welfare*, i.e., sum of utilities of agents. This goal corresponds to finding a feasible exchange that maximizes the number of items exchanged, and no agent has $-\infty$ utility. Notice that feasibility implies that the total number of items collected from agents must be at least the total number of items received by agents. Therefore, in order to find an exchange with non-negative total social welfare, we should make sure that the number of items each agent receives *and* is in his wish list is not more than the number of items he gives away, i.e., we must have $C_1(a) \subseteq W_a$, and also $|C_1(a)| = |C_2(a)|$.

An *exchange mechanism* extracts the private information, i.e, I_a and W_a of each agent a, and maps it to an exchange. We are interested in designing *truthful* (or strategyproof) mechanisms in

which it is a dominant strategy for each agent a to report his true private information (I_a, W_a). Our goal is to design truthful exchange mechanisms maximizing the social welfare. We will give a more formal definition of the problem after defining a bipartite graph representation of the problem.

5.2.1 Bipartite Graph Modeling

The above formulation of the problem allows for a clean representation using a directed bipartite graph. This representation will help in deriving a polynomial-time algorithm for the unconstrained problem and also argue about the constrained-length problem. This representation will be used throughout the rest of the chapter. Here, we first define this representation, and then formally state our problem in terms of this graph representation.

Given an instance $(A, I, \{(I_a, W_a)|a \in A\})$ of the exchange mechanism problem, define a bipartite directed graph $G = (A \cup I, E)$, where $E = \{(a, i)|i \in I_a\} \cup \{(i, a)|i \in W_a\}$. That is, there is an edge from an agent to an item if the agent owns the item, and from an item to an agent if the agent needs the item. A (directed) cycle in graph G is a sequence of directed edges in G where each edge appears at most once. This graph representation helps in arguing about the exchange market problem, since *any feasible exchange in the exchange market problem corresponds to a set of edge-disjoint directed cycles in G and vice versa.*

Proposition 18 *In an exchange market problem* $(A, I, \{(I_a, W_a)|a \in A\})$*, any feasible exchange corresponds to a set of* edge-disjoint directed cycles *in its graph representation* $G(A \cup I, E)$*, and vice versa.*

Proof. First of all, we can interpret a simple directed cycle in this graph as a feasible exchange as follows: any agent in this cycle gives away the item immediately after him in the cycle, and receives the item immediately before him. More generally, any set of edge-disjoint cycles can be interpreted as a feasible exchange. Conversely, any feasible exchange corresponds to a subgraph in which the in-degree of each vertex is equal to the out-degree ($|C_1(a)| = |C_2(a)|$), and therefore it can be decomposed to a set of edge-disjoint cycles. The utility of an agent in each cycle is equal to the number of cycles to which his corresponding vertex belongs and in which he receives an item in his wish list. ■

We can now formally state the definition of *truthful exchange mechanisms* using our graph

representation of the problem. Truthfulness states that by misreporting, an agent will either be asked to provide an item he does not own or receive fewer items that he wants.

Definition 2 *Consider a bipartite graph $G(A \cup I, E)$ and let $a \in A$ be a vertex where I_a is the outgoing neighbors of a and W_a is incoming neighbors for a. Consider any other subsets $I'_a \subseteq I$ and $W'_a \subseteq I$, $I'_a \cap W'_a = \emptyset$, and let G' be the graph representation of the problem where we replace (I_a, W_a) with (I'_a, W'_a) in G. An exchange mechanism is* truthful *if for any such G and G', in the set of cycles produced by the mechanism on G' there is either a cycle with an edge from a to $I \backslash I_a$, or the number of cycles with edges from W_a to a is no more than the number of cycles including a in G.*

We can now define the *unconstrained exchange market* problem as follows:

Definition 3 *Given a graph representation $G(A \cup I, E)$ of an instance $(A, I, \{(I_a, W_a) | a \in A\})$ of the exchange market problem, the goal of the* unconstrained exchange maximization *problem is to find a set E of edge-disjoint (directed) cycles with the maximum number of edges.*

We also define the *length-constrained* exchange market problem, or equivalently, *the k-exchange market* problem as follows:

Definition 4 *Given a graph representation $G(A \cup I, E)$ of an instance $(A, I, \{(I_a, W_a) | a \in A\})$ of the exchange market problem and a constant k, the goal of the* length-constrained exchange maximization *problem is to find a set E of edge-disjoint cycles, each of size at most $2k$, with the maximum number of edges.*

Note that by restricting the size of the cycles by $2k$, rather than k, we are limiting the number of agents (equivalently, items) in each cycle by k. For example, the case of $k = 2$ corresponds to swapping two items among two agents, and thus a cycle of size 4 in the bipartite graph.

Let $OPT(G)$ be the maximum social welfare of a set of feasible exchanges given a graph G. In this chapter, we are interested in designing truthful mechanisms to approximate $OPT(.)$ on every instance. We say that an algorithm f is α-approximation if for all G, $S(f(G), G) \geq \alpha OPT(G)$. Notice that any approximate mechanism avoids selecting an infeasible exchange, since the optimal social welfare is always at least 0, achieved by not picking any exchanges.

5.3 Length-Constrained Exchange Markets

In this section, we study the length-constrained exchange market problem. We show several impossibility results from truthful mechanism design and computational complexity point of views for any $k \geq 2$, and one approximately optimal mechanism for the length-constrained problem with $k = 2$.

5.3.1 Inapproximability of truthful mechanisms

In this section, we show the inapproximability of truthful mechanisms for length-constrained market exchange problem for $k \geq 2$. First we show a result for deterministic mechanisms and then extend it to randomized mechanisms.

Theorem 19 *No deterministic truthful mechanism for the k-constrained problem can have an approximation ratio better than $\frac{3k+1}{3k+2}$.*

Proof. Consider an instance of the k-exchange market problem with $k+1$ agents $a, b, c_1, \ldots, c_{k-1}$ and $3k+3$ items. Each agent owns 3 items (exclusively), and each item is in the wish list of one or two other agents. Each item is coded by a pair, where the first element is the agent owning it, and the second element is the agent(s) wishing for it. The agents and items are,

- Items owned by a: $(a, b), (a, c_1), (a, bc_1)$.
- Items owned by b: $(b, a), (b, c_1), (b, ac_1)$.
- Items owned by c_i for $1 \leq i \leq k-2$: $(c_i, c_{i+1}), (c_i, c_{i+1})', (c_i, c_{i+1})''$.
- Items owned by c_{k-1}: $(c_{k-1}, ab), (c_{k-1}, ab)', (c_{k-1}, ab)''$.

For example, (b, ac_1) is an item that is owned by agent b, and is wished for by agents a and c_1. Agents a and b are symmetric in this instance. Notice that any cycle involving any agent c_i for $1 \leq i \leq k-1$, also involves all such agents. In particular, any cycle involving c_1 involves all c_i, $1 \leq i \leq k-1$, and therefore has size at least $k-1$. We conclude that no feasible cycle can involve a, b, and c_1 all together (otherwise it will have size $k+1$). Also notice that no feasible cycle can involve exactly one of a, b, or c_1. As a result, the sum of the utilities of a, b, and c_1 for any feasible cycle is an even number. Since there are 9 items that a, b, and c_1 want, the sum of the utilities of

a, b, and c_1 in any feasible exchange is at most 9. But since the sum of their utilities is an even number, it can be at most 8. We therefore conclude that at least one of these three agents will have utility at most 2 in any feasible solution to this instance, regardless of the approximation factor. We next show by considering cases that that agent will benefit from misreporting. Since a and b are symmetric, there are two cases:

1. Agent a's utility is at most 2 (the case for agent b is similar). Assume that agent a removes item (b, a) from his wish list. The following exchange is still feasible, and has social welfare $3k + 2$:

   - $(a, b), (b, ac_1)$.
   - $(a, c_1), (c_i, c_{i+1})$ for $1 \leq i \leq k - 2$, and (c_{k-1}, ab).
   - $(a, bc_1), (c_i, c_{i+1})'$ for $1 \leq i \leq k - 2$, and $(c_{k-1}, ab)'$.
   - $(b, c), (c_i, c_{i+1})'$ for $1 \leq i \leq k - 2$, and $(c_{k-1}, ab)''$.

2. Agent c_1's utility is at most 2. Assume that agent c_1 removes item (a, c_1) from his wish list. The following exchange is still feasible, and has social welfare $3k + 2$:

   - $(a, b), (b, a)$.
   - $(a, bc_1), (c_i, c_{i+1})$ for $1 \leq i \leq k - 2$, and (c_{k-1}, ab).
   - $(b, c_1), (c_i, c_{i+1})$ for $1 \leq i \leq k - 2$, and $(c_{k-1}, ab)'$.
   - $(b, ac_1), (c_i, c_{i+1})$ for $1 \leq i \leq k - 2$, and $(c_{k-1}, ab)''$.

Notice that in each case an agent removed an item from his wish list that was exclusively wanted by him. Therefore, in each instance the social welfare can be at most $3k + 2$, and the specified exchange is optimal. In the specified exchange the agent removing an item has utility 3. In fact, we show that in each the agent removing his item will have utility 3 in any exchange with social welfare $3k + 2$.

1. Assume for contradiction that a has utility at most 2 in an exchange with social welfare $3k + 2$. Since there are k other agents and the utility of each agent is at most 3, all other agents must have utility 3. But since a participates in at most 2 exchanges, the number of

items that either b and c_1 wants that are offered is at most 5. Therefore b and c_1 can not both have utility 3.

2. This case is similar. Assume that c_1 has utility at most 2 in an exchange with social welfare $3k + 2$. Since c_1 participates in at most 2 exchanges, the number of items that either a and c_1 wants that are offered is at most 5. Therefore a and c_1 can not both have utility 3.

We conclude that the agent removing his item will have utility 3 in any exchange with social welfare $3k+2$. Since any algorithm with approximation factor $\frac{3k+1}{3k+2}$ must choose such an exchange, it can not be truthful. ■

A randomized mechanism may choose exchanges at random. In this case, we assume that agents are risk-neutral and try to maximize their *expected utility*. The question is if it is possible to design a truthful mechanism that does not give any incentive to agents to misreport their private information in order to increase their expected utility.

Theorem 20 *No randomized truthful mechanism for the k-exchange problem can have an approximation factor better than $\frac{3k+\frac{17}{9}}{3k+2}$.*

5.3.2 Truthful $\frac{1}{8}$-approximation for the 2-exchange problem

In this section, we present a $\frac{1}{8}$-approximation truthful mechanism for the length-constrained exchange problem with $k = 2$. The algorithm is as follows:

1. Partition agents into sets A and B by placing each agent independently at random with probability $1/2$ into set A (and otherwise in B).

2. Let $a_1, \ldots, a_k$ be the agents in A, and $b_1, \ldots, b_{n-k}$ the agents in B.

3. Visit every pair of agents in $A \times B$ in order $(a_1, b_1), (a_1, b_2), \ldots, (a_1, b_{n-k}), (a_2, b_1), \ldots, (a_k, b_{n-k})$.

4. When visiting a pair of agents, consider exchanging all pairs of items in an arbitrary order, and add that exchange if feasible.

First, we show the above algorithm is a $\frac{1}{8}$-approximation algorithm, and then we show it corresponds to a truthful implementation.

Lemma 21 *The above algorithm is a* $1/8$ *approximation algorithm for the* 2*-exchange market problem.*

Proof. Consider an optimum set of exchanges OPT. Let $OPT(A)$ be the subset of OPT consisting only of exchanges between an agent in A and an agent in B. Since every element of OPT will be in $OPT(A)$ with probability $1/2$, we must have $E_A[|OPT(A)|] \geq |OPT|/2$. Fixing A, the algorithm heuristically considers adding exchanges in $A \times B$. Since each possible exchange intersects with at most 4 exchanges in $OPT(A)$, and also every element of $OPT(A)$ has intersection with at least an exchange picked by the algorithm (otherwise it would have been picked), this algorithm picks at least $OPT(A)/4$ exchanges. This implies that the algorithm is a $1/8$-approximation. ■

We next prove the truthfulness of the algorithm by showing that it satisfies a property, which we call *interaction-freeness*, that is a sufficient condition for truthfulness of greedy algorithms. Equivalently, we say that the algorithm is interaction-free. A greedy algorithm fixes an ordering over a subset of all exchanges, visiting them one by one, adding a cycle whenever it is feasible (according to the item lists and wish lists of the two agents), and it does not intersect with a cycle that is already picked. At any time during the process of the algorithm, and for any agent a, define the *interaction set* $S(a) \subseteq A$ to be the set of agents that are already affected (possibly indirectly) by a as follows: At the start of the process, let $S(a) = \{a\}$ for all a. Assume that (b, c) are the agents currently considered for an exchange. Then for any agent a such that $b \in S(a)$ (respectively for c), update $S(a)$ by adding c (respectively b). Intuitively, a greedy algorithm is interaction-free if for any two agents that are being considered for the first time in the algorithm, they have not previously interacted, i.e., they are not in each other's interaction set. More formally, a greedy algorithm satisfies the interaction-free property, if for any two agents a and b, whenever an exchange involving agents a and b is considered in the process of the greedy algorithm, one of the following is true:

- $a \notin S(b)$ and $b \notin S(a)$, or
- the only exchanges that are already considered for agents a and b involve both a and b.

Lemma 22 *Any greedy algorithm that satisfies the interaction-free property is truthful.*

Proof. Fix an agent a. For any $b \neq a$ such that an exchange between b and a is ever considered, let $\hat{I}_b \subseteq I_b$ and $\hat{W}_b \subseteq W_b$ be the sets of items that are not used in any exchange before the first time an exchange between a and b is considered. Directly from the definition of interaction-freeness, it follows that $\hat{I}_b$ and $\hat{W}_b$ are independent of the strategy of a. That is, from the perspective of agent a, the algorithm ranks a subset of other agents with $\hat{I}_b$ and $\hat{W}_b$, and then considers greedily adding a subset of all possible exchanges according to the ranking. To prove truthfulness, we only need to show a does not benefit by misreporting in this simple procedure.

Consider an order σ on a subset of exchanges each involving a. Assume for contradiction that a benefits from reporting I'_a and W'_a. That is, a's utility when reporting truthfully is k, and his utility when reporting I'_a and W'_a is $k' > k$. Since all the exchanges involve a, his utility when reporting truthfully is equal to the number of cycles picked by the algorithm. Let $C = \{c_1, \ldots, c_k\}$ be such a set. Let $C' = \hat{C} \cup \bar{C}$ be the set of cycles picked when a misreports, where $\hat{C}$ is the set of exchanges in which a does not receive an item that he wants, and $\bar{C}$ is the set of exchanges in which a receives and item he wants (and therefore $|\bar{C}| = k'$). Assume that the set of items a gives away in C' is a subset of I_a, since otherwise he will have a large dis-utility. Let σ' be the projection of σ on $C \cup C'$. The outcome of the greedy algorithm on σ and σ' is the same, and therefore i will still benefit by misreporting in σ'. We can further assume that $C \cap C' = \emptyset$, by removing any element of $C \cap C'$ from σ'. Removing such an element will decrease i's utility of being truthful and his utility by misreporting by the same amount. Notice that for any element $\bar{c}$ of $\bar{C}$ there must be an element of C that has non-empty intersection with $\bar{c}$ (otherwise $barc$ would have been picked by the algorithm when a reports truthfully). We say that c covers $\bar{c}$ in this case. Let $C_1 \subseteq C$ be all the elements that cover at most one element of $\bar{C}$, and let $\bar{C}_1 \subseteq \bar{C}$ be all the elements that are covered by C_1. Remove C_1 and $\bar{C}_1$ from σ'. Notice that i would still benefit from misreporting since the number of elements removed from C is at least that of $\bar{C}$. So we can assume that any element of C now covers at least two elements from $\bar{C}$. Consider the first element of $C \cup \bar{C}$ according to σ'. Such a cycle must be in C, since otherwise it would have been picked when a is truthful. Let c be this element. Since c is not picked when a misreports, there must be an element $\hat{c}$ of $\hat{C}$ that appears before c in σ', and has intersection with it. That is, there are 3 elements of C' that have intersection with c. Recall that all these cycles involve a. Let items i and j be the items a gives and receives in c, respectively. For a cycle to have intersection with c, either a must give i or receive

j. This implies that there are at most two non-intersecting cycles that intersect with c. This is a contradiction to the feasibility of C'. ■

Lemma 23 *The above algorithm is a truthful algorithm.*

Proof. Consider the time when the algorithm first visits pair (a_i, b_j). It can be shown inductively that $S(a_i) = \{b_1, \ldots, b_{j-1}\}$, and $S(b_j) = \{a_1, \ldots, a_{i-1}, b_{j+1}, \ldots, b_n\}$. This shows that $a_i \notin S(b_j)$ and $b_j \notin S(a_i)$. ■

We next show by an example that a very simple violation of interaction-free-ness property by a greedy algorithm can make it not truthful.

Example 2 *Assume that we have 3 agents and 5 items, and consider the instance shown in Figure 5.1. Consider an order σ such that $\{(a,2),(b,1)\} \succ_\sigma \{(a,3),(b,1)\} \succ_\sigma \{(a,3),(c,5)\} \succ_\sigma \{(b,4),(c,5)\}$ (what σ does on the rest of the exchanges is irrelevant and thus not represented here for simplicity). This means that the algorithm first considers an exchange in which agent b gives item 1 to agent a and receives 2, and so on. The greedy algorithm parameterized by σ chooses $\{(a,2),(b,1)\}$ and $\{(a,3),(c,5)\}$. The valuation of agent b for this set of exchanges is 1. Now assume that b misreports his wish list as $W'_b = \{3,5\}$, instead of the true set which is $W_b = \{2,3,5\}$. The outcome of the algorithm on this instance is going to be $\{(a,3),(b,1)\}$ and $\{(b,4),(c,5)\}$. The valuation of b for this set is 2. Therefore, the static algorithm parameterized by σ is not truthful.*

Notice that this algorithm violates the interaction-free-ness property because at the time when the last exchange is considered between b and c, we have $S(b) = \{a,b,c\}$ and $S(c) = \{a,c\}$. Since $c \in S(b)$, thus, the algorithm is not interaction-free.

5.3.3 Computational Complexity

In this section, we show that the length-constrained market exchange problem is APX-hard for any $k \geq 2$. First, we discuss the case of $k = 2$. To prove APX-hardness for $k = 2$, we use the fact that there exists a factor-preserving reduction from the edge-disjoint 3-cycle partitioning of 3-partite graphs to the market exchange problem with $k = 2$ [13]. Here, we show that edge-disjoint 3-cycle partitioning of 3-partite graphs is APX-Hard, which in turn, implies our desired

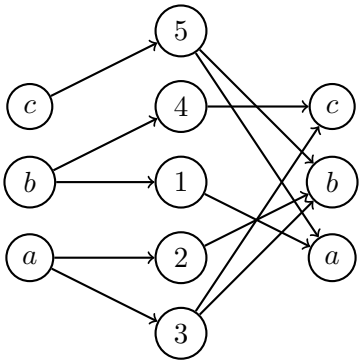


Figure 5.1: The exchange market of Example 2. For the ease of visualization, each vertex representing an agent is split into two vertices.

result. Formally, the problem edge-disjoint 3-cycle partitioning of 3-partite graphs problem, given a tripartite graph $G(V_1, V_2, V_3; E)$ where V_1, V_2 and V_3 are disjoint sets of vertices, and $E \subseteq \{V_1 \times V_2\} \cup \{V_1 \times V_3\} \cup \{V_2 \times V_3\}$, the goal is to find the maximum number of edge-disjoint triangles in G. A factor-preserving reduction from EdgeDisjTrianglePar to the 2-exchange market problem is given in [13]. To show the APX-hardness of 2-exchange market problem, it remains to prove that EdgeDisjTrianglePar is APX-hard.

Holyer [61] proved that edge-partitioning of general graphs into edge-disjoint triangles is NP-complete. A more careful analysis of this proof shows that the edge-partitioning of general graphs is APX-hard [63]. We first note that the set of graphs that Holyer used in his NP-hardness proof for edge-partitioning triangles is in fact tripartite. To show this, we define some notations from [61]. Let graph $H_{3,n}$ be a graph with n^3 vertices $V = \{(x_1, x_2, x_3) \in \{0, 1, 2\}^n \mid \sum_{i=1}^{3} x_i = 0 (\text{mod } n)\}$. Let $((x_1, x_2, x_3), (y_1, y_2, y_3))$ be an edge in $H_{3,n}$ if there exists $i, j \in \{1, 2, 3\}$, $i \neq j$, such that $x_k = y_k (\text{mod } n)$ for $k \neq i, j$ and $y_i = (x_i + 1) (\text{mod } n)$ and $y_j = (x_j + 1) (\text{mod } n)$. The resulting graph reduced from any 3SAT instance in Holyer's proof is a result of combining and joining $H_{3,p}$'s. It is not hard to verify that $H_{3,n}$ is 3-vertex-colorable and any combination and joint of these graphs is also 3-vertex-colorable. As a result, Holyer's proof of NP-hardness [61] and its extension for APX-hardness [63] of edge-partitioning of general graphs implies the APX-hardness of EdgeDisjTrianglePar. This, in turn, implies that 2-exchange market problem is APX-hard.

The APX-hardness proof for the k-exchange market problem where $k > 2$ is very similar to

that of the 2-exchange market problem. First, one can give a similar factor-preserving reduction from the problem of edge-partitioning of a k-partite graph to k-cycles to the k-exchange market problem. Now the APX-hardness of the k-exchange market problem boils down to the APX-hardness of edge-partitioning of k-partite graphs to k-cycles, which can be shown by giving a reduction from EdgeDisjTrianglePar to edge-partitioning of a k-partite graph to k-cycles. To see this, given an instance $G(V_1, V_2, V_3; E)$ of EdgeDisjTrianglePar, we construct a k-partite graph $G'(U_1, U_2, \ldots, U_k; E')$ where $U_1 = V_1, U_2 = V_2, U_3 = V_3$, and for $i \geq 4$, $U_i = E_3$ where $E_3 = E \cap \{V_2 \times V_3\}$, i.e., each node in U_i corresponds to an edge $e \in E_3$ from V_2 to V_3. Denote by $e_4, e_5, \ldots, e_k$ the $k-3$ nodes in G' corresponding to edge $e \in E_3$. We also form the edges of $E' = \cup\{(u, v)|u \in V_2, v \in V_3$, as follows:

- include all edges from nodes in U_1 to nodes in U_2 and U_3 for each pair of nodes whose corresponding nodes in G are connected, i.e., add $\{(u, v)|u \in V_1, v \in V_2 \cup V_3, (u, v) \in E(G)\}$.
- for each edge $e = (u, v) \in E_3$, add the following edges to E': $(u, e_4), (e_4, e_5), \ldots, (e_{k-1}, e_k), (e_k, v)$.

It is not hard to see that any triangle (w, u, v) where $(u, v) = e \in E_3$ in graph G corresponds to the k-cycle $(w, u, e_4, e_5, \ldots, e_k, v)$ in G' and vice versa. G is a tripartite graph and G' is a k-partite graph. As a result, the above is a factor-preserving reduction from EdgeDisjTrianglePar to the problem of edge-partitioning a k-partite graph to k-cycles. Therefore, APX-hardness of k-exchange market problem follows from APX-hardness of EdgeDisjTrianglePar.

5.4 Unconstrained Exchange Market Problem

In this section, we give a polynomial-time algorithm for the unconstrained exchange market problem. As we stated earlier, we would like to maximize the number of edges by picking a set of edge-disjoint cycles. One can write this problem as a maximum circulation problem (or minimum cost circulation problem with negative cost on the edges [3]), and solve it in polynomial time us-

[3]The minimum cost circulation problem is as follows: Given a graph G and capacities and costs u_e and c_e for each edge $e \in E(G)$, find a circulation flow f with capacities $f(e) \leq u_e$, and the minimum total cost $\sum_{e \in E} c_e f(e)$. Our maximum circulation problem can be modeled by setting $c_e = -1$ for each edge in $E(G)$, and applying the algorithm for the minimum cost circulation problem [51].

ing an algorithm that interactively improves the current solution with a cycle in an augmenting graph [51]. We will present a variant of this algorithm that satisfies a desired set of properties (e.g, a specific monotonicity property). We start by a high-level description of the solution for the maximum circulation problem [51].

Given a directed unweighted anti-symmetric graph G, a flow is a function $f : G \to \mathbb{Z}$. Flow f is feasible in G if it satisfies:

- $\forall e \in G, f_e \leq 1$, and $\forall e \notin G, f_e \leq 0$,
- $f_{(u,v)} = -f_{(v,u)}$,
- For any vertex v, $\sum_{e \sim v} f_e = 0$, where $e \sim v$ if v is an endpoint of e.

The goal is to find a feasible flow that maximizes *weight* $w_G(f) = \sum_{e \in G} f_e$. Given a flow f, we now define the residual graph G_f corresponding to f:

Definition 5 *Given a graph G and a circulation f, we define the residual graph corresponding to f to be $G_f(V, E_f)$ with $E_f = \{(u, v) \in E(G) | f(u, v) = 0\} \cup \{(v, u) \in E(G) | f(u, v) = 1\}$.*

We say G_f admits a flow f' if f' is a feasible circulation flow in G_f. The following is a well-known result which connects optimality of the flows to absence of cycles with positive weight in the residual graph.

Lemma 24 *[51] A flow f is optimal if and only if its residual graph G_f does not admit any flow f' with $w_G(f') \geq 0$.*

The above lemma suggests the following algorithm for the exchange market problem:

1. Initialize flow $f := 0$, and maintain a feasible flow.
2. Construct G_f.
3. If G_f admits a flow f' with positive weight, augment f with f' by setting $f := f + f'$, and go back to Line 2. Otherwise, terminate.

Notice that since the weight of the integral flow increases by at least 1 (since optimal flows are integral) in each step, the pseudo-algorithm terminates after polynomial number of steps. Thus, if the selection of an augmenting flow is done in polynomial time in each step, the algorithm runs in polynomial time.

Chapter 6

Identifying/characterizing social phenomena in online networks

6.1 Social Ratings: Friends or the Crowd?

6.1.1 Introduction

It is generally believed that friends are similar to each other [80]. In fact, many social psychology studies have shown this [74, 82, 23] and report that people all over the world find similarity as a desirable quality in a friend (e.g.[113]). Although this is a well-established sociological premise, this phenomenon is not well studied in the context of *online* social media. When making choices, people use information from a number of sources including friends, family, experts, media, and the general public. Two sources that are particularly relevant in an online setting are the opinions of friends and ratings from the general public. Friends are believed to influence choices of their friends. In many cases, however, recommendations from one's friends are in contrast to opinions of individuals in the general public who are not one's friends.

Specifically, focusing on friends and the general public as two components of social influence is important because these sources of social information are already used in a variety of algorithms and applications online. Social recommender systems take into account the actions of a user's friends and make recommendations accordingly. Social search is also gaining more attention. Google recently launched its +1 button for search results and ads in order to improve its search

algorithm. If a user thinks that a search result or an ad is useful she can click on the +1 button. The +1 will be displayed along with the user's name in the search results to all her friends who subsequently search a similar query. For users who are not friends, only the number of +1's will be displayed. Facebook uses a similar approach for business pages with the intention of getting higher click-through rates. The model that we suggest can be leveraged to design better algorithms for these and other similar applications.

In the first part of this work we study how an online user's decision is influenced by recommendations from friends and ratings from the general public, particularly when these two sources of information are in conflict with each other. This question is interesting for two reasons. First, understanding how people trade off friends' opinions with ratings from the general public helps to determine the weight assigned by consumers to these two sources when they are uncertain about choosing one of two possible options. Second, this information can be used when designing algorithms that display these two sources of information in order to increase the probability of a user selecting one of the options. For example, an online social network platform that has information about how a user's friends and the general public have rated two different items can display to the user the item that she is more likely to select. On the other hand, if users tend to disregard some source of information, this source need not be shown to the user. Finally, an advertiser that wishes to make the user choose a certain item may strategically choose which pieces of information to show.

In the second part of this research, our main research question is:

To predict future ratings of an individual, how useful is it to know the ratings of their friends? In other words, for applications such as recommender systems that currently do not have an underlying social network (e.g. Netflix and Amazon) how beneficial is it to invest in social data to increase the accuracy of their predictions?

To find the predictive value of average ratings vs. friends' average ratings, we formulated our problem as the prediction of ratings for users of social rating networks. We included different features such as friends' average rating (computed by us) and the average rating by all the users who have rated the book (given by Goodreads) in the predictions incrementally. Specifically, we applied ordered logistic regression, support vector machines and matrix factorization for the task of prediction and evaluated our models by comparing their accuracy and errors. For this work we

studied two social rating networks: Goodreads and Flixster. Goodreads is an online social book cataloging network that features a rating system. Since the dataset was not available online, we crawled the Goodreads website and gathered the relevant data. Flixster is a social movie rating website. This dataset was available [66].

This is important to study since predicting future behavior is crucial for many online applications such as recommender systems, online ad placement/
allocation, and web search ranking. Applications like viral marketing and social recommender systems, in fact rely on the premise that the behavior of one's social contacts is a good predictor of one's behavior.

Extensive studies have been done to show the effect of social influence on an individual's decision to *adopt* or *select* a new product, for example in viral marketing and in the area of recommender systems [58, 66, 67]. However, our goal in this study is to determine how much users actually rate items similarly to their friends as opposed to the general public (regardless of the underlying reason why the item was chosen in the first place).

6.1.2 Swayed by Friends or the Crowd? [8]

In particular, in the first part of the work we ask the following questions:

1. How much are one's choices influenced by the opinions of her friends compared to ratings from the general public? What mathematical model predicts this?

2. Do friends' negative opinions have a stronger or weaker effect than friends' positive opinions about an item?

3. Do friends' opinions have the same effect on one's decision in higher risk situations versus lower risk situations?

To answer the above questions, we performed user studies on Mechanical Turk involving around 350 participants using positive and negative opinions from friends, as well as ratings from the general public; the latter was represented by the average number of stars. We find that the choice between two options fits a logit model. Our major contributions are (1) Our model is able to predict the probability of selection of an item by a user given two choices when recommendations from friends and star ratings from the general public is displayed, (2) We find that negative opinions from friends are more influential than positive opinions, and (3) We observe that people exhibit

more random behavior in their choices when the decision involves less cost and risk. Our results are quite general in the sense that people across different demographics trade off recommendations from friends and ratings from the general public in a similar fashion.

6.1.3 Method

Our goal is to study how people trade off information from friends and the general public when choosing between two items. Moreover, our experiments allow us to compare a setting where the information from friends consists of positive recommendations to a setting where the information from friends consists of negative opinions. Furthermore, we compare people's choices with respect to two types of decisions: one that involves a monetary cost (booking a hotel) and a low risk decision that involves no monetary cost (watching a movie trailer). We chose booking hotels because the user cannot go and check it out before deciding and should rely on the information she gets from others. Similarly, a user may not have any information about a movie trailer before she watches it. We can think of the setting with the movie trailers as a less serious decision, since it involves less cost (just a couple of minutes of one's time) and risk. Users often make choices of this type online, e.g., when watching Youtube videos, clicking on a link or ad, etc.

In total, we conducted three user studies: booking a hotel with positive recommendations from friends (Study 1), booking a hotel with negative opinions from friends (Study 2) and watching a movie trailer with positive recommendations from friends (Study 3).

To collect the data we conducted the three studies with 350 participants each in the form of surveys on Amazon's Mechanical Turk (MTurk) during July and August 2011.

We asked each worker to put herself in the following hypothetical situation: she is about to book a hotel (resp. watch a movie trailer) an on e-commerce site (resp. online), and among the options, she has come down to two between which she is indifferent. The website has an underlying social network of friends (or it runs on top of an online social network). For each of the two options, we provide the following information:

(i) the overall rating (in terms of stars on the scale of 1 to 5) based on ratings from a large number of previous customers (resp. users) in the case of selecting which hotel to book (resp. which movie trailer to watch)

(ii) the number of friends who recommend (resp. have negative opinions about) the option in the case of positive (resp. negative) recommendations

For each question, the option that has more stars is the one that is less recommended by friends; that is, we did not use a pair of options where one clearly dominated the other.

A sample question from each survey is shown in Figure **??**. We incorporated a few tricks in order to make sure the workers were not cheating. We discarded answers from the workers who did not pass our test.

Overall, we rejected 33% of the responses across all 3 studies because they were invalid. The average completion time for each valid HIT was 174.8 seconds while the average completion time for the invalid HITS was 153.3; this suggests that the workers that were rejected had not taken the task as seriously as the rest of the workers.

6.1.4 Results

For each question, there are two options that the worker can select from, which we refer to as option 1 and option 2. We denote the number of stars S_i and the number of friends' recommendations by F_i, for $i = 1, 2$. To predict the probability that option 1 is selected, we conduct a logistic regression[1] on our dataset of choices with the difference in the number of friends (i.e., $F_1 - F_2$) and the number of stars (i.e., $S_1 - S_2$) for each question as the predictor variables. We denote the corresponding coefficients by α_f and α_s respectively.

We also ran the logistic regression with an intercept, but the intercept was not statistically significant. This is good news, since a statistically significant intercept in this case would imply bias (that is, the position in which an option is presented would affect the probability that it is

[1]We note that a number of other empirical studies also use the logit choice function to model social influence [88, 49, 103].

Table 6.1: Study 1: Positive Opinions for Hotels

Predictor	Estimated Coefficients	z-value
α_f	0.20471*** (0.027)	7.597
α_s	0.73549*** (0.050)	14.307

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo-R^2 = 0.95

selected). We next report the results from each survey separately.

6.1.5 Study 1: Positive Opinions for Hotels

6.1.5.1 Model 1

We first only considered the difference in the number of stars and friends as predictor variables. The estimated coefficients along with other parameters are shown in Table 6.1, and as can be seen both are statistically significant. Observe that both coefficients are positive; this is intuitive, since more stars (resp. more positive recommendations) indicate that the option is better and thus the worker is more likely to select it. Finally, the pseudo -R^2 for this model[2] is 0.95, indicating that the fit is very good.

Interpretation of the coefficients We first interpret the coefficients for our model in terms of marginal effects on the odds ratio. The odds ratio measures the probability that the dependent variable is equal to 1 relative to the probability that it is equal to zero. For the logit model, the log odds of the outcome is modeled as a linear combination of the predictor variables; therefore, the

[2]We computed Efron's pseudo -R^2 which is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where, N is the number of observations in the model, y is the dependent variable, $\bar{y}$ is the mean of the y values, and $\hat{\pi}$ is the probabilities predicted by the logit model. The numerator of the ratio is the sum of the squared differences between the actual y values and the predicted π probabilities. The denominator of the ratio is the sum of squared differences between the actual y values and their mean [55].

Table 6.2: Cross validation for study 1

Left out question	Actual	Predicted	|Difference|
Q1	0.54	0.53	0.01
Q2	0.58	0.55	0.03
Q3	0.71	0.62	0.09
Q4	0.74	0.74	0.00
Q5	0.77	0.77	0.00
Q6	0.74	0.72	0.02
Q7	0.82	0.83	0.01
Q8	0.54	0.53	0.01

odds ratio of a coefficient is equal to $\exp$(coefficient). Since $\alpha_s = 0.735$, we conclude that a unit increase in $S_1 - S_2$, multiplies the initial odds ratio by $\exp(0.735) = 2.07$. For the friends predictor variable, the odds ratio is equal to $\exp$(0.204) = 1.22. Another way to interpret the coefficients is in terms of relative change in the probability when there is one unit of change in one of the predictor variables while other parameters remain the same. In this case the relative probability increases by at most 10% with a unit change in $F_1 - F_2$ and by at most 35% with a unit change in $S_1 - S_2$.

To further assess the predictive power of the model, we performed **cross validation**. We left out one question at a time and estimated the coefficients using the remaining questions. Then, we predicted the probabilities for the question that was left out. The predicted values were very close in all cases with absolute mean difference of 0.021. The actual values and their differences can be found in Table 6.2.

Finally, we used one of the questions of this survey twice in Amazon's Mechanical Turk (in two separate HITS) in order to see whether workers would react to the question in similar ways. We found that the percentage of workers that chose the first option of the question was similar in both cases (26% versus 24%), further validating our approach.

Table 6.3: Study 2: Negative Opinions for Hotels

Predictor	Estimated Coefficients	z-value
α_f	-0.281*** (0.030)	9.378
α_s	0.503*** (0.050)	10.018

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo-R^2 = 0.95

6.1.5.2 Model 1'

In Model 1', we included all self reported demographic information as predictor variables in addition to the stars and friends' recommendation variables. This information includes: gender, age, and education level. More specifically, we coded the following variables as dummy variables. We found that these extra coefficients are not statistically significant. This suggests that people in different demographics trade off ratings from the public and friends' recommendations similarly.

6.1.6 Study 2: Negative Opinions for Hotels

6.1.6.1 Model 2

In this section we look at negative opinions from friends — instead of positive recommendations. In particular, each option is characterized by the number of stars (based on information from the general public) as well as the number of friends who have negative opinions about it. We run a logistic regression and report the results in Table 6.3. As can be seen in the table both variables are statistically significant and the pseudo-R^2 measure for this model is 0.95 which implies that the model is a good fit. Moreover, as we would expect, the friends coefficient is negative in this case, as more negative opinions from friends decrease the probability that the worker selects an option.

Table 6.4: Cross validation for study 2

Left out question	Actual	Predicted	|Difference|
Q1	0.30	0.25	0.05
Q2	0.39	0.41	0.02
Q3	0.43	0.44	0.01
Q4	0.54	0.53	0.01
Q5	0.58	0.60	0.02
Q6	0.38	0.39	0.01
Q7	0.40	0.42	0.02
Q8	0.45	0.50	0.05

Interpretation of the coefficients Similarly to Study 1, we interpret the coefficients for our model in terms of marginal effects on the odds ratio. For the present model (negative recommendations), the fact that $\alpha_s = 0.503$ means that one unit increase in $S_1 - S_2$, multiplies the initial odds ratio by $\exp(0.503)$ = 1.65. In other words, the odds of choosing option 1 increases by 65%. For the friends predictor variable, the odds ratio is equal to $\exp(-0.281) = 0.75$, which means that the odds of selecting option 1 decreasing by 25%. Equivalently, the relative odds of selecting option 1 when $F_1 - F_2$ decreases by one unit is $(\exp(0.281) - 1) \approx 32\%$. Another way to interpret the coefficients is by looking at the relative changes in the probability of choosing each option. In this case the relative probability decreases by at most 14% with a unit change in $F_1 - F_2$ and by at most 26% with a unit change in $S_1 - S_2$.

For this study we did **cross validation** as well to test the predictive power of our model. The results are shown in Table 6.4. The predicted and actual values are very close (mean absolute difference = 0.231).

Table 6.5: Study 3: Positive Opinions for Movie Trailers

Predictor	Estimated Coefficients	z-value
α_f	0.167^{***} (0.049)	7.101
α_s	0.349^{***} (0.027)	6.014

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo-R^2 = 0.61

6.1.7 Study 3: Positive Opinions for Movie Trailers

6.1.7.1 Model 3

Our third study considers the effect of positive recommendations from friends in a low risk decision: choosing which movie trailer to watch. We perform a logistic regression and report the estimated coefficients in Table 6.5. The estimated coefficients are statistically significant; however, in this case pseudo-R^2 is 0.61 which is lower than the pseudo-R^2's for previous models Models 1 and 2 (0.95). The coefficients for stars and friends are $\alpha_s = 0.349$ and $\alpha_f = 0.167$. As for Model 1, both coefficients are positive, since people are more likely to select an option if it has more stars and/or more positive recommendations from friends. Therefore, the odds ratio for the number of stars is 1.41 and for the number of friends' recommendations is 1.18. By computing the relative probability changes, we conclude that an additional star increases the probability of selecting that option by 18%, whereas an additional recommendation from a friend increases the probability by 8%.

6.1.8 Results and Discussion

This work studies how positive and negative opinions from friends affect our decisions compared to ratings from the crowd for different types of decisions. Our three user studies result in some interesting conceptual findings about the tradeoff between these two types of social influence.

First, negative opinions from friends are more influential on one's decision than positive opinions. We can see this by comparing the odds ratios of Study 1 and Study 2, in which the number

of positive and negative friends' opinions are shown respectively: the odds ratio for the friends variable is higher in Study 2 (1.32 versus 1.22 where the difference is statistically significant with $p = 0.046$), whereas the odds ratio for the stars variable is higher in Study 1 (2.07 versus 1.65 where the difference is statistically significant with $p = 0.001$). In other words, one less negative opinion from a friend has a larger effect than one more positive opinion, whereas one more star increases the odds of an option being chosen less in the case that negative opinions from friends are present. Such an asymmetry between the effect of negative and positive actions and opinions have been studied in the social psychology literature [27, 89, 107, 32, 54]. The *positive-negative asymmetry effect* has been observed in many domains such as impression formation [17], information-integration paradigm [18] and prospect theory for decision making under risk [68]. The finding in all the above cited work is that negativity has stronger effects than equally intense positivity. Our results confirm this finding in online settings.

Second, people exhibit more random behavior when the decision involves less cost and less risk. We can see this by comparing the results from Study 1 and Study 3, where the decisions are "which hotel to book" and "which movie trailer to watch" respectively. Booking a hotel clearly involves a monetary cost and some risk, whereas the worse thing that can happen with a movie trailer is to waste a couple of minutes of one's time. The odds ratios are lower in Study 3 than Study 1 (1.18 versus 1.22 for friends with $p = 0.345$ which is not statistically significant, and 1.41 versus 2.07 for stars where the difference is statistically significant with $p < 0.0001$). This implies that one added star has a smaller influence on one's decision in the case of movie trailers. However, one added friend has basically the same influence as in the case of hotels. Moreover, the fraction of respondents choosing either option is closer to half compared to the hotel booking surveys. This implies that the choices were more random in this case, which may be explained by the fact that choosing which movie trailer to watch is a less important/serious decision than booking a hotel.

Third, we observe that in all three studies one more star increases the probability of selecting that option more than one more (resp. less) friend in the case of positive (resp. negative) recommendations. Equivalently, the odds ratio of the stars' coefficient is larger than the odds ratio of the friends' coefficient (2.07 versus 1.22, 1.65 versus 1.32 and 1.41 versus 1.18 for studies 1, 2 and 3 respectively where for all three studies the differences are very statistically significant.) This does not mean that the number of friends' positive or negative recommendations does not influence

decisions; on the contrary, an additional recommendation (resp. one less negative opinion) from friends changes the probability by at least 18% across all three studies. The fact that an additional star has a larger effect that an additional friend opinion is reasonable if we consider that the number of stars is bounded between 1 and 5, whereas the number of friends' recommendations may take values from a larger range.

Forth, for all of our user studies, we find out that the demographic variables (gender, age, and education level) do not significantly impact the choice that is made, implying that people across different demographics trade off recommendations from friends and ratings from the crowd in a similar way. It also implies that our predictive model and results are generalizable across different demographics.

6.2 Predicting Social Ratings in Online Networks: Friends or the Crowd?

Our main research question in this part is:[3]

To predict future ratings of an individual, how useful is it to know the ratings of their friends? In other words, for applications such as recommender systems that currently do not have an underlying social network (e.g. Netflix and Amazon) how beneficial is it to invest in social data to increase the accuracy of their predictions?

6.2.1 Data

The analyses in this work are based on two different social rating networks: Goodreads.com and Flixster.com. While data from Flixster was available online [66], we had to crawl the Goodreads website in order to gather data that was relevant to our research question[4].

6.2.2 Prediction Methods

We formulate our main research question as follows: how beneficial is it to know one's friends' ratings to predict one's future ratings in online social applications? In order to investigate this

[3]This work is in submission.

[4]Our dataset is available upon request.

question, we model the problem as a prediction task and explore the predictive value of the public's average ratings vs. friends' average ratings on two social rating networks. We employ three different prediction methods: ordered logistic regression, support vector machines and matrix factorization. We include features in the prediction model incrementally: the first model does not include the average rating from the public, the second model excludes friends' average ratings, and the third includes both of the previously excluded features (i.e. all features). Other than the average rating variables, we included the subject areas of the books and the number of times each book is rated as predictor variables to increase the accuracy of our results. Our choice of additional features was based on the available data.

In most cases, average rating and friends' average rating are close to each other (i.e. their difference is less than 1). This might suggest that including both average ratings (friends' and the public) will be redundant. Therefore, we consider two other cases were these two were at least 1 and 2 units apart. In other words cases where there is a divergence between friends' opinions and the general public's opinions.

6.2.3 Evaluation Setting

To evaluate and compare our prediction techniques we use root mean squared error (RMSE) as our evaluation metric. RMSE is commonly used as a metric for evaluating recommender system accuracy [57]. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{(u,i)\in \text{Test}} \left(r_{u,i} - \hat{r}_{u,i}\right)^2}{|\text{Test}|}}$$

where Test is the test data, (u, i) is the pair of user u and item i, $r_{u,i}$ is the actual rating of user u for item i and $\hat{r}_{u,i}$ is the estimated rating predicted by our model. We performed cross validation in order to assess the accuracy of our prediction models. The training, test and validation sets were comprised of 60%, 20%, and 20% of the data respectively.

6.2.4 Results

In this section, first, we present the results of applying the methods described in 6.2.2 on all data and then turn to cases where there is a divergence between friends' ratings and the public's ratings.

Predictors	Model 1	Model 2	Model 3
Friends' Rating	3.8694^{**} (0.03)	–	2.9202^{**} (0.03)
The Public's Rating	–	8.8029^{**} (0.07)	6.3313^{**} (0.07)
RMSE	**1.07**	**1.04**	**1.00**

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6.6: **GoodReads:** Logit coefficients for models M1, M2 and M3. Standard errors are indicated in parentheses.

6.2.4.1 Results over All data

Ordered Logistic Regression. In addition to RMSE, ologit yields coefficients that show how important each independent variable is in predicting the value of the dependent variable. We applied ordered logistic regression (ologit) to the Goodreads data to three different cases: 1) included all dependent variables (prediction features) except for the public's ratings (Model 1) 2) included all dependent variables (prediction features) but friends' average ratings (Model 2) 3) included all prediction features.

Table 6.6 depicts the coefficients from each model along with their standard errors. The last row in Table 6.6 shows the RMSE values obtained for each model. As can be seen, the public's rating has a higher coefficient when both variables are included as predictive features in the model. This implies that one's ratings have a higher correlation with the public's rating. Furthermore, the RMSE values confirm this observation. RMSE is the lowest in Model 3 (the difference between RMSE's of Model 3 and Model 2 is statistically significant with $p < 0.001$) when both variables are part of the prediction. Model 2 has a lower RMSE than Model 1 (the difference is statistically significant with $p = 0.0377$), suggesting that on this dataset using the public's rating would result in a more accurate prediction.

Support Vector Machines (SVM) and Matrix Factorization (MF). We applied SVM and MF to the Goodreads and Flixster datasets incrementally same as above. The RMSE values for the Goodreads and Flixster datasets are summarized in Tables 6.7 and 6.8 respectively. RMSE values from Flixster and Goodreads validate each other: for all cases, Model 3 performs better than Model

2 followed by Model 1. The p-values of these differences show that they are statistically significant.

	Ologit	SVM	MF
All but public's ratings	1.07	1.07	1.03
All but friends' ratings	1.04	1.04	1.00
All features	1.00	0.98	0.95

Table 6.7: **GoodReads:** RMSE values: Logistic regression, SVM and Matrix Factorization applied to all data

	SVM	MF
All but public's ratings	1.16	1.12
All but friends' ratings	1.13	1.08
All features	1.07	1.04

Table 6.8: **Flixster:** RMSE values: SVM and Matrix Factorization applied to all data

6.2.4.2 Results over Divergent Data

In this part we describe our results on cases where the average ratings among friends and the public were at least 1 unit (Model 4) and 2 (Model 5) units apart.

Ologit. Similar to the previous case, we fitted ologit to the Goodreads data including all predictive features in order to compare the coefficients. The summary of ologit for these two additional cases are summarized in Table 6.9. We can see that for Model 4 (second column) the public's average rating has a lower coefficient value compared to Table 6.6. The case is different for Model 5 (third column of Table 6.9). In this case, friends' average rating has a larger coefficient than the public's average rating (unlike previous cases). This implies that in cases where the public's opinions is different from friends' opinions, friends' are better predictors of one's rating. This observation can be confirmed by the RMSE values of these two models summarized in Table 6.12. The RMSE value for the case where the difference is at least 2 and friends' ratings are included is less than the case where only the public's ratings are considered.

Variables	M4: Diff ≥ 1	M5: Diff ≥ 2
Friends' Rating	2.599^{**} (0.0338)	1.900^{**} (0.268)
The Public's Rating	5.385^{**} (0.1162)	1.387^{**} (0.298)

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6.9: GoodReads: Logit coefficients for models M4 and M5: where friends' average rating and the public's average rating are at least 1 and 2 units apart respectively.

SVM and MF. We ran SVM and MF on the subset of data with convergence between ratings from friends and the public. Our observations are consistent with what our findings from fitting ologit. In particular, we report RMSE values for cases where the difference is at least 1 in Table 6.10. RMSE values for Model 1 and Model 2 are very close – differences are not statistically significant for any of the technics. The case is different for the subset of data with difference of at least 2. The corresponding RMSE values can be seen in Table6.12. In these cases, the difference between the RMSE values of all models is statistically significant and in all cases Model 3 performs the best, followed by Model 1 and then Model 2. Notice the difference with Subsection 6.2.4.1. This shows the importance of friends when there is a divergence between their ratings and public's ratings.

	Ologit	SVM	MF
All but public's ratings	1.09	1.08	1.03
All but friends' ratings	1.09	1.07	1.02
All features	1.03	1.05	0.98

Table 6.10: **GoodReads:** RMSE values: Logistic regression, SVM and Matrix Factorization applied to cases where the difference is at least 1 unit

6.2.5 Discussion

Our study results in some interesting findings about the value of friends' average ratings vs. the public's average ratings in predicting ratings. First, all three methods (ologit, SVM and MF) show that surprisingly average rating is a better predictor of an individual's rating compared to

	SVM	MF
All but public's ratings	1.21	1.17
All but friends' ratings	1.18	1.15
All features	1.14	1.12

Table 6.11: **Flixster:** RMSE values: SVM and Matrix Factorization applied to cases where the difference is at least 1 unit

	Ologit	SVM	MF
All but public's ratings	1.06	1.05	1.03
All but friends' ratings	1.09	1.10	1.06
All features	1.01	0.98	0.97

Table 6.12: **GoodReads:** RMSE values: Logistic regression, SVM and Matrix Factorization applied to cases where the difference is at least 2 units

	SVM	MF
All but public's ratings	1.25	1.24
All but friends' ratings	1.28	1.26
All features	1.23	1.18

Table 6.13: **Flixster:** RMSE values: SVM and Matrix Factorization applied to cases where the difference is at least 2 units

her friends' average rating in general. This was observed by including these two variables incrementally and comparing the coefficients of the variables, and RMSE values. Second, as can be seen, the RMSE errors are at their minimum when both variables are included in the prediction task, suggesting that these two predictor variables are not redundant and including both improves prediction.

We then turned to cases where friends' average rating and the public's average rating where at least 1 and 2 units apart. These are cases where the general public either liked or did not like a book but friends' of that user had different opinions. In these cases, the public's average ratings is not as

predictive as the previous cases. In fact, where the difference is at least two units, friends' average rating is a better predictor of the rating variable. This implies the importance of the availability of friends' ratings to discover niche audiences. This is our third interesting finding.

In general, comparing these three different prediction methods across all the above experiments shows that MF always achieves better accuracy than SVM and ologit. SVM never performs worse than ologit (i.e. most of times better with some cases ties between them.)

Another interesting observation is that although the dataset is smaller when we only look at cases where there is a considerable difference between what friends think about a book or movie and what the public thinks, RMSE values do not always increase.

6.2.6 Practical Implications

The present work offers insights that can be helpful in various online domains such as referral and viral marketing in online social networks, social network advertising, recommender systems, e-commerce websites and in general in any application which determines a ranking of items. Our work shows that the presence of an underlying social network will improve the predictions of users' interests in different items.

As a specific example, let's say a user is looking for an item on Amazon.com. Amazon has a huge set of items that match this query. Amazon would like to present a ranking which results in a purchase, therefore, they should predict the user's interest (i.e. rating) for each item. Our work shows that utilizing friends' ratings will improve this prediction.

Moreover, our work shows that leveraging the social network of friends can help online platforms – e-commerce website or recommender systems – discover items' niche audience. For instance, consider a case where the average rating for an item is much lower than a user's friends' ratings. Our studies show that friends are much better predictors of an individual's ratings in that case.

Furthermore, it is worth noting that this study is particularly helpful in solving the cold-start issue [72]. In a cold-start situation, a user does not have a history of ratings in the system (e.g. she has just joined the network). However, users usually import contacts from other services such as email services when they join. Therefore, information from her friends will be very valuable, as shown by our work.

Limitations. The results discussed above may be affected by sparsity of information about friends in the online domain. The fact that the nature and meaning of friendships are different between the online and offline worlds [112] may be another contributing factor to the reason why our findings are not completely consistent with sociological studies.

The features included in the models are not complete. Obviously there are many other features influencing one's ratings that were not obtainable by us.

6.2.7 Conclusion and Future Directions

This work shows that the sociological premise that friends are similar to each other does not necessarily entail that online behavior of one's friends' are good predictors of one's future behavior although it improves predictions when combined with other features. More interestingly, in cases where the opinion of friends and the public diverge from each other, friends' ratings become more significant. For future work, it would be interesting to run randomized experiments over time to distinguish the impact of homophily and social influence on ratings. It would be interesting also to include more predictor features such as the strength of the friendship link to the predictions in order to make them more accurate and robust.

6.3 Social Exchange

6.3.1 Introduction

Nowadays, hundreds of millions of people spend a significant part of their social lives on the web. As a result, they reveal various types of social interactions by their actions such as likes, favorite markings, and comments. The increasing amount of data from these actions in online social networks has opened new opportunities to analyze social interactions and social behavior of individuals at macro and microscopic levels. Such analysis could also be used for social network simulations by developing user models. In addition, such studies can help designers of social media platforms increase user engagement in their websites. This is important because more engagement could lead to increased revenue for the platform.

In this work, we aim to analyze the motivations behind users' actions on online social media over an extended period of time. We look specifically at users' likes, comments and favorite mark-

ings on their friends' posts and photos. These actions are analogous to what social scientists and economists call prosocial giving or in general prosocial behavior. Prosocial behavior is defined as "voluntary behavior intended to benefit another person" [75] and has been a puzzle to researchers of different fields because while this type of behavior benefits others, it is often costly for the person performing it [102]. Most theories of why people exhibit prosocial behavior isolate two distinct motivations: Altruism and reciprocity.

In our work, we focus on identifying the underlying motivations behind users' prosocial giving on social media. In particular, our goal is to identify if the motivation is altruism or reciprocity. For that purpose, we study two datasets of sequence of users' actions on social media: a dataset of wall posts by users of Facebook.com, and another dataset of favorite markings by users of Flickr.com. The main difference between these two datasets is that in the Flickr dataset, a user can mark any other user's photo as favorite but in the Facebook dataset, interactions are limited to friends only. We study the sequence of users' actions in these datasets and provide several observations on patterns related to their prosocial giving behavior. We do so by first dividing the interactions into two categories of one-way vs. reciprocal interactions (see Section 6.3.2); and then by computing two scores for each interaction by any pair of users: a total activity score and a relative activity score, called the *net score*, which is computed as the difference between the number of social actions between two users in each direction (See Section 6.3.3). We further take into account the magnitude of reciprocity by computing the number of social givings in each direction and by considering the persistence of interactions by observing the duration of their relationship (See sections 6.3.4 and 6.3.5). Moreover, we explore the dynamics of users' interactions over time, e.g., by computing the sequence of ups and downs of interactions between two users. We cluster users' interactions based on their characteristics summarized above and derive insights from those clusterings (see Section 6.3.6). Finally, we define a prediction task to further study the importance of different factors in determining the length of an interaction. We show that there is a positive correlation between persistence and reciprocity. As the main result of our study, we show a clear trend that more reciprocal interactions are more persistent and longer. More specifically, our findings indicate that for most users in both datasets reciprocity is the primary motivation for sustaining online relationships with others. While the results are very similar for both datasets, we do observe more persistent and more reciprocal interaction in the Facebook dataset. Finally, we study the impact

of friendship on the level of reciprocity among users on Flickr. We examine persistence and reciprocity of pairwise interactions between pairs of friends and pairs of users who are not friends on Flickr but have interacted with each other. Most notably, in this study, we observe higher levels of altruism between friends and higher levels of reciprocity among non-friend pairs.

6.3.2 Datasets and initial Pruning

Before discussing our results, in this section we first elaborate on our two datasets. We will also discuss characteristics of these datasets and define *one-way interactions and reciprocal interactions.*

6.3.2.1 Flickr Dataset

We obtained a Flickr dataset from the Max Planck Institute for software systems[5] containing documents with approximately 7.5 million favorite markings over a 104 day period starting on November 2, 2006. The data contains the list of all photo favorite markings from users with their timestamps. From this input, for each pair of users (A, B), we extract social interactions of (A, B) as a sequence of events with their timestamps. This allows us to easily analyze the social interactions of pairs of individuals on the network.

For example, using this data, we classify interactions between any pair of users into two categories: *one-way interactions and reciprocal ones*. Reciprocal interactions are the ones in which each of the users in the pair gives at least one favorite marking to the other user in the pair. The rest of the interactions are one-way interactions; for these pairs, only one of the users does a favorite marking for the other one. In total, there were 6,327,024 pairwise interactions in the dataset out of which 6.7% of them were reciprocal interactions and 93.3% of them were one-way interactions.

We first removed what we determined to be "super-users"; i.e., super-users are users who receive huge amounts of user feedback but have very little overall outward activity of their own. We define super-users as users who have more than a hundred reactions received than given. Super-user behavior is abnormal and not of interest to this study. We purged these users from the original dataset. In the Flickr dataset, 561 of a total of 482,458 users were identified as super-users. We

[5] http://socialnetworks.mpi-sws.org/data-www2009.html

remove interactions related to these users that are around 2 million interactions, after which around 7% of interactions were reciprocal and 93% were one-way.

6.3.2.2 Facebook dataset

We acquired a dataset from wall postings of users of Facebook.com from the New Orleans regional network that is similar in form to our Flickr dataset [111]. This dataset consists of 838,092 wall posts from September 2006 to January 2009. The details of this dataset can be found in [111]. One of the key differences between the two social networks is that, unlike Flickr, Facebook strongly limits the amount of interaction users can have with other users who are not on their friends list. We observe that out of 183,413 pairwise interactions in this dataset, almost 44% of them are reciprocal interactions and 56% of them are one-way interactions. After removing super-users that account for around 9,000 interactions, 174,595 interactions remain, out of which 45.5% are reciprocal ones.

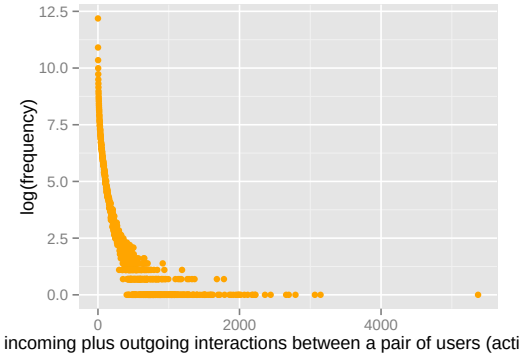
6.3.3 Initial Analysis

In order to present our intial analysis of users' social interactions, in this section, we first define a couple of notions. Let's assume that user A posts a photo on Flickr or a status on Facebook. If user B likes the post or comments on it, we call this behavior *reacting*. We define *activity score* for user u as the number of times u reacts to a post or receives a reaction on her posts. For example, in the above scenario both A and B's activity scores are 1.

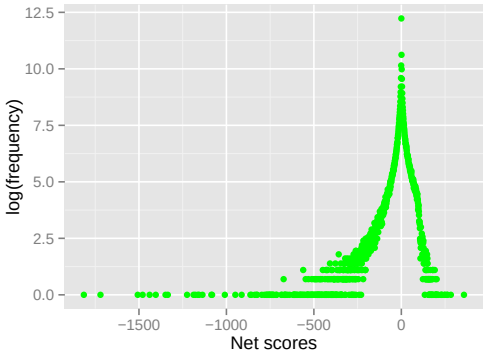
To quantify this "reaction" behavior more closely, we set up a credit/debit-like structure and define a *credit score* or simply a *net score* as follows: For every comment or like spent by user B we add a point to user A's credit score and subtract one point from user B's score. So in effect, user A is earning points and B is spending points. We apply these scores to compute a net score for each user and a net score for each pair of users. Credit score for a pair (A, B) of users is defined as the number of reactions A has for B minus the number of reactions that B has for A. To calculate an individual user's credit score, referred to as the *net score*, we define it as the number of reactions received from others minus the number of reactions given to other users' posts. This means a user who gives more than he receives will have a negative score, as if he had used more credit than he had received.

6.3.3.1 User Score Distributions

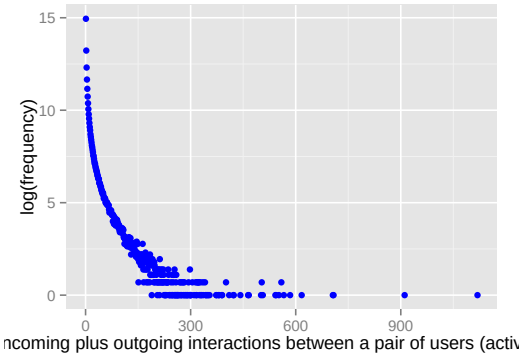
Figure 6.2a shows the by-user distribution of activity scores with a log-scale on the y-axis and Figure 6.2b shows the same distribution for net scores.



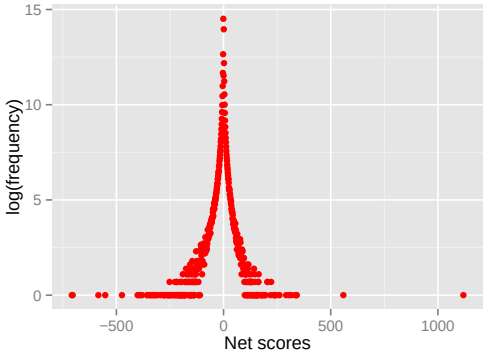
(a) Distribution of user activity scores.



(b) Distribution of user net scores.



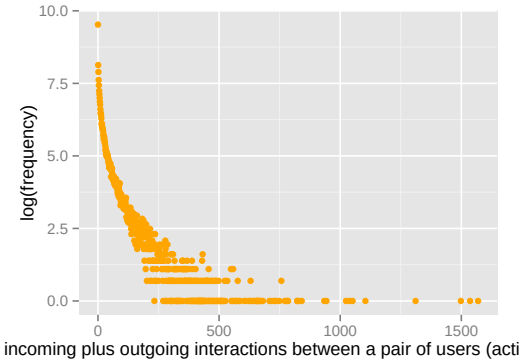
(c) Distribution of pair activity scores.



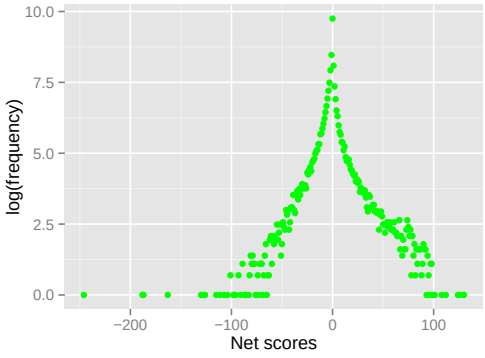
(d) Distribution of pair net scores.

Figure 6.1: Flickr scores

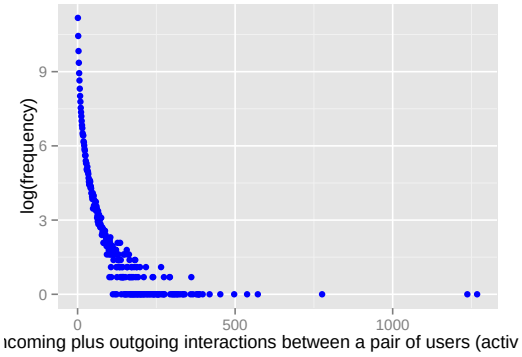
What can be gleaned from these graphs is, first of all, that the majority of users have very low total action and roughly balanced action. This means that the majority of users give or receive very little and that, if they do, it is usually balanced. What can also be seen is that with the super-users removed, interactions are slightly biased, in the net-sense, toward more giving than receiving. We have a greater distribution of negative scores than positive. The net picture confirms we have successfully thrown out our super-users. It also shows us that even for values less that one hundred, there seems to be a faster trail-off of users with positive scores than users with negative scores.



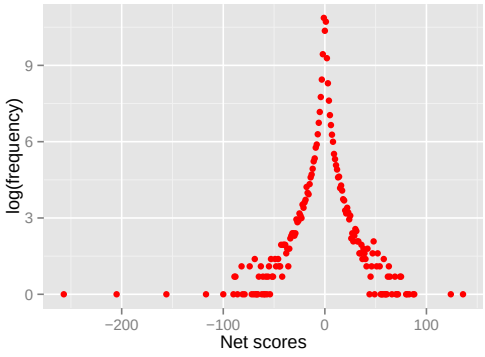
(a) Distribution of user activity scores.



(b) Distribution of user net scores.



(c) Distribution of pair activity scores.



(d) Distribution of pair net scores.

Figure 6.2: Facebook scores

6.3.3.2 User-pair Score Distributions

We now present the same types of distributions but on a user-pair by user-pair basis. Figure 6.2c shows the by-pair distribution for activity scores, again with a log-scale on the y-axis. Figure 6.2d shows the same distribution for by-pair net scores.

One of the things that can be seen from these graphs is that the number of user interactions trails off very quickly, meaning the majority of users are not giving or receiving very many favorite markings at all. What can also be seen is that a big portion of interactions are roughly equal in the net score criteria. It is also important to note that even though we have thrown out the extreme (celebrity-like) super-users, there do seem to be some relationships that are celebrity-like, where one user is receiving a lot more likes than they are giving in the net-sense. One thing that this

phenomenon shows is that there are users that are giving high numbers of favorite markings to a particular user, but receiving enough from other users to offset this behavior.

6.3.4 One-Way vs. Reciprocal Interactions

In the previous section, we introduced the concept of one-way vs. reciprocal interactions and provided some basic statistics about them. In this section, we present an in-depth comparison of these types of interactions, and introduce the concept of "magnitude of reciprocity" for further analysis.

6.3.4.1 Score distributions for each interaction type

As discussed earlier, we partitioned the non-super user interactions into two types: one-way interactions and reciprocal interactions. Note that any user may appear in both types of one-way and reciprocal interactions, since the partitioning for interactions is performed over each interaction pair.

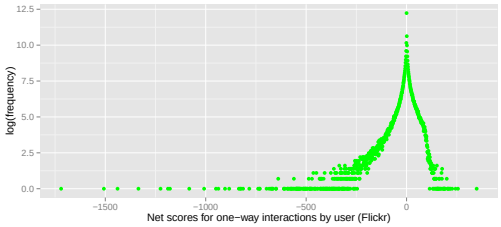
We first calculated both an *activity score* and a *net score* for the user interaction pair. We plotted distributions of scores both by user and by pair for each of these partitions.

By-User Net Scores. While the by-user activity scores provide little useful data, the net score distributions from this method provide us with insight into the differences between reciprocal and one-way interactions. Below we present two graphs. First, we present net score distributions by-user for one-way interactions (Figures 6.3a and 6.4a). The y-axis is on a logarithmic scale. Figures 6.3b and 6.4b shows the same distribution for reciprocal interactions.

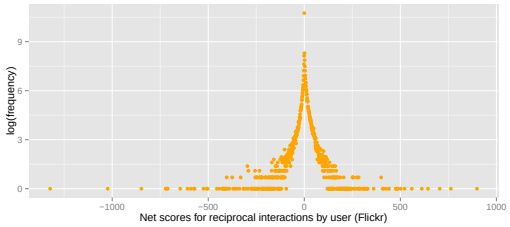
An interesting finding from the results of by-user interactions is that most users' one-way interactions are biased against them, meaning that when interactions are one-way most users give more than they receive. Reciprocal interactions by users however tend to be more symmetrically distributed. On average then for most users, one-way interactions tend to be more lopsided than reciprocal ones.

By-Pair Activity Scores. Looking at scores on a pairwise basis allows us to get into the data on an interaction-by-interaction level.

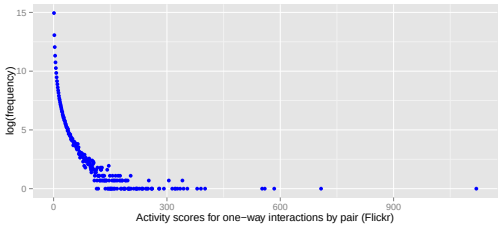
First, we present the activity score distribution by-pair for one-way interactions with a log scale on the y-axis (Figures 6.3c and 6.4c). We also present the same plot for reciprocal interactions



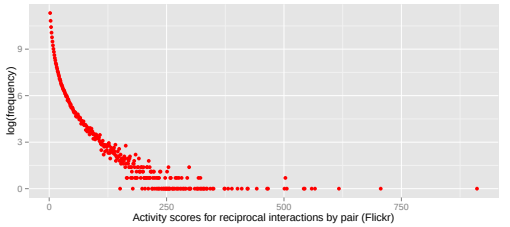
(a) Distribution of net scores by user for one-way interactions.



(b) Distribution of net scores by user for reciprocal interactions.



(c) Distribution of activity scores by pair for one-way interactions.



(d) Distribution of activity scores by pair for reciprocal interactions.

Figure 6.3: Distribution of activity and net scores for one-way vs. reciprocal interactions for the Flickr dataset.

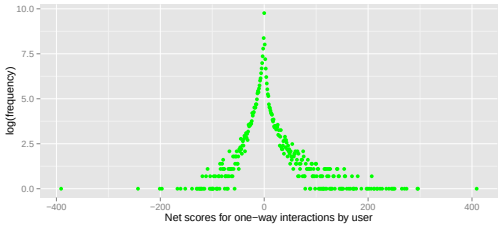
(Figures 6.3d and 6.4d).

In this case, looking at the total number of interactions by pair instead of the net number, allows us to see that for reciprocal interactions the total number of interactions per pair bows out much further than the same data for one-way interactions. This observation points to the manner in which reciprocal interactions tend to last longer than one-way interactions.

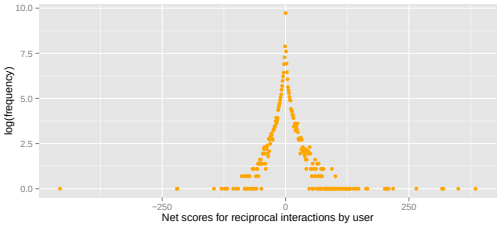
6.3.4.2 Magnitude of Reciprocity and Persistence of Interactions

Partitioning social interactions to only one-way vs. reciprocal interactions is restrictive as it does not distinguish amongst different levles of reciprocity. In this section, we define a new measure called the magnitude of reciprocity. This measure defines how reciprocal an interaction is between two users. We define this measure because intuitively, an interaction where users A and B give each other three photo markings seems more evolved than a relationship in which there is one marking in each direction, even though each of these interactions is assigned the same net score.

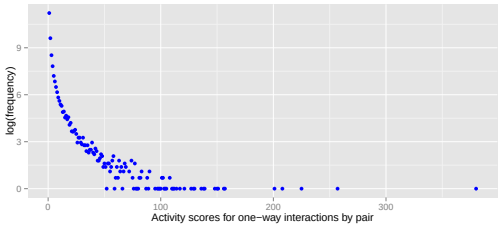
We define magnitude as the minimum number of interactions in either direction. For example,



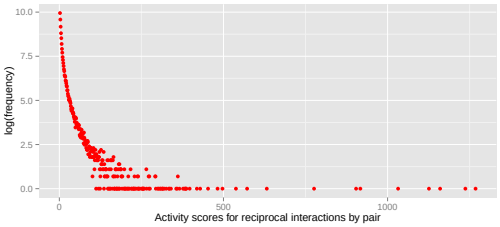
(a) Distribution of net scores by user for one-way interactions.



(b) Distribution of net scores by user for reciprocal interactions.



(c) Distribution of activity scores by pair for one-way interactions.



(d) Distribution of activity scores by pair for reciprocal interactions.

Figure 6.4: Distribution of activity and net scores for one-way vs. reciprocal interactions for the Facebook dataset.

an interaction with three likes from user A to B and three likes from user B to A would be of magnitude three. Likewise an interaction with three likes from user A to B and five likes from user B to A is of magnitude three. As we increase the magnitude threshold, the number of users decreases. We first plotted distributions for magnitudes of three, five and ten, but they were similar to the plots we had for the magnitude of one which was discussed earlier. On the other hand, as discussed in the following paragraphs, we found out that combining the magnitude of reciprocity of interactions together with localized vs. persistent interactions brings new insights.

Next we delved into the differences between localized and persistent interactions: in order to study persistence of interactions, we first identified a notion of a *week* in each dataset We can then call relationships that occur entirely within the timeframe of a week localized and the ones that go beyond a week persistent.

We then performed these calculations on our data for one-way and reciprocal interactions as well as for interactions of magnitudes three, five and ten. These results for both datasets are shown in Tables 6.14, 6.15, 6.16 and 6.17.

As can be seen, for the Flickr dataset, there is a noticeable shift from localized to persistent interactions as the relationships increase in reciprocity. Interactions that are even at all reciprocal have a much greater chance of being persistent: 62% vs. 7%. These longer lasting interactions may be more meaningful than the short-lived localized ones. And, as can be seen, even a slight increase in magnitude drastically increases the likelihood of persistent relationships. At magnitude three, 94% of relationships are persistent and by magnitude 10, practically 100% are. The results for the Facebook dataset follow a similar trend, however, the percentage of reciprocal interactions is higher.

Type of pairs	one-way pairs	reciprocal pairs
# Persistent pairs	291299	184417
% Localized pairs	92.68%	38.02%
% Persistent pairs	7.32%	61.98%

Table 6.14: Localized vs. persistent data for one-way and reciprocal interactions for Flickr dataset.

Magnitude	3	5	10
# Total pairs	58416	28423	10266
% Localized pairs	5.82%	1.74%	0.35%
% Persistent pairs	94.18%	98.26%	99.65%

Table 6.15: Localized vs. persistent data for magnitude 3/5/10 interactions for Flickr dataset.

The results are generally similar to the that of the Flickr dataset with a difference that even one-way interactions were more persistent in this dataset.

Overall we observe very similar results for the Facebook dataset compared to the results for the Flickr dataset, and the percentage of persistent ones goes up a lot as we increase the magnitude cutoff.

Type of pairs	one-way pairs	reciprocal pairs
# Total pairs	102820	80592
% Localized pairs	76.89%	37.35%
% Persistent pairs	23.10%	62.64%

Table 6.16: Localized vs. persistent data for one-way and reciprocal interactions for Facebook dataset.

Magnitude	3	5	10
# Total pairs	23703	12225	4581
% Localized pairs	7.52%	2.86%	0.63%
% Persistent pairs	92.4%	97.13%	99.36%

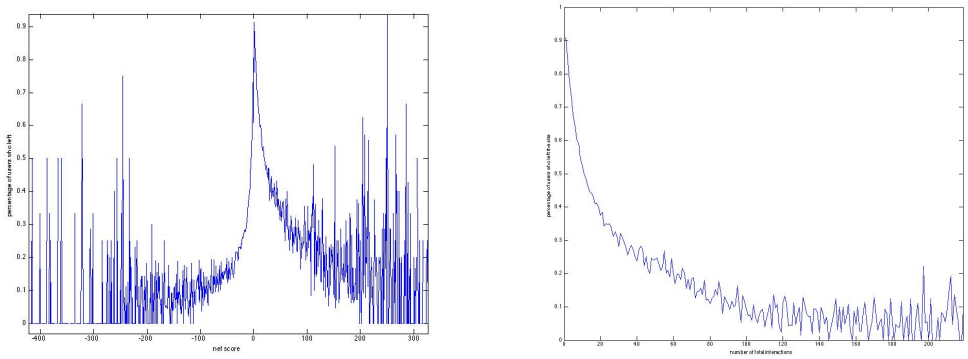
Table 6.17: Localized vs. persistent data for magnitude 3/5/10 interactions for Facebook dataset.

6.3.5 Interactions Over Time and User Retention

In this part, in order to have a better understanding of the dynamics of prosocial giving, we analyzed social interactions and their correlation to user retention over time. More precisely, we studied the correlation between users' retention and their scores, reciprocity, persistence, and magnitude of the interactions over time. To this end we used a sliding timeframe to determine user retention levels.

In order to calculate user retention rates, we used training and testing sliding windows to measure user activity. In the training window which accounted for the first $\frac{2}{3}$ fraction of the total timeframe, we calculated net score for each user. In the testing part, which accounted for the last $\frac{1}{3}$ fraction of the total time, we checked for outgoing activity from that user. A user who had no activity during this test window was considered to have left the social network. Figure 6.5a shows the percentage of users who left the site for a distribution of net scores clustered around zero.

Interestingly, what can be seen from this distribution is that a greater percentage of users who are net givers remain active on the social network compared to those who are net receivers. In fact, the users most likely to leave are those who receive slightly more than they give. In each



(a) Percentage of users who left for a given net score.

(b) Percentage of users who left for a given activity score.

Figure 6.5: Percentage of users leaving for activity and net scores.

case, users with high activity scores are less likely to leave compared to those with scores clustered around zero.

To confirm the above observation, we also analyze the leaving patterns for users based on their total activity scores shown in Figure 6.5a. We again see that users who are more active are much less likely to leave the site. This may indicate that, overall, user activity level is a better predictor of how users behave than net scores.

6.3.6 Insights from Clustering

In this section, we study the dynamics of interactions between each pair of users in further details by clustering all interaction pairs into a small number of interaction types. For this purpose, in this section we only consider persistent interactions, since those were the ones from which we could derive meaningful insights.

6.3.6.1 Clustering Methodology

We first compute cumulative net scores for each interaction pair by taking into account the ”sequence” associated with ups and downs of two users’ interactions, i.e., we construct a sequence of numbers for each interaction as follows: the sequence starts out at 0, adding a value of one for each interaction from user A to user B and subtracting a value of one for each interaction from user B

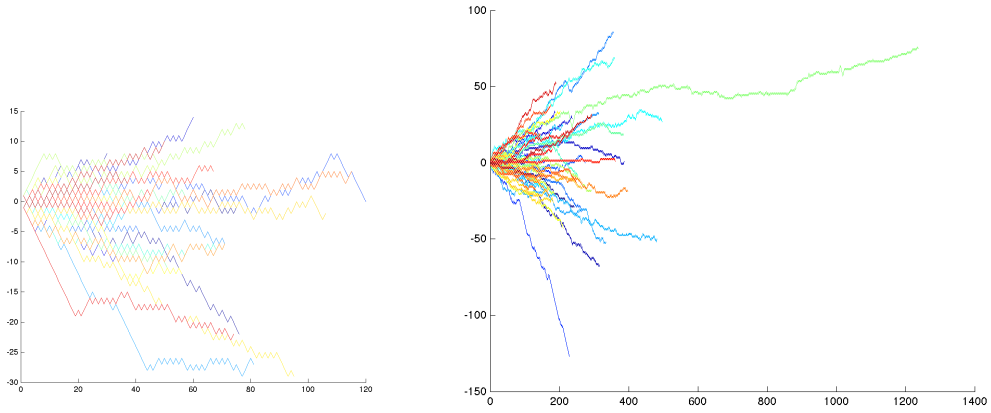
to user A (similar to computing net scores for pairs, presented in Section 6.3.2).

We then use these variable-length sequences and analyze them to extract certain features for each interaction pair. Using these features, we perform k-means clustering using the R SciPy package on the dataset in order to discern different types of interaction pairs. This package uses Euclidean distance as its distance function.

Feature set for clustering. We performed k-means clusterings on the dataset using different sets of features. We then revised our feature set based on the quality of the output of the clustering. We measured the quality of the clustering by observing the inter-cluster and intra-cluster distances. After trying a variety of subsets of features, we decided to focus no the following six categories of features:

1. The normalized number of sign changes, i.e., the number of times that the sign of the prefix net score changes in the sequence of net scores computed for this interaction.
2. The normalized difference between the maximum and minimum points on the sequence.
3. The normalized slope of the overall sequence.
4. The number of sign changes in the first half of the sequence.
5. The number of sign changes in the second half of the sequence.
6. The length of the sequence.

Choosing the number of clusters k. While running the k-means algorithm we had to determine the desirable number k. We followed the following methodology to choose an appropriate k: Let the inter-cluster distance for a clustering be the average distance between centroids and the intra-cluster distance for the same clustering be the average distance from points in a given cluster to their centroid. We computed intra- and inter-cluster distances for each clustering to glean better insight into which number of clustering was the best. More specifically, we chose a number k at which the intra-cluster distance is low and the inter-cluster distance is high. Based on our analysis for both data sets, the clusters seemed to be the furthest apart at $k = 6$ and the points were closest to their centers at $k = 6$. Therefore we chose the solution corresponding to $k = 6$ for both datasets.



(a) Cluster of long reciprocal interactions. (b) Cluster of very long reciprocal interactions.

Figure 6.6: Two of the clusters for the Facebook dataset.

6.3.6.2 Clustering Results

We analyzed each cluster in the output by visualizing the cluster and found out a description for the cluster based on the value of the features for interactions in that cluster. Here we describe these clustering outputs and insights from them.

For the Flickr dataset, we cluster a total number of $475,716$ persistent social interactions, and came up with the following six clusters:

- 1) a cluster of length-2 one-way interactions with 23.6% of interactions,
- 2) a cluster of short one-way interactions of length 3 to 5 with 38.5% of interactions,
- 3) a cluster of short reciprocal interactions of length 2 to 5 with 20.3% of interactions,
- 4) a cluster of medium-size one-way interactions of length 5 to 20 with 11.3% of interactions,
- 5) a cluster of medium-size reciprocal interactions of length 5 to 20 with 4.9% of interactions, and
- 6) a cluster of long reciprocal interactions of length 20 and above with 1.1% of interactions.

One can easily observe the following from the above: We first note that even inside the persistent interactions, the overall number of one-way interactions is much larger than the number of

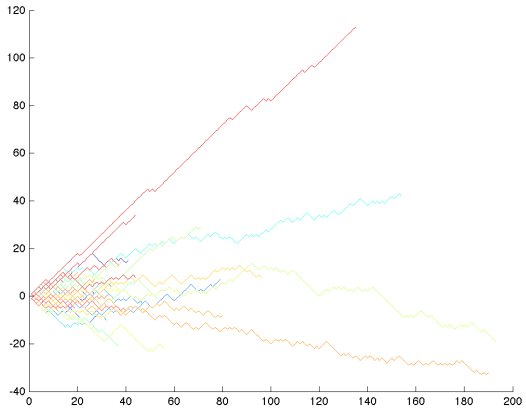


Figure 6.7: A Flickr cluster of long reciprocal interactions.

reciprocal ones and therefore most clusters with one-way interactions are much larger. However, despite having many more overall one-way interactions, the algorithm did not find a cluster for long one-way interactions indicating that the long persistent interactions are almost all reciprocal. There exists a cluster corresponding to the long reciprocal interactions. A random sample of 20 sequences for these two clusters are depicted in Figures 6.7 and 6.8.

For the Facebook dataset, we cluster a total number of $74,241$ persistent social interactions, and came up with the following six clusters:

- 1) a cluster of short one-way interactions of length 2 to 5 with 17.2% of interactions,
- 2) a cluster of short reciprocal interactions of length 2 to 5 with 14.3% of interactions,
- 3) a cluster of medium-size one-way interactions of length 5 to 20 with 14.7% of interactions,
- 4) a cluster of medium reciprocal interactions of length 5 to 20 with 37.7% of interactions,
- 5) a cluster of long reciprocal interactions of length 20 to 200 with 15.3% of interactions, and
- 6) a cluster of very long reciprocal interactions of length 200 to 1400 with 0.6% of interactions.

The results for the Facebook dataset differ from the Flickr dataset in a couple of ways but the general trend of the results remain similar. First of all, the percentage of persistent interactions

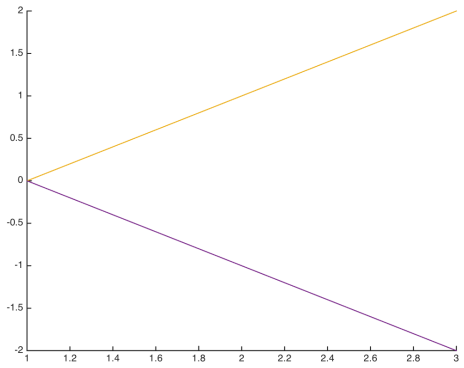


Figure 6.8: A Flickr cluster of one-way interactions.

is higher overall. More notably, the percentage of reciprocal interactions is much higher in this dataset compared to the Flickr dataset. Moreover, we get very long reciprocal interactions and as a result, the clustering puts them into two different clusters. The general trend for the length of one-way vs. reciprocal interactions remains the same, i.e., the percentage of long reciprocal interactions is much higher than that of one-way interactions. The clustering algorithm does not find a specific cluster for long one-way clusters and instead finds two clusters of long and very long reciprocal interactions. A random sample of 20 sequences for these two clusters are depicted in Figures 6.6a and 6.6b.

After choosing the best clustering for each dataset, we also examined the percentage of users for which all their interactions with other users were clustered into one cluster only. Interestingly, we observe that more than 50% of the time, users would be clustered into the same cluster.

6.3.6.3 Prediction

Another way to study the correlation and impact of different parameters on the persistence of an interaction is to define a prediction task of estimating the length of an interaction as a function of a set of features associated with that interaction. Here, we aim to predict the length of pairwise interactions using the same set of features that we used for clustering:

1. The normalized number of sign changes, i.e., the number of times that the sign of the prefix net score changes in the sequence of net scores computed for this interaction.

2. The normalized difference between the maximum and minimum points on the sequence.

3. The normalized slope of the overall sequence.

4. The number of sign changes in the first half of the sequence (SC1)

5. The number of sign changes in the second half of the sequence (SC2)

6. The length of the sequence.

We use linear regression and multivariate adaptive regression splines (MARS) [46] to study the impact of each of the above variables in predicting the length of an interaction. We first present the result for the Flickr dataset. Linear regression and logistic regression were not good fits on this dataset, so we opted to explore other methods. In particular, using multivariate adaptive regression splines resulted in a good fit on this dataset. For this purpose, we applied package earth [84]. We utilized a 10-fold cross validation on this data, and report the mean of out-of-fold R-Squareds as the final cross validation R-Squared value (cvRs). cvRs for our model is 0.90. This model also reports how important each predictor variable is in predicting the dependent variable. For this dataset, MARS determined that SC1 and SC2 were not important, and as a result, they were not used. More notably, MARS determined that the overall number of sign changes, absolute difference, and the slope were the most important predictors, respectively. We observe a similar behavior for the Facebook dataset with cvRs = 0.95. For this dataset, however, even linear regression models results in a good fit, and we get similar insights regarding the significance of different features in performing this prediction task. In particular, the fact that the number of sign changes is the most important factor in predicting the length of a pairwise interaction, implies that there is a strong positive correlation between length and reciprocity (which validates what we had seen in previous sections).

6.3.7 Friends vs. Non-friends

One implication of our studies presented in this work is that there is a positive correlation between persistent interactions and being more reciprocal. An interesting and related question is to study the level of altruism vs. reciprocity based on other characteristics of pairwise interactions among

users. In particular, it is interesting to study the impact of friendship among pairs of users on the reciprocity of their interactions.

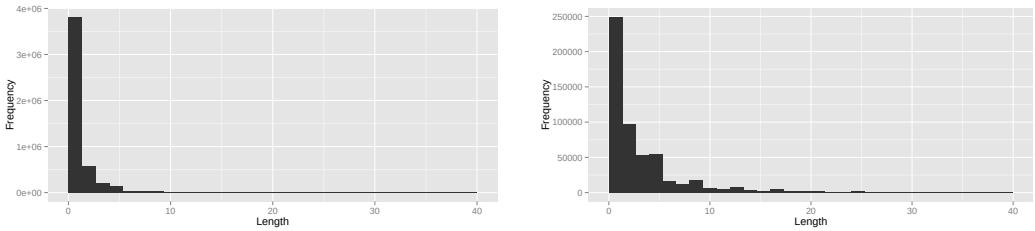
To achieve this goal, we performed another data analysis on the Flickr dataset. On Flickr, users can declare a friendship relationship with other users on the website. Our goal is to study the impact of friendship on the altruism and reciprocity of interactions. We first note that only 0.8% of the pairwise interactions correspond to pairs of users who had declared their friendship from the beginning of the crawl, and 1.4% of all pairwise interactions correspond to pairs of users who became friends during the crawl. We first note that the average length of these two categories of friend relations are 4.6 and 3.7 respectively. These are much larger than the average length of any pairwise interaction on the whole Flickr dataset which is around 2.1. On the other hand, the average length of non-friend interactions is around 1.6. Next, we compute the distribution of absolute net score for each category of interactions and compare the average and standard deviation of these distributions for each type of interactions. The distribution of length and net score of friends and non-friends are shown in Figures 6.9 and 6.10. The average and standard deviations of the net score distributions are summarized in Table 6.18. Interestingly, we find out that even though the interactions between friends are longer (i.e., more persistent), both average and standard deviation of the absolute value of the net score between these pairs are larger. In other words, while interactions between friends are more persistent, the level of altruism (captured in the net score) is higher for these friends. At a first glance, this is contradicting the observed positive correlation between the length of the interactions and their reciprocity however it can be interpreted as follows: For friendship-based relationships the underlying reason for an interaction may be beyond what can be seen on a social media site. On the other hand for non-friend interactions, the response rate and reciprocity may play a more significant role in continuing the relationship.

6.3.8 Related Work

There has been work done on the motivations of participation in online user generated communities – cases where a user contributes to a group without obvious/immediate rewards. For instance, Lampe et al [73] have turned to the theory of "uses and gratifications and organizational commitment" to explain why users create content online. They find that users may continue generating content in the site for reasons that may not be the same as those that led them to the site in the first

Data Set	Avg. Length	Avg. Net Score	Std. Dev.
All	2.1	1.84	3.89
Friend	4.6	3.17	7.45
Friend during	3.7	2.74	5.30
Non-friend	1.6	1.52	2.60

Table 6.18: Average Length and Net Score for friend versus non-friend interactions.



(a) Length of interaction between users who were never friends.

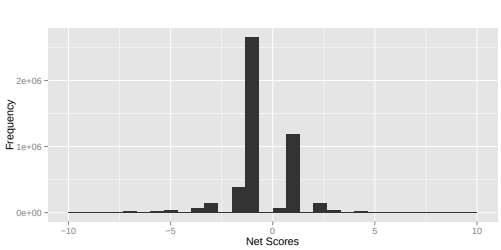
(b) Length of interaction between users who were friends from the beginning.

Figure 6.9: Distribution of lengths of friends and non-friends.

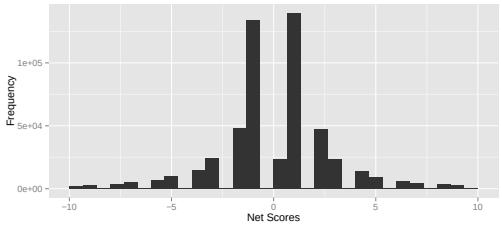
place. They also find that a ”sense of belonging” is an important determinant of user participation.

In other work on question-answering websites (Mathoverflow and Yahoo Answers), users reported getting information, giving information, reputation building, relationship development, recreation, and self discovery as their main motivation to answer questions. However, motivations vary based on the community type [106]. In this work, they found that users may continue participating in a site for reasons other than those that drew them to the site in the first place. Additionally, a sense of belonging to the site was important to all types of users across all types of users. Our data analysis shows that the predictors of contribution among these users seemed to not be associated with how easy the site was to use for them, but may instead have social or cognitive factors.

Social scientists and economists have explained prosocial giving in two main ways: altruism and expecting future reciprocal exchange [28, 83, 45]. For example, in one study of underlying motivations of prosocial giving, researchers ran a dictator game experiment among college students asking them to allocate a sum of money among their peers, friends or strangers one time.



(a) Net scores of users who were never friends.



(b) Net scores of users who were friends from the beginning.

Figure 6.10: Distribution of net scores of friends and non-friends.

Controlling for social exchange, they found that a significantly higher proportion of the money was allocated by the students to their friends compared to the amount allocated to strangers. In a recent paper [104], the authors study other aspects of social exchanges such as power and status in online networks, however they do not study reciprocity or related social giving behavior in online social media. Finally, we note a related work [114] in which the authors study the presence of a notion of reciprocity known as Guanxi in a Chinese forum (Mitbbs). However, the context of interactions in this Mitbbs is based on predefined virtual points (credits). As a result, the social giving behaviors studied in this paper are not the subject of study in [114]. Moreover, while Guanxi is related to reciprocity, it captures a different aspect of reciprocity motivated by a sense of obligation or indebtedness in human relations [115]. While other similar research has been performed in similar settings, to the best of our knowledge, our work is the first that studies the distinction between altruism and reciprocity as social givings in online social sites.

6.3.9 Concluding Remarks

In this work, we present a study of users' social interactions in two social media sites: Flickr and Facebook. Our goal is to determine the level of reciprocity vs. altruism in all pairwise social interactions among users. Our study shows that there is a positive correlation between persistence and reciprocity . Finally, we study the difference between altruism and reciprocity between friends and non-friends and observe that while interactions between friends are more persistent, the level of altruism is higher in friendship-based interactions. This may seem surprising, however, this phenomenon can be explained as follows: while non-friend interactions are motivated by the prospect

of future reciprocal favorite markings, the underlying reasons for user interactions between friends is beyond their limited interactions on the site therefore, more altruism appear in these interactions.

We note that our study does not imply causality between reciprocity and persistence of social interactions (we show a positive correlation exists between them). As a future research direction, it would be interesting to run real-world experiments on a social network in order to study causality of reciprocal behavior on the persistence of social interactions. Finally we note that, from the practical point of view, our study offers insights in improving user engagement in online social platforms. For example, the positive correlation between reciprocity and persistence of social interactions can be leveraged in newsfeed ranking algorithms to encourage users to increase levels of reciprocity in their interactions. Moreover, such analysis could also be used for social network simulations by developing better user behavior models in online social sites.

Bibliography

[1] Darpa networking challenge. *https://networkchallenge.darpa.mil/Default.aspx*.

[2] Darpa network challenge project report. *http://www.eecs.harvard.edu/cs286r/papers/ProjectReport.pdf*, 2010.

[3] *sec.gov/Archives/edgar/data/1326801/000132680114000007/fb-12312013x10k.htm*, 2013.

[4] IABinternet advertising revenue report. *http://www.iab.net/media/file/IABInternetAdvertisingRevenueReportHY2013FINALdoc.pdf*, 2013.

[5] Facebook doubleclick for publishers (dfp) optimization website. *https://www.facebook.com/business/a/online-sales/ad-optimization-measurement*, 2014.

[6] Google doubleclick bid manager website. *http://www.thinkwithgoogle.com/products/doubleclick-bid-manager.html*, 2014.

[7] Google doubleclick for publishers (dfp) optimization website. *http://static.googleusercontent.com/media/www.google.com/en/us/doubleclick/pdfs/optimization.pdf*, 2014.

[8] Zeinab Abbassi, Christina Aperjis, and Bernardo A Huberman. Swayed by friends or by the crowd? In *Social informatics*, pages 365–378. Springer, 2012.

[9] Zeinab Abbassi, Sepehr Assadi, and Mina Tahmasbi. Predicting ratings in online social networks: Friends or the crowd? *WIN Workshop 2015*.

[10] Zeinab Abbassi, Aditya Bhaskara, and Vishal Misra. Optimizing display advertising in online social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1–11. International World Wide Web Conferences Steering Committee, 2015.

[11] Zeinab Abbassi, Nima Haghpanah, and Vahab Mirrokni. Exchange market mechanisms without money.

[12] Zeinab Abbassi, Nidhi Hegde, and Laurent Massoulié. Distributed content curation on the web. *ACM Transactions on Internet Technology (TOIT)*, 14(2-3):9, 2014.

[13] Zeinab Abbassi and Laks V. S. Lakshmanan. On efficient recommendations for online exchange markets. In *ICDE*, pages 712–723, 2009.

[14] Zeinab Abbassi and Vishal Misra. Multi-level revenue sharing for viral marketing. *Proceedings of ACM NetEcon*, 2011.

[15] David J. Abraham, Avrim Blum, and Tuomas Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *ACM Conference on Electronic Commerce*, pages 295–304, June 13-16 2007.

[16] Hessameddin Akhlaghpour, Mohammad Ghodsi, Nima Haghpanah, Vahab S Mirrokni, Hamid Mahini, and Afshin Nikzad. Optimal iterative pricing over social networks. In *Internet and Network Economics*, pages 415–423. Springer, 2010.

[17] N.H. Anderson. Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, 70(4):394, 1965.

[18] N.H. Anderson. *Foundations of information integration theory*. Academic Press New York, 1981.

[19] I. Ashlagi and A. Roth. Individual rationality and participation in large scale, multi-hospital kidney exchange. In *ACM Conference on Electronic Commerce*, pages 321–322, 2011.

[20] Itai Ashlagi, Mark Braverman, and Avinatan Hassidim. Matching with couples revisited. In *ACM Conference on Electronic Commerce*, pages 335–336, 2011.

[21] Itai Ashlagi, Felix A. Fischer, Ian A. Kash, and Ariel D. Procaccia. Mix and match: A strategyproof mechanism for multi-hospital kidney exchange. In *ACM Conference on Electronic Commerce*, pages 305–314, 2010.

[22] Baruch Awerbuch, Yossi Azar, Amir Epstein, Vahab S. Mirrokni, and Alexander Skopalik. Fast convergence to nearly optimal solutions in potential games. In *ACM Conference on Electronic Commerce*, pages 264–273, 2008.

[23] A.J. Bahns, K.M. Pickett, and C.S. Crandall. Social ecology of similarity: Big schools, small schools and social relationshipsp. *Group Processes & Intergroup Relations*, 1:13, 2011.

[24] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. Social influence in social advertising: evidence from field experiments. In *ACM Conference on Electronic Commerce*, pages 146–161, 2012.

[25] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[26] Hongji Bao and Edward Y Chang. Adheat: an influence-based diffusion model for propagating hints to match ads. In *Proceedings of the 19th international conference on World wide web*, pages 71–80. ACM, 2010.

[27] R.F. Baumeister, E. Bratslavsky, C. Finkenauer, and K.D. Vohs. Bad is stronger than good. *Review of general psychology*, 5(4):323, 2001.

[28] Roland Bénabou and Jean Tirole. Incentives and prosocial behavior. Technical report, National Bureau of Economic Research, 2005.

[29] Smriti Bhagat, Amit Goyal, and Laks VS Lakshmanan. Maximizing product adoption in social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 603–612. ACM, 2012.

[30] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: An o($n^{1/4}$) approximation for densest k-subgraph. In *Pro-*

ceedings of the Forty-second ACM Symposium on Theory of Computing, STOC '10, pages 201–210, New York, NY, USA, 2010. ACM.

[31] Anna Bogomolnaia and Herve Moulin. Random matching under dichotomous preferences. *Econometrica,*, 72:257–279, 2004.

[32] C.M.K. Cheung and M.K.O. Lee. Online consumer reviews: Does negative electronic word-of-mouth hurt more? *AMCIS 2008 Proceedings*, page 143, 2008.

[33] Yang Cheung, Dah-Ming Chiu, and Jianwei Huang. Can bilateral isp peering lead to network-wide cooperative settlement. In *Computer Communications and Networks, 2008. ICCCN'08. Proceedings of 17th International Conference on*, pages 1–6. IEEE, 2008.

[34] Peter Clifford and Aidan Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[35] Charles J Colbourn. *The combinatorics of network reliability*. Oxford University Press, Inc., 1987.

[36] M. Peralta D. L. Hoffman, Th. P. Novak. Building consumer trust online. *Communications of the ACM*, 42(4):80–85, 1999.

[37] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[38] Shaddin Dughmi and Arpita Ghosh. Truthful assignment without money. In *ACM Conference on Electronic Commerce*, pages 325–334, 2010.

[39] Eyal Even-Dar and Asaf Shapira. A note on maximizing the spread of influence in social networks. *Internet and Network Economics*, pages 281–286, 2007.

[40] Uriel Feige, László Lovász, and Prasad Tetali. Approximating min sum set cover. *Algorithmica*, 40(4):219–234, 2004.

[41] Joan Feigenbaum, Arvind Krishnamurthy, Rahul Sami, and Scott Shenker. Hardness results for multicast cost sharing. *Theoretical Computer Science*, 304(1):215–236, 2003.

[42] J. Feldman, M. Henzinger, N. Korula, V. Mirrokni, and C. Stein. Online stochastic packing applied to display ad allocation. In *ESA*, 2010.

[43] J. Feldman, N. Korula, V. Mirrokni, S. Muthukrishnan, and M. Pal. Online ad assignment with free disposal. In *WINE*, 2009.

[44] Lisa Fleischer, Michel X. Goemans, Vahab S. Mirrokni, and Maxim Sviridenko. Tight approximation algorithms for maximum separable assignment problems. *Mathematics of Operations Research*, 36(3):416–431, 2011.

[45] Bruno S Frey and Stephan Meier. Pro-social behavior in a natural setting. *Journal of Economic Behavior & Organization*, 54(1):65–88, 2004.

[46] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

[47] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *Americal Mathematical Monthly*, 69:9–15, 1962.

[48] Arpita Ghosh and Preston McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 137–146, New York, NY, USA, 2011. ACM.

[49] E.L. Glaeser, B. Sacerdote, and J.A. Scheinkman. *Crime and social interactions*. National Bureau of Economic Research Cambridge, Mass., USA, 1995.

[50] Michel X. Goemans, Erran L. Li, Vahab S. Mirrokni, and Marina Thottan. Market sharing games applied to content distribution in ad-hoc networks. In *MobiHoc*, pages 55–66, 2004.

[51] Andrew V. Goldberg and Robert Endre Tarjan. Finding minimum-cost circulations by canceling negative cycles. *J. ACM*, 36(4):873–886, 1989.

[52] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 5(1):73–84, 2011.

[53] Mangesh Gupte, MohammadTaghi Hajiaghayi, Lu Han, Liviu Iftode, Pravin Shankar, and Raluca Ursu. News posting by strategic users in a social network. *Internet and Network Economics*, pages 632–639, 2009.

[54] Y.Y. Hao, Q. Ye, Y.J. Li, and Z. Cheng. How does the valence of online consumer reviews matter in consumer decision making? differences between search goods and experience goods. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[55] J.W. Hardin, J.M. Hilbe, and J. Hilbe. *Generalized linear models and extensions*. Stata Corp, 2007.

[56] Jason Hartline, Vahab Mirrokni, and Mukund Sundararajan. Optimal marketing strategies over social networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 189–198. ACM, 2008.

[57] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[58] Shawndra Hill, Foster Provost, and Chris Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

[59] D. Hochbaum and E. Olinick. The bounded cycle-cover problem. *INFORMS Journal on Computing*, 13(2):104–109, 2001.

[60] Richard A Holley and Thomas M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.

[61] Holyer. The NP-completeness of some edge partitioning problems. *SIAM journal of Computing*, 10(3):713–717, 1981.

[62] N. Immorlica and M. Mahdian. Marriage, honesty, and stability. In *Symposium on Discrete Algorithms (SODA)*, 2005.

[63] N. Immorlica, V. S. Mirrokni, and M. Mahdian. Cycle cover with short cycles. In *STACS*, pages 641–653, 2005.

[64] Interactive-Advertising-Bureau:. social advertising best practices. *http://www.iab.net/media/file/Social-Advertising-Best-Practices-0509.pdf*, 2009.

[65] Kamal Jain and Vijay Vazirani. Applications of approximation algorithms to cooperative games. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 364–372. ACM, 2001.

[66] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems (RecSys '10)*, pages 135–142. ACM Press, 2010.

[67] Mohsen Jamali, Gholamreza Haffari, and Martin Ester. Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns. In *WWW*, pages 527–536, 2011.

[68] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[69] Richard M Karp and Michael Luby. Monte-carlo algorithms for the planar multiterminal network reliability problem. *Journal of Complexity*, 1(1):45–64, 1985.

[70] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[71] Jon Kleinberg, Christos H Papadimitriou, and Prabhakar Raghavan. On the value of private information. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, pages 249–257. Morgan Kaufmann Publishers Inc., 2001.

[72] Le T Lam X, Vu T and Duong A. Addressing cold-start problem in recommendation systems. In *ICUIMC*, pages 208–211, 2008.

[73] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. Motivations to participate in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1927–1936. ACM, 2010.

[74] P Lazarsfeld and R Merton. Friendship as a social process: a substantive and methodological analysis. *Freedom and Control in Modern Society, ed. M Berger*, pages 18–66, 1954.

[75] Stephen Leider, Markus M Möbius, Tanya Rosenblat, and Quoc-Anh Do. Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics*, 124(4):1815–1851, 2009.

[76] P. Lu, X. Sun, Y. Wang, and Z.A. Zhu. Asymptotically optimal strategy-proof mechanisms for two-facility games. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 315–324. ACM, 2010.

[77] P. Lu, Y. Wang, and Y. Zhou. Tighter bounds for facility games. *Internet and Network Economics*, pages 137–148, 2009.

[78] Richard TB Ma, Dah Ming Chiu, John Lui, Vishal Misra, and Dan Rubenstein. Internet economics: The use of shapley value for isp settlement. *IEEE/ACM Transactions on Networking (TON)*, 18(3):775–787, 2010.

[79] Richard TB Ma, Dah Ming Chiu, John Lui, Vishal Misra, and Dan Rubenstein. On cooperative settlement between content, transit, and eyeball internet service providers. *Networking, IEEE/ACM Transactions on*, 19(3):802–815, 2011.

[80] W.A. Mason, F.R. Conrey, and E.R. Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, 11(3):279–300, 2007.

[81] Avner May, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Game in the newsroom: Greedy bloggers for picky audience. In *Workshop on Social Computing and User-Generated Content, 2013*.

[82] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[83] Stephan Meier. A survey of economic theories and field evidence on pro-social behavior. 2006.

[84] S. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. *earth: Multivariate Adaptive Regression Splines*, 2011. R package.

[85] Vahab S. Mirrokni, Sebastien Roch, and Mukund Sundararajan. On fixed-price marketing for goods with positive network externalities. In *WINE*, pages 532–538, 2012.

[86] Vishal Misra, Stratis Ioannidis, Augustin Chaintreau, and Laurent Massoulié. Incentivizing peer-assisted services: a fluid shapley value approach. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 215–226. ACM, 2010.

[87] Elchanan Mossel and Sébastien Roch. Submodularity of influence in social networks: From local to global. *SIAM J. Comput.*, 39(6):2176–2188, 2010.

[88] A. Paez, D.M. Scott, and E. Volz. A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design*, 35(6):1055–1069, 2008.

[89] G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European review of social psychology*, 1:33–60, 1990.

[90] Galen Pickard, Iyad Rahwan, Wei Pan, Manuel Cebrián, Riley Crane, Anmol Madan, and Alex Pentland. Time critical social mobilization: The darpa network challenge winning strategy. *arXiv preprint arXiv:1008.3172*, 2010.

[91] Ariel D. Procaccia and Moshe Tennenholtz. Approximate mechanism design without money. In *ACM Conference on Electronic Commerce*, pages 177–186, 2009.

[92] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *The Semantic Web-ISWC 2003*, pages 351–368. Springer, 2003.

[93] Robert W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

[94] A. E. Roth and M. Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis*. Cambridge University Press, 1991.

[95] A.E. Roth. The national resident matching program as a labor market. *JAMA. Journal of the American Medical Association,*, 275:1054–1056, 1996.

[96] Alvin E Roth. Repugnance as a constraint on markets. Technical report, National Bureau of Economic Research, 2006.

[97] Alvin E. Roth, Tayfun Snmez, and Utku Unver. Kidney exchange. *Quarterly Journal of Economics*, 19, 2004.

[98] Alvin E. Roth, Tayfun Snmez, and Utku Unver. Pairwise kidney exchange. *Journal of Economic Theory*, 125:151–188, 2005.

[99] Tim Roughgarden. Intrinsic robustness of the price of anarchy. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 513–522. ACM, 2009.

[100] H. Zhang S. Ba, A. B. Whinston. Building trust in online auction markets through an economic incentive mechanism. *Decision Support Systems*, 35(3):273–286, 2003.

[101] Lloyd S Shapley. A value for n-person games. Technical report, 1952.

[102] Brent Simpson and Robb Willer. Altruism and indirect reciprocity: The interaction of person and situation in prosocial behavior. *Social Psychology Quarterly*, 71(1):37–52, 2008.

[103] A.T. Sorensen. Social learning and health plan choice. *The RAND Journal of Economics*, 37(4):929–945, 2006.

[104] Bogdan State, Bruno Abrahao, and Karen Cook. From power to status in online exchange. In *Proceedings of the 4th ACM International Conference on Web Science (WebSci 2012), Evanston, IL, USA, 2012*. ACM, 2012.

[105] Tayfun Snmez and Utku Unver. House allocation with existing tenants: An equivalence. *Games and Economic Behavior*, 52:153–185, 2004.

[106] Yla R Tausczik and James W Pennebaker. Participation in an online mathematics community: differentiating motivations to add. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 207–216. ACM, 2012.

[107] S.E. Taylor. Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, 110(1):67, 1991.

[108] Catherine Tucker. Social advertising. *Available at SSRN: http://ssrn.com/abstract=1975897 or http://dx.doi.org/10.2139/ssrn.1975897*, 2012.

[109] Erik Vee, Sergei Vassilvitskii, and Jayavel Shanmugasundaram. Optimal online assignment with forecasts. In *ACM EC*, 2010.

[110] Adrian Vetta. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *FOCS '02*, page 416. IEEE Computer Society, 2002.

[111] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[112] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *EuroSys*, pages 205–218, 2009.

[113] Philip F Yabrudi and Lutfy N Diab. The effects of attitude similarity-dissimilarity, religion, and topic importance on interpersonal attraction among lebanese university students. *The Journal of Social Psychology*, 106(2):167–171, 1978.

[114] Jiang Yang, Mark S Ackerman, and Lada A Adamic. Virtual gifts and guanxi: supporting social exchange in a chinese online community. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 45–54. ACM, 2011.

[115] Mayfair Mei-hui Yang. *Gifts, favors, and banquets: The art of social relationships in China*. Cornell University Press, 1994.