Project Recommendation Using Heterogeneous Traits in Crowdfunding

Vineeth Rakesh
Dept. of Computer Engr
Wayne State University
vineethrakesh@wayne.edu

Jaegul Choo
Dept. of Computer Sci. and Engr.
Korea University
jchoo@korea.ac.kr

Chandan K. Reddy
Dept. of Computer Science
Wayne State University
reddy@cs.wayne.edu

Abstract

Crowdfunding has gained widespread popularity in recent years. By funding entrepreneurs with creative minds, it is gradually taking over the role of venture capitalists who provide the much needed seed capital to jump start business ventures. Despite the huge success of the crowdfunding platforms, not every project is successful in reaching its funding goal. Therefore, in this paper, we intend to answer the following question "what set of features determine a project's success?". We begin by studying the dynamics of Kickstarter, a popular reward-based crowdfunding platform, and the impact of social networks on this platform. Contrary to previous studies, our analysis is not restricted to project-based features alone; instead, we expand the features into four different categories: temporal traits, personal traits, geo-location traits, and network traits. Using a comprehensive dataset of 18K projects and 116K tweets, we provide several unique insights about these features and their effects on the success of Kickstarter projects. Based on these insights, we build a supervised learning framework to learn a model that can recommend a set of investors to Kickstarter projects. By utilizing features from the first three days of project duration alone, we show that our results are significantly better than the previous studies.

1 Introduction

For several years, entrepreneurs had to seek the help of banks, brokers, and other financial intermediaries to acquire the necessary funds for starting a business venture. Such financial constraints were a huge bottleneck to people with innovative ideas. However, this scenario has changed drastically with the emergence of *crowdfunding* platforms. Thanks to the widespread use of internet, entrepreneurs can effectively post their ideas on crowdfunding websites and gain the attention of people all over the world. The concept of crowdfunding is analogous to micro-financing or crowdsourcing (Morduch 1999), where the seed capital is collected by soliciting funds from a large group of people, rather than a single individual (venture capitalist).

Crowdfunding can be characterized into four different types: equity-based, lending-based, reward-based, and donation-based. In equity-based crowdfunding, the investors receive some form of stake from the company. The donation-based is similar to a charitable venture, while in lending-based, the investors are repaid for their investment. Finally, the most popular form of crowdfunding is the reward-based, where users receive some form of gift in return of their investment. Kickstarter, one of the popular crowdfunding sites, mainly adopts this reward-based crowdfunding mechanism while raising over 480 million dollars in pledged amount and 19,911 successfully funded projects in 2013. This domain follows the "all or nothing" policy, which means that the pledged money is collected only if the goal amount is reached; if not, the entire money is returned back to the investors. Kickstarter terms the investors as *backers*, and the founders of projects as *creators*. The creators post their ideas by providing a detailed description of their project which includes the scope of the project, video description, reward details, topical categories, location, updates, FAQs, etc. The backers then invest in the project based on its quality and their personal interests.

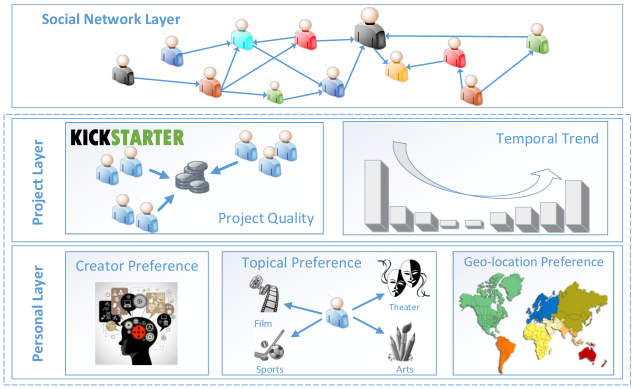


Figure 1: Impact of project-, personal-, and social network-based features on Kickstarter users.

Despite being a valuable platform for crowdfunding ventures, statistics show that only 43% of the projects succeed in reaching their pledged goal (KickStarterStats). Additionally, the margin by which successful projects exceed their pledged goal is extremely narrow (Kuppuswamy and Bayus 2014). Being a relatively new domain, very few studies have explored the crowdfunding domain from a data mining perspective (An, Quercia, and Crowcroft 2014; Xu et al. 2014;

Etter, Grossglauser, and Thiran 2013). Although innovative in their approach, these studies restrict themselves to the standard set of features that are available readily from the Kickstarter domain. Therefore, in this paper, we divide our goal into two parts: first, we leverage diverse information sources to construct a set of features that can play a significant role in determining the success of Kickstarter projects; second, we utilize these features to develop a model that can recommend backers to Kickstarter projects.

Recommendation in Kickstarter poses a number of challenges. First, Kickstarter is a highly heterogeneous domain where a diverse set of features collectively influence the users' decision to back a project. Hence, recommendation cannot be based on some simple set of straightforward features that are directly available from the projects. Figure 1 shows the different *layers* of features that affect a user's interest in a particular project. In this paper, the social network *layer* is derived from the Twitter domain, and the project *layer* is based on the quality of Kickstarter projects and their temporal trends. The personal *layer* is decomposed into the user's preference over the creator of a project, the topical preference of this user, and finally the influence of geolocation. The second challenge is mainly due to the transient nature of projects. More precisely, in conventional recommendation such as movies or books, it is reasonable to apply collaborative filtering techniques since the recommended items are reusable, i.e., an item can serve many users for several years. This is not the case for our setting; in Kickstarter, once a project is expired after its posting period, we cannot recommend the project to any user.

In the first part of our work, we perform a comprehensive study on the set of features and their effect on Kickstarter projects. In the second part, we propose a supervised learning approach that effectively utilizes these features to tackle this unique recommendation problem. We formulate our recommendation problem as a binary classification/regression problem, where given a backer-project pair, the trained model computes the score that represents the likelihood of funding. Utilizing the proposed approaches together with a gradient boosting tree, a state-of-the-art learner model, we achieve a practically useful level of performance up to 0.89 AUC (area under the curve) value. Additionally, we perform an in-depth evaluation of our model using over 795K backer-project relations and a wide variety of other data sources like backer profiles, tweets and profile information of twitter users. Our analysis reveals various interesting knowledge about the behaviors of Kickstarter users with respect to their backing frequency, social network, geolocation, and other personality-based traits. The major contributions of this paper are summarized as follows:

1. We perform an exhaustive study of the crowdfunding domain from the project, backer, social network, and geolocation perspectives to provide several unique insights on inter- and intra-domain factors that affect the success of Kickstarter projects.
2. Our analysis is based on diverse data sources such as: (1) content information of projects, (2) profile information of backer and creators and (3) heterogeneous information from the Twitter network.
3. We build a robust predictive model for recommending backers in crowdfunding domain that achieves an AUC of 0.89, and a precision up to 0.8.

The rest of this paper is organized as follows. We begin by presenting the related work on this topic in Section 2. We then explain the characteristics of the Kickstarter domain in Section 3. Section 4 describes the data collection methodology. Section 5 analyzes the features of Kickstarter. Section 6 describes the model, and shows the results of our experiments. Finally, the conclusions obtained through this study are presented in Section 7.

2 Related Work

In this section, we discuss studies that analyze the Kickstarter and other crowdfunding domains.

Crowdfunding and Kickstarter: Since crowdfunding is still an emerging platform, most studies on this domain are relatively new. One of the most comprehensive studies on Kickstarter can be seen in (Kuppuswamy and Bayus 2014) and (Mollick 2014). In (Kuppuswamy and Bayus 2014), the authors examine the dynamics of kickstarter domain, and (Mollick 2014) explains various types of crowdfunding platforms. The authors of (Gerber, Hui, and Kuo 2012) and (Hui, Greenberg, and Gerber 2014) perform a real-world analysis on crowdfunding platforms. Their study is based on a real-time survey that aims to learn the motivation behind users who create and invest crowdfunding projects. In (Mitra and Gilbert 2014), the authors use natural language processing techniques to analyze the textual content of Kickstarter projects, while (Xu et al. 2014) leverages the updates of projects to determine their success rate. There are very few papers that study the role of *twitter* for Kickstarter projects. In a recent study, Lu *et al.* (Lu et al. 2014) delineate the impact of social network on Kickstarter projects.

Studies on other Crowdfunding platforms: Apart from Kickstarter, there are many other crowdfunding platforms. In our previous work, we analyzed the micro-financial activities in *Kiva.org* (Choo et al. 2014a; 2014b). Few research works (Bruett 2007), (Andreoni 1990) and (Ashta and Assadi 2009) explored the effects of the internet on microfinancing, and peer-to-peer lending transactions.

Recommender Systems Essentially, recommender systems can be divided into two main categories: content-based methods and collaborative filtering methods. For a comprehensive summary of collaborative filtering techniques, the reader is referred to these survey articles (Su and Khoshgoftaar 2009; Adomavicius and Tuzhilin 2005). In addition to these techniques, graph-based algorithms like the PageRank can also be used for recommendation (Rakesh et al. 2014). Such algorithms are effective in social networks such as Twitter and Facebook where the relationship (or linkage) between the users can be mined to construct social graphs. In this research, a collaborative filtering method is inapplicable due to the challenges that were discussed in Section 1. Therefore, we use a content-based approach where the personality-based features can be viewed as user's profile and the project-based features represents the prod-

uct content. However, it should be noted that the typical content-based approach, mainly originating from information retrieval literature (Baeza-Yates, Ribeiro-Neto, and others 1999), focuses only on textual information. In order to integrate all the other information available, our approach extends it in the context of ad-hoc information retrieval (Manning, Raghavan, and Schütze 2008), which considers various information as features and trains a learner model for predicting a relevance score of an item.

The paper closest to our research is characterized by a similar goal as that of ours (An, Quercia, and Crowcroft 2014). In their paper, the authors adopt a hypothesis-driven approach to analyze features from Kickstarter. Despite a novel approach, their analysis is based on very basic set of features such as number of updates, comments, facebook friends, etc.; such features are readily available from the Kickstarter platform. Additionally, our recommendation is based on the state-of-the-art gradient boosting tree model that achieves a much higher accuracy than the SVM classifier used in this paper. To the best of our knowledge our work is the first to perform an extensive analysis of the Kickstarter domain by utilizing project-, user profile-, geo-location- and social network-based attributes and to build an effective recommendation system for Kickstarter projects.

3 Characteristics of Kickstarter Campaign

Before exploring the features of Kickstarter, we investigate the general characteristics of this crowdfunding domain. Figure 2(a) shows the overall trend of successful and failed projects. We observe that a majority of projects exceed their funding goal by a very marginal amount. Additionally, projects that exceed their target goal by over 150% are extremely few. This suggests that people are not interested in supporting the projects once the project goal amount is received. Figure 2(b) shows the success ratio of top-10 categories of Kickstarter projects, where the top three project categories are dominated by film & video, music and games. Furthermore, the success ratio of these categories ranges roughly between 35%-65%, with Theater being the highest (about 65%) and technology being the lowest (about 35%).

The relationship between backers and the pledged amount is an essential component of Kickstarter. The backing pattern of Kickstarter users follows a power-law distribution, as depicted in Figure 2(c). We see that a large number of people tend to back just one or two projects; people who back more than 100 projects are extremely few. On the other hand, Figure 2(d) shows a strong correlation (about 0.68) between the number of backers and the pledged amount. Therefore, in this paper, we assume that backers impact the success of a project, and our goal is to investigate the effect of various features that impact the backing count.

4 Dataset Description

Kickstarter Database: For our experiments, we obtained six months of Kickstarter data from *kickspy*.[1] Our dataset spans from 12/15/13 to 06/15/14, which consists of 27,270 projects characterized by 30 project-based attributes. These

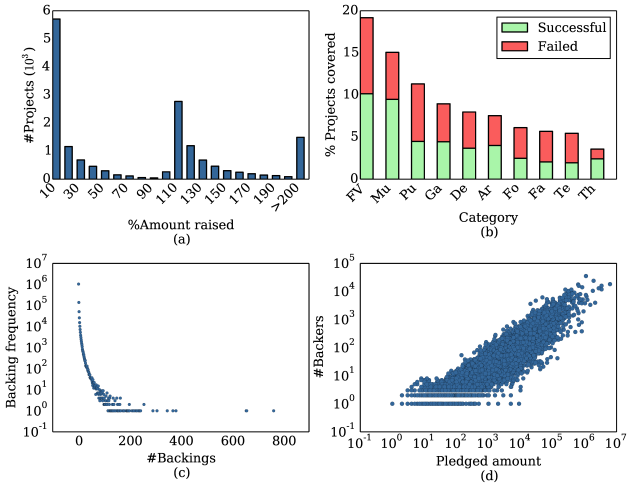


Figure 2: (a) the percentage of the total goal amount raised by Kickstarter projects; (b) the category-wise success ratio of projects: FV-film & video, Mu-music, Pu-publishing, Ga-games, De-design, Ar-art, Fo-food, Fa-fashion, Te-technology, and Th-theater; (c) the backing frequency of Kickstarter users; and (d) the relationship between the pledged amount and the number of backers.

attributes include a number of static features such as project goal amount, duration, textual content, etc., and two dynamic features: per-day increase in the number of backers and pledged amount. To prepare our dataset, we removed canceled or suspended projects. We also removed projects with less than one backer and $100 as a pledged amount. In this manner, we obtained 18,143 projects with over 1 million backers. We denote our projects database by $\mathcal{K}$ and backers database by $\mathcal{B}$. The statistics of our database are given in Table 1.

Table 1: Kickstarter data statistics for 18,143 projects collected from Dec 2013 - Jun 2014.

Attribute	Mean	Min	Max	StdDev
Goal Amt	26,531.2	100	100,000,000	758,366.5
Pledged Amt	11,023.6	100	6,224,955	78,550.8
backers count	138	1	35,383	633.7
Duration(days)	31	1	60	10.05

Twitter Database: Twitter is often used as a means to promote Kickstarter projects. To explore the effects of these promotions, we built our database by retrieving tweets containing URLs that begin with *http://kck.st* [2]. By expanding these short URLs, we eliminated tweets that did not map to our database $\mathcal{K}$. Using this method, we obtained 106,738 unique tweets, which covered 55% of our projects. The remaining 45% were never promoted using Twitter. Additionally, we also retrieved the complete profile information of the promoters who tweeted these tweets; we denote this database by $\mathcal{S}$.

5 Analyzing Kickstarter Features

The success of an entrepreneurial venture is heavily dependent on its quality from the project contents. Therefore, in

[1] www.kickspy.com

[2] we used the query API available at www.topsy.com

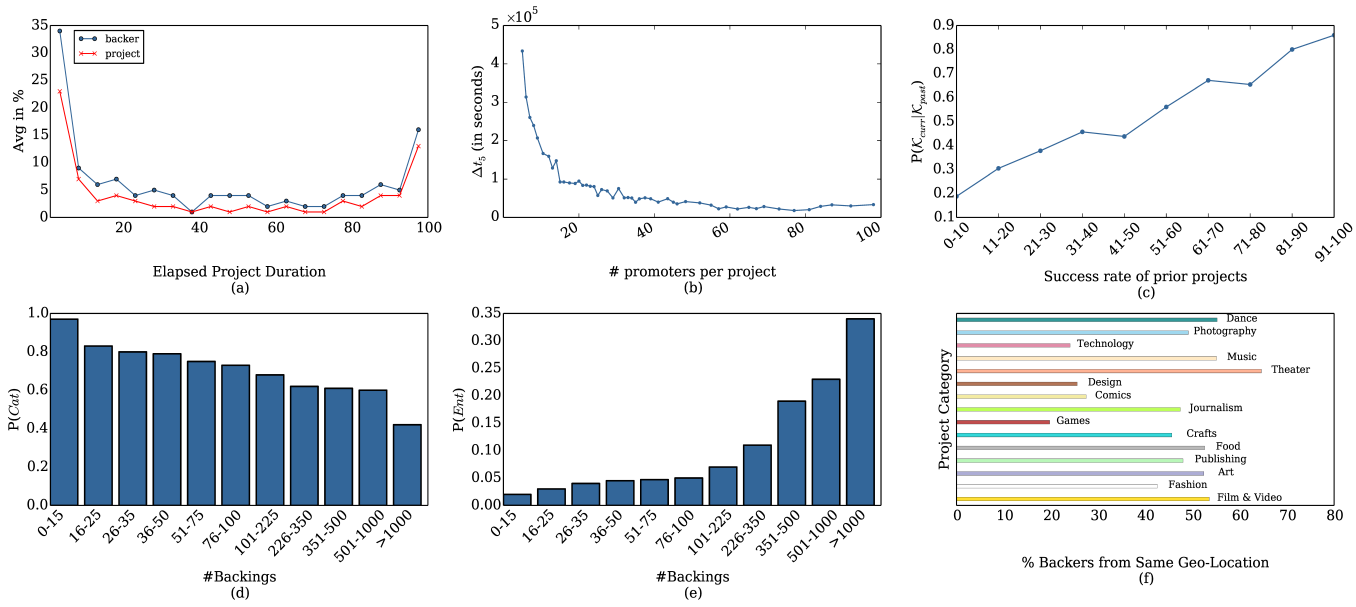


Figure 3: Analysis of features from Kickstarter domain. (a) Temporal progression of funds and backers in Kickstarter; (b) Adoption rate of project promotions in Twitter; (c) effect of *creator's* prior success rate on the success of his current kickstarter project; (d) topical preference of users towards projects; (e) trust relationship between the backers and creators; and (f) effect of geo-location on Kickstarter projects.

this section, we explore the set of factors that initiate a user to invest in Kickstarter ventures. We begin by categorizing the features into four main groups: (a) project-based traits, (b) personality-based traits, (c) location-based traits, and (d) network-based traits. The project-based traits are purely those derived from the qualities of Kickstarter projects. The personality-based traits are further divided into creator personality and backer personality. It represents the characteristics of creators who host the Kickstarter project, and the backers who invest in these projects. In the location-based traits, we intend to understand the role of geo-location [3] on Kickstarter projects. Lastly, the network-based traits are derived from the social media (Twitter) domain. In the following sections, we explain these features in detail.

5.1 Project-based Traits

Static features: We use 16 different features, which include generic features such as the duration of project, the goal amount, the number of facebook shares, the main topical category, the sub-categories, the number of backers, the number of updates about the project progress, the pledged amount, comments, location, currency and rewards. The content-based features include the total number of words in the project description, the number of risks and challenges, the number of FAQs, the number of images, and the presence of videos.

Temporal features: One of the most interesting aspects of Kickstarter campaign is the U-shaped distribution of fund progression over time. As shown in Figure 3(a), a large percentage of the pledged goal is accumulated in the first few days of the project duration. The progression then tapers out during the mid-phase, which can be considered as a dormant period in the project funding cycle. However, unlike the progression of news stories, the funding activity does not decay monotonically towards the end; instead, we suddenly see an increase in the pledged amount during the final phase of the project funding cycle (few days before the project end date). A popular term for this phenomenon is the *deadline effect* (Webb and Weick 1979). It is also important to note that the accumulation of backers follows closely with the pledged amount, where a majority of backing activity happens during the first and last weeks of the funding cycle. This classic behavior of Kickstarter data has also been shown in recent studies (Kuppuswamy and Bayus 2014; Lu et al. 2014). Another important temporal dynamics is related with the spread of Twitter promotions over time. If the first few tweets about a kickstarter project are tweeted within a short time frame, the number of Twitter users who adopt and promote these tweets are much higher. In social science, this phenomenon is widely known as the *Herding instinct* (Nadeau, Cloutier, and Guay 1993). This effect is depicted in Figure 3(b), where Δt_5 denotes the average time delay between the first 5 consecutive tweets. From this figure, we can conclude that *rapid early promotions correlate with more promotions overall*. Later, in this paper, we will show that these promotions are crucial for the success of projects.

5.2 Personal Traits

Backer personality: To begin with, we retrieve the backing history of all the users in $\mathcal{B}$, and obtain the list of categories and creators for every project in their backing history. The history of categories and creators are denoted by the sets $\mathcal{H}(C)$ and $\mathcal{H}(E)$, respectively. The personality of backers are analyzed using these two sets.

[3] we consider city and state as geo-location

- **Topical preference:** Topical preference plays an important role in determining the interests of users (Welch et al. 2011). In our setting, we define this as the tendency of users to continuously back projects in the same topical category. We examine this by calculating the conditional probability of a user u to back a category c, given c is present in the backing history of this user. We represent this probability by $P(M(u,c)|c{\in}\mathcal{H}_u(C))$, where $\mathcal{H}_u(C)$ indicates the set of categories from user u's backing history. $P(M(u,c)|c{\in}\mathcal{H}_u(C))$ is calculated by Bayes' theorem as follows:

$$P(M(u,c)|c{\in}\mathcal{H}_u(C)) = \frac{P(c{\in}\mathcal{H}_u(C)|M(u,c))P(M(u,c))}{P(c{\in}\mathcal{H}_u(C))} \quad (1)$$

  Figure 3(d) shows the outcome of this experiment. It can be seen that irrespective of the number of backings, the probability of users backing the same category is very high. Although this probability decreases with the increase in backing count, this reduction is significant only for users with very large backings (i.e. over 1000). This signifies that *backers have strong topical preference over Kickstarter projects*.
- **Mutual trust:** It is shown that the investors do not just randomly choose projects for backing; instead, they look for a long-term connection to the creator (Gerber, Hui, and Kuo 2012). We call this attribute as the *mutual trust*. To validate this claim, we calculate the conditional probability of a user u to back a creator e, given that e is in the backing history of this user. This probability is represented by $P(M(u,e)|e{\in}\mathcal{H}_u(E))$, where $H_u(E)$ indicates the set of creators from user u's backing history and $P(M(u,e)|e{\in}\mathcal{H}_u(E))$ is calculated similar to Equation (1). In Figure 3(e), we see that this probability increases with the increase in backing count of users. In other words, when users start backing more and more projects in Kickstarter, they *tend to develop an inclination towards creators whom they have backed in the past*. This inclination leads to a stronger relationship with the creator thereby creating a mutual trust.

Creator personality: The personality of project creators is measured using three features: a) the number of projects hosted by the creator, b) the number of projects backed, and c) the expertise of the creator. The first two features are obtained from the profile information, while the third feature is analyzed as follows:

- **Creator expertise:** We claim that a creator having a high success ratio in his past projects is more likely to succeed in his current project. We evaluate this notion by calculating the probability $p(k_{curr}|k_{past})$, where k_{curr} denotes success of current project, and k_{past} denotes the success ratio of prior projects. From Figure 3(c) we can distinctly see that this probability increases with the increase in the creator's success ratio. Hence, *experienced creators have a greater chance to succeed in the crowdfunding domain*.

5.3 Location-based Traits

To understand the role of geo-location, for every project in our database $\mathcal{K}$, we calculate the percentage of backers whose geo-location matches with the project's location. The result of this study is depicted in Figure 3(f), which clearly shows that geo-location does impact the success of projects. Nonetheless, it is interesting to note that *the impact of geo-location is not uniform for all the categories of projects*; for instance, projects on games, comics, and technology are relatively less dependent on their geo-location, while projects on theater, food, and dance are highly dependent. A reasonable explanation for this trend can be attributed to the rewards that are provided by the projects. For example, the rewards offered by theatrical projects mostly include items such as movie tickets, tickets to the premier shows or personal interaction with the cast members. Such rewards are extremely dependent on the proximity to the project's geo-location since people from distant geo-locations might not travel to see the performances. Contrary to this, rewards offered by technical projects can be shipped to people all over the world.

5.4 Network-based Traits

One of the main reasons for a project's failure is the lack of publicity (Hui, Greenberg, and Gerber 2014). Therefore, before we examine the social network features, we will see whether Twitter-based promotions impact the success of Kickstarter projects. Table 2 shows that projects with promotions have 63% chance to succeed in their funding goal, while those without promotions have a mediocre success rate of 34%. This shows that Twitter-based promotions are crucial for determining the success of projects. Hence, we divide our analysis into two parts: first, we examine the impact of various network measures over the success of Kickstarter projects; second, we build the communities of Kickstarter users from Twitter to examine the effect of these communities over the backing habits of individuals.

Table 2: Success rate of projects with promotional activities. w/o-promo: without promotional activities; w-promo: with promotional activities.

# Projects w/o-promo	Success w/o-promo	# Projects w-promo	Success w-promo
8207	34%	9935	63%

Impact of Twitter network on Kickstarter projects: For this analysis, we construct a network using the Twitter database $\mathcal{S}$, which contains the set of users who tweeted about Kickstarter projects. Each user is a node, and a directed link exists between nodes A and B based on the following conditions: 1) if A is a follower of B; 2) if A mentions B in his tweet. In the first case, we assign a link weight of 1, and for the second case, the link weight depends on the number of times A has mentioned B in his past tweets. By constructing this graph, we study the effect of various network-based measures over the backer count of Kickstarter projects. Figure 4(a) shows that the number of backers increases with the number of nodes (i.e. promoters in Twitter). However, the accumulation of backers is not only based on the number of promoters, but it also depends on the connectivity between these promoters. This notion is conveyed in Figures 4(b) and 4(c), where the former shows that *the stronger tie strength between the*

promoters results in greater accumulation of backers, while the latter captures the same notion in terms of the number of bi-connected components. In social network analysis research, the concept of bi-connectivity is utilized to measure the *structural cohesion* in social groups (Moody and White 2003). Lastly, Figure 4(d) shows that *projects promoted by influential twitter users have the potential to attract many backers*; where the influence is determined by the PageRank scores. To plot the y-axis of this figure, we first calculate the PageRank scores of all the nodes (i.e. promoters in Twitter) in the directed graph. For each project, we then calculate number of promoters in the top-100 PageRank scores.

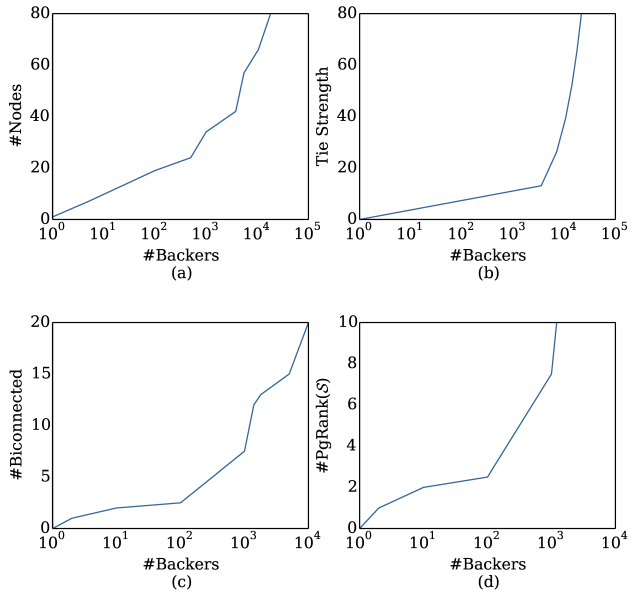


Figure 4: Impact of Twitter network on the backers of Kickstarter projects. (a) shows the impact of a promoter count, (b) and (c) shows the importance of connectivity in the network of promoters, and (d) shows the importance of influential promoters.

Community influence on investors' backing habits: Many studies have shown that communities from social network play an important role in influencing the actions of individuals (Asur et al. 2011; Bakshy et al. 2011). Applying the same analogy, we say that *the backing habits of investors in Kickstarter are influenced by their social circle (or community)*. To validate this statement, we begin by explaining the procedure for creating Kickstarter communities from Twitter. Next, we describe how we calculate the influence score between these communities and the individuals who backed the projects.

To construct the communities, we use the promoters (e.g., twitter users) in $\mathcal{S}$ and create a bipartite graph of projects and users where each edge denotes the action of a user $s \in \mathcal{S}$ tweeting about a project $k \in \mathcal{K}$. These tweets can be simply promotions, or it can signify the action of backing. The bipartite graph is then projected into a unipartite graph resulting in a network that consists only of the users s. The edge weight between the users (s_1, s_2) is computed using Jaccard index, which is given by:

$$W = \frac{|\mathcal{K}(s_1) \cap \mathcal{K}(s_2)|}{|\mathcal{K}(s_1) \cup \mathcal{K}(s_2)|} \tag{2}$$

where $\mathcal{K}(s_1)$ and $\mathcal{K}(s_2)$ denote the set of projects that are tweeted by users s_1 and s_2, respectively. To form the communities from this network structure, we use the modularity metric, which is defined as:

$$M = \frac{1}{2W} \sum_{i,j} \left[W_{ij} - \frac{n(i)n(j)}{2W} \right] \delta(C_i, C_j) \tag{3}$$

where W_{ij} is the edge weight between vertices i and j, and W is the summation of all the edge weights. $n(i)$ and $n(j)$ are obtained by summing up all the edge weights of nodes i and j, respectively. C_i denotes the community that i belongs to, and δ is the Kronecker delta. Lastly, we use the Louvain method of community detection (Blondel et al. 2008) over this unipartite network to obtain 160 communities. A snapshot of this procedure is shown in Figure 5.

To calculate the influence of these communities over the backing habits of the users, we use our database $\mathcal{B}$ and retrieve a subset of backers who have their Twitter account information embedded in their Kickstarter profiles. We call this set as $\mathcal{B}_{tw}$, where $|\mathcal{B}_{tw}| = 9,266$. To measure the influence of the community $c \in C$ over the user $b_{tw} \in \mathcal{B}_{tw}$, we calculate their *Affinity* score as:

$$Affinity(b_{tw}, c) = |F(b_{tw}) \cap F(c)| \tag{4}$$

where $F(b_{tw})$ and $F(c)$ denote the set of all followers and followees of b_{tw} and c, respectively; $|F(b_{tw}) \cap F(c)|$ indicates the number of mutual friends between this backer and the members of the community. Figure 6(a) shows the outcome of this analysis, where $P(M(b_{tw}, k)|M(c, k))$ indicates the probability of the user $b_{tw} \in \mathcal{B}_{tw}$ backing the project k, given that k is backed by the members of the community c. Figure 6(b) is similar to 6(a) except that $\mathcal{T}$ denotes the project category. From these figures, one can see that the stronger *affinity* of a user towards a community leads to the greater chance for this user to back the same project (or project category) that was backed by the community.

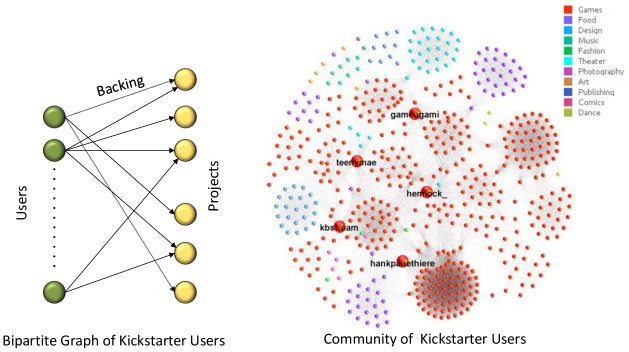


Figure 5: A Twitter-based community formed by the backers of Kickstarter (right) by projecting the bipartite graph of user and projects (left). The colors represent different topical categories of communities, and the names denote the top users of the community.

6 Recommending Backers

Considering the complexity and heterogeneity of our data and the problem, it is important to use the most suitable and powerful prediction model that are available. To this end, we have employed a gradient boosting tree (GBtree)[4] (Hastie, Tibshirani, and Friedman 2009; Friedman 2001). A GBtree is an ensemble method where an individual learner is a decision tree (Breiman 1993).

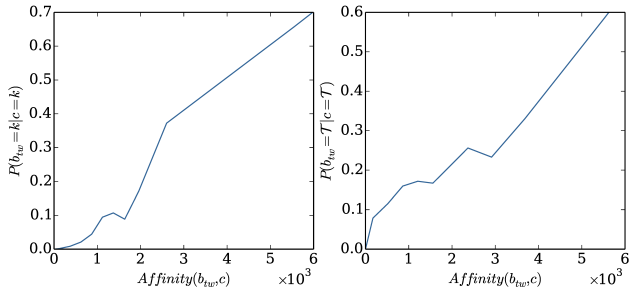


Figure 6: Influence of Twitter-based Kickstarter communities over the backing habits of users.

The reason for choosing a GBtree for our problem is as follows: First of all, an ensemble method is known for its superior generalization capability for unseen data. Furthermore, a decision tree, our base learner, uses one variable at each node when it is trained/constructed as well as when it is applied to test data. This characteristic prevents us from worrying about how to properly consider heterogeneity in the features we generated. The drawback of using other learners, such as logistic regression and support vector machines, is that heterogeneous features have to be normalized via, say, standardization of their distributions by transforming each feature to have zero mean and unit variance. Such normalization does not always make sense for binary and integer features, and it also removes the nonnegativity of our feature representation that offers intuitive interpretation of them. It should be noted that the key contribution of this work is more about extracting the important features and understanding the domain by providing novel insights, but not necessarily about building a new predictive modeling algorithm.

6.1 Experimental Setup

We formulate the task of recommendation as a binary classification/regression problem. That is, every backer-project pair (b, k) is an individual data item, and given such a pair, our task is to predict how likely a user will back a project. Our aim is not only to show the superior prediction performance of our model, but also to conduct an in-depth analysis on the features discussed in the earlier sections. To achieve this, we create a dataset, $\mathcal{D}$ using a subset of backers from $\mathcal{B}$, defined as:

$$\mathcal{D} = \{(b, k)|k \in \mathcal{K}(\mathcal{T}), (b \in \mathcal{B} \cap prof(b) = 1)\}$$

[4]The GBtree implementation we used is available at https://sites.google.com/site/carlosbecker/resources/gradient-boosting-boosted-trees

where $\mathcal{K}(\mathcal{T})$ is the set of all the project in $\mathcal{K}$ that are covered by our tweet database $\mathcal{T}$, and $prof(b) = 1$ indicates the existence of the *complete* profile information of the backer b. In total, the cardinality of set $\mathcal{D}$ is 795,347. To train and test our model, *we only consider the features available within the first three days of project duration.* This setting is much more realistic when compared to previous studies that use the complete set of features that are available only at the end of the project duration. This includes the features from project-, temporal-, personal-, geo-location-, and network-based traits that were discussed in the previous sections. However, *we eliminate comments, updates, and the number of Facebook shares* from the project-based feature since these features are generally not present in the initial stages of a project. It should be noted that $\mathcal{D}$ consists of only the positive samples, which indicates the action of user $b \in \mathcal{B}$ backing a project $k \in K$. Therefore, to create a balanced dataset, we augment $\mathcal{D}$ with $795,347$ randomly selected negative instances. Out of this entire dataset, we test the following cases by filtering out the data instances matching the conditions for each case.

Case 1: *Evaluating the influence of social network.* The influence from social network (Section 5.4) is much stronger on backers who have their Twitter profile. Additionally, the community-based influence is applicable only for backers who are connected to the Twitter network. Therefore, to evaluate this feature, we create a dataset $\mathcal{D}_{tw}$, which is defined below:

$$\mathcal{D}_{tw} = \{(b, k)|(b, k) \in \mathcal{D} \cap (b \in \mathcal{B}_{tw})\}$$

where $\mathcal{B}_{tw}$ is the set of backers who have their Twitter profiles.

Case 2: *Evaluating the impact of geo-location.* In Section 5.3, we showed that the geo-location does not affect every category to a similar extent. To further support this result, we use the dataset $\mathcal{D}$ and retrieve only those projects which have the following categories: 1. Theater, 2. Music, 3. Games, and 4. Technology. We chose these categories because theater and music strongly depend on their geo-locations, while games and technology have very weak geo-location dependency (Figure 3(f)). We term this dataset as $\mathcal{D}_g$.

The datasets $\mathcal{D}$, $\mathcal{D}_{tw}$, and $\mathcal{D}_g$ are used for our evaluation which was performed using the standard 10-fold cross validation strategy.

6.2 Performance Evaluation

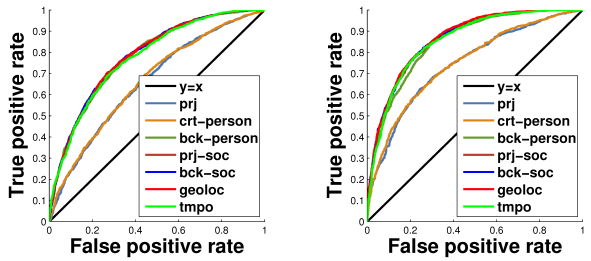
Using our evaluation methodology, we want to understand how much each feature contributes towards the recommendation performance. To achieve this, we use the set of attributes from Section 5 and categorize them into the following groups:

(a) `prj` (13 dimensions): All the features from the project-based traits (Section 5.1) except for updates, comments, and the number of facebook shares.

(b) `crt-person` (3 dimensions): Features from creator personality, which includes number of projects created, projects backed by the creator, and success ratio of the creator.

(c) `bck-person` (4 dimensions): Features from the backer personality such as the number of backings, categories of backed projects, topical preference, and creator preference.
(d) `prjsoc` (4 dimensions): Social network features on Kickstarter projects such as the number of promoters, the tie strength, the bi-connected components, and the PageRank of promoters from the first three days of the project duration.
(e) `bcksoc` (1 dimension): The influence score of community over the backers.
(f) `geoloc` (1 dimension): The influence score of geolocation over projects.
(g) `tmpo` (9 dimensions): The accumulation over the first three days in terms of the number of backers, the funding amount, and the number of tweet promotions.

Therefore, every object in our dataset (i.e., a backer-user pair) is represented by a 35-dimensional vector. In addition to these feature groups, we also split the Kickstarter users based on their backing frequency to study the performance of recommendation depending on various funding experiences. We begin by reporting the overall performance of our model, followed by the analysis of the variable importance for various feature groups. We conclude the evaluation by reporting the ranking performance of the recommendation model.



(a) BF = 2-10 (b) BF = 10-100

(c) BF = 101-500 (d) BF = 501-1000

Figure 7: The ROC curve results for different backing frequency (BF) values.

6.3 Predictive Performance

Overall performance: We test the performance of our model by gradually incorporating more features described in Section 6.2 to the experiments for different backer types. Figure 7 shows the performance results as the receiver operating characteristic (ROC) curve, and their AUC values are summarized in Table 3.[5] One can see that more features generally lead to better performances, and the best AUC values ranges from 0.79 to 0.88 when using all the available features, indicating the efficacy of the features inspired by our in-depth analyses.

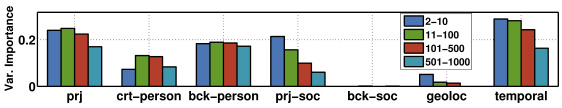
Table 3: Cumulative AUC values obtained in the plots shown in Fig. 7.

	Backing frequency			
Feature	2-10	11-100	101-500	501-1000
`prj`	0.736	0.744	0.707	0.659
`+crt-person`	0.744	0.748	0.719	0.664
`+bck-person`	0.882	0.849	0.830	0.780
`+prjsoc`	0.883	0.862	0.832	0.780
`+bcksoc`	0.882	0.862	0.834	0.784
`+geoloc`	0.886	0.864	0.836	0.783
`+tmpo`	**0.886**	**0.871**	**0.838**	**0.792**

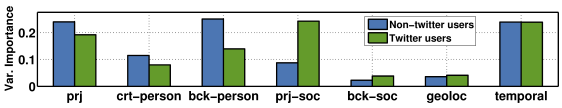
Analysis of feature groups: The analysis on the variable importance of each feature group is shown in Figure 8. We highlight the following insights about backing behaviors:

1. *The temporal progression of funds, backers, and tweet promotions have the strongest variable importance.* This claim is supported by high variable importance of the temporal features for all the different types of backers, as shown in Figure 8(a).
2. *Backers strongly depend on their personal preferences to fund a project*. This variable, which is denoted by `bck-person` includes the topical preference, and the mutual trust that were discussed in Section 5.2. In Figure 8(a), the inclusion of this feature has a significant effect over all the backer types.
3. *The impact of social network monotonically decreases with the increase in backing frequency.* This effect is shown by the `prjsoc` feature in Figure 8(a). From this trend, we infer that experienced investors do not solely rely on social network-based promotions, but instead they probably consider various other aspects of the projects for their backing decisions. Contrary to this, inexperienced investors are easily attracted to fund projects which have large promotional activity.
4. *Social network has stronger influence over backers who have their Twitter profile.* From Figure 8(b), we can see that the variable importance of `prjsoc` is distinctly higher for Twitter users (i.e. backers with Twitter profile) when compared to the non-Twitter users. This is because the non-Twitter users are not exposed to the activities in social media and therefore they seldom notice the tweets about Kickstarter projects. We also see that such users rely more on project and personal features. This trend is similar for community-based influence `bcksoc`. Nonetheless, it should be noted that, the very low variable importance of this feature is due to the fact that users with this feature are extremely fewer in number.

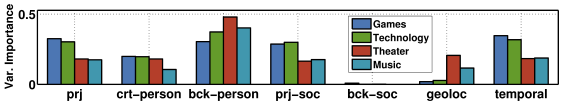
[5]The AUC value is computed using the trapezoidal approximation (Bradley 1997).



(a) Dataset $\mathcal{D}$



(b) Dataset $\mathcal{D}_{tw}$ (**Case 1**)



(c) Dataset $\mathcal{D}_g$ (**Case 2**)

Figure 8: Variable importance of various Kickstarter features. (a)-(c) shows the AUC value improvements over 0.5 when using only a particular feature set.

5. *The influence of geo-location strongly depends on the topical category of the project.* Figure 8(c) confirms our analysis in Section 5.3 where projects belonging to theater and music categories have a greater dependency on the `geoloc` feature when compared to games and technology.

6.4 Ranking the Backers

Although our experimental setting is a binary classification, the desired capability from learning the function $f(b, k)$ by a GBtree is to compute the likelihood of funding, which allows us to rank the most appropriate backer for a particular project. Therefore, to evaluate the performance of ranking, we use the standard information retrieval measures. For every project, we compute: 1) P@k: The *precision at rank k* for our task is defined as the fraction of rankings in which the true backers are ranked in the *top-k* positions, 2) MRR: The *mean reciprocal rank* is the inverse of the position of the first true backer in the ranked set of backers produced by our model, 3) S@k: The success at rank k is the probability of finding at least one true backer in the *top-k* ranked set, and 4) DCG: The discounted cumulative gain (Järvelin and Kekäläinen 2002) is based on the simple idea that highly relevant backers are more important than marginally relevant ones.

Table 4: Performance comparison between different sets of features using MRR and Precision metrics.

Features	MRR	P@1	P@10	P@20
`prj`	0.5	0.314	0.322	0.321
`+crt-person`	0.505	0.324	0.329	0.323
`+bck-person`	0.816	0.707	0.684	0.623
`+prjsoc`	0.828	0.728	0.688	0.626
`+bcksoc`	0.834	0.722	0.71	0.62
`+geoloc`	0.89	0.818	0.706	0.618
`+tmpo`	**0.892**	**0.824**	**0.708**	**0.627**

Table 4 shows the results of MRR and precision, while Table 5 reports the results of success at k and DCG. We clearly see that the addition of features results in a performance boost for all the measures. There is a clear increase in precision and MRR after the addition of backer personal traits, and this increase is further boosted with the addition of social network, geo-location, and temporal features. This trend is similar for all the other performance measures. It should be noted that, unlike the previous research (An, Quercia, and Crowcroft 2014), our recommendation is purely based on the features from the first three days of the project duration. Despite this fact, we can achieve a high precision value of 0.82.

Table 5: Performance comparison between different sets of features using Success at k and DCG metrics.

Features	S@1	S@10	DCG
`prj`	0.314	0.902	8.24
`+crt-person`	0.324	0.894	8.371
`+bck-person`	0.707	0.996	17.297
`+prjsoc`	0.728	0.998	17.424
`+bcksoc`	0.71	0.998	17.209
`+geoloc`	0.818	0.998	18.232
`+tmpo`	**0.824**	**0.998**	**18.388**

7 Conclusion

In this paper, we performed a rigorous analysis of the Kickstarter crowdfunding domain to reveal several unique insights about project-, social-, temporal-, and geo-location-based features that affect the success of its project campaigns. We showed that backers are strongly influenced by their topical preference and the trust relationship towards the creator of projects. In the analysis of network-based features, we used the network of promoters from Twitter to show that the success of projects depends on the connectivity between the promoters. Additionally, we created Twitter-based communities of Kickstarter users to study its impact on the backing habits of individuals. Our analysis revealed that the backing habits are influenced by their social circle (or community). Lastly, we reported that the effect of geo-location is not uniform for all the projects; instead, it depends on their topical category. In the second part of this paper, we used the analyzed set of features to build a model that recommends a set of backers to Kickstarter projects. Using the gradient boosting tree, a state-of-the-art learner model, and the features from only the first three days of project duration, we were able to achieve an AUC of 0.89, and a precision up to 0.8.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation grants IIS-1231742 and IIS-1242304.

References

Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-

of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on* 17(6):734–749.

An, J.; Quercia, D.; and Crowcroft, J. 2014. Recommending investors for crowdfunding projects. In *Proceedings of the 23rd international conference on World wide web*, 261–270. International World Wide Web Conferences Steering Committee.

Andreoni, J. 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. *The economic journal* 464–477.

Ashta, A., and Assadi, D. 2009. Do social cause and social technology meet? impact of web 2.0 technologies on peer-to-peer lending transactions. *Cahiers du CEREN* 29:177–192.

Asur, S.; Huberman, B. A.; Szabo, G.; and Wang, C. 2011. Trends in social media: persistence and decay. In *ICWSM*.

Baeza-Yates, R.; Ribeiro-Neto, B.; et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74. ACM.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159.

Breiman, L. 1993. *Classification and regression trees*. CRC press.

Bruett, T. 2007. Cows, kiva, and prosper. com: How disintermediation and the internet are changing microfinance. *Community Development Investment Review* 3(2):44–50.

Choo, J.; Lee, C.; Lee, D.; Zha, H.; and Park, H. 2014a. Understanding and promoting micro-finance activities in kiva. org. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 583–592. ACM.

Choo, J.; Lee, D.; Dilkina, B.; Zha, H.; and Park, H. 2014b. To gather together for a better world: understanding and leveraging communities in micro-lending recommendation. In *Proceedings of the 23rd international conference on World wide web*, 249–260.

Etter, V.; Grossglauser, M.; and Thiran, P. 2013. Launch hard or go home!: predicting the success of kickstarter campaigns. In *Proceedings of the first ACM conference on Online social networks*, 177–182. ACM.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 1189–1232.

Gerber, E. M.; Hui, J. S.; and Kuo, P.-Y. 2012. Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms. In *CSCW Workshop*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hui, J. S.; Greenberg, M. D.; and Gerber, E. M. 2014. Understanding the role of community in crowdfunding work. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 62–74. ACM.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)* 20(4):422–446.

KickStarterStats. Kickstarter stats. *Kickstarter* https://www.kickstarter.com/help/stats. accessed 07-23-2014.

Kuppuswamy, V., and Bayus, B. L. 2014. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. *UNC Kenan-Flagler Research Paper* (2013-15).

Lu, C.-T.; Xie, S.; Kong, X.; and Yu, P. S. 2014. Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 573–582. ACM.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press, Cambridge.

Mitra, T., and Gilbert, E. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 49–61. ACM.

Mollick, E. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing* 29(1):1–16.

Moody, J., and White, D. R. 2003. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review* 103–127.

Morduch, J. 1999. The microfinance promise. *Journal of economic literature* 1569–1614.

Nadeau, R.; Cloutier, E.; and Guay, J.-H. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review* 14(2):203–213.

Rakesh, V.; Singh, D.; Vinzamuri, B.; and Reddy, C. K. 2014. Personalized recommendation of twitter lists using content and network information. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Su, X., and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009:4.

Webb, E., and Weick, K. E. 1979. Unobtrusive measures in organizational theory: A reminder. *Administrative Science Quarterly* 650–659.

Welch, M. J.; Schonfeld, U.; He, D.; and Cho, J. 2011. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 327–336. ACM.

Xu, A.; Yang, X.; Rao, H.; Fu, W.-T.; Huang, S.-W.; and Bailey, B. P. 2014. Show me the money!: an analysis of project updates during crowdfunding campaigns. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 591–600. ACM.