Empirical Studies of Online Crowdfunding

EMPIRICAL STUDIES OF ONLINE CROWDFUNDING

by

Qiang Gao

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Qiang Gao, titled Empirical Studies of Online Crowdfunding and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

____ Date : 07/14/2016
Paulo Goes

____ Date : 07/14/2016
Mingfeng Lin

____ Date : 07/14/2016
Jesse Bockstedt

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date : 07/14/2016
Dissertation Director: Paulo Goes

____ Date : 07/14/2016
Dissertation Director: Mingfeng Lin

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Qiang Gao

ACKNOWLEDGEMENTS

I am grateful too for the support and advise from my dissertation committee, Jesse Bockstedt, Wei Chen, Paulo Goes, and Mingfeng Lin in Department of Management Information Systems at Eller College of Management. I need to thank especially my advisors, Paulo Goes and Mingfeng Lin. I have been amazingly fortunate to have advisors who offered unflagging support and wise advice. They taught me how to question thoughts and express idea. Their patience and supports helped me overcome many crisis situations and finish this dissertation. I hope that one day I would become as good an advisor to students as they have been to me.

DEDICATION

None of this would have been possible without the love and patience of my family to whom this dissertation is dedicated. They have been a constant source of love, concern, support and strength all these years. I would like to express my heart-felt gratitude to my family.

Table of Contents

List of Tables

List of Figures

Abstract

Online crowdfunding, an emerging business model, has been thriving for the last decade. It enables small firms and individuals to conduct financial transactions that would previously been impossible. Along with unprecedented opportunities, two fundamental issues still hinder crowdfunding ability to fulfill its potentials: the information asymmetry and the understanding of the impact of crowdfunding. Both are actually exacerbated by the "virtual" nature of these marketplaces. The success of this new market therefore critically depends on both improving existing mechanisms or designing new ones to mitigate the issue of unobservable fundraiser quality, which can lead to adverse selection and market collapse; and better understanding the impact of crowdfunding, and particularly its offline impact, which will allow the effective allocation of scarce resources.

My dissertation includes three essays around these topics, using data from *debt-*, *reward-* and *donation-based* crowdfunding contexts, respectively. My first two essays focus on two popular but understudied components in crowdfunding campaigns, texts and videos, and aim at predicting fundraiser quality by quantifying texts and videos. In particular, the first essay focuses on developing scalable approaches to extracting linguistic features from texts provided by borrowers when they request funds; and on using those features to explain and predict the repayment probability of the problematic loans. The second essay focuses on videos in reward crowdfunding, and preliminary results show excellent predictive performance and strong associations between multi-dimensional video information and crowdfunding campaign success and quality. The last essay investigates the impact of educational crowdfunding on school performance, using data from a crowdfunding platform for educational purposes. The results show that educational crowdfunding plays a role far beyond simply a financial source. Overall, my dissertation identifies the non-financial impact of crowdfunding as well as potential opportunities for efficiency improvement in the crowdfunding market, which have thus far not been documented in the literature.

Chapter 1 Overview

Crowdfunding has been thriving for the last decade. It provides opportunities for individuals and small or medium-sized companies to raise funds from a large number of online investors (i.e., the “crowd”); each investor typically provides a small amount for a variety of purposes such as arts, education, and startups. Unlike in traditional financial markets, there is no intermediary between funders (i.e., people who provide funds) and fundraisers (i.e., people who request funds) in online crowdfunding, which results in a decreased transaction cost (Belleflamme, Lambert, & Schwienbacher, 2014). In addition to reducing cost, this new business model conducted via Internet overcomes many constraints of traditional markets such as geographic boundaries, access to capital, and time constraints (Agrawal, Catalini, & Goldfarb, 2011, 2013). All of these benefits brought by crowdfunding have allowed it to grow at an astonishing speed. According to Massolution’s annual crowdfunding industry reports[1], this market started from $2.7 billion in 2011 and has been growing at an annual rate of over 81%, reaching $ 34 billion in 2015. The growth for next decade is forecast to continue at similar rate.

Along with its rapid growth, crowdfunding market has developed into four major categories depending on the returns that investors can obtain (Mollick, 2014) (see Belleflamme et al. (2014), Mollick (2013), Mollick (2014), and Schwienbacher and Larralde (2012) for more detailed information about mechanisms for each category and crowdfunding as a whole). Some crowdfunding projects such as arts or humanitarian projects fall into the category of donation-based crowdfunding where investors donate to the projects without seeking for any financial returns. The second category is debt-based crowdfunding where investors act as lenders and receive interests from borrowers (i.e., fundraisers who request funding). The third category of crowdfunding is the equity-based crowdfunding where fundraisers provide equity to investors who invest in companies, mostly startups. The last category of crowdfunding is the reward-based crowdfunding where investors (i.e., “backers”) contributes to a variety of projects for reason that is similar

[1] http://reports.crowdsourcing.org/index.php?route=product/product&product_id=54&tracking=5519a8852227b

to pre-sales – receiving some forms of rewards such as earlier access to an event or discount to products.

Despite these opportunities, two fundamental issues in the new market remain. The first issue is information asymmetry since crowdfunding is ultimately a two-sided market. It occurs because funders have limited and often unverified information about fundraisers (i.e., people who request money) (Lin, Prabhala, & Viswanathan, 2013) due to legal framework regarding privacy and the nature of the virtuality of the Internet. For example, we have no other verified information regarding borrowers in debt-based crowdfunding except of credit information such as credit score and debt-to-income-ratio from credit report agency. In equity-based crowdfunding, fundraisers provide very limited information about themselves and their business plans due to concerns about leaking business advantages. It is much worse for both reward- and donation-based crowdfunding where fundraisers almost disclose no verified information about themselves and their campaigns unobserved quality. As a result, funders in this market are hesitate to invest because of concerns about the competitiveness of fundraisers and fraud (Agrawal et al., 2013). Faced with these challenges, the long-term viability of crowdfunding markets therefore critically hinges on the ability of all stakeholders in the market to create, identify, and utilize "signals" that can help to mitigate such information asymmetry and correctly evaluate the quality of crowdfunding campaigns.

The second issue is a lack of understanding of the impact of crowdfunding, particularly offline (Mollick & Kuppuswamy, 2014). For donation-based crowdfunding, it is more obvious. Although contributions to disaster relief and help to third-world hunger certainly have impacts on the life of victims, the majority of donations goes to creative ideas such as films, arts, education, and music[2]. There is little information about the impact of contributions on fundraiser behaviors, particularly their offline, after they receive funds: do fundraisers only treat the crowdfunding as a financial source? Or will the fact that donations come from a pool of warm-hearted strangers encourage or empower them to better use the funds? For other crowdfunding contexts, although we can possibly infer the impact of having crowdfunding to certain extent, it is still unclear. One possible reason for

[2] http://reports.crowdsourcing.org/index.php?route=product/product&product_id=54

this lack of information is that since the outcomes of crowdfunding projects (i.e., the results after projects were funded) may not look good, the project creators are reluctant to reveal them due to concerns future funding (Mollick & Kuppuswamy, 2014). Another possible explanation is that because this market is in its early stage, the long-term impact of crowdfunding has not emerged. Like other resources, crowdfunding resources are limited. Because of lack of understanding about the impact of crowdfunding on different aspects of our society and the extent of such impact, we are facing difficulties in effectively allocating capital, an issue that may lead to a market failure.

My dissertation includes three essays that are developed around these two topics, using data from debt-, reward-, and donation-based crowdfunding. In the first essay, we examine the role of text as a potential mitigating mechanism for information asymmetry in debt-based crowdfunding contexts. Text is prevalent in the online market, but its economic value is far from certain. This paper investigates how the linguistic styles of text (i.e., loan descriptions provided by borrowers) can help mitigate the information asymmetry, affect the success of loan requests, and reveal the quality of borrowers. The values of text, if any, in debt-based crowdfunding is relatively smaller due to the availability of a large amount of financial information, comparing to that in other types of crowdfunding. If we still can find that text indeed possesses value for mitigating information asymmetry. That value should be much larger in other types of crowdfunding. Thus, I choose debt-based crowdfunding as the study context. The results show that borrowers indeed take the text into consideration for their funding decisions and the well-established features related to creditworthiness (readability, positivity, objectivity, and deception cues) are all meaningfully related to loan repayment. However, investors do not correctly interpret all of these linguistic styles. This suggests opportunities for efficiency improvement by leveraging the texts that are not yet documented in literature.

My second essay studies the information value of videos in reward-based crowdfunding. Video is a major channel through which crowdfunding campaign creators disclose information about the campaigns and themselves. This paper examines the impact of videos on contributors' funding decisions, the values of multi-dimensional video information in predicting campaign quality, and the interpretation of the values of this information by contributors. I choose reward-based crowdfunding as the study context because video is the main informational disclosure channel there and is relatively more important than videos in other types of crowdfunding. The results show that the multi-dimensional video information does indeed possess significant predictive power about campaign quality. We further find that contributors consider the videos in their funding decisions, and can only correctly part of the information. This study has important implications for both academics and practices.

My last essay investigates the offline impact of donation-based crowdfunding. I choose education as the study context for two main reasons. First, funding has been continuously dwindling for public education, donation-based crowdfunding has become a viable financing source. The second reason is the "public good" nature of education. We can have a better understanding about the impact of charitable giving on public goods. Thus, this study examines the impact of online educational crowdfunding on school performance by exploring a geographical expansion of a crowdfunding site for educational purposes. The initial results show that funds raised through crowdfunding have a positive impact on classroom performance, especially when teachers are required to disclose more verifiable personal information and therefore become more accountable. The findings not only show the offline impacts of online crowdfunding, but also have implications for the management of traditional donations for education purposes.

This dissertation is structured as follows. Chapters 2 gives a brief review about related literature. Chapter 3 includes three essays of my dissertation. Essay 1 studies the role of text as a mitigating mechanism in debt-based crowdfunding. Essay 2 is the study of informational values of videos in crowdfunding. Essay 3 examines the impact of the educational crowdfunding on school performance. Chapter 4 provides insights from this dissertation and discuss possible future research. Chapter 5 concludes.

Chapter 2 Related Literature

In this section, I provide a high-level review of the two main streams of literature related to the major issues I study. More detailed literature review will be presented in each study later in Chapter 3. The first stream of literature examines mechanisms in mitigating asymmetric information in crowdfunding and another investigates the impact of crowdfunding.

2.1 Mechanisms Mitigating Information Asymmetry in Crowdfunding

The majority of existing literature on how to mitigate information asymmetry in crowdfunding contexts has focused on the roles of various mechanisms that can potentially overcome asymmetric information. Some studies used demographical information as potential mitigating mechanism (e.g. race, beauty, and culture) (Burtch, Ghose, & Wattal, 2014; Duarte, Siegel, & Young, 2012; Ravina, 2012). Burtch et al. (2014) examined offline information and demonstrated that cultural similarities can encourage lenders' lending actions. Others studied the demographical information posted on online campaign pages and found that the appearance of people in the photos attached to the lending requests had unexpected impacts on investor funding decisions. While borrowers who appear more trustworthy and more beautiful have higher probabilities of having their loans funded (Duarte et al., 2012; Ravina, 2012), only the appearance of trustworthiness indeed reveal the true quality of borrowers, who default less often (Duarte et al., 2012).

In addition to the demographical information, social information (e.g. friendship, group membership, and endorsement) is a major source of mitigating mechanisms (Freedman & Jin, 2008; Hildebrand, Puri, & Rocholl, 2016; Lin et al., 2013; Liu, Brass, & Chen, 2015). While all these studies show that friend endorsements and friend bids are good signals for loan quality and relate to significantly higher rates of repayment, group loans have significantly lower return of rates than non-group loans (Freedman & Jin, 2008). Furthermore, some studies examined the role of financial information (e.g. debt-to-income-ratio, home ownership, and credit grade)(Iyer, Khwaja, Luttmer, & Shue, 2015; Zhang & Liu, 2012). They all found that soft and non-standard information is more important than credit information alone in revealing borrower quality.

However, these studies adopted a manual information processing approach that cannot be easily extended to different types of crowdfunding[3]. In addition, the mitigating mechanism used in those studies can be only applied to certain types of crowdfunding. Therefore, my dissertation aims to fill these gaps and to provide an understanding of what universal factors of all types of crowdfunding can effectively contribute to the funders' decisions and better indicate the campaign's quality; This is done by implementing scalable approaches such as machine learning, which in turn can help mitigate information asymmetry.

2.2 Impact of Crowdfunding

Few studies examine the impact of crowdfunding, particularly offline, and they have mainly focused on two topics. The first topic is the factors affecting online contributions (Agrawal et al., 2011; Burtch, Ghose, & Wattal, 2013; Burtch et al., 2014; Lin & Viswanathan, 2015; Zhang & Liu, 2012). The relevant research investigates both offline and online elements that affect investors' funding decisions. Agrawal et al. (2011) and Lin and Viswanathan (2015) showed that geographical constraints existing in traditional financial market still have detrimental effects on contributions in online environment. That is, contributors in crowdfunding are still more likely to fund crowdfunding projects that are geographically closer to them. On the other hand, online social influence, specifically prior contribution behaviors from other contributors, greatly affects others to contribute (Burtch et al., 2013; Zhang & Liu, 2012).

The second topic is related to mechanisms affecting fundraiser behaviors and campaign outcome (i.e., outcome after crowdfunding campaigns were funded) (Hildebrand et al., 2016; Mollick & Kuppuswamy, 2014). Group is a common mechanism in crowdfunding to both increase funding success and improve campaign quality, but Hildebrand et al. (2016) found that group leader bids, a mechanism to ensure fundraiser quality, have no impact on borrowers' future repayment. Mollick and Kuppuswamy (2014) did a survey to examine the effects of various factors on the outcomes of successfully funded crowdfunding projects. They found that several factors, such as featured by crowdfunding site, having outside

[3] Mollick (2014)provided a more compressive description about the four types of crowdfunding: reward-based, debt-based (i.e. Peer-to-Peer lending (P2P)), patronage-based, and equity-based.

endorsement, and many social network friends, improve campaign outcomes in several aspects. For instance, raising outside funds after crowdfunding campaigns, getting additional benefits from a campaign, and delivering products on time. However, common features of all aforementioned studies are that they either mainly focus on online behaviors or explore the outcomes from a small sample of survey. That is, we have little understanding about the true impact of crowdfunding on the offline behaviors of fundraisers.

For what have been discussed above, both information asymmetry and the understanding of the impact of crowdfunding, have not been systematically studied. Following chapter, which includes three essays targeting these issues, specify the approaches and findings from addressing them.

Chapter 3 Essays Addressing Information Asymmetry and the Impact of Crowdfunding

This chapter includes three essays addressing issues of either information asymmetry or the impact of crowdfunding. The first two essays examine whether unstructured data – text or video – in crowdfunding can be used as potential mitigating mechanisms for asymmetric information in either debt- or reward-based crowdfunding. The last essay explores the geographic expansion of a crowdfunding site for educational purpose to investigate the impact of donation-based crowdfunding projects on school performance improvement.

3.1 Economic Value of Texts: Evidence from Online Debt Crowdfunding

Abstract

Texts are prevalent in online markets, but its economic value is far from certain. This paper examines whether linguistic styles of texts can help mitigate issues of information asymmetry, and more importantly, whether investors can “correctly” interpret the economic value of texts. Using data from online debt crowdfunding, we first show that investors indeed take into account the “loan purpose” descriptions that borrowers provide in their loan requests, even though these texts are not verified or legally binding. We then analyze the linguistic features of these descriptions, and show that well-established features related to creditworthiness (readability, objectivity, negativity, and deception cues) all meaningfully relate to loan repayment. Interestingly however, investors do not correctly interpret the economic values of all linguistic features, most notably deception cues. Finally, we show that these automatically extracted features can improve the predictive accuracy of loan defaults. This suggests that even though “texts” are often considered “soft” or “non-standard” information in finance, it can be quantified and standardized into credit risk modeling. Our study points to opportunities for efficiency improvement by leveraging texts that are not yet documented in the literature.

Key words: texts, peer-to-peer lending, crowdfunding, predictive analysis, sentiment analysis, subjectivity analysis, readability analysis, deception detection, machine learning

1. Introduction and Background

We study the role of texts in online debt crowdfunding, especially as a potential mechanism to mitigate information asymmetry in this nascent but extremely fast growing market.

Online crowdfunding has been growing at a fascinating speed for the past decade. It broadly refers to the phenomenon that individuals or organizations can now use the internet to raise funds for a variety of causes. Depending on what the investors receive in return, it can take the form of donation (e.g., Kiva.org), debt (e.g., Zopa.com, Prosper.com), rewards (e.g., Kickstarter.com or Indiegogo.com), or equity (e.g., Seedrs.com) crowdfunding. As an example of how large this market has become, by one estimate, US-based debt crowdfunding platforms, in 2014 alone, facilitated the funding of more than $8.9 billion of loans, and attracted over $1.32 billion of investment from venture capital firms[4]. More than that, the debt crowdfunding market size in China in 2014 is estimated to be about 3.8 times that of the US[5].

It is therefore not surprising to see increasing research interests in crowdfunding in the past few years. Due to their popularity, many studies focus on rewards and debt crowdfunding. Published studies include Lin et al. (2013), Burtch et al. (2013, 2014), Zhang and Liu (2012), Mollick (2014), Pope and Sydnor (2011), and Lin and Viswanathan (forthcoming). For debt crowdfunding, researchers have examined a wide range of mechanisms that address information asymmetry or affect investor behaviors, including social networks (Lin et al. 2013), rational herding (Zhang and Liu 2012), geography (Lin and Viswanathan forthcoming), demographics (Pope and Sydnor 2011) and appearance (Duarte et al. 2012).

Text is the most common feature across all crowdfunding types and platforms. In almost any crowdfunding platform, those seeking funding almost always write a "pitch" in an effort to convince potential investors. Yet, text is also the most fuzzy and least understood aspect of the crowdfunding process. Some existing studies touched upon texts, however they mostly treat text as a control variable, and rely on manual coding of small

[4] http://cdn.crowdfundinsider.com/wp-content/uploads/2015/04/P2P-Lending-Infographic-RealtyShares-2014.jpg

[5] Calculated based on data estimated by LendAcademy.com, AltFi Data, and PCAC; see http://www.lendacademy.com/webinar-chinese-p2p-lending/

samples. Examples include Lin et al. (2013), Michels (2012), and Sonenshein et al. (2011). Given the prevalence and prominence of texts in crowdfunding campaigns, it is highly unlikely that investors will completely ignore them. In addition, research in many fields suggests that linguistic features, or how texts are written, can in fact carry highly valuable information about the writer (Ghose, Ipeirotis, & Li, 2012; Hancock, Curry, Goorha, & Woodworth, 2007; Liu, 2012; Tetlock, 2007). If texts can reflect some otherwise unobservable qualities of the fundraiser, and they can also affect investor behaviors, then they have the potential to become a useful mechanism in mitigating information asymmetry. It is therefore imperative that we systematically investigate the role of texts in crowdfunding so as to ensure market efficiency. Recent development of text analytics and its successful applications in many fields suggest that we can go beyond manual coding of small samples—we should be able to extract text features in a scalable fashion, and investigate their impact on a large scale. We therefore examine whether texts can serve as a useful mechanism to mitigate asymmetric information by asking the following specific research questions:

(1) Do investors consider texts provided by fundraisers?

(2) How do text characteristics relate to the creditworthiness of fundraisers? And if yes,

(3) Can investors correctly interpret the informational value of these texts?

We choose debt crowdfunding as our context to answer these research questions for the following reasons. First, data from debt crowdfunding should be able to provide a very conservative estimate of the impact of texts. Unlike rewards crowdfunding where there is very little standardized quantitative information across different campaigns, borrowers in debt crowdfunding are almost always required to disclose a large amount of information from their personal or business credit files. Such quantitative financial information is the basis for loan underwriting in offline debt finance, so it should already explain a large amount of variation in the data, be it the quality of loans or the probability of successful funding. If characteristics of texts can further explain or predict outcomes after controlling for the financial information, it will suggest that texts can indeed be a powerful tool in online crowdfunding. Second, the "quality" information is well-recorded and objective in debt crowdfunding. When a loan reaches maturity, we can unequivocally determine if the investment opportunity is as good as it appears at the time of funding, and if the borrower is as creditworthy as they claim. By contrast, the outcome can be dramatically different, and therefore hard to compare, across different types of campaigns in rewards crowdfunding (e.g., the success of a charity event vs. a new web service). Finally, even though debt crowdfunding is just one form of crowdfunding, it shares many similarities with other types of crowdfunding: it is still a two-sided market, with similar incentives for each side of the market. That is, fundraisers try to convince investors to allow them to use funds, and investors try to distinguish between different fundraisers to generate returns. Therefore, debt crowdfunding is an ideal context to study these research questions for the broader phenomenon of crowdfunding.

2. Empirical Context

Our data comes from Prosper.com, one of the largest P2P lending sites in the US with more than 2 million members and over $2 billion in funded loans by 2014. We describe a typical lending process, especially features directly related to our study. Additional details of this site are available in published studies using data from the same website, such as Pope and Sydnor (2011), Zhang and Liu (2012) and Lin et al. (2013).

To become either a borrower or lender, one must first provide and verify a valid email address. Then they must then go through an identity verification process by providing information such as their social security number, bank account number, driver's license, and street address. Then the borrower creates a loan request (known as a "listing") that contains information about the borrower and their request. The website anonymizes the borrower's identity to protect their privacy, but extracts information from the borrower's credit reports and displays it on the listing page. Also on the listing page is the information about the borrower's loan request such as the amount of loan requested, the maximum interest rate that they are willing to borrow at (which could be bid down during the auction process), and the format of the auction. The auction format can be "immediate funding" when listings will end as soon as 100% of the requested funds are received; or "open" for a specified number of days so that additional funds can come in to help lower the interest rate. These are typically 36-month loans.

Also on the listing page, borrowers usually provide several paragraphs of free-form texts, where they describe information about themselves, why they need the loan, what they intend to do with the loan, and why they are trustworthy candidates for a loan. This is essentially a "pitch" to all potential lenders. Prosper.com does not verify these texts, and borrowers are free to write anything that they would like to increase their chances of receiving funds.

Before lenders can place bids on listings, they have to first transfer funds from their bank accounts to their non-interest-bearing Prosper.com accounts. These funds do not generate returns until invested in a loan. They can bid as little as $25 and need to specify the minimum interest rates associated with the bid. Prosper.com automatically aggregates all bids in the manner of a second-price proxy auction and the same interest rate applies to all lenders on the same loan.

Successful listings that attract 100% of requested funds become loans after the website's verification. Every month during the life cycle of the loan (usually 36 months), Prosper.com will automatically debit the borrower's bank account and repay the investors after deducting fees. As long as the borrower makes the monthly payment in time, the loan status is marked as "current." Otherwise, the status will incrementally change to "late," "1 month late," "2 months late," etc. If borrower fails to make payment for more than two months, the loan will be marked as "defaulted". Defaulted loans will be transferred to third-party collection agencies designated by Prosper.com, and will negatively affect the borrower's personal credit score.

3. Do Lenders Consider Texts in their Decisions?

Before we study how texts relate to loan repayment or investor behavior, we need to verify if the presence of texts matter at all in this market. The presence of texts does not guarantee that investors will consider them. Even though debt crowdfunding platforms typically allow borrowers to provide texts, the platforms do not verify the contents of these texts. Borrowers could claim one thing in the text, but do something completely different once they receive the loan. Nonetheless, borrowers are legally obliged to repay these personal loans, just as they would repay a loan from the bank. Since these texts are neither verified nor legally binding, lenders do not necessarily have to consider texts for their investment decisions. On the other hand, texts are what makes these requests more personal, and lenders could very well be looking beyond hard credit data in their decision. Therefore, it is an empirical question whether investors look at texts.

To answer this question we exploit two exogenous policy changes on Prosper.com. The first one occurred between May 3, 2010 and June 1, 2010, when Prosper.com unexpectedly switched off the prompts for text descriptions for borrowers at AA and A credit grades[6]. Those borrowers were not able to include texts in their loan requests, whereas other borrowers never noticed any difference. This policy change therefore provides an ideal opportunity to study if lenders really care about texts in a "difference-in-differences" manner: if lenders do not consider texts at all in their decisions, the difference in funding probabilities between the treatment and control groups should be largely the same before and after the policy change.

We use data from April 1st, 2010, to July 1st, 2010, and divide them into three periods: one month before the policy change (T1), one month after the policy change (T2), and one month after the description section reinstated for treatment group (T3). We use propensity score to match borrowers on observable information provided by Prosper.com when borrowers are requesting loans but exclude information extracted from their credit report highly correlated to their credit grades. Then, we calculate the average treatment effects on treated group (ATET) and compare the differences cross these three periods.

Our results in Figure 3-1-1 suggest that the treatment group's funding probability decreased by over 4 percentage points in T2 (compared to T1) when the loan descriptions were removed. This probability reverted to about the same level as T1 after the loan description section had been reinstated in T3. During these periods, the funding probability remained largely the same for the control group. This suggests that investors indeed consider texts provided by borrowers.

[6] This change was initially unannounced but later reversed when investors started complaining, which further confirms that investors do value these texts provided by borrowers.

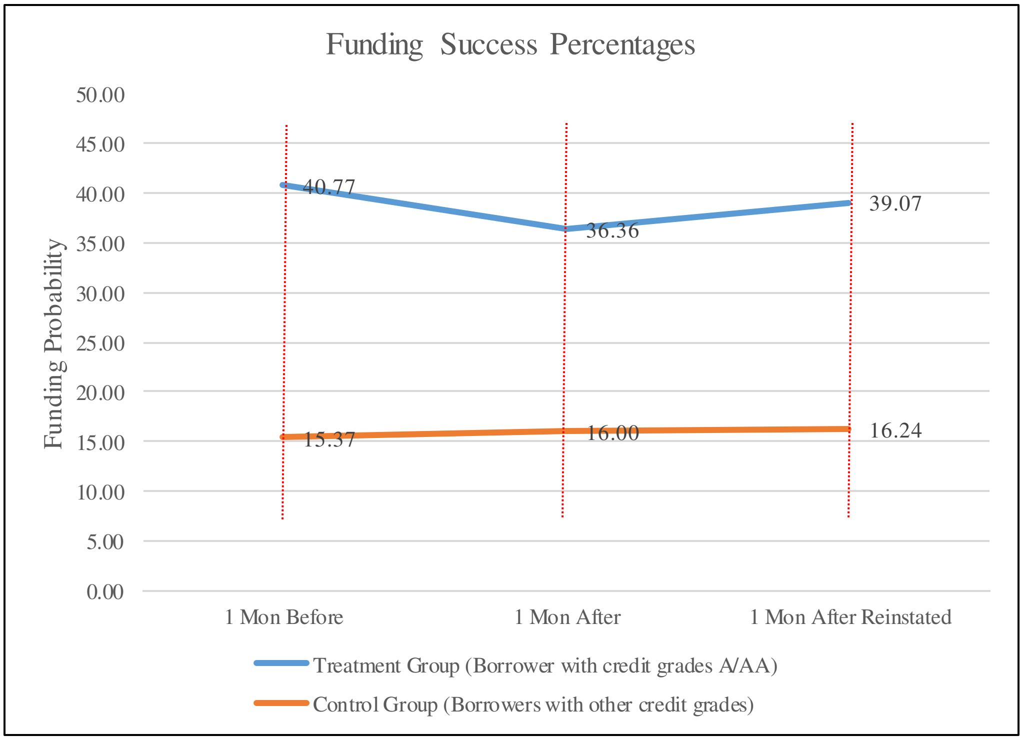


Figure 3-1-1 Funding Probability Before and After Policy Change

A second policy change occurred on September 6, 2013, when Prosper.com removed the text section from all loan requests[7]. This site-wide policy change provides another opportunity to further examine lenders' reaction to texts: if lenders indeed value texts in their decision process, the removal of texts should discourage them from investing. We obtain all loan requests created two months before and after this date. For each day, we calculate the average number of bids that listings received, and check how this number changed in response to the removal of texts. On average, we find that a listing receives 34.16% fewer bids after the removal of texts (Figure 3-1-2). It is easy to see that lenders made fewer bids per listing when texts were removed. These results again show that lenders value texts in their lending decisions.

[7] We contacted the Prosper.com management in May 2014 about this change. In their response, they said that they were in the process of restructuring that part of the request, and expected to restore that feature in a different format. It should be noted that many other P2P lending sites such as LendingClub, or those in UK and China, still include texts with loan requests; therefore this change on Prosper.com does not suggest that texts are becoming less relevant.

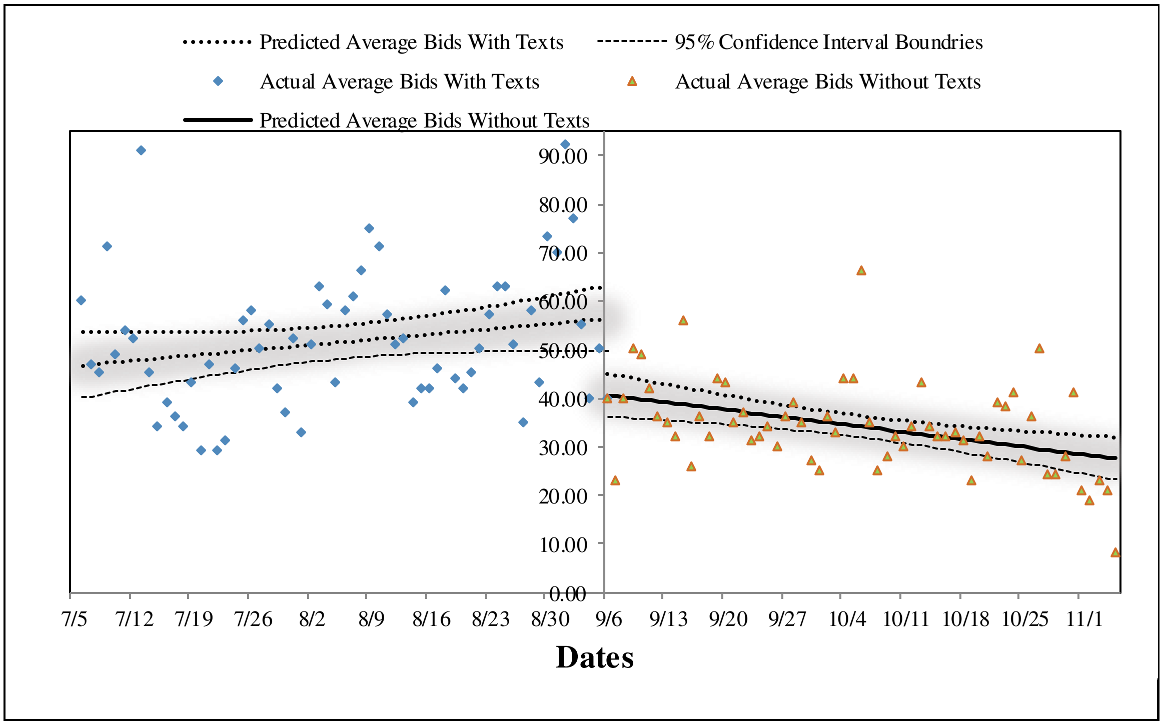


Figure 3-1-2 Average Number of Bids Per Listing, Before and After Policy Change

In addition to the natural experiments, we also examine borrowers who submitted multiple loan requests but not all included texts. This resulted in 25,709 requests from 8,419 borrowers, of which 7.12% of requests with texts were funded, more than double the 3.15% ratio for requests without texts. This and a more detailed borrower panel data model on funding probability (see Appendix A) both further confirm that investors indeed care about texts provided by borrowers.

All the above results show that even though texts on Prosper.com are neither verified nor legally binding, lenders still consider them a valuable and important piece of information in their lending decision process. A natural question, therefore, is whether such behavior can be justified, or if it is rational. This is equivalent to two of our research questions: (1) How are texts related to, and can they predict, loan repayment? (2) Do lenders act on the informational value of texts correctly in the loan funding process? We answer the first question using loan repayment as the outcome of interest, as a function of linguistic features that we extract from the texts. Then we use the funding success of listings as the outcome variable to answer the second question: If lenders correctly interpret the informational value of texts, then linguistic variables that predict higher likelihood of default should also be associated with lower likelihood of funding success. If not, then it suggests opportunities to improve market efficiency.

4. Texts and Loan Repayment

4.1 Characterizing Texts, and the Choice of Linguistic Features

We now investigate if there is actually a link between texts provided in a loan request and the outcome of that loan (i.e. whether or not the loan is repaid[8]). Following common practice in existing literature that examines the role of texts in contexts ranging from financial statements to online messages, we focus on linguistic features of the text, or how texts are written[9].

There is virtually an infinite number of ways to characterize linguistic features of texts. Since our context is online debt crowdfunding where lenders will only lend to someone who can and will repay the debt[10] (Duarte et al., 2012), we focus on linguistic features that have been well established in the literature to reflect the writers' willingness to repay (Flint, 1997) or ability to repay (Duarte et al., 2012). In addition, to ensure that our paper's scope is manageable, we focus on linguistic features that are not only frequently applied (therefore validated) in multiple domains, but also have well-accepted methods or algorithms for extraction.

After an extensive literature search, we identified the following four linguistic features that satisfy all the above requirements: (1) *Readability;* (2) *Positivity*; (3) *Objectivity*; and (4) *Deception Cues*. As will soon become obvious, all four linguistic features have been

[8] In a robustness test we investigate an alternative outcome variable, i.e. the percentage loss of principal. Results are highly consistent.

[9] To the best of our knowledge, few existing studies simultaneously address the content of texts and linguistic features of texts. This is partly due to the fact that text content analysis is still a developing field. We therefore focus on linguistic features. In fact this is reasonable on Prosper.com because the contents of texts is not verified and not legally binding, so its informational value is lower anyway. By contrast, linguistic features are much more difficult to hide or misrepresent, and is independent of context (Zhou et al. 2004). Nevertheless, in one of our robustness tests we control for text contents as well, and the results are overwhelmingly consistent.

[10] Altruistic motivations are possible. However, such motivations are highly unlikely to be predominant on Prosper.com because (1) there are easier places to donate online, such as Kiva or DonorsChoose; and (2) Prosper.com's mechanism dictates that even if all but one lender on a loan is not altruistic (assuming everyone else is willing to demand 0% interest rate), the borrower still has to pay the non-zero interest rate to everyone. In addition, all published studies such as Lin et al. (2013) and Zhang and Liu (2012) show that financial information such as credit grades matters in persuading investors.

well documented in the literature to reflect writer's willingness and ability to repay a debt, and there are also generally accepted methods of quantifying them from texts. However, each of these four features has only been separately examined in non-crowdfunding contexts, raising the question of which feature may dominate when all four are incorporated in the same model. Our context of online debt crowdfunding provides a unique opportunity to understand the relative impacts of these linguistic features.

We now formally develop our hypotheses on how these linguistic features are related to borrowers' ability and willingness to repay and therefore loan repayment likelihood.

4.2. Explanatory Analysis

4.2.1. Hypotheses

Readability of texts refers to the simple idea of how accessible the texts are. Research in a wide range of disciplines has shown that readability is a simple but effective indicator for the writer's capabilities—even companies' financial capabilities. For example, Li (2008) shows that firms with more readable annual financial reports have higher earnings. Former SEC Chairman Christopher Cox suggests that the readability of financial information provided by firms can reveal whether firms are capable of achieving better stock performance and higher earnings (Loughran & McDonald, 2014). In practice, investors can indeed infer the potential performance of firms from the readability of their financial disclosure (Rennekamp, 2012).

If the readability of texts written by public companies when addressing their investors can reflect their financial capabilities, it is only logical to infer that the readability of texts written by individuals when raising funds through crowdfunding will similarly reflect their financial capabilities, or their abilities to repay the principal of the loan plus interest. This is further supported by the literature: less readable texts are more likely to be written by someone who have less education; all else equal, those who have less education are less likely to be gainfully employed, have stable income, and be able to repay their debts. More specifically, Tausczik and Pennebaker (2009) suggests that readability reflects "the education, social status, and social hierarchy of its authors," since those with more years of schooling are more likely to write more readable texts than those who have fewer years of education (Hargittai, 2006). In turn, Card (1999) showed through an expansive literature survey that education causally increases earnings. It is hence not surprising that mortgage lenders use level of education to predict default (Munnell, Tootell, Browne, & McEneaney, 1996). A study even concludes that an additional year of education for household heads results in a predicted decline of 8 percent in the probability of bankruptcy (Fay, Hurst, & White, 2002). Taken together, all else equal, loan descriptions that are more readable are more likely to have been written by borrowers who are better educated and have more stable and higher income (Gregorio & Lee, 2002); they will have higher capabilities to repay and be less likely to default (Campbell & Dietrich, 1983). We therefore hypothesize,

H1: All else equal, loans with more readable text descriptions are less likely to default.

We next turn to positivity. Studies in finance and accounting have consistently found that, firms that show a positive attitude in their financial documents—reflecting their optimism and confidence—typically have better performance (Cardon, Wincent, Singh, & Drnovsek, 2009; Chen, Yao, & Kotha, 2009; Davis, Piger, & Sedor, 2006; Li, 2010). Positive attitude in firms' reports not only indicates better current performance (Li, 2010), but also reflects the firm's optimism. More optimistic businesses, in turn, typically have higher future performance (Davis et al., 2006). In the venture capital literature, entrepreneurs' positive attitude is a strong indicator of their passion and motivation in building their businesses, which in turn are important predictors of how likely they are to be successful when faced with difficulties (Chen et al. 2009; Cardon et al. 2009). In other words, confidence typically reflects both the capability as well as a strong willingness to accomplish goals. On the other hand, the literature also suggests that the positive relationship between positive attitude and success may be curvilinear: Entrepreneurs who are overly confident (suffering from "overconfidence") may tend to fail as well (Dushnitsky, 2010). Since borrowers in online lending markets are also faced with at least some degrees of uncertainty for the duration of a loan, the positivity of texts written by these borrowers should also similarly reflect the capability and willingness to repay their debt with interest in the future. We thus hypothesize,

H2: All else equal, loans with more positive textual descriptions are less likely to default. This relationship should exhibit a curvilinear relationship.

Before we move on to the next linguistic feature, let us address a potential counter-argument to H2 (this will apply to the next few hypotheses as well). One may argue that unscrupulous borrowers may pretend to be optimistic. However, good borrowers will never find it profitable to imitate bad borrowers. As long as not all bad borrowers are all successful in imitating good borrowers, our hypothesis holds. Most important, the reason that linguistic analysis has become so popular is precisely that it is very hard to imitate and misrepresent oneself.

We now turn to *Objectivity*, which captures the extent to which texts are about describing objective information. Information objectivity has long been established as an indicator of quality and trustworthiness (Archak, Ghose, & Ipeirotis, 2007; Chen et al., 2009; Ghose et al., 2012; Ghose & Ipeirotis, 2011; Metzger, 2007). The more objective a piece of text is, the more likely that the writer's claim is based on facts (Metzger 2007), and the more likely that the writer is credible and trustworthy (Chen et al. 2009). As an example, when reviews of a hotel contain more objective information, readers are more likely to trust those reviews and their authors, and are more likely to use those hotels (Ghose and Ipeirotis 2011; Ghose et al. 2012). For the same reason, venture capitalists value objective information over subjective information (Chen et al. 2009) when evaluating entrepreneurs. In our context therefore, all else equal, loans with more objective texts are likely to have been written by someone who is more trustworthy, i.e. good borrowers who are more willing to make an effort to repay debt even when faced with challenges to do so. We therefore hypothesize,

H3: All else equal, loans with <u>more objective</u> textual descriptions are <u>less</u> likely to default.

We now turn to the last but by no means the least linguistic feature, *Deception Cues*. As the name suggests, deception cues refer to "red flags" in the way that texts are written, which may be indicative of intention to deceive or defraud. Deception cues emerge because people who are trying to lie or hide information tend to say or write in a particular manner. Specifically, research in psycholinguistics has shown that fabricated stories are linguistically different from true stories, and contain rich deception cues (Pennebaker, Mehl, & Niederhoffer, 2003). For this reason, deception cues have seen applications in a wide range of contexts, from communications and psychology to forensic science and criminology, as well as information systems (Abbasi & Chen, 2008; Abbasi, Zhang, Zimbra, Chen, & Nunamaker Jr, 2010; Burgoon, Buller, Guerrero, Afifi, & Feldman, 1996; Hancock et al., 2007; Loughran & McDonald, 2011; Pennebaker et al., 2003; Porter & Yuille, 1996; Vrij, 2008; Zhou, Burgoon, Nunamaker, & Twitchell, 2004).

For our context, the intention to deceive is unambiguously related to borrowers' willingness to repay a debt, so deception cue is an important linguistic feature to consider in understanding borrowers' texts. All else equal, if the text provided by a borrower has rich deception cues, that may indicate that the borrower is trying to provide false or exaggerated information to convince potential lenders. Their actual capability or willingness to repay will be likely much lower than they hope the investors would believe. We thus hypothesize,

H4: All else equal, loans with more deception cues in their textual descriptions are more likely to default.

4.2.2. Data and Variables

We gathered all loans funded on Prosper.com between January 1st, 2007 and May 1st, 2012. All loans during this time were three-year loans; therefore as of the time of our study, we can gather objective information on whether these loans were repaid or not at the end. This is essentially the ultimate resolution of the quality uncertainty about borrowers that lenders were faced with at time of lending. Hence, our primary outcome variable is whether a loan is defaulted at the end of its life cycle. Our dataset includes 34,110 loans, of which 11,899 were defaulted. Next, we discuss how we extract and quantify linguistic features from texts provided by borrowers at the time of their requests[11].

[11] We attempted factor analysis to see if different dimensions of linguistic features can naturally load onto several factors, which would have enabled us to use fewer variables to capture most key linguistic features. The results show that many variables do not meet the reliability criterion, which requires four or more loadings of at least 0.6 (Field, 2009). Factor analysis is therefore not appropriate here, so we retained the original variables.

4.2.2.1. Readability[12]

Drawing on existing literature (Archak, Ghose, & Ipeirotis, 2011; Ghose et al., 2012; Ghose & Ipeirotis, 2007, 2011; Li, 2008; Loughran & McDonald, 2014; Mollick, 2014; Rennekamp, 2012), we measure readability in three dimensions: *spelling errors, grammatical errors, and lexical complexity*. The first two dimensions are based on the simple idea that if a text contains more spelling and grammatical errors, it is less readable (Archak et al., 2011; Ghose et al., 2012; Ghose & Ipeirotis, 2011; Li, 2008; Loughran & McDonald, 2014). As is common in natural language processing, we use the spelling error corpus to identify *spelling errors* (Jurafsky & James, 2000). The spelling error variable is measured by the total number of spelling errors in a loan description. We use Stanford statistical Probabilistic Context Free Grammar (PCFG) parser to measure *grammatical errors* (Klein & Manning, 2003) by quantifying probabilistically how far the text is from correct grammatical structures in the parser's large, hand-coded database[13].

The third dimension or metric of readability that we use is the well-known *Gunning-Fog Index (FOG).* The *FOG* index was first constructed by Gunning (1969) to evaluate the lexical complexity of texts and is the most commonly used metric to measure the complexity of financial reports in financial market (Li, 2008). Based on DuBay (2004), the formula for FOG is:

$$FOG\ Score = 0.4 \times (ASL + 100 \times AHW)$$

Here, *ASL* denotes the Average Sentence Length in the text; *AHW* (Average Hard Words) denotes the proportion of words containing more than two syllables ("hard words") for every 100 words in a loan description.

4.2.2.2. Positivity and Objectivity[14]

Since the texts in our dataset are in a specific domain (lending), we use a machine learning approach rather than lexicon-based approach for our *positivity* and *objectivity* analyses. This will maximize the accuracy of linguistic feature extraction (Pang & Lee, 2008). We first created a stratified random sample of loans (1% from each credit grade). Then we put

[12] Details about how we construct readability variables are in Appendix C.
[13] More details for the measurement of grammatical error is in Appendix C.
[14] More details for the construction of positivity and objectivity variables are in Appendix C.

70% of the sample into a training dataset, and the rest into a testing dataset. Two research assistants manually coded the textual descriptions of these loans.

To measure the positivity of each loan description, we follow the supervised approach as described in Pang, Lee, and Vaithyanathan (2002). Specifically, our classifier uses a combination of unigram (single word) and the word's corresponding part-of-speech tag (or POS tag, i.e., whether that word serves as a subject, verb, or object in its sentence). This classifier estimates a probability that each sentence (of each loan description) is positive; rather than "binning" it into positive, negative, or neutral. This is the same procedure as used in Ghose and Ipeirotis (2011). The positivity score of a loan description is then defined by the average of those probabilities across sentences.

We now turn to objectivity. We build our classifier using features similar to those used in Barbosa and Feng (2010), including the numbers of polarity words (negative, positive, and neutral words), strong or weak subjective words, modal words (such as "can" and "would"), numbers, unigrams, unigram-POS tag combinations, bi-char (two adjacent characters), adjectives, and adverbs. We then use this classifier to classify all texts. Similar to positivity, we extract the objectivity probability of each sentence, and average these probabilities to obtain the objectivity score for each loan description.

4.2.2.3. Deceptive Cues[15]

For deception cues, we closely follow established methods of deception detection (Hancock et al., 2007; Toma & Hancock, 2012; Zhou, Burgoon, Twitchell, Qin, & Nunamaker Jr, 2004) and measure deception cues in four dimensions: *cognitive load*, *internal imagination*, *negative emotion*, and *dissociation*.

Linguistic deception cues fall into two general categories: *nonstrategic* linguistic cues, which reflect psychological processes experienced by deceivers and are likely to be present without the deceiver's intention to show them; and *strategic* cues that are strategically and purposely used by deceivers (Toma & Hancock, 2012). A common nonstrategic cue, *cognitive load,* is based on cognitive theory for nonverbal contents (Vrij, Fisher, Mann, & Leal, 2008): Deceivers need to invest more cognitive resources, because they not only have

[15] More detailed discussions on deception detection can be found in Hancock et al. (2007) and Zhou et al. (2004). Appendix C provides more details about quantification of these cues.

to fabricate information or stories that do not exist or never happened, but also to avoid detection, a process that generates higher cognitive load (Vrij, 2000) and leads to less complex stories (Newman, Pennebaker, Berry, & Richards, 2003). To measure cognitive load, we use *concreteness*, which will be higher for fabricated stories due to heightened cognitive burden (Newman et al., 2003). The concreteness value of each loan description is calculated as the mean value of the concreteness of all content words, using the MRC Psycholinguistic Database described in Wilson (1988).

The second dimension of deception cues is *internal imaginations.* Reality monitoring theory states that events from external experience (i.e. real experience) contain more temporal and spatial information than events generated from internal imaginations (Johnson & Raye, 1981). The literature therefore uses *temporal* and *spatial information* to capture internal imaginations, which is lower in fabricated stories. We measure temporal information by combining results derived from two well-known tools, the Stanford SUTime parser (Chang & Manning, 2012) and the time component from LIWC (Linguistic Inquirer and Word Count) (Pennebaker, Francis, & Booth, 2001). We similarly measured spatial information by combining results from two approaches: Stanford name entity recognizer (Finkel, Grenager, & Manning, 2005) and LIWC space words (Pennebaker et al., 2001).

A third common nonstrategic deception cue is *negative emotion.* The act of lying leads to a wide range of negative emotions (e.g., anxiety, sadness, and guilt) (Toma & Hancock, 2012; Vrij, 2000). The literature routinely quantify negation emotion from two sources: *content negation word* (Hancock et al., 2007) and *functional negation word* (Toma & Hancock, 2012)[16]. We consider both.

The last deception cue dimension is a strategic deception cue: *dissociation*. Deceivers tend to use more non-first person pronouns (e.g., "he", "him", or "her") in their writings in order to *dissociate* themselves from their fabricated stories (Hancock et al., 2007; Newman et al., 2003; Toma & Hancock, 2012). To measure this, we follow the literature and compute the percentage of non-first person pronouns in a text (Hancock et al., 2007).

[16] Content negation word are negating words such as "not" and "never". Functional negation words are words that are semantically negative.

4.2.2.4. Control Variables

Following prior studies that use data from Prosper.com (Freedman & Jin, 2008; Lin & Viswanathan, 2015; Michels, 2012), we use three groups of control variables: hard information (e.g., credit grade and debt-to-income-ratio), auction information (e.g., loan amount and loan category), and social information (i.e., group membership and friend investment). These groups of information include almost all information lender can observe on loan request web page when borrowers are requesting loans. We call it "standard information". We also include monthly dummies. Detailed variable descriptions and summary statistics are provided in Appendix C.

4.2.3. Empirical Models

As discussed in H1-H4, if linguistic features can help predict loan default, then all else equal, loans with text descriptions that are less readable, less positive, less objective, and show more "red flags" as evidenced by deception cues, should be more likely to default. We now empirically verify these conjectures by testing the relationship between linguistic features of the loan request descriptions, and the probability that the loan defaulted. Our unit of analysis is each loan. The dependent variable is whether or not a loan defaulted at the end. The main independent variables are the linguistic feature metrics that we described in previous section. In addition, we include a large number of control variables as described in the previous section. Since the dependent variable is binary, we use probit models with robust standard errors (results from logistic models are highly consistent). To address selection issues, we estimate a Heckman model with the two-step procedure (Heckman, 1979). To better illustrate the impact of individual linguistic features, we incrementally add different features and estimate the following models[17]:

Model 1: (Readability)

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times\ ControlVariables\ _i + \varepsilon_i$

Model 2: (Model 1 + Positivity)

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times\ ControlVariables$

[17] As discussed later in the paper, the sequence of these linguistic features does not matter.

$$+ \varepsilon_i$$

Model 3: (Model 2 + Objectivity)

$$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times Objectivity_i$$

$$+ \alpha_4 \times\ ControlVariables\ _i + \varepsilon_i$$

Model 4: (Model 3 + Deception Cues)

$$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times Objectivity_i$$

$$+ \alpha_4 \times\ Deception_i + \alpha_5 \times\ ControlVariables\ _i + \varepsilon_i$$

In these models, Readability, Positivity, Objectivity and Deception are all vectors of metrics for linguistic features in each category[18].

4.2.4. Explanatory Model Results

We report estimation results and marginal effects (holding other variables at their mean) in Table 3-1-1. Table 3-1-2 provides a summary, and we discuss these findings in detail next.

Table 3-1-1 Coefficients and Marginal Effects of Default Probability Models

Variables	Model 1		Model 2		Model 3		Model 4	
	Coeff	*ME*	*Coeff*	*ME*	*Coeff*	*ME*	*Coeff*	*ME*
Spelling errors	0.0714***	0.0235***	0.0433**	0.0143**	0.0697***	0.0231***	0.0700***	0.023***
	(0.0180)	(0.00590)	(0.0178)	(0.00589)	(0.0179)	(0.00591)	(0.0181)	(0.00592)
Grammatical errors	5.67E-4***	1.86E-4***	5.87E-4**	1.94E-4**	5.28E-4***	1.75E-4***	5.77E-4***	1.89E-4***
	(9.37e-05)	(3.08e-05)	(8.99e-05)	(2.97e-05)	(9.04e-05)	(2.98e-05)	(9.44e-05)	(3.09e-05)
Lexical Complexity (FOG)	0.103***	0.0339***	0.0560**	0.0185**	0.0831***	0.0275***	0.109***	0.036***
	(0.0288)	(0.00946)	(0.0247)	(0.00817)	(0.0261)	(0.00862)	(0.0294)	(0.00963)
Positivity			-0.101**	-0.0882**	-0.144***	-0.0506***	-0.627***	-0.206***
			(0.0494)	(0.0166)	(0.031)	(0.00260)	(0.222)	(0.0729)
Overconfidence			0.0714***	-0.006***	-0.00827	-0.00273	0.00591	0.012
			(0.0180)	(0.00834)	(0.0239)	(0.00791)	(0.0217)	(0.0219)
Objectivity					-0.0112***	-0.0037***	-0.012***	-0.004***
					(0.00308)	(0.00102)	(0.00311)	(0.00102)
Concreteness							0.160***	0.0523***
							(0.0287)	(0.00941)
Non-first-person pronouns							0.503***	0.319***
							(0.155)	(0.186)

[18] Our results of multicollinearity analysis show no strong correlations among these variables. Please see details in Appendix C. We will also discuss the validity of the nested structure of these models in Section 4.2.4.5.

Spatial Information							-0.043***	-0.042***
							(0.0123)	(0.0109)
Temporal Information							-0.049***	-0.016***
							(0.0124)	(0.00406)
Negative Emotion							0.016***	0.005***
							(0.00428)	(0.00140)
Observations	32,052		32,052		32,052		32,052	

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.010)
2. *HI*: Hypothesized impact; *ME*: marginal effects; *Coeff*: Coefficients.
3. For credit grades, AA is the baseline.
4. Control variables were included but not reported for brevity.

Table 3-1-2 Key Findings of Explanatory Analyses

H	Relation	Finding	Comments
H1	Readability vs. Default Rate	Supported	Requests that are less lexical ease of read and have less spelling and grammatical errors are less likely to default.
H2	Positivity vs. Default Rate	Partially supported	Positive requests are less likely to default, though we did not find evidence of a curvilinear relationship
H3	Objectivity vs. Default Rate	Supported	Objective requests are less likely to default.
H4	Deception vs. Default Rate	Supported	Requests that contain more non-1st person pronouns, more negation words, less spatial and temporal information and that are higher in concreteness are more likely to defaults.

4.2.4.1. Readability

We measure readability using three metrics: spelling error, grammatical error, and lexical complexity (FOG score). Results in all models show that all three metrics are statically significant and consistent with our hypothesis H1: Loans with descriptions that contain fewer spelling and grammatical errors, and less complicated, are less likely to default. As shown in Table 3-1-1, in the full model (Model 4) which includes all linguistic features, one unit decrease in spelling error is associated with a 2.30% reduction in the loan’s default probability, whereas one unit increase in lexical complexity (FOG) is associated with 3.56% higher probability of default.

4.2.4.2. Positivity

Consistent with the main effect hypothesized in H2, we find that loans with more positive descriptions are indeed less likely to default, since the coefficient on "positivity" is negative and statistically significant. From Model 4 of Table 3-1-1, we see that when the positivity score increases by one unit, the corresponding loan is 20.6% less likely to default. On the other hand, there does not seem to be sufficient support for the "overconfidence" hypothesis (second part of H2): The quadratic term of positivity is statistically significant only when other linguistic features are not considered (Model 2, not in Model 4). H2 is therefore partially supported.

4.2.4.3. Objectivity

We hypothesize that requests with more objective information should be less likely to default (H3), and our results indeed support it. The coefficient on objectivity is negative and significant, suggesting that all else equal, loans with more objective information are indeed less likely to default. In terms of effect size, one unit increase in objectivity score is associated with 0.396% reduction in loan default (the full model).

4.2.4.4. Deception Cues

We measure deception cues along four distinct dimensions: *cognitive load*, *internal imagination*, *disassociation,* and *negative emotion*. We discuss each dimension in turn below.

Cognitive load. Loan requests higher in cognitive load, as measured by higher concreteness measures, are more likely to default. This result is consistent with existing deception detection research and with our hypothesis. The marginal value of concreteness in Model 4 of Table 3-1 shows that all else equal, the default probability of a loan will be 5.23% higher when concreteness is one unit higher.

Internal imagination. Our hypothesis is that descriptions low in spatial and temporal information should be more likely to default, indicating a negative coefficient. Results on both spatial and temporal information are indeed consistent with this hypothesis.

Disassociation and *negative emotion*. While the coefficient for the number of negative emotion words is statistically significant, its marginal effect is relatively small. On the other hand, *dissociation* shows stronger association with default. A 1% increase in non-first person pronouns in the loan description is associated with a 31.9% increase in default probability.

The above results show that all deception cues identified in prior literature are still valuable cues in identifying "lemons" in online debt crowdfunding. Therefore, H4 is supported.

4.2.4.5. Control Variables and the Nested Structure of Our Explanatory Models

Variables on non-text information in our models yield consistent results as prior literature. For example, loans with higher credit grades are less likely to default, whereas loan requests that use immediate funding are more likely to default. This also attests to the validity of our model.

In addition, it is easy to see that our models (Models 1-4) are nested within each other. We conduct likelihood ratio tests for each adjacent pairs of models (e.g. Models 1 vs. 2, 2 vs. 3, and then Models 3 vs. 4.), and find statistically significant support that it is valid to add additional linguistic features. This is not surprising since these linguistic features capture different aspects of texts that are largely independent of each other. For that same reason, the sequence of nesting is in fact immaterial. That is, we could use deception cues as the first feature, then positivity, then objectivity, and so on; the likelihood ratio tests results are still supportive.

We now verify the robustness of our findings through a series of additional tests, which we describe in the next section.

4.2.5. Robustness and Generalizability of the Explanatory Model

4.2.5.1. Concerns for Unobservable Variables

To ensure that our findings concerning the linguistic feature variables are not driven by unobserved variables, we conduct two additional tests that are independent of each other.

In our first test, we construct a panel dataset by focusing on borrowers who had multiple loans in our sample. By using a borrower fixed effect model we will be able to account for borrower-level unobservables. This dataset contains 9,809 loans from 2,419 borrowers, and the number of loans for each borrower ranges from 2 to 7. We use the sequence of loans within borrowers as the time variable. Results from the model, reported in Table 3-1-3, are consistent with our findings in the main analysis. We also test for autocorrelation in this panel data model using the approach suggested by Wooldridge (2010), and find no evidence of autocorrelation.

Table 3-1-3 Borrower Panel Model Results

Variables	Borrower Panel Coeff	Odd Ratios
Spelling errors	0.425** (0.215)	1.530** (0.328)
Grammatical errors	0.00248*** (8.24e-4)	1.002*** (8.26e-4)
Lexical complexity (FOG)	0.0301* (0.0183)	1.031* (0.0188)
Positivity	-0.227* (1.220)	-1.254* (1.530)
Overconfidence	-0.160 (0.389)	-0.852 (0.332)
Objectivity	-0.182*** (0.0404)	-0.833*** (0.0337)
Concreteness	-0.104* (0.226)	0.902* (0.204)
Non-first-person pronouns	0.0867* (0.0518)	1.091* (0.0565)
Spatial information	-0.136* (0.0819)	-0.873* (0.0715)
Temporal information	-0.311** (0.143)	-0.733** (0.105)
Negation emotion	0.0617 (0.0479)	1.064 (0.0509)

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.010)
2. Control variables were included but not reported for brevity.

In a second test, we use an instrumental variable approach. More specifically, we use the linguistic feature metrics of the focal borrower's friends on Prosper.com (c.f. Lin et al. 2013) who are also borrowers themselves. Linguistic features of borrower friends should be valid instruments for a borrower's own linguistic features, for two reasons. First, the expansive literature on homophily suggests that friends are likely to have similar backgrounds, so their linguistic features should be correlated with each other. Second, a borrower friend's linguistic features cannot directly affect a focal borrower's loan repayment. This is consistent with an auxiliary finding in Lin et al. (2013) that the number of friends who are borrowers themselves has no significant bearing on any outcome of loans. Therefore, we compute each of the linguistic feature values for each of the focal borrower's friends who are borrowers, and use the median value across a focal borrower's friends and use that value as an instrument for the focal borrower's linguistic feature metric[19]. Due to the large number of variables and since our goal is to check for robustness of our findings, we instrument for one dimension of linguistic features at a time (i.e., testing the robustness of one hypothesis at a time). Results, as reported in Table 3-1-4, are mostly consistent with our main findings.

Table 3-1-4 Instrumental Variable Model Coefficients

Variables	R-IV	P-IV	O-IV	D-IV
Spelling errors	0.0821***	0.0539***	0.0148***	0.0333**
	(0.393)	(0.0175)	(0.0286)	(0.0240)
Grammatical errors	0.000416***	0.000571**	0.000136*	0.000468**
	(7.53e-05)	(0.000252)	(0.000142)	(0.000221)
Lexical complexity (FOG)	0.00543**	0.0421**	0.0283*	0.000893*
	(0.00217)	(0.00345)	(0.00272)	(0.00432)
Positivity	-0.568***	-0.408**	-0.225**	-0.400**
	(0.115)	(1.617)	(0.108)	(0.164)
Overconfidence	0.144***	0.0130**	0.0153	0.0163
	(0.0442)	(0.0325)	(0.0895)	(0.0370)
Objectivity	-0.00463*	-0.0333**	-0.0203*	-0.0136***
	(0.00419)	(0.00040)	(0.000105)	(0.00496)
Concreteness	0.168***	0.0208**	0.0138***	0.0979**
	(0.0219)	(0.287)	(0.0362)	(0.0423)
Spatial information	-0.0459***	-0.0508**	-0.0720***	-0.0753**
	(0.00722)	(0.0198)	(0.00978)	(0.0293)
Temporal information	-0.0396	-0.0250***	-0.0311***	-0.0350**
	(0.0249)	(0.0244)	(0.0378)	(0.0152)
Non-first-person pronouns	0.0252***	0.0186**	0.0109*	0.0284***
	(0.00654)	(0.0176)	(0.00795)	(0.0221)
Negative emotion	0.0229***	0.0484**	0.0117***	0.0168**
	(0.00398)	(0.0214)	(0.0428)	(0.142)

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.010).
2. R-IV: readability instrumental variable model; O-IV: objectivity instrumental variable model; P-IV: positivity instrumental variable model; D-IV: deception instrumental variable model.
3. Control variables are included but not reported for brevity.

[19] To further ensure that linguistic features of friends are correlated with those of the focal borrowers, if the borrower has friends who are in the same credit grade, we focus on those borrowers only in this calculation. However this refinement is not critical; results are similar when we do not impose this.

4.2.5.2. Generalizability: Replication using data from LendingClub.com

One possible critique of our explanatory model is that we only use data from Prosper.com. While Prosper.com data has been used widely in many published studies, to further ensure the generalizability of our findings, we replicate our explanatory model using data from another major peer-to-peer lending site in the US, LendingClub.com. We obtained information on all 46,290 LendingClub loans that were originated between Jan 1, 2008 and December 31, 2013. We then use the same method to extract linguistic features and construct explanatory models.

Results reported in Table 3-1-5 show that the majority of our findings still hold in this new dataset. The only exception is grammatical errors, which is insignificant. However, this can be attributed to website policies: Texts on LendingClub.com are much shorter (an average of 46 words per loan) than Prosper.com (an average of 135 words). Under such space constraints, it is less likely for borrowers to make grammatical errors. The empirical variation of this variable is therefore smaller and less likely to be statistically significant.

Table 3-1-5 LendingClub.com Repayment Probability Model

Variables	LC (Coeff)	LC(ME)
Spelling errors	0.0713*** (0.0224)	0.0170*** (0.00532)
Grammatical errors	-0.000415 (0.000315)	-9.88e-05 (7.50e-05)
Lexical complexity (FOG)	0.0167*** (0.00489)	0.00398*** (0.00116)
Positivity	-0.0167123* (0.00488)	-0.0262929* (0.00713)
Overconfidence	0.006563 (0.000273)	0.01192 (0.000120)
Objectivity	-0.0113238** (0.00393)	-0.036001** (0.00997)
Concreteness	0.000656** (0.000273)	0.000156** (6.50e-05)
Non-first-person pronouns	0.0230*** (0.00664)	0.00547*** (0.00158)
Spatial Information	-0.0113*** (0.00393)	-0.00269*** (0.000936)
Temporal Information	-0.0175*** (0.00422)	-0.00417*** (0.00100)

Negation Words	0.0272*** (0.00560)	0.00647*** (0.00133)
Observations	46,280	46,280

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.010)
2. LC (Coeff): LendingClub model coefficients; LC(ME): LendingClub marginal effects.
3. Control variables were included but not reported for brevity.

4.2.5.3. Robustness: Loan Loss Percentage as an Alternative Outcome Variable

Our main model uses a binary indicator (defaulted or not) as the outcome variable. We now examine an alternative dependent variable, i.e., the percentage of principal lost on a loan. If there is no default, this number will be zero. We therefore estimated a Tobit model with the same set of independent variables as our repayment probability model. We estimated the standard errors using the robust Huber-White sandwich estimator. As shown in Table 3-1-6, linguistic feature results are again all qualitatively consistent with our main model. Additionally, the squared term of *positivity* is now statistically significant, showing that overconfident borrowers are indeed more likely to default when we use more fined-grained measurement of loan performance (H2).

4.2.5.4. Robustness: Controlling for Text Contents

Published papers that focus on linguistic features rarely, if ever, control for contents of texts. This is partly because automated text content coding is still a developing field. As mentioned earlier, the contents of texts that borrowers write on Prosper.com are neither verified by the platform, nor legally binding on borrowers. Nevertheless, using small samples of manually coded texts, several studies on peer-to-peer lending suggest that the content, or *what* the borrowers write, can still be indicative of loan quality (Herzenstein, Sonenshein, & Dholakia, 2011; Michels, 2012)[20]. We now investigate if our results still hold when we account for contents of texts.

Due to the scale of our dataset, manual coding is not possible. To automate the process of content detection and extraction, we implement the Latent Dirichlet Allocation (LDA) topic modeling approach to extract major topics of loan texts, following Blei, Ng, and Jordan (2003). We identify six major topics—*expense and income, education, employment,*

[20] It should be also noted that these studies do not consider linguistic features.

business, family, and *credit history*—and classify loan description texts into corresponding topics[21].

We add these content variables to our loan repayment model, and report the results in Table 3-1-6. It should be noted that linguistic features and linguistic contents capture different things, but to be cautious, we checked for multicollinearity in this new model, and found no correlations between linguistic features and text contents. Most important, our results remain mostly consistent after controlling for the contents of texts.

Table 3-1-6 Results of Robustness Tests

Variables	Principle Percentage Loss Model	Repayment Model With Topics	
		Coeff	Marginal Effects
Spelling errors	1.620***	0.0451**	0.0146**
	(0.495)	(0.0183)	(0.00592)
Grammatical errors	4.6e-3*	0.000439***	0.000142***
	(0.00251)	(9.58e-05)	(3.10e-05)
Lexical complexity (FOG)	0.518*	0.0999***	0.0324***
	(0.793)	(0.0300)	(0.00970)
Positivity	-8.77**	-0.748***	-0.243***
	(3.888)	(0.225)	(0.0728)
Overconfidence	9.587***	0.0108	0.0230
	(1.187)	(0.0217)	(0.0218)
Objectivity	-0.453***	-0.0100***	-0.00326***
	(0.0895)	(0.00319)	(0.00103)
Concreteness	1.217**	0.0937***	0.0303***
	(0.553)	(0.0293)	(0.00949)
Spatial information	-0.617***	0.027***	0.0877***
	(0.209)	(0.00490)	(0.00159)
Temporal information	-1.167***	-0.0316***	-0.0103***
	(0.339)	(0.00756)	(0.00245)
Negation emotion	0.199*	-0.0392***	-0.0127***
	(0.114)	(0.0126)	(0.00409)
Non-first-person pronouns	0.621***	0.0103**	0.00334**

[21] Latent Dirichlet Allocation (LDA) is a generative probabilistic model that is commonly used for topic modelling. It models documents as a mixture of latent topics, where each topic is defined to be a distribution over words and has different probabilities (Blei, 2003). In our study, we first used LDA model to derive the possible topics of loan descriptions, then defined the topic of a description by selecting one topic that possesses the highest probability. LDA is among the most successful recent learning models, but can be heavily domain-specific due to the bag-of-words used for topics (Blei, 2002).

	(0.128)	(0.00433)	(0.00140)
Topic 1 (education)		-0.0514	-0.0166
		(0.0474)	(0.0153)
Topic 2 (employment)		0.127***	0.0423***
		(0.0444)	(0.0146)
Topic 3 (Business)		0.114***	0.0379***
		(0.0441)	(0.0145)
Topic 4 (family)		0.105**	0.0349**
		(0.0420)	(0.0138)
Topic 5 (credit history)		-0.151***	-0.0476***
		(0.0397)	(0.0129)

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.010)
2. For credit grades, AA is the baseline.
3. Control variables were included but not reported for brevity.

4.2.5.5. Discussion: Can We Trust These Texts?

One possible concern for the explanatory model (and perhaps also the predictive model later) is that these texts may not have been written by the borrower themselves. Borrowers may ask others to "package" their descriptions to increase the chances of receiving loans. In addition, borrowers may try to match the language of funded loans from the past. While these may appear plausible, they are not the case in our study. The most important reason is that: a direct test of the above scenarios is exactly what we do in this paper. If texts are not written by borrower themselves or are the result of successfully imitating someone else, then the linguistic features that we study will not be significant in explanatory models, and adding these variables will not improve the prediction models that we will discuss in the next section. But we have shown that linguistic features do explain loan outcome, and we will show that incorporating them improves the performance of prediction models. These scenarios therefore, are unlikely to be a first order concern in our study.

There are some other minor but still important reasons why these two scenarios are not likely. For the first scenario, note that debts on Prosper.com are personal loans, which is usually a very private matter (this is the reason that Prosper.com does not allow personally identifiable information on listings). So even if the borrower is comfortable asking someone else to edit, that someone else is most likely from the borrower's immediate family or social circle, who will still share similar economic and social status (and therefore capability and willingness to repay) as the borrower themselves. In that case, the results will not change. For the second scenario, if it were true, we should observe that the average values of the linguistic feature variables should all be improving over time. That did not turn out to be true either; we do not observe any clear upward or downward trend in any of the linguistic feature variables. And finally, perhaps addressing both scenarios, one most likely venue where the borrower may have had help with the descriptions is when they are part of a group on Prosper.com. It is most likely when the group leader can financially benefit from the successful funding of the loan. Prosper.com allowed this "commission" for a short while, which is documented in Hildebrand et al. (2013). We create a dummy variable to capture listings where group leaders have a non-zero reward, and interact that variable with linguistic feature variables. Results show that these interaction terms are not statistically significant. Therefore, even when the borrower was most likely to have received direct help in the writing process (due to financial rewards for the group leaders), we find no evidence that linguistic feature variables have lower explanatory power.

All above analyses demonstrate the economic value of linguistic features. A natural question then is: can investors correctly interpret these linguistic features in their investment decisions?

5. Can Investors Correctly Interpret Linguistic Features?

If investors were able to correctly interpret the information value of linguistic features as we have found in this paper, then we should observe that linguistic features that predict higher likelihood of default should also predict lower probability of funding *before* the request became a loan. This is essentially a linguistic "*efficient market hypothesis.*" If not, it will suggest that there are still arbitrage opportunities in the market that are not fully taken advantage of. Whether our data confirms or refutes this hypothesis, testing the above conjecture has important implications for practitioners and policymakers.

We therefore examine how linguistic features of loan requests are associated with their likelihood of successful funding. To this end, we expand our dataset to include all loan requests posted by borrowers on Prosper.com during the same period of time, regardless of whether the requests were funded and became loans or not. We then use the same method to extract linguistic features for all these requests. Our unit of analysis is each loan request, and the main dependent variable is whether or not a request was successfully funded. Our independent variables remain the same as our previous explanatory model[22], including the linguistic features. We report summary statistics of data used in this model in Appendix D. Due to the binary nature of outcome variable (funded = 1), we estimate the following probit model (logit model yields very similar results):

$$\textit{Probability (Funded=1)} = \beta_0 + \beta_1 \times \textit{Readability}_i + \beta_2 \times \textit{Sentiment}_i + \beta_3 \times \textit{Subjectivity}_i$$

$$+ \beta_4 \times \textit{Deception}_i + \beta_5 \times \textit{ControlVariables}_{\ i} + \zeta_i$$

Results and a summary of findings of models 5 are reported in Tables 3-1-7 and 3-1-8, respectively.

Table 3-1-7 Funding Success Results

Variables	Probit Funding Coeff	Funding success Marginal Effects
Spelling errors	-0.0700***	-0.0142***
	(0.00735)	(0.00149)
Grammatical errors	-0.000213***	-4.31e-05***
	(1.25e-05)	(2.52e-06)
Lexical complexity (FOG)	-0.0196**	-0.00397**
	(0.00930)	(0.00188)
Positivity	0.715***	0.145***
	(0.0944)	(0.0191)
Overconfidence	-0.131***	-0.0265***
	(0.0281)	(0.00568)
Objectivity	-0.0199***	-0.00404***
	(0.00136)	(0.000275)
Concreteness	-0.0358***	-0.00725***
	(0.0106)	(0.00214)
Spatial information	0.0209***	0.00422***
	(0.00181)	(0.000365)
Temporal information	0.000344***	6.97e-05***
	(9.41e-05)	(1.90e-05)
Non-first-person pronouns	0.0289***	0.00585***
	(0.00464)	(0.000939)
Negative emotion	0.00442***	0.000894***

[22] We use the same set of variables so that we can easily compare the findings from both funding success and loan default, and better interpret whether investors have fully utilized these valuable "linguistic signals." Our approach that includes the same set of variables for both funding success and default is consistent with previous studies (Lin et al. 2013; Lin and Viswanathan 2015; Pope and Sydnor, 2011).

	(0.00166)	(0.000335)
Observations	317,692	317,692

Notes:

1. Robust standard errors in parentheses (* p<0.10 ** p<0.05 *** p<0.01)
2. Control variables and intercept were included but not reported for brevity.

Table 3-1-8 Key Findings of Funding Probability Model

Relations	Comments
Readability vs. Funding Probability	Requests that contain less spelling and grammatical errors are more likely to be funded.
Positivity vs. Funding Probability	Positive requests are more likely to be funded. Overconfident ones are less likely to be funded
Objectivity vs. Funding Probability	Requests that contain more objective information are less likely to be funded.
Deception vs. Funding Probability	Investors can only correctly interpret spatial and temporal information.

Some interesting patterns emerge from this analysis. We find that our linguistic "efficient market hypothesis" is supported for some features, but not all. The features that investors are generally able to correctly interpret include part of *readability* and *positivity*. For example, loans that are easier to read (fewer spelling and grammatical errors), which are less likely to default (H1), are indeed more likely to be funded.

On the other hand, for objectivity and deception cues, investor behaviors are often contrary to the repayment outcome of the loans. Most remarkably and quite unfortunately, some of the *deception cues* are indeed able to successfully "deceive" investors. Specifically, investors are *more* likely to fund loans that have more non-first-person pronounces and negation words, even though these deception cues are well established in the literature (Pennebaker et al., 2003). Out of the many deception cues, one saving grace is that investors are indeed able to interpret the value of temporal and spatial information, as they are less likely to fund a loan when it is lower on those metrics. We also find that investors are less likely to invest in loans that have higher *objectivity* scores, even though our previous tests show that loans that score higher on this aspect are less likely to default. This appears to suggest a behavioral bias among investors, which means that they can be swayed by subjective and emotional appeals, a finding that echoes Lin and Viswanathan (forthcoming).

To ensure the robustness of these findings, we also estimate an instrumental variable model for funding probability, using the same instrument (borrower friend linguistic features) as we did in Section 4.2.5.1. Results again remain highly consistent[23].

To sum up, while investors are able to correctly interpret the value of some linguistic features such as readability and sentiment, they are more susceptible to manipulations by or ignore some other features, especially deception cues. If we juxtapose results in this section on investor behaviors and the prior results on repayment probabilities, at least *some* market efficiency gains could have been achieved if the platform or investors exploit linguistic features in a scalable fashion. To address this discrepancy however, the explanatory framework will not be sufficient. If we can use information available at the time of the loan request to *predict* the likelihood of default, and show that linguistic features can *improve* such abilities to predict, then it will suggest that such market efficiencies can indeed be improved. For this reason, we now examine if linguistic features can improve the prediction of loan default using predictive analyses.

6. Predictive Analyses

If the linguistic features that we extracted can improve prediction accuracy of loan defaults, the predictive analysis will further demonstrate the economic value of texts. This is also consistent with the suggestions from Shmueli and Koppius (2011). Moreover, it will show that texts, even though they are highly unstructured and "non-standard," can indeed be quantified and utilized in a scalable fashion using well-established methods of extraction.

6.1 Methods, Variables, and Models for Predictive Analyses

The goal of our model is to predict a binary outcome (default or not). Following Iyer et al. (2015), we use regression rather than classification methods because we have already demonstrated the validity of regression model in our explanatory analysis.

We draw on the existing literature to build several different prediction models. These models are (1) credit grade variables only; (2) linguistic features only; (3) baseline model which includes all non-linguistic feature variables; (4) baseline model plus readability; (5)

[23] Due to space constraints we do not report the detailed results here, but they are available from the authors.

baseline model plus objectivity; (6) baseline model plus positivity; (7) baseline model plus deception cues, and (8) a full model that includes all variables. This incremental approach is similar to that used in Ghose and Ipeirotis (2011) and Ghose et al. (2012).

For all models, we set a randomly 70/30 split. 70% of samples are used for estimating our models and the rest as out-of-sample. We use a stratified 10-fold cross-validation for model evaluation. We run each model 10 times and all results are based on the 10-run average. To measure the quality of our linguistic screen mechanism and compare model performance, we use *area under the ROC curve (AUR)* because the number of repaid loans is larger than the number of those defaulted, following prior literature (Ghose et al., 2012; Ghose & Ipeirotis, 2011; Iyer et al., 2015; Lobo, Jiménez-Valverde, & Real, 2008).

6.2 Predictive Results and Discussion

Results for our predictive analysis are presented in Table 3-1-9. The first analysis, which only includes linguistic variables and nothing else, are in the upper right corner. The AUC value of 0.59 compares favorably to the benchmark value of 0.5 (purely random prediction), indicating that linguistic features, even if used alone, possess some predictive power.

All models with linguistic feature variables have higher AUC values and larger predictive accuracies than the baseline model. The best single-linguistic-dimension model is baseline plus deception cues. This echoes our results in the explanatory models, in that variables for deception cues have the largest marginal effects and explanatory power. If we could only choose one category of linguistic features to focus on, the best candidate is deception cues. Not surprisingly, the full model that incorporates all linguistic features performs the best.

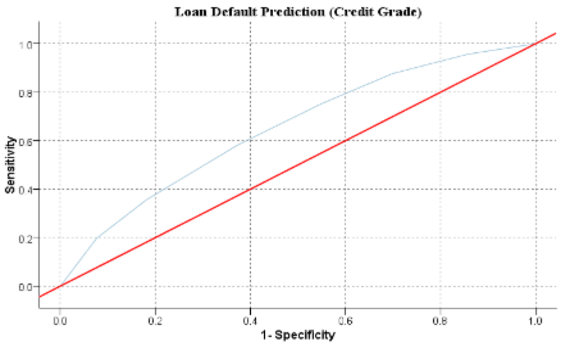
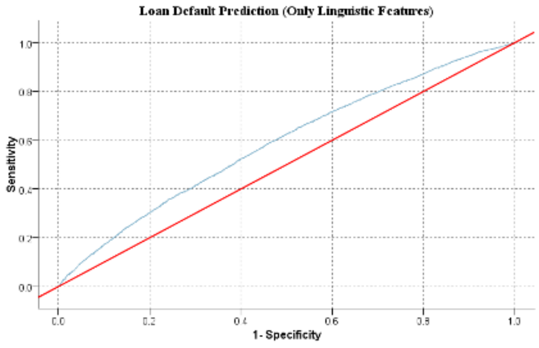
We can also compare the predictive power of three possible screening mechanisms. These three mechanisms are credit grades alone, standard information at the time of loan requesting, and our full linguistic model. While lenders who use standard information can achieve default prediction with 34% (((0.682-0.5)-(0.635-0.5))/(0635-0.5)%=34.81%) greater accuracy than what is possible by using just borrower's credit grade, they can further increase the margin to 65% by including linguistic features. In an industry where a 0.01 improvement in AUC for loan repayment prediction is considered noteworthy (Iyer et al., 2015), this result is not trivial.

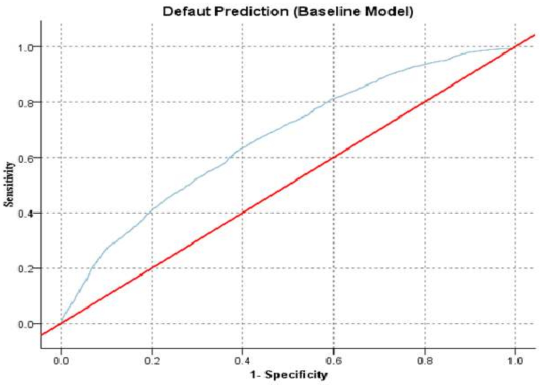
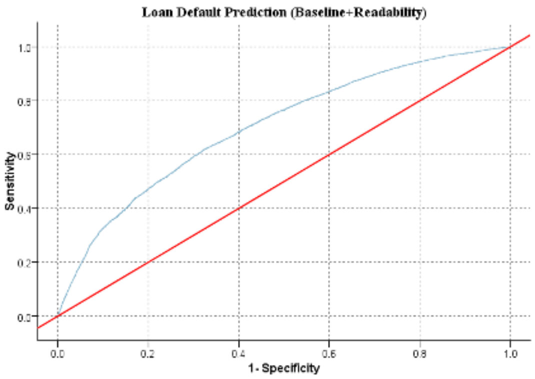
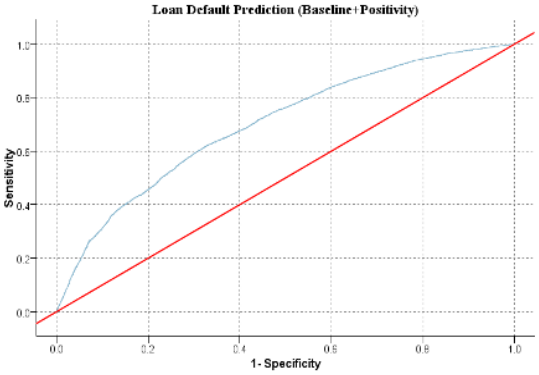
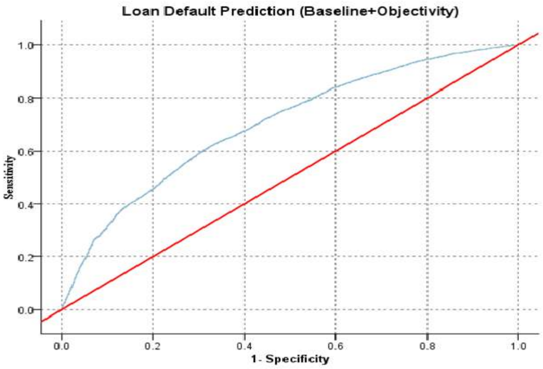
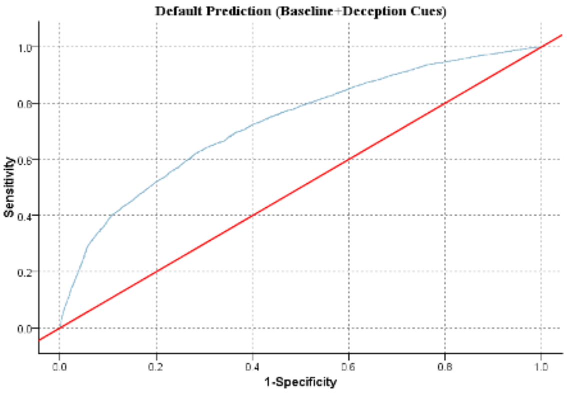
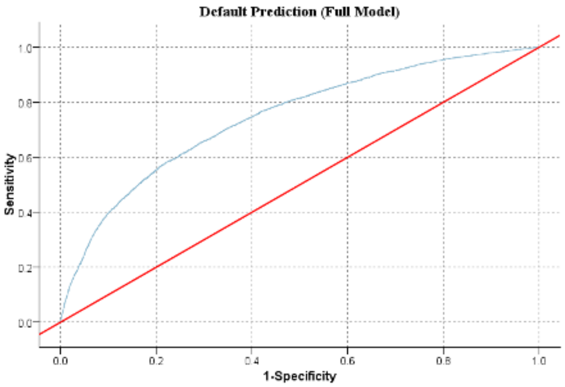
We further conduct t-tests to compare the performance of baseline against other models, following prior literature (Abbasi, Albrecht, Vance, & Hansen, 2012; Abbasi & Chen, 2008; Abbasi et al., 2010). The performance gains are significant (p<0.043).

An additional benefit of our prediction model, especially the one involve deception cues, is to detect potentially fraudulent loan requests. In our sample, 5.19% of all defaulted loans started to default in the very first month after origination. It is reasonable to assume that these borrowers had no intention to repay when requesting loans. Even when we use only baseline plus deception cues prediction model, we achieved 0.814 AUC. These results further attest to the validity of our prediction models, especially the deception cues features.

These results from our predictive model have significant economic value due to the size of the crowdfunding market. As mentioned earlier, debt crowdfunding in the US originated more than $8.9 billion in loans in the year 2014 alone. If linguistic features could help us avoid even one percent of bad loans, that could translate into substantial benefits for all stakeholders.

Table 3-1-9 Predictive Analysis Results

Credit Grade	**Linguistic Dimension Only**
	
AUC:0.635	AUC:0.59
Baseline Model	**Baseline + Readability**

	
AUC:0.682	AUC:0.702
Baseline + Positivity	**Baseline + Objectivity**
	
AUC:0.707	AUC:0.7
Baseline + Deception Cues	**Full Model**
	
AUC:0.718	AUC:0.724

7. Discussions, Implications, and Future Research

Our study is the first to systematically study the role of linguistic styles of texts (i.e., *how* loan requests are written) in an online debt crowdfunding (also known as peer-to-peer lending) market. Using a unique dataset from a leading platform, Prosper.com, we draw on existing literature that mostly examines linguistic features one at a time, and jointly investigates their relationship to loan quality and investor behavior. We use machine learning and text mining techniques to quantify and extract linguistic features in a scalable fashion, then build both explanatory econometric models and predictive models surrounding linguistic features. We find that the readability, positivity, objectivity and deception cues embedded in texts written by borrowers can indeed reflect borrowers' creditworthiness and predict loan default.

The main findings of this paper are as follows. Using natural experiments and borrower panel datasets, we first confirm that investors indeed value texts written by borrowers. Then, we find that if a loan's textual descriptions are less readable, less optimistic, less objective, or richer in deception cues, it is less likely to be repaid and more likely to lose more of its principal. These results are robust when we use a borrower panel data model, or use the borrower's friend linguistic features in an instrumental variable framework. They are also robust when we control for automatically extracted text contents. Almost all results still hold when we use data from another website, LendingClub.com. Next, we test a linguistics-based "efficient market hypothesis" and find that investors interpret the information value of most linguistic features correctly, but not all. For example, they still fall for some well-known deception cues that are indicative of "lemons" among loans. Therefore, there are indeed opportunities for arbitrage, or room for market efficiency improvement if we exploit these linguistic features in a scalable fashion. We verify this by showing, using a predictive modeling approach, that incorporating linguistic features helps improve the performance of prediction models for loan default. Among the four individual linguistic features, *deception cues* provide best predictive powers. A prediction model that integrates all four categories performs even better.

Our study makes the following contributions. First, our study attests to the economic value of texts in debt crowdfunding, and the feasibility of automating the extraction of linguistic features. In an effort to speed up the funding process of loans, platforms are increasingly more reluctant to provide loan descriptions. Our analysis in this paper shows that not only can linguistic features be automatically quantified using standard methods, they are indeed meaningfully related to loan repayment. Texts provide useful and usable information about borrower creditworthiness that should be taken advantage of.

Second, whereas prior studies in social science that examine linguistic features only focus on a particular dimension at a time, it is not clear whether these individual linguistic feature dimensions are still impactful when used in conjunction with other features. Our analysis, both predictive and explanatory, incorporates multiple relevant dimensions simultaneously.

Third, our study also contributes to an understanding of investor behaviors when it comes to nascent financial products such as online peer-to-peer lending. We show that investors can interpret some, but not all, potentially useful information from texts. This, in turn, represents opportunities for efficiency improvement, and should be of interest to all stakeholders.

Lastly, our study contributes to the growing literature in crowdfunding by examining texts, an almost universal "tool" and a potential signaling mechanism for all types of crowdfunding.

There are important implications from this paper for both researchers and practitioners. First, we show that it is possible to extract economically meaningful linguistic features in a scalable fashion. Platforms or third parties can potentially find opportunities to improve market efficiency by better leveraging linguistic features, rather than ignoring them. Second, we show that investors do not always interpret the information value of linguistic features correctly. This provides opportunities for investor education, especially retail investors who may be less sophisticated. Alternatively, the platform can implement fraud detection techniques *before* allowing those listings to be posted publicly. This will improve the loan performance in their portfolio and have long-term benefits.

Some limitations of our study can be fertile grounds for future research. First, there are always other linguistic features that are not yet identified in the literature but can be tailored to the crowdfunding context. It is beyond the scope of our paper to do this, so we focused on features that are not only relevant to trustworthiness, but also have been frequently applied and have well-established methods of automatic extraction. Second, for researchers interested in crowdfunding, our study and our methods can be readily adapted or extended to other types of crowdfunding, be it donation, rewards, or equity. It is reasonable to assume that texts should be even more important in other crowdfunding types, because those types of crowdfunding do not have the luxury of extensive hard financial information that a loan context would provide. However, whether texts matter, how they matter, and which aspects matter in other crowdfunding types are all important empirical questions. Despite these limitations however, the comprehensive set of models and results in this paper fills an important gap in the crowdfunding literature, and provides a solid first step in understanding the economics value of texts and other unstructured information in crowdfunding and online markets.

3.2 The Impact and Informational Value of Videos in Rewards Crowdfunding

Abstract

Video has become a major information disclosure channel in online markets. However, few studies examine its informational values, particularly the values of its multi-dimensional information. This study is the first to examine the role of videos in both affecting contributors' funding decisions and revealing the campaign quality by exploring video data from a reward-based crowdfunding website. The results show that crowdfunding campaigns with videos are almost 244% more likely to be funded than those without them, and that the multi-dimensional information in videos predicts both funding success and crowdfunding campaign quality. Our findings further show that investors can only correctly interpret part of the disclosed information in the videos, and it is possible to improve market efficiency using such information. Our study has important implications for both academic research and practices.

Keywords: video, crowdfunding, technical properties, video contents, predictive analysis, voluntary information disclosure, visual and aural cues

1. Introduction

"Rule #1 for Kickstarter videos: make one! There are few things more important to a quality Kickstarter project than video. Skipping this step will do a serious disservice to your project"

-- Kickstarter School 2011

Video has become one of the major components of crowdfunding campaigns. Fundraisers (i.e., people who request funds) use it as a major channel to disclose all kinds of information regarding themselves and campaigns they create with the goal of convincing potential contributors (i.e., people who provide funds) to contribute.

Few studies on crowdfunding examine the impact of video on campaign success, only using it as a control variable and producing conflicting conclusions. Mollick (2014) found that if a campaign includes no video, its funding success rate decreases about 26% but Frydrych, Bock, Kinder, and Koeck (2014) did not observe this relationship. In addition to the conflicting findings about the impact of videos on funding success, these studies only treat video information as one-dimensional (i.e., either existence or non-existence). As we known, video contains rich information and that information should be multi-dimensional (Kumar & Tan, 2015). Thus, the informational value of videos is unknown. Furthermore, no study investigates whether information in videos created by fundraisers can serve as a signal to reveal the campaign quality (i.e., outcomes of crowdfunding campaigns after successfully funded) – one of the most important factors that affects the healthy development of this new emerging market (Mollick, 2014). Therefore, we conduct this study to answer the following three specific questions:

1) *How does the presence of videos affect campaign funding success?*
2) *How does the information in the videos predict the funding success?*
3) *Can we predict campaign quality using information in the videos?*

We structure our paper as follows to answer these research questions. Section 2 provides a brief review about related literature. Section 3 describes the empirical context, data, and variables. We then address the first research question in section 4 by comparing the differences in funding success between campaigns with and without videos, after matching them on observed campaign characteristics. Our results clearly show that, all else being equal, campaigns with videos are almost 90% more likely to be funded than those without videos. Section 5 addresses the second research question. We implement predictive analysis and find that the multi-dimensions of video information indeed can improve the prediction about funding success. Sections 4 and 5 provide insights about the impact of video and its embedded information's effects on contributors' funding decision. In section 6, we take it a step further, studying whether the information in videos can reveal the quality of crowdfunding campaigns and answer the last research question. Our findings show that the multi-dimensional information in videos can greatly improve predictions of campaign quality. We finally conclude our paper in section 7 by summarizing our main findings, discussing their implications, and pointing to several future research directions.

2. Related Literature

We draw from several streams of literature to understand how the presence of videos in crowdfunding can affect contributors' funding decisions and why information disclosed in videos may reveal the campaign quality.

2.1 The Impact of Videos on Viewers

Two arguments support that videos affect viewer behavior. Drawing from literature in psychology and communication, the first argument is that the "social presence" of fundraisers in the videos wins the trust of contributors and then their contributions (Elliott, Hodge, & Sedor, 2011; Short, Williams, & Christie, 1976). Video as a communication venue provides significant sensory information about its creators such as personality, experience, and knowledge to its viewers (Chaiken & Eagly, 1983). This rich sensory information attracts viewers' attention, engages them, and elicits their responses by reflecting the video creators' "social presence" (Short et al., 1976). Similarly, videos provided by crowdfunding fundraisers reveal sensory information about their markers. Information includes who they are, why they need the funds, how they will spend these funds, and why they are trustworthy. This information significantly reflects fundraisers' "social presence".

This social presence has a positive relationship with trust (Basoglu & Hess, 2014). Previous empirical studies in many fields such as finance, IS, and marketing show that “social presence” influences viewers’ trust of websites (Awad & Ragowsky, 2008; Chiu, Hsu, & Wang, 2006; Cyr, 2008; Cyr, Head, Larios, & Pan, 2009), online vendors (Gefen, Benbasat, & Pavlou, 2008; Gefen, Karahanna, & Straub, 2003), recommendation systems (Hess, Fuller, & Campbell, 2009), and firms (Pennington, Wilcox, & Grover, 2003). This trust, caused by the “social presence” in the videos, further affects the behaviors of individuals such as consumers and investors. For example, investors recommend larger investments after viewing financial restatement announcements online via video (Elliott et al., 2011). People in videoconferences tend to be more influenced by heuristic cues —such as how likeable they perceive the speaker to be — than by the arguments presented by the speaker (Ferran & Watts, 2008).

Another argument is that the inherent characteristics of videos affect reviewers’ behaviors. Videos possess rich presentation formats such as color, dynamic movement, visual cues, and sound (Xu, Chen, & Santhanam, 2015). These information formats have a significant impact on reviewers’ perceptions and behaviors (Lim & Benbasat, 2000; Watson-Manheim & Bélanger, 2007). Among these formats, visual presentation is especially important and provides vivid information (Kumar & Tan, 2015). The vividness, which refers to the presentational richness of a medicated environment (Steuer, 1992), provides more substantial information through multiple sensory channels such as visual cues, dynamic motion, and nonverbal language (Lim, Benbasat, & Ward, 2000), and helps viewers better evaluate the given information (Jiang & Benbasat, 2007; Klein, 2003; Li, Daugherty, & Biocca, 2001). As a result, people tend to positively respond to information delivered through videos (Short et al., 1976). For example, Kumar and Tan (2015) found that introducing a video resulted in an increase in sales, a finding confirmed in other research (Daugherty, Li, & Biocca, 2008; Jiang & Benbasat, 2007).

All aforementioned studies support the arguments that videos can affect people’s perceptions and behaviors. Following these arguments, fundraisers in crowdfunding vividly describe their campaigns and show their personality and trustworthiness. Therefore, the impact of videos on contributors is likely to be significant.

In addition to the positive relationship between including videos and funding success, another important aspect of videos is that fundraisers can voluntarily disclose important information about themselves and their campaigns in order to persuade potential contributors.

2.2 The Values of Voluntarily Disclosed Information

Whether unverified voluntary disclosure can really reveal the quality of information givers is a significant and ongoing debate in many fields such as IS, finance, and security (Beyer, Cohen, Lys, & Walther, 2010; Gao, 2010; Mitra & Ransbotham, 2015)[24].

Proponents of voluntary disclosure frequently cite the best well-known theory that supports voluntary disclosure – "unravelling result" (Viscusi, 1978). This theory states that if the information holders possess better information about a subject than the information receivers do, and there is zero cost to disclose it, they will always disclose it. The underlying reason is that rational information receivers will always consider the non-disclosure as having the lower quality (Grossman, 1981; Milgrom, 1981). Empirical studies frequently show that individuals and firms may disclose extensive information about themselves truthfully to gain trust (Beyer et al., 2010; Dranove & Jin, 2010; Jiang, Heng, & Choi, 2013). As a result, this closure increases both investors' welfare (Gao, 2010) by reducing capital costs, and firms' values, by eliminating undeniable non-disclosure consequences such as increased security risks (Ransbotham, Mitra, & Ramsey, 2012; Wang, Kannan, & Ulmer, 2013). In contrast, less disclosure is associated with decreased sales (Forman, Ghose, & Wiesenfeld, 2008; Jin, 2005; Lewis, 2011). These studies all show that voluntary disclosure reveals the true nature of information holders and benefit the market as a whole.

On the other hand, many researchers who consider the lesser value of such voluntary disclosure mainly focus on the costs incurred by the information discloser and the role of information receivers that may lead to the failure of unravelling (Beyer et al., 2010; Dranove & Jin, 2010). They believe that information givers may only serve their own interest (Xu & Zhang, 2013). While these information holders have more completed

[24] (Beyer et al., 2010) and (Dranove & Jin, 2010), and (Verrecchia, 2001) provide detailed discussion about information disclosure.

information than information receivers, such as investors and consumers, about focal subjects, they tend to only disclose favorable information when the cost of disclosure has to be zero or comparably lower than the reward of disclosure (Dye, 1985; Fung, Graham, & Weil, 2007; Milgrom, 1981). In addition to the costs of disclosure, previous studies also show that because of the unverified nature of disclosed information, information receivers do not pay attention to the available information (Beyer et al., 2010; Dranove & Jin, 2010; Fishman & Hagerty, 2003). This results in less truthful information disclosed (Dranove & Jin, 2010). Therefore, the literature about the value of voluntary information disclosure in fields excluding crowdfunding has mixed conclusions.

Similarly, voluntary information disclosure is a major approach that fundraisers use to persuade potential investors that they are trustworthy in crowdfunding contexts. These fundraisers, similar to information givers in all aforementioned literature, have better knowledge about their projects and mostly incurs no cost during voluntarily disclosure. They possibly disclose the truthful information about themselves. Several studies examine the information values of some of these disclosures, particularly the values reflecting the quality of fundraisers, but show mixed results about the true values of such disclosure. While Herzenstein et al. (2011) examined six different identity claims in borrowers' loan requests and found that these claims have no impact on loan performance, Duarte et al. (2012) and Pope and Sydnor (2011) respectively show that disclosed trustworthy faces indeed link to good quality of crowdfunding projects. Michels (2012) eventually concludes that the total quantity of disclosure is more important to reflect the campaign quality, consistent with the well-known theory that more soft information equals lower risk (Petersen, 2004).

Videos in crowdfunding are the exact channel through which fundraisers disclose a larger amount of information, particularly rich sensory information beyond those in aforementioned studies (Chaiken & Eagly, 1983). They contain visual and verbal cues reflecting additional information such as the personality of their creators unobserved from other information channels (Elliott et al., 2011). Previous studies in economics and finance show that this visual and verbal information predicts the quality of their owners (Hamermesh & Biddle, 1993; Hobson, Mayew, & Venkatachalam, 2012; Mayew &

Venkatachalam, 2012; Mobius & Rosenblat, 2006). Therefore, we expect that information disclosed in crowdfunding videos predicts the campaign quality.

3. Research Context, Data, and Variables

We chose reward-based crowdfunding as our study context. The main reason is that the informational values of videos in reward-based crowdfunding may be more apparent and salient. In other types of crowdfunding the availability of financial and other "hard information" grants potential contributors the opportunities to reduce the extent of information asymmetry, and to at least infer the possible quality of fundraisers. Unlike these, reward-based crowdfunding provides no such information and fundraisers can only present themselves as trustworthy by disclosing information through either video or text (i.e., project descriptions) embedded on the campaign web pages. Therefore, we can have a better understanding of the informational values of videos by exploring the role of videos in reward-based crowdfunding rather than in other contexts.

3.1 Indiegogo.com and Its Data

Our data is from Indiegogo.com, one of the leading reward-based crowdfunding sites, created in 2008 in the US, which allows people to solicit funds for an idea, charity, or startup. At Indiegogo, fundraisers can create web pages for their campaigns in which they have the option to embed videos. By using videos, they describe what their campaigns are, why they are trustworthy, and what perks (i.e., rewards) they provide as a return for contributions. We select this website rather than other reward-based platforms for two major reasons. One main reason is that this site vigorously promotes the inclusion of videos in the campaign pitches, and more than half of campaigns posted on this site incorporated videos up until our study, when a much higher ratio occurred. Another main reason is that, unlike most crowdfunding sites, it allows the inclusion of global participation, both potential fundraisers and funders, a policy that will improve the generalization of our results.

The collected data covers all campaigns at this site from January 14th, 2008 to October 14th, 2013. The total number of projects is 48,791, of these, 15,446 of them were successfully funded, and 24,313 of them had video embedded. The lengths of most videos are less than three minutes. All campaigns in collected dataset have passed their possible maximum active duration, 60 days, and ceased to be active. I have all objective information about all campaigns that have concluded. The basic statistics of our dataset is in Table 3-2-1.

Table 3-2-1 Summary Statistics

Data periods	Jan 4th, 2008-Oct 14th, 2013
# of campaigns	48,791
# of successfully funded campaigns	15.446
# of campaigns with videos	24,313

3.2 Independent Variables

3.3.1 Presence of Videos

The variable *presence of videos* is used to indicate whether a crowdfunding campaign includes videos or not. It is a binary variable, and equals 1 if campaigns contain videos, otherwise 0.

3.3.2 Multi-dimensional Quantification of videos

In addition to treating videos as binary variables (i.e., either existence or not), we measure information in the videos in two major dimensions that may affect contributors' decision making (Elliott et al., 2011). The first dimension of information is inherent *technical properties* of videos and another dimension is the *contents of videos* related to campaign properties and entrepreneurs' characteristics.

3.3.2.1 Technical Properties

We measure the technical properties in two sub-dimensions: visual and audio dimensions (Ma, Hua, Lu, & Zhang, 2005). The visual sub-dimension includes two types of information. The first type is the inherent basic properties (e.g., bit rate, width, frame rate, resolution, duration, and definition) that reflect the quality and visual appeal of videos. This is because previous studies show that visual properties can affect viewers' affective states and induce emotion from viewers (Chaiken & Eagly, 1983; Short et al., 1976). This emotion changes viewers' behaviors. For example, the color and graphic layout affect consumer consumption (Mathwick, Malhotra, & Rigdon, 2001).

In addition to visual properties, previous research in the financial market frequently finds that voice reflects the mental states of speakers and affects the perception of listeners about the speakers (Hobson et al., 2012; Mayew & Venkatachalam, 2012). Our audio dimension includes such voice information such as the pitch range, and other audio properties, such as jitter, shimmer, and signal-to-noise ratio (SNR), which measures the voice quality.

We extract these technical properties using three approaches. Because all videos posted on Indiegogo during our study period are from YouTube.com, we use YouTube API[25] to obtain part of the visual properties. The rest of visual properties are extracted by using MediaInfo[26], an open-source program that displays technical information about media files. We extract audio properties using Praat software (Boersma, 2002), a free scientific computer software package for the analysis of speech in phonetics. The detailed explanations for each technical properties are listed in Table 3-2-2.

Table 3-2-2 Video Technical Properties

Variables	Explanations
Video Technical Properties	
Video duration	The length of videos
Video definition	Sharpness of the image
File size	Size of video files
Video bit rate	Bit per second
Width	Width of image
Height	Height of image
Video frame rate	The frequency at which an imaging device displays consecutive images
Audio Technical Properties	
Audio bit rate	Bit per second
Audio channel	Audio signal communications channel
Mean of pitch	The quality of a sound governed by the average rate of vibrations producing it
Jitter rate	The undesired deviation of a periodic signal from the ideal timing
Shimmer rate	Amplitude variations of consecutive voice signal periods
Average of harmonic to noise	The ratio of the sum of the powers of all harmonic components to the power of the fundamental frequency

[25] We access YouTube API from http://code.google.com/apis/youtube/overview.html in June, 2015.

[26] We access MediaInfo software from https://mediaarea.net/en/MediaInfo

3.3.2.2 Contents of Videos

The second major dimension of videos is the *contents of videos*. Drawn on literature in crowdfunding, psychology, finance and venture capital, we further divide these contents into three sub-dimensions. These three sub- dimensions cover topics of campaign founders' personality (Judge, Higgins, Thoresen, & Barrick, 1999), management team characteristics (Kaplan, Klebanov, & Sorensen, 2012; Kaplan, Sensoy, & Strömberg, 2009), and campaign properties (Mollick, 2014).

We represent each sub-dimension with several indicators. Personality of campaign funders is defined based on the widely accepted five personal traits (Judge et al., 1999). We only include *Extraversion* (i.e., positive emotions), *Dependable* (i.e., *Conscientiousness*), and *Passion* for variety of experience (i.e., *Openness to experience*) factors, which can be reflected in the short videos, rather than *Agreeableness* (i.e., cooperative rather than suspicious and antagonistic), *Neuroticism* (i.e., experience unpleasant emotions easily), which can only be detected through interaction and long-term observations.

In addition to personality, previous studies show both management teams and campaign characteristics are important factors that affect investors' decision and campaign future success (Mollick, 2014). Among management team characteristics, we include past experience or relevant expertise of team members, and team size, the factors that are crucial for the success of ventures (Kaplan et al., 2012; Kaplan et al., 2009). Because most crowdfunding campaigns are unique and relate to various topics, we only consider their uniqueness and general characteristics in this study.

To gain higher accuracy in identifying these contents, we used a human-coding approach for this task. We first used stratified sampling approach to extract 8% of all projects with embedded videos based on project categories. Then, two experienced research assistants coded extracted videos for their indicators and have over 90% consensus over all coded questions. By using this method, we derived all necessary variables for this study[27]. The coding questions are available in Table 3-2-3.

[27] There is no commonly accepted method of video content mining we can use to retrieve the targeted information.

Table 3-2-3 Video Contents Questions

Question#	Questions	Options	Explanations
1	Overall, the quality of this video is:	Very bad	The extent to which this video is made and edited in professional way.
		Bad	
		Neither bad and good	
		Good	
		Very good	
2	The overall feeling about this video	Strong negative	After watching this video, your overall feeling about this video
		Negative	
		Neither negative nor positive	
4	The overall disclosures of project founders?	Very little	Sufficient disclosure should show project founders' personal information such as name, background, financial, educational, and demographic information.
5	Did this video mention the founders' experiences or expertise?	Yes	The founders showed or talked about their past experiences or relevant expertise for this project.
6	Did this video demonstrate any primary functional outcome?	Yes	The primary functional outcome might be a prototype of product, part of movie or music, or section of book.
7	Will the outcome be tangible?	Yes	Tangible outcome can be a book, a piece of music, or a product.
		No	Intangible outcome will not be in a physical shape.
8	Did this video mention third part endorsement?	Yes	Third party endorsement can be third party testimonies, certification, or protection. For example, it showed fan supports, patents, or business partnership.
9	Is this project is one-off project?	Yes	This project is one-off and not extensible.

10	Did this video mention the uniqueness of this project?	Yes	Uniqueness means that the outcome of this project is non-existing.
11	Please rate the founders in following personality dimensions in scale 1-5?	Dependable (1,2,3,4,5)	"Dependable" measures the possibility of whether the founders can deliver the project. Score 5 means that you strongly believe that the founders will deliver promised outcome.
		Extravert (1,2,3,4,5)	"Extravert" measures the extent to which the founders are talkative, sociable, gregarious, assertive, or active. Score 5 means that you strongly believe that the founders possess the aforementioned characteristics.
12	Is this project founded by team?	Yes	The video or narratives mentioned whether the project is founded by team or individual.
13	Rate the founders' passion towards this project in scale 1-5	Passion : 1,2,3,4,5	Score 5 means that the founders show strong emotional fondness of this project.

3.3 Dependent Variables

There are two different dependent variables. One is the *funding success* (i.e., result of requesting funds). Another is the *campaign quality* (i.e., whether fundraisers deliver their promises after they are successfully funded).

3.3.1 Funding Outcome: Funding Success

Funding success is the outcome of requesting funding. It measures whether a campaign has reached its funding target after a certain period defined by the crowdfunding platform. It equals to 1 if a campaign reaches the target, or to 0 if a campaign fails to reach its goal after 60 days at Indiegogo.

3.3.2 Delivery of Promises: Campaign Quality

Campaign quality is used to measure whether fundraisers have fulfilled their promises made during the funding period after they are funded. However, no standard measurement for campaign quality is defined in previous studies, because neither standard calibrating criteria nor campaign outcomes are available. In our study, we define our own approach to measure it with higher accuracy. When fundraisers post their campaigns, they always post web links to them. We use the existence of campaign links 6 months after campaign closure as the measurement of campaign quality, excluding campaign categories in theater and films that are one-time and short lived. The outcome of this approach is a binary variable, either it exists or not. We designed a three-layer approach to measure the campaign quality as described in Table 3-2-4.

Table 3-2-4 Three-layers Approach to Identify Campaign Quality

Layers	Approach	Evaluation Criteria
1	Search whether external links provided by fundraisers exist.	a. Webpage updated time. b. If updated time is not available, check reply status code:200
2	We exclude certain types of campaigns, such as theater, its existing time period is very short. Because web pages normally disappear after certain period, we use Google search approach (keyword: campaign name+ campaign location).	Matching campaign names.
3	For campaigns that can't be confirmed, using manual approach	

We had two research assistants to validate the results after the implementation of this method and have over 93% agreement about the results.

3.4 Control Variables

To control for other variables that might affect the results, we quantify other information used in previous crowdfunding studies, including social (Lin et al., 2013; Liu et al., 2015), image (Duarte et al., 2012), rewards (Frydrych et al., 2014), and featured information (i.e., specifically mentioning campaigns at the front page of the Indiegogo website). We only include variables that can be observed when fundraisers are requesting funding. Table 3-2-5 shows the basic statistics of our variables, excluding video content described in a later section.

Table 3-2-5 Basic Statistics of Dataset

Variables	Observation	Mean	SD	min	max
Funding option	48,791	0.954	0.210	0	1
Requested amount	48,791	79,481	8.810e+06	1	1.800e+09
Percentage funded	48,791	1.455	15.95	0	731
Raised amount	48,791	4,427	62,333	0	1.281e+07
Funding success	48,791	0.317	0.465	0	1
Featured	48,791	0.0229	0.150	0	1
Presence of video	48,791	0.498	0.500	0	1
Having image	48,791	0.569	0.495	0	1
Number of rewards	43,629	7.223	3.342	1	72
Team size	48,492	2.175	2.257	1	125
Funding option	48,791	0.954	0.210	0	1
Campaign categories	48,786	11.61	6.128	1	24
Number of verified friends	21,892	414.7	283.1	0	999

Our study aims to reveal the role of videos in affecting funding success and predicting campaign quality by using the aforementioned video information. We will first describe our analyses for the impact of videos on contributors' funding decisions.

4. To Video Or Not? How Presence of Video Affects Funding Success

4.1 Method

We first examine the impact of videos on contributors' funding decisions. Our goal is to understand the relationships between the presence of video and the funding success.

Our initial analysis is an overall view of how the presence of videos (either existence or not) affects the funding success rates. Since both the number of campaigns and the percentages of campaigns containing videos are gradually increasing year by year shown in Figure 3-2-1, we compare the funding success percentages between campaigns with videos and without videos across these multiple years.

However, we cannot avoid the biases caused by fundraisers' self-selection and other omitted variables. That is, because we do not assign the videos to campaigns randomly, we are unable to control for observed and unobserved variables that drive fundraisers to include videos in their campaigns. It is also a reasonable assumption that there are some unobserved reasons that might influence contributors to contribute to campaigns when fundraisers simultaneously include the videos in their endeavors.

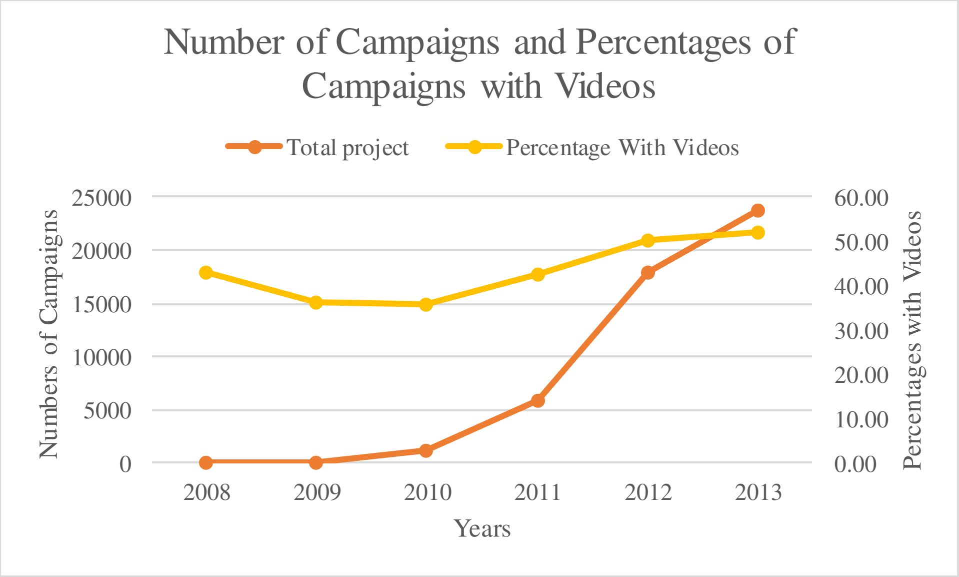


Figure 3-2-1 Number of Campaigns and Percentages of Campaigns with Videos

To address this issue, we follow the approach implemented by Aral, Muchnik, and Sundararajan (2009), Oestreicher-Singer and Zalmanson (2013), Seamans and Zhu (2013), and Bapna and Umyarov (2015) by using a propensity score matching the approach (see Rosenbaum and Rubin (1985)), which will allow us to mitigate aforementioned concerns (Oestreicher-Singer & Zalmanson, 2013). We consider the group of campaigns with videos as the "treatment" group and the group of campaigns without videos as the "control" group. After matching campaigns in the treatment group with campaigns in the control group based on control variables, we implement the following Logit model since the dependent variable is binary (i.e., successfully funded or not):

$$Y_i = \alpha + T_i + \beta \times WithVidoe + Controls_i + \ \varepsilon_i$$

Where Y_i is a vector of funding results, either fully funded (i.e.,1) or failed (i.e.,0); T_i is a vector of time when campaigns are posted; $WithVidoe$ = 1 if campaign i contains videos, otherwise 0; $Controls_i$ is a vector of control variables for campaign i; and ε_i is a standard error term.

4.2 Findings

Figure 3-2-2 first shows that campaigns with videos have much higher funding success rates than those without videos every year during our study period. Then, we examine the impact of video after matching campaigns with or without videos on control variables.

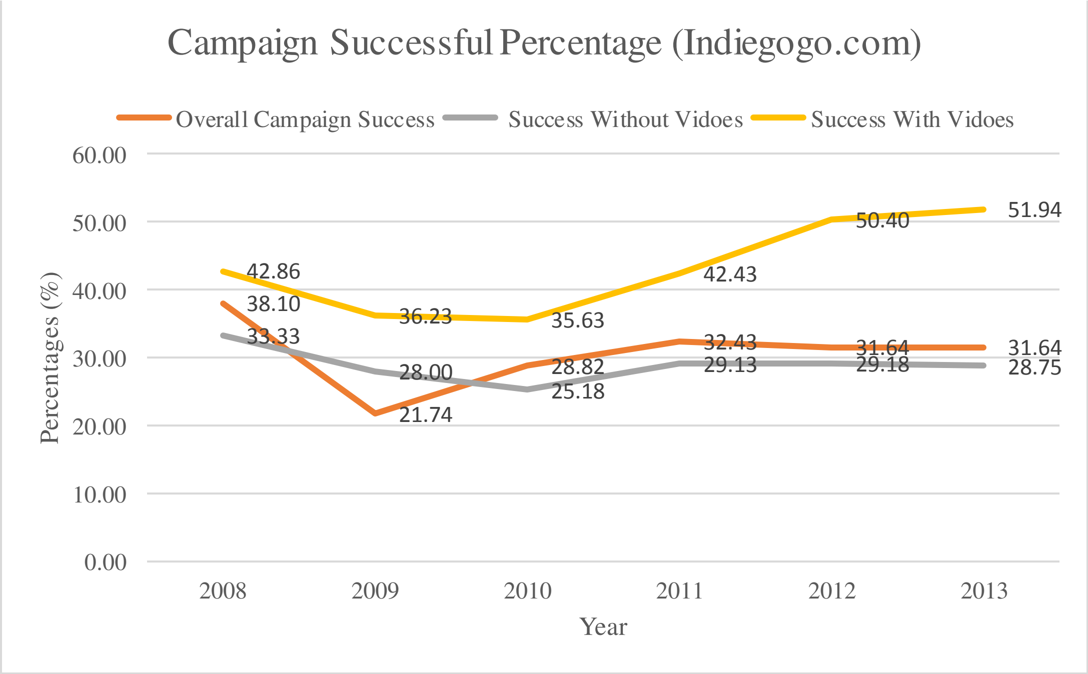


Figure 3-2-2 Funding Success Percentages Comparisons Between Campaigns with and without Videos

Table 3-2-6 presents the main results of the Logit model after propensity score matching. The second column shows the odds ratio from the model. Our initial insight confirms our expectation that campaigns, which include videos in their web pages, are 244% more likely to be fully funded than campaigns that contain no video. That is, contributors prefer to fund campaigns with videos. This finding is consistent with that of Mollick (2014) but has larger magnitude after controlling some observed factors when fundraisers are requesting funds. Our result indicates that although the information disclosed in the videos is neither verified and non-legally binding, the “social presence” reflected from this disclosure indeed wins the trust of potential contributors and eventually leads to their contributing behaviors, a finding consistent with those in offline contexts (Basoglu & Hess, 2014; Elliott et al., 2011; Ferran & Watts, 2008; Short et al., 1976).

Additionally, our results show other interesting findings. The first finding is that campaigns featured at the home page of Indiegogo are six times more likely to be fully funded than those that are not. This shows that getting featured by the crowdfunding sites is an effective tool that enhances funding success. Second, the number of verified friends is important for funding success. Fundraisers who have more friends from Facebook and Twitter are more likely to be funded. If we convert the change in log odds to the change in odds (we use log of friends in our model), a one-member increase in friend size, we expect to see about 10% increase in the odds of being funded. Last, other interesting findings are: more reward levels and having images embedded in the campaign pages are good approaches to increase funding success odds.

Table 3-2-6 Results of Funding Success Logit Model

Variables	**Odds ratio**
Presence of video	2.44***
	(0.0197)
Funding options	0.303***
	(0.0157)
Requested amount	-0.446***
	(0.00489)
Starting date	1.000
	(2.75e-05)
Featured	6.523***
	(0.438)
Having image	0.904***
	(0.0207)
Number of rewards	1.028***
	(0.00381)
Team size	1.056***
	(0.00467)
Number of verified friends	1.000***
	(5.67e-05)
Campaign categories	Yes
Constant	789.0***
	(103.3)
Observations	48,378

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

4.3 Robustness

Given that a binary dependent variable (i.e., funding success or not) in our main analysis may not capture the granular effects of video on funding outcome, we use a more detailed dependent variable – funding percentage – as the dependent variable. Many crowdfunding sites including Indiegogo use a flexible funding option, in other words, fundraisers still keep the contributions even though a campaign is not fully funded, but pays a higher fee to crowdfunding sites.

The new dependent variable, funding percentage, is continuous with value of at least 0 rather than binary. We repeat previous analyses using the new dependent variable and Table 3-2-7 shows the results. *Presence of video* is still statistically significant and positively related to funding success. This is consistent with our initial findings. Other results are also consistent with our initial findings, but vary in magnitudes.

Table 3-2-7 Results of Robustness Test

Variables	**Coefficients**
Presence of video	1.582***
	(0.119)
Funding options	0.839***
	(0.284)
Requested amount	-2.241***
	(0.0410)
Starting date	0.000966***
	(0.000148)
Featured	1.548***
	(0.381)
Having image	0.709***
	(0.125)
Number of rewards	0.215***
	(0.0195)
Team size	0.154
	(0.252)
Number of verified friends	0.000358
	(0.000313)
Campaign categories	Yes
Constant	17.92***
	(0.665)
Observations	48,380

Note:

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

All above analyses confirm that fundraisers can increase their funding success probabilities by including videos. However, videos contain multi-dimensional information, and which dimensions of information can affect their future success are unknown.

5. What Kind of Videos to Include? How Multi-dimensional Video Information affects Fundraising Success

In this section we investigate what information in the videos affects the future success of campaigns. We conduct this analysis with two main goals: (1) understanding the associations between various dimensions of video information and the funding probabilities. (2) Finding the effective new measurements that predict future funding success. According to Shmueli (2010) and Shmueli and Koppius (2011), the most appropriate approach to reach our goals is predictive rather than explanatory analysis (see Shmueli (2010) and Shmueli and Koppius (2011) for more detailed explanations).

5.1 Method

We now examine how the multi-dimensional information in the videos predicts funding success. To the best of our knowledge, no study in crowdfunding context can provide references about what models are fit for the targeted problems. Remarkably however, a wide range of different models is available for predictive purpose, ranging from regression techniques to machine learning techniques (Cui, Wong, & Lui, 2006).

In this study, we choose decision tree, a machine learning model, for the prediction of campaign quality due to the following reasons. Decision tree offers inherent transparency and interpretability, which help users follow the path of the tree and understand the classification rules (Wang et al., 2013). Moreover, studies such as Karhade, Shaw, and Subramanyam (2015), Perols, Chari, and Agrawal (2009), and Schwartz, Bradlow, and Fader (2014) show that decision tree, which is nonparametric in nature, makes it more effective in handling both categorical and numerical variables that prevail in our dataset. Furthermore, decision tree has better performance for biased and stratified samples (Long, Griffith, Selker, & D'agostino, 1993; Zadrozny, 2004).

We draw on the existing literature to build several different prediction models (Eliashberg, Hui, & Zhang, 2007; Ghose & Ipeirotis, 2011). The first model contains only video information (Eliashberg et al., 2007). This will reveal the true predictive power of video information without the "help" of other control variables. Then, we start with a model that includes only control variables (i.e., baseline model), then examine whether incorporating one additional dimension of video information (i.e., technical properties or video contents) will improve the prediction performance of our baseline model. Thus, we build two individual video dimension models. Since video contents contain more information about fundraisers and campaigns, we further build three individual video contents models, each incorporating control variables, and one and only one sub-dimension of video contents. Finally, we build a "full" model that contains all control and video variables.

For all models, we set a random 70/30 split. 70% of samples are used for estimating our models and the rest as out-of-sample. We use a stratified 10-fold cross-validation for model evaluation. We run each model 10 times and all results are based on the 10-run average. To compare model performance, we use *prediction accuracy* and *area under the ROC curve (AUR)* because the number of unfunded campaigns is larger than the number of funded campaigns[28], following prior literature (Ghose et al., 2012; Ghose & Ipeirotis, 2011; Iyer et al., 2015; Lobo et al., 2008).

To this end, conditioning on the availability of videos, we shrink our dataset to include only campaigns about which we have information in terms of technical properties and contents of videos. Our unit of analysis is each campaign, and the main outcome variable is *funding success*. Table 3-2-8 shows the basic statistics of this new dataset.

Table 3-2-8 Basic Statistics of Funding Success Dataset

Variables	Obs	Means	Std	Min	Max
Video Information					
Video duration	1,531	208.5	172.0	13	2,580

[28] AUC is more appropriate when classes have different size since AUC count for the false positive rate as well. A ROC curve is a technique for visualizing, organizing and selecting classifiers based on their performance and has been increasingly used in machine learning communities (Fawcett, 2006) and prediction modeling. For a review, please see Fawcett (2006) and Iyer et al. (2015). For a prior application of ROC curves, please see Bradley (1997).

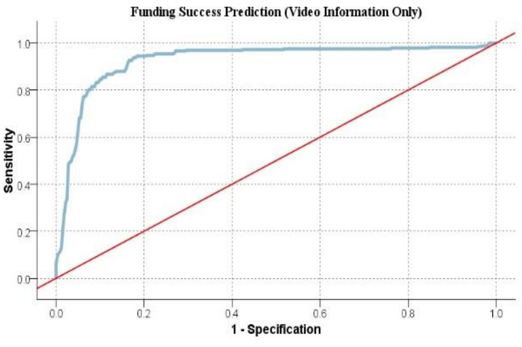
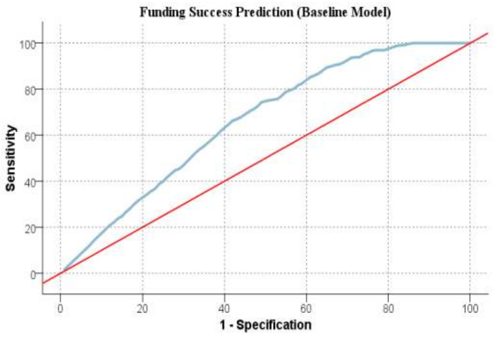
Video definition	1,531	1.585	0.493	1	2
File size	1,531	3.205e+07	3.710e+07	306,867	5.590e+08
Video bit rate	1,531	1.087e+06	762,543	22,256	3.673e+06
Width	1,531	945.3	367.7	204	1,280
Height	1,531	554.9	186.7	128	720
Video frame rate	1,531	27.54	3.591	6	30
Audio Information					
Audio bit rate	1,531	144,257	47,073	3,096	192,000
Audio channel	1,531	1.961	0.194	1	2
Mean of pitch	1,505	575.4	440.8	94.81	4,988
Jitter rate	1,505	3.381	1.406	0.903	10.79
Shimmer rate	1,505	18.06	3.054	6.881	31.21
Average of harmonic to noise	1,505	6.589	2.119	0.193	15.91
Video Contents					
Question 1	1,531	2.451	1.001	1	4
Question 2	1,531	3.674	0.688	1	5
Question 3	1,531	3.634	0.958	1	5
Question 4	1,531	3.285	1.089	1	5
Question 5	1,531	2.135	1.099	1	5
Question 6	1,531	0.430	0.495	0	1
Question 7	1,531	0.389	0.488	0	1
Question 8	1,531	0.406	0.491	0	1
Question 9	1,531	0.141	0.415	0	2
Question 10	1,531	0.573	0.495	0	1
Question 11	1,531	0.370	0.483	0	1
Dependable	1,531	3.152	0.776	1	5
Extravert	1,531	3.376	0.891	1	5
Passion	1,531	3.542	0.909	1	5
Question 13	1,531	2.116	1.115	1	4
Campaign quality	1,531	0.858	0.349	0	1
Control Variables					
Campaign categories	1,531	12.03	6.222	1	24
Number of verified friends	1,531	30.36	169.1	0	2,310
Having image	1,531	0.553	0.497	0	1
Number of rewards	1,531	7.165	3.499	1	20
Team size	1,531	1.128	0.772	1	17
Campaign category	1,531	12.03	6.222	1	24
Funding option	1,531	1	0	1	1
Percentage funded	1,531	1.123	9.376	0	215.7
Funding success	1,531	0.287	0.452	0	1
Requested amount	1,531	22,194	144,284	1	5.000e+06
Featured	1,531	0.0235	0.152	0	1

5.2 Findings

Table 3-2-9 shows prediction results of all models. Results of our first analysis, in which we only include video information and nothing else, are in the upper left corner of Table 3-2-9. The AUC value of 0.912 compares favorably to the benchmark value 0.5 that is purely random prediction and the accuracy with a satisfactory value of 80.53%, indicating that video information indeed possesses predictive power for funding success.

The remaining graphs in Table 3-2-9 show several important findings. The full model that includes all variables achieves the highest values both in AUC and prediction accuracy, comparing with other models. Built on the baseline model, the addition of video information increases AUC value and prediction accuracy about 17 % and 25%, respectively, a notable improvement in financial market for funding success prediction (Iyer et al., 2015). In addition, contents of video possess slightly stronger predictive power than technical property. Furthermore, among video contents, mentioned project information such as third party endorsements, the tangibility of outcomes, and the uniqueness of projects is the best predictor for funding success, and *management team* and *personality* possess similar predictive power in terms of both prediction accuracy and AUC values.

Table 3-2-9 Results of Funding Success Prediction

Video Information Only	**Baseline Model**
	
Accuracy: 80.53% AUC:0.912	Accuracy: 75.48% AUC:0.841
Baseline + Technical Properties	**Baseline + Contents**

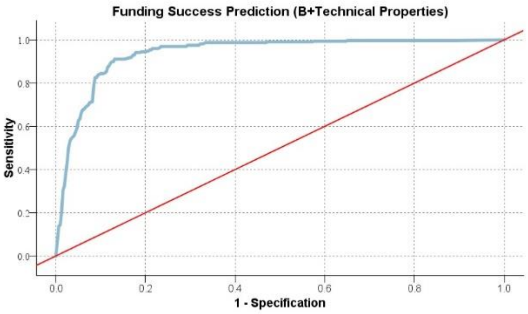
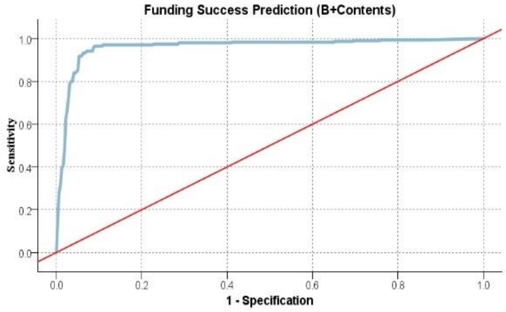
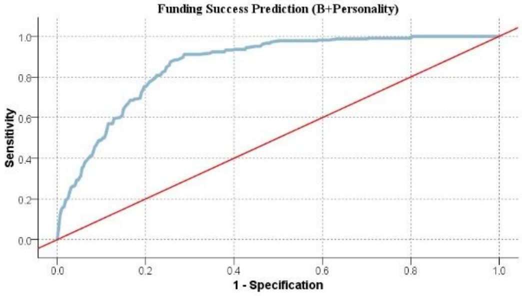
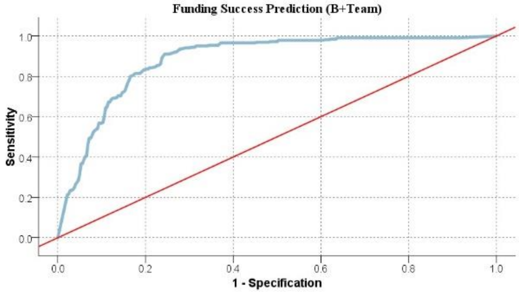
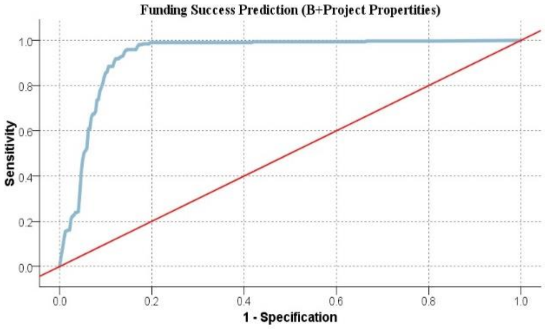
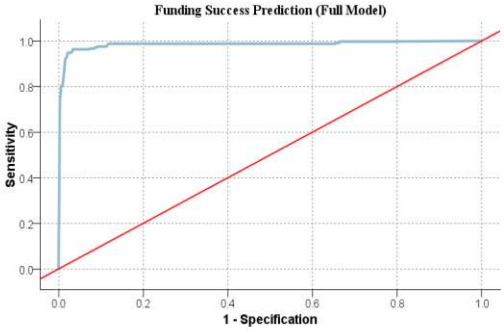
	
Accuracy: 86.57% AUC:0.940	Accuracy: 85.83% AUC:0.957
Baseline + Personality	**Baseline + Management Team**
	
Accuracy: 80.66 % AUC:0.871	Accuracy: 82.12 % AUC:0.873
Baseline + Project Properties	**Full Model**
	
Accuracy: 85.18% AUC:0.93	Accuracy: 94.39 % AUC:0.984

Table 3-2-10 shows the top 10 classification rules from the full model. The most important finding is that both contents and technical properties of videos are important predictors for funding success. In addition, the personality and experience of fundraisers shown in the videos play an important role in affecting contributors' funding decisions. They are more likely to contribute to fundraisers who are more passionate, positive, dependable, and experienced, characteristics that consistently link to entrepreneurs' success in offline financial market (Cardon et al., 2009; Chen et al., 2009; Li, 2010; Unger, Rauch, Frese, & Rosenbusch, 2011). However, on the contrary to previous finding (Baum & Silverman, 2004), smaller team size receive relatively more contributions in crowdfunding context. Our explanation is that contributors may not view larger management teams as possessing human capital (Ahlers, Cumming, Günther, & Schweizer, 2015) but the experience and management skills are among their most important selection criteria (Zacharakis & Meyer, 2000). Furthermore, the technical properties of videos, such as video bit and frame rates, play a larger part in funding success.

Table 3-2-10 Results of Classification Rules

Ranks	**Variable Names**	**Type of Variables**	**Rule**	**Interpretations**
1	Team size	Control	=1	Campaigns with team size larger than 1 are more likely to be funded
2	Means of pitch	Audio (technical properties)	≤ 1433.64	Fundraisers who have lower voice are more likely to be funded
3	Passion	Personality (video Contents)	≥4	Fundraisers who look extreme passionate are more likely to be funded
4	Bit rate	Visual (technical properties)	≥ 1916,570	Campaigns that has larger video file size are more likely to be funded
5	Frame rate	Visual (technical properties)	≥ 23,976	Campaigns with video frame rate at least 24 HZ have higher funding success rate
6	Image	Control	≥1	Campaigns with images are more likely to be funded
7	Third party endorsement	Project Characteristics (video contents)	≥1	Campaigns receiving third party endorsements are more likely to be funded
8	Extravert	Personality (video contents)	≥4	Fundraisers who look extreme positive are more likely to be funded
9	Experience and expertise	Management team (video contents)	≥1	Campaign team members who have previous experience or are experts in funded areas are more likely to be funded

10	Dependable	Personality (video Contents)	≥3	Fundraisers who look at least trustworthy are more likely to be funded

Our results show that the multi-dimensional video information can affect contributors' funding decisions and contents of videos are the most important predictors for funding success. However, the disclosure information in videos is neither verified nor legally binding; in other words, the fundraisers can disclose anything they want without consequences. As a result, whether the disclosed information in videos reflects the quality of fundraisers is unknown and we will examine it in the next section.

6. Can We Predict Campaign Quality Using Information About Videos?

If videos indeed reveal the quality of fundraisers, then we should expect that the disclosed information predicts the campaign quality before the fundraisers fulfill the promises they made when they were requesting funds. Additionally, we should be able to identify which dimensions of information we extract from the videos can better predict the campaign quality. This will provide valuable suggestions for reducing information asymmetry in crowdfunding market. We therefore examine how video information predicts the campaign quality.

6.1 Method

Because the goal of our model is to predict the campaign quality, the most appropriate analytic approach to answer our research question is predictive analytics (Shmueli & Koppius, 2011).

We implement decision tree classification models for the same reasons described in previous analysis. In addition to including all models built for predicting funding success in previous analysis, we add a new model that contains both control variables and a variable that indicates whether campaigns include videos. Thus, we can better understand whether video discloses true values about campaign quality beyond mere presence in the campaigns by comparing prediction performance between models, which include information of videos, and this new model. To test this new model, we use full dataset.

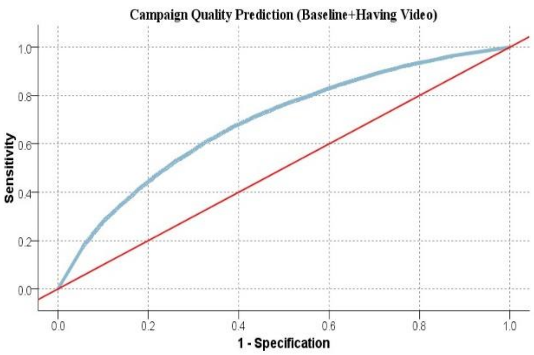
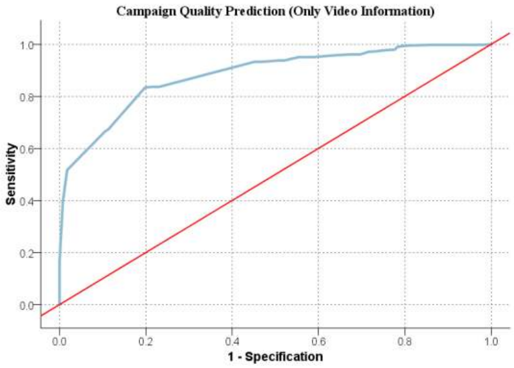
For all models, we choose the same model settings as previous analysis and set a random 70/30 split. 70% of samples, used for estimating our models and the rest as out-of-sample. We use a stratified 10-fold cross-validation for model evaluation, and *prediction accuracy* and *area under the ROC curve (AUR)* for prediction evaluations following prior literature (Ghose et al., 2012; Ghose & Ipeirotis, 2011; Iyer et al., 2015; Lobo et al., 2008).

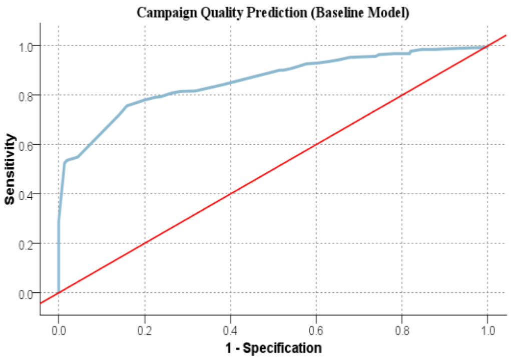
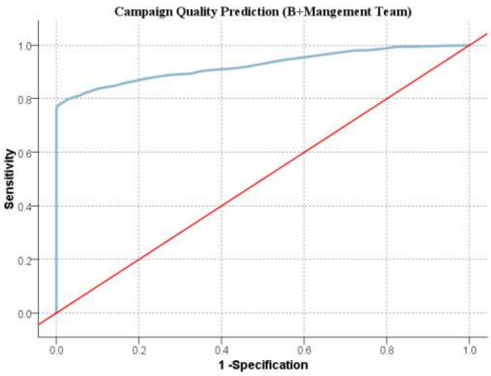
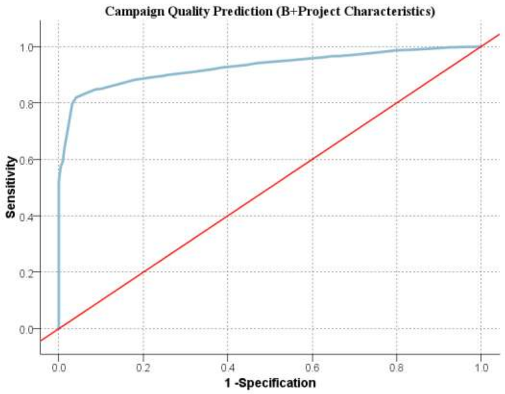
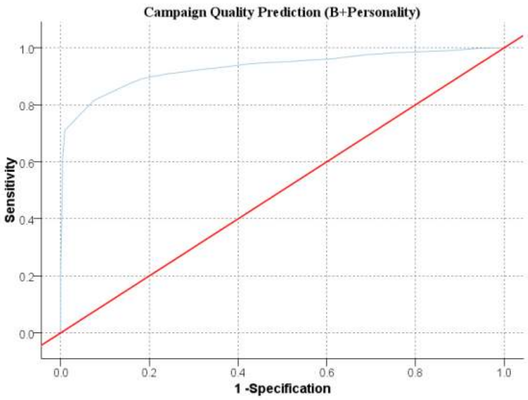
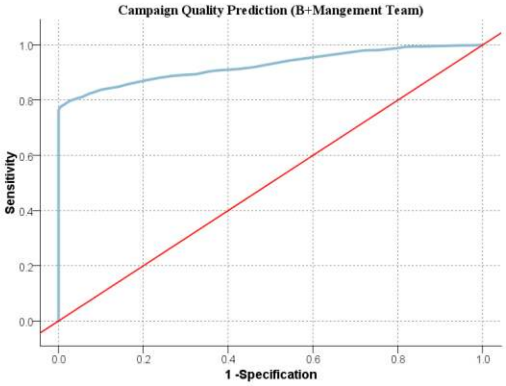
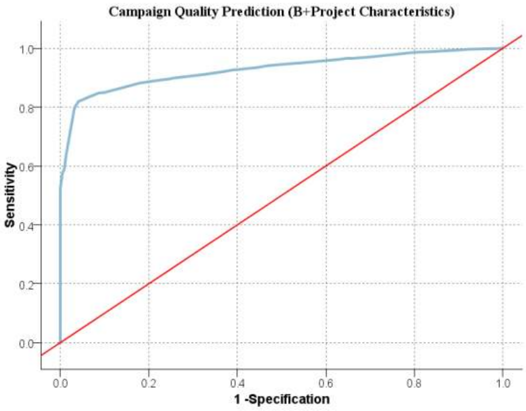
6.2 Findings

Table 3-2-11 shows the prediction results of all models. The inclusion of videos indeed reveal the campaign quality as the AUC values in the first row of table are all larger than 0.5 that is purely random prediction. In addition, models containing video information all have better prediction performance for campaign quality than the model having video as a binary predictor.

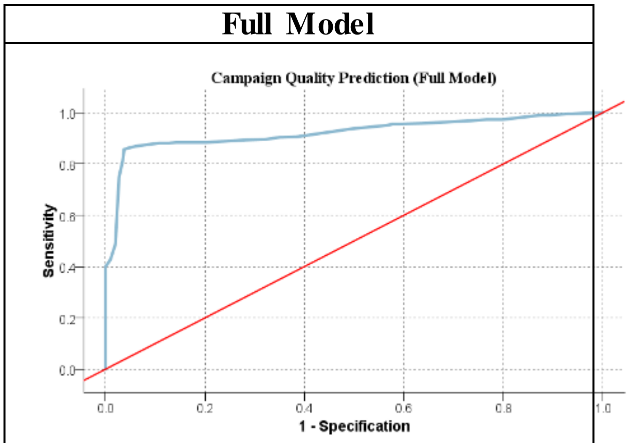
Table 3-2-11 shows several important findings. First, the *baseline model* obtains prediction accuracy of 84.85% and AUC value of 0.789. After including additional video information, the *full model* achieves a significant improvement in both prediction accuracy and AUC values. Second, the full model that includes all variables achieve the highest values both in AUC and prediction accuracy, comparing with other models. Third, the video contents possess slightly stronger predictive power than technical property. Last, the three sub-dimensions of video contents possess similar predictive power for campaign quality in terms of both prediction accuracy and AUC values.

Table 3-2-11 Campaign Quality Prediction Results

Presence of Videos	**Video Information Only**
	
Accuracy:64.14% AUC:0.689	Accuracy:81.1% AUC:0.882

Baseline Model	**Baseline + Technical Properties**
	
Accuracy:84.85% AUC:0.789	Accuracy:88.32% AUC:0.917
Baseline + Video Contents	**Baseline + Personality**
	
Accuracy:90.84% AUC:0.921	Accuracy:87.87% AUC:0.927
Baseline + Management Team	**Baseline + Project Properties**
	
Accuracy: 88.29% AUC:0.924	Accuracy: 90.96% AUC:0.932

Full Model



Accuracy: 94.14% AUC:0.959

Our results in Table 3-2-12 list the top 10 most important rules in the classification process that predict the campaign quality. Among video information, if fundraisers look more dependable, have more social friends, have more experience, disclose more information about themselves, have a lower voice, and receive third party endorsements, they create better quality campaigns.

Table 3-2-12 Results of Classification Rules

Ranks	Variable Names	Type of Variables	Rule	Interpretations
1	Categories	Control	= 1,3,4,5,8, 16, 17	Campaigns categories in Animals, Comic, Community, Environment, Politics, and Religion have higher quality
2	Level of perks	Control	≥6	Campaigns that have reward types exceed 6 have higher quality
3	Dependable	Personality (video Contents)	≥3	Fundraisers who look at least trustworthy create higher quality campaigns
4	One-time project	Project Characteristics (video contents)	≥1	Sequential campaigns have higher quality
5	Number of friends	Control	≥343	Campaigns creators who have at least 343 online friends create higher quality campaigns

6	Experience and expertise	Management team (video contents)	≥1	Campaign team members who have previous experience or are experts in funded areas create higher quality campaigns
7	Overall disclosure	Project Characteristics (video contents)	≥1	Fundraisers who disclose more personal information create higher quality campaigns
8	Means of pitch	Audio (technical properties)	≤ 1988.57	Fundraisers who have lower voice create good quality campaigns
9	Requested amount	Control	≥$7,250	Campaigns with requested amount exceeding $7,250 have better quality
10	Third party endorsement	Project Characteristics (video contents)	≥1	Campaigns receiving third party endorsements have better quality

6.3 Juxtaposing Prediction Findings with Findings from Fundraising Success Analyses

6.3.1 Is the Presence of Videos Informative?

Results from campaign quality prediction show that the presence of videos in a campaign is informative. The additional disclosed information in videos, such as management team, campaign creator' personality, and even technical properties of videos, is indeed a good indicator for campaign quality. Such information truly inform the campaign contributors about the trustworthiness of the fundraisers, and the "unravelling result" (Viscusi, 1978) is supported in crowdfunding context. We will next discuss whether investors correctly interpret this information. If they do not, this will provide opportunities to improve market efficiency using such information.

6.3.2 Do Investors Interpret Video Information Correctly?

If we lay results from funding success analysis and campaign quality prediction side by side, we will infer how the investors interpret the informative video information.

On the one hand, investors most of the time can correctly interpret certain information. For example, contents of videos are very indicative of campaign quality, and investors correctly interpret some of these contents. Specifically, personality (e.g., dependable) of fundraisers, experience or expertise of fundraisers, and third party endorsement of campaigns are among the most important predictors for campaign quality. The funding success prediction shows that they are also top predictive rules for funding success. That is, investors are more likely to fund campaigns with videos including such information. Similarly, investors can correctly interpret some technical properties of videos, such as lower voice, which are good indicators for campaign quality.

On the other hand, investors have not utilized all information that reveals the campaign quality. For example, non-one-time campaigns (i.e., campaigns raise funds multiple times for sequential products) have good quality but investors have not seriously considered this information in their funding decisions. Investors are more likely to be affected by technical properties but this information is less important in revealing campaign quality. This leaves rooms for improving marking efficiency using this information.

6.4 Additional Analyses

We further conduct t-tests to compare the performance of baseline against other models that include control and either part or all video information, following prior literature (Abbasi et al., 2012; Abbasi & Chen, 2008; Abbasi et al., 2010). The performance gains are significant with p-values of 0.057 or lower. We further test the predictive validity by performing a predictive test on a percentage of requested amount that is funded, and obtain similar results. For brevity we do not report these results here.

All results so far confirm that video in crowdfunding plays an important role in both affecting contributors' funding decision and revealing campaign quality. However, investors can only correctly interpret part of this multi-dimensional video information.

7. Conclusions, Implications and Limitations

Videos have become a widely used tool in online markets such as product and crowdfunding markets to disclose important information, but there is little understanding about its values, particularly the values of its multi-dimensional information. We conducted this study to examine the role of videos both in affecting contributors' funding decisions and revealing the campaign quality by using video data from an online reward-based crowdfunding website. Our initial results show that if fundraisers include videos in their fund requests, they have an almost 90% more probability of being fully funded than they do not. Then, we find that though the disclosed multi-dimensional information in videos is unverified, investors still consider it in their decision marking. Finally, we examine whether disclosed information in videos reveals the campaign quality and how investors interpret such information. The results show that the disclosed information can predict the campaign quality. Specifically, we find that the disclosed information in the videos, such as fundraisers' voice, their personality, and campaign team characteristics, indeed reveal the campaign quality. These empirical findings are consistent with those in previous literature. We find that contributors can correctly interpret part of this information but not all, and this creates opportunities to improve market efficiency.

Our study contributes to both academic research and practices. First, we contribute to the growing interests in videos from finance, IS, and marketing (Elliott et al., 2011; Ferran & Watts, 2008; Kumar & Tan, 2015; Xu et al., 2015), but this study is different from them in that to the best of our knowledge, it is the first that treats video information as multi- instead of single-dimensional. Second, we enrich the growing literature on reducing information asymmetry in crowdfunding in IS (Burtch et al., 2013, 2014; Burtch, Ghose, & Wattal, 2015; Hildebrand et al., 2016; Lin & Viswanathan, 2015; Liu et al., 2015; Zhang & Liu, 2012) by suggesting videos as an effective mitigating mechanism, a supplement to well-known "soft information" pool (Petersen, 2004). Third, we add findings to long-lasting debates about the effectiveness of voluntary information disclosure (Dranove & Jin, 2010; Lewis, 2011; Loewenstein, Sunstein, & Golman, 2014) by proving its effectiveness in crowdfunding contexts. Last, our study contributes to literature in finance, accounting, and marketing about the impact of visual and aural cues on viewers' behaviors.

Our study has direct managerial implications for crowdfunding practitioners, website management, and policy makers. Fundraisers can use our findings to increase their funding success while contributors can utilize our results to improve their investment efficiency. In addition, website management can implement our approach and findings to both better screen crowdfunding campaigns and design mechanisms to facilitate this process.

Furthermore, policy makers can use our findings to better educate potential investors and to enact policies for more voluntary information disclosure.

Several limitations in this study may pave the way for future research. First, we only use a manual approach to code video contents because there is no automatic approach available. With the rapid development in techniques, we expect the emergence of such technologies to improve our approach. Second, because of the limited video data from crowdfunding, we could not verify our results on other crowdfunding sites. Last, we only extract general information about campaigns and their creators in this study because of the various purposes of crowdfunding campaigns. In other future studies, we can focus on more specific information for certain types of crowdfunding campaigns, when this emerging market is exponentially expanding and more information is available.

3.3 More than Just Money: Educational Impact of Online Charitable Crowdfunding

Abstract

Public funding for education has been dwindling in recent years in the US. An interesting emerging source of education funding is online donation-based crowdfunding, where teachers can raise funds for class-related projects. Unlike traditional donations or public sources of funding, online crowdfunding requires teachers to exert significant efforts to persuade strangers online to donate. In addition, the teachers' identities are directly associated with the projects; therefore, they are more likely to feel personally accountable for those funds. We posit that these differences can lead to better fund utilization, resulting in positive impacts on student performance more than just the financial aspect of funds. We empirically test this conjecture by exploiting the geographical expansion of DonorsChoose.org, the largest education-purpose donation crowdfunding site. Our results show a positive impact of these donations on classroom performance, especially when teachers are required to disclose more information about themselves and are therefore more accountable. These findings not only show the offline impacts of online crowdfunding, especially in the domain of public goods, but also have implications for the management of traditional donations for education purposes.

Keywords: charitable giving, public goods, donation crowdfunding, efforts of justification, mandatory disclosure.

1. Introduction

Crowdfunding has emerged as a revolutionary and promising approach to fund charitable causes (Özdemir, Faris, & Srivastava, 2015) such as arts, education, environment, and even scientific research (Wheat, Wang, Byrnes, & Ranganathan, 2013). Several well-known crowdfunding platforms for charitable causes have frequently attracted media and public attention. Watsi, a site for crowdfunding thousands of surgeries around the globe, for example, is on Fast Company's[29] list of world's most innovative companies in 2016, abreast of names like Apple and Facebook. DonorsChoose.org, a donation-based website for education purpose, has an impressive list of supporters--Bill and Melinda Gates, Sheryl Sandberg, and Stephen Colbert, among others, with President Obama praised it as "strengthening America's leadership in the 21st century by improving education in science, technology, engineering and math".

However, despite its popularity, relatively few studies examine the offline impact on behavior of fundraisers (i.e., people who receive money). Do the fundraisers merely care about obtaining money, but carelessly spend it after funding success? Will the fact that donations come from a large "crowd" of either strangers or acquaintances motivate or empower those who receive the funds to make better use of them? The answers to these questions not only shed light on the impact of crowdfunding, but also provide suggestions for effective allocation of capital.

We seek answers for these questions in an educational context for two reasons. First, we have a more objective and consequential measurement for the impact of crowdfunding in this context – school performance – than other contexts such as arts and creative ideas. Second, educational crowdfunding only makes up a tiny fraction of the budget for public education. Hence, if we still detect a positive relationship between crowd donations and school performance improvement, this will suggest that crowdfunding provides values far beyond just monetary contributions because the funding amount is almost negligible compared to school budgets. Therefore, we study the following research question in this paper: ***How does online crowdfunding affect the school performance?***

[29] http://www.fastcompany.com/company/watsi

We draw on theories from psychology and finance to hypothesize a positive relationship between educational crowdfunding and student performance. This positive relationship is based on two separate streams of theories. First, both action- and emotion-based psychological theories provide suggestions for how efforts spent on obtaining funding and reflection on receiving funds (e.g. positive and negative affects) may motivate educators to improve student performance (Aronson & Mills, 1959; Baek, Yoon, & Kim, 2015; Brown & Peterson, 1994; Carlson, Charlin, & Miller, 1988; Cheema & Bagchi, 2011; Grant & Dutton, 2012; Kivetz, 2003; Kivetz & Simonson, 2002; Shah, Eisenkraft, Bettman, & Chartrand, 2015). Second, identifying information about the teachers who request funds will be permanently associated with the fundraising campaign. Therefore, teachers will feel personally accountable. Both arguments suggest that educational crowdfunding should have a positive impact on student performance. Furthermore, when more personally identifiable information about teachers is required in the campaigns, the effect should be even stronger.

We use data from a geographical expansion of a donation-based crowdfunding website for education purpose, DonorsChoose, to test our hypotheses. By using the entry of this site into a State, we empirically examine the impact of crowd-funded charitable giving on school performance by implementing both difference in difference (DID) with propensity score matching and relative time model, which addresses the concern of parallel path (Angrist & Pischke, 2008). We further utilize the policy changes at this site, which mandate the disclosure of verifiable information, as exogenous shocks to validate such impact by comparing the school performance changes before and after these policies.

Our findings are two-fold. First, our analysis shows that the psychological states of fundraisers after receiving funds indeed motivate them to achieve the end goal of these funds – improving student performance. Second, we find that mandatory disclosure reinforces such effects, but effectiveness depends on the types of disclosure. The more identifiable and verifiable a disclosure, the more effective it is.

However, the positive impact of crowdfunding and effectiveness of mandatory disclosure may not be exogenous; rather, it may be highly related to other factors such as omitted variables, reverse causality, and selection biases that may simultaneously affect our findings. We adopted three approaches to address these potential endogenous issues. Our first approach is to utilize an external policy change that affect school funding as an instrumental variable. The second approach is the Heckman 2-stage selection procedure. The last is through propensity score matching. These approaches help us control the aforementioned concerns and the results of these tests confirm our findings.

The rest of the paper is organized as follows. Section 2 lays out theoretical foundations for hypotheses development. Section 3 describes our research context and data sources. We explains our empirical strategies in Section 4. Section 5 reports our results and provides discussions. Section 6 concludes by summarizing our findings, outlining their potential implications, and pointing out future research direction.

2. Related Literature and Hypotheses Development

2.1 Crowdfunding and Charitable Giving

Previous crowdfunding studies inform our analysis, but do not provide a complete answer to our research question. Existing crowdfunding literature that studies the impact of crowdfunding mainly focuses on two areas: factors affecting contributions (Agrawal et al., 2011; Burtch et al., 2013, 2014; Lin & Viswanathan, 2015; Zhang & Liu, 2012) and mechanisms determining the quality of fundraisers (Hildebrand et al., 2016; Iyer et al., 2015; Lin et al., 2013; Liu et al., 2015). A few studies are also related to charitable causes, seeking explanations for donors behaviors by using newly available information from online social media such as online social influence (Burtch et al., 2013; Koning & Model, 2014; Saxton & Wang, 2013; Smith, Windmeijer, & Wright, 2015) and price of giving (Meer, 2014). A common feature of these studies is that they focus on online behaviors.

Beyond crowdfunding, a rich pool of literature on charitable giving can be found in economics, psychology, sociology, and marketing. Almost all of them tried to answer two questions: why do people give? And who will give?[30] Some studies aim at identifying important mechanisms that drive donors' charitable giving. Their results show that people donate for charitable causes because of both pure altruism (Andreoni, 2006) and other

[30] see Bekkers and Wiepking (2010), Bekkers and Wiepking (2011), and Wiepking and Bekkers (2012) for more detailed review about charitable giving

impure motivations (Andreoni, 1989) such as good reputation and recognition among their peers (Alpizar, Carlsson, & Johansson-Stenman, 2008; Soetevent, 2005), joy of giving (Strahilevitz & Myers, 1998), and "warm glowing" (Andreoni, 1989, 1990). Others investigate the factors that affect aforementioned mechanisms, studying individual and household characteristics such as age (Sargeant, 1999), education (Wiepking & Maas, 2009), gender (Einolf, 2011), income and wealth (Karlan & List, 2006), and marital status (Andreoni, Brown, & Rischall, 2003). However, donors in all these studies contribute directly to nonprofit organizations and we lack information about the direct impact of charitable giving on the end recipients[31].

2.2 Effort Justification and Reflecting on Benefits Received

We eventually draw on two separate streams of psychological theories to infer impact of crowd-funded charitable giving on school performance. The first stream of theory is action-based theory. In our study we only focus on the theory of effort justification, which states that when people make more effort to complete tasks, they often attempt to justify their added efforts by attaching greater values to the outcome that required more efforts than the outcome that required less or none (Aronson & Mills, 1959; Kruger, Wirtz, Van Boven, & Altermatt, 2004). Such effects of effort justification can be found in many aspects of our lives and have been confirmed in many different contexts (Baek et al., 2015; Brown & Peterson, 1994; Cunha Jr & Caldieraro, 2009; Norton, Mochon, & Ariely, 2011). All of these studies conclude that efforts elevate our valuations of goals and products produced by those efforts. Furthermore, effort justification can even lead people to *raise* their goals and challenge themselves beyond the attainment of preset goals (Harmon-Jones, Amodio, & Harmon-Jones, 2009). This has also been confirmed in various studies (Axsom & Cooper, 1985; Cheema & Bagchi, 2011; Kivetz, 2003; Kivetz & Simonson, 2002; Olivola & Shafir, 2013; Shah et al., 2015; Thaler, 1980; Wan & Chiou, 2010), including charitable giving context where Olivola and Shafir (2013) provide evidence that effort justification indeed functions in offline charitable giving and promote the continuity of contributions to charitable causes.

[31] In this study, we study long-term and repeated charitable events for pubic goods instead of one-time events such as disaster donation in crowdfunding context.

Educators in the charitable crowdfunding context are likely to exhibit a similar pattern of behavior, because they have to make significant efforts to obtain funding from potential donors. Whichever platform they post their funding requests; they need spend the time and energy to create the project proposal, publicly promote their projects, repeatedly explain projects and answer questions, and patiently wait for the completion of the fundraising process, with no guaranteed success in obtaining the requested funds. They therefore develop a sense of ownership (Morewedge, Shu, Gilbert, & Wilson, 2009; Norton et al., 2011) regarding successfully funded projects and are motivated to make sincere efforts to achieve the goals of these projects, i.e., to improve student performance. According to previous arguments, this effort expenditure motivates the educators to hold higher valuations for the projects and their goals — improving student performance.

Another line of theories is emotion-based theory. Reflection on receiving donations causes emotional feeling in beneficiary (i.e., educators in our context), either positive or negative. Psychology research has long shown that receiving a benefit will cultivate a positive effect (Carlson et al., 1988; Isen, Clark, & Schwartz, 1976) in the beneficiary, which in turn encourages them to take more contributing behaviors (Bartlett & DeSteno, 2006), in our context, behaviors to improve student performance. Some may argue that being beneficiary can also cause people to have negative feelings about themselves such as helpless, indebtedness, and incompetence (Fisher, Nadler, & Whitcher-Alagna, 1982; Flynn & Brockner, 2003). Thus, receiving benefits may damage beneficiary's senses of self-efficacy and control (Chow & Lowery, 2010), and may demotivate them to contribute less to others.

However, previous studies regularly show that beneficiaries tend to reduce these negative effects by viewing themselves as benefactors. That is, they turn their roles from beneficiary to benefactor – reciprocally giving back to others (Grant & Dutton, 2012). Psychologist Daryl Bem in his well-known self-perception theory (Bem, 1972) states that people perceive their identities through their own behaviors. When beneficiary starts contributing to others, they are likely to view themselves as benefactor instead of beneficiary. This experience of giving will enhance a beneficiary's sense of self-efficacy, which further motivates them to contribute more (Alessandri, Caprara, Eisenberg, & Steca, 2009; Grant & Gino, 2010). In addition, being a benefactor (e.g., educators in our context), promoting the well-being of surrounding people (e.g., students), is an universally shared value and belief (Schwartz & Bardi, 2001). Hence, no matter how educators reflect on being funded, they are more likely to contribute to others – to improve student performance.

Together, we base on both action- and emotion-based theories and hypothesize:

Hypothesis IA: All else equal, crowdfunding improves school performance.

Beyond educators' self-perceptions about their efforts and being funded, social connections built on the donation relationship will possibly impact on educator behaviors. Although the number of studies regarding the social connection in crowdfunding is still relatively small, all of them conclude that it has an important impact on participant behaviors (Freedman & Jin, 2008; Lin et al., 2013; Liu et al., 2015). In educational crowdfunding, the educators normally observer a list of visible donors, either acquaintance or not. This fact may make educators feel more socially connected with the donors and the donations become more personal. Many empirical studies demonstrate that the more intensive the social connection is, the larger impact it has (Bandiera, Barankay, & Rasul, 2009; Karlan, 2007). We therefore hypothesize that:

Hypothesis IB: All else equal, the more donors a school has, the better it performs.

2.3 Mandatory Information Disclosure

One increasingly common feature at crowdfunding sites is information disclosure. Disclosure can be either voluntary or mandatory and may expose identifiable information about the fundraisers. Existing literature on information provision has long considered this approach as a potentially effective tool for enhancing accountability (Lewis, 2011). Similarly, the disclosure of educators' identifiable information may enhance their effort-induced sense of accountability. However, previous research shows mixed results about the effectiveness of voluntary disclosure due to both external and internal factors (Dimoka, Hong, & Pavlou, 2012; Dranove & Jin, 2010; Jin, 2005; Jin & Kato, 2006; Lewis, 2011) such as the uncertainty about information giver, disclosure cost, competition, and heterogeneous preference[32]. Jin and Leslie (2003) empirically show that mandatory

[32] See (Beyer et al., 2010), Dranove and Jin (2010), Jin (2005), and Jin and Kato (2006) for detailed discussion about voluntary disclosure.

disclosure is more effective than voluntary disclosure in motivating people to be more accountable. Therefore, we focus on the effects of mandatory disclosure in our study.

Many empirical studies show that mandatory information provision can effectively enhance information givers' sense of accountability even if information receivers do not pay attention to disclosed information (Dranove & Jin, 2010; Fung et al., 2007). For example, Jin and Leslie (2003) find that the disclosure of hygiene ratings of restaurants in LA motivated restaurants to be more accountable – improving their sanitation practices – though less people paid attention to them. These patterns are possibly related to a psychological phenomenon -- spotlight effect (Gilovich, Medvec, & Savitsky, 2000), in which people tend to have an exaggerated expectation of attention and believe that they are constantly being watched. Another similar explanation about the impact of mandatory disclosure on information givers is from Loewenstein et al. (2014) – "the telltale heart effect"[33].

For these reasons, if educators are mandated to disclose identifiable and verifiable information about themselves, they are more likely to believe that they are in the "spotlight" of the donors when they receive the funds. They will be motivated to be more accountable and work harder to better use those funds, which in turn will improve student performance. We therefore hypothesize:

Hypothesis II: Conditioning on having crowdfunding projects, mandatory disclosure of personally identifiable information about teachers will lead to even further improvements in education performance.

3. Research Context and Data Sources

3.1 DonorsChoose.org

We obtained our crowdfunding data from DonorsChoose.org, the largest online crowdfunding site for education purposes in the United States. DonorsChoose.org, founded in 2000, focuses on helping public school teachers across US improve their teaching by

[33] The Telltale Heart is from Edgar Allen Poe's (1843) famous short story in which a protagonist imagines that the police can hear the heartbeat of the man he has killed and buried beneath the floorboards of his apartment.

allow them to raise funds online for classroom projects. We will first briefly describe the funding process about which Meer (2014) and Smith et al. (2015) provide more detailed information.

Teachers first submits an informal, written proposal to DonorsChoose. This proposal describes the motivation for this project, the total amount of money needed for the project, and a detailed list of resources requested. It also includes other information such as school demographic information (e.g., the poverty information of students) and the number of students this project supports. The volunteers appointed by the platform will vet this proposal by calling the principal of the school where the teacher is teaching. After verification, the teacher will be able to post the project at DonorsChoose. The webpage includes not only the aforementioned information but also normally a photograph of the classroom and the students.

Potential donors will browse and choose exactly which posted projects they will fund. DonorsChoose uses an all-or-nothing rule: if the total amount raised does not reach the requested amount by the deadline (normally 6 months from the posted date), the raised fund will be returned to the donor's account as credits. Donors can either repurpose the funds for other projects or send a DonorsChoose.org gift card to the same teacher for his next project. DonorsChoose will allocate this fund to an urgent project in need if donors do not choose an option after 30 days. If a project is fully funded, DonorsChoose does not directly transfer the fund to the teachers; rather, they use the donations to purchase the requested resources and deliver them to the teachers. Donors will receive pictures and thank you notes from the teachers and students once the project is completed.

DonorsChoose began operating only in New York when it was first established, but gradually expanded to all 50 states by September of 2007. As of May 2016[34], this platform has raised almost $440 million from 2 million donors, for about 300 thousands teachers in 69 thousand schools. Over 71% of public schools in US have at least one teacher who has posted one project on the site. Around 85% of total posted projects are fully funded. Among posted projects, 36% seek classroom supplies, 30% ask for technology, and 21% request

[34] https://www.donorschoose.org/about/impact.html

books. In the next section, we will describe other data sources and summary statistics of our datasets.

3.2 Data Sources and Summary Statistics

To investigate the impact of DonorsChoose.org on education improvement, we first collect data from this platform. Data span from September 2004 to June 2015, including all information about the projects, donors, resources, and essays.

Our education data are from California Department of Education (CDE)[35]. CED aggregates reports from California schools and learning support resources and creates a dynamic education data collection. This collection contains a wide variety of information including school accountability scores (e.g., California's Academic Performance Index (API)), enrollment, graduates, dropouts, course enrollments, staffing, and English learners. Data are available at state, county, district, and school level. The available information is from 2000 to 2013.

To account for potential confounding factors that may affect school performance and funding requests, we also collect local school district information such as demographic characteristics, socioeconomic factors, and school financing. This information is from the US Census Bureau. The income and poverty information is retrieved from *Small Area Income & Poverty Estimates for School District, Counties, and States reports*[36]. The school district financing information is extracted from *Local Education Agency (School District) Finance Survey Data*[37]

We aggregate data from above resources by using several standard IDs. California has its own County-District-School (CDS) code. We first convert CDS code to NCED code (i.e., National Center for Education Statistics) and then match data from CDE, DonorsChoose, and Census Bureau. Table 3-3-1 presents the summary statistics of our dataset.

Therefore, our final sample consists of 9,292 public schools from California. Out of these 6,899 posted projects at DonorsChoose during our study period, and 5,457 have been

[35] Collected from http://www.cde.ca.gov/ds/
[36] https://www.census.gov/did/www/saipe/
[37] https://nces.ed.gov/ccd/f33agency.asp

successfully funded at least once. As many as 52,033 teachers from these schools created 161,946 projects, 70% of which were successfully funded. Next, we will describe our empirical strategies in answering our research questions.

Table 3-3-1 Summary Statistics

Information	Statistics
Total projects from California	161,949
Total fully funded projects from California	124,079
Total teachers who requested funding	52,033
Total teachers with successfully funded projects	40,751
Total schools requested funding	6,899
Total schools with successfully funded projects	5,457
Total Public Schools in California	9,292

4. Methods

Our study examines the impact of crowdfunding on education improvement and effects of mandatory identity disclosure on such impact. The key empirical challenge is to identify two causal relationships: 1. having DonorsChoose projects and school performance. 2. Conditioning on having projects, mandatory identity disclosure and school performance. Many factors may intertwine to affect these relationships and create endogeneity issues. For example, teachers who are better in nurturing learning in students may also have DonorsChoose projects. Under such circumstance, if school improves performance after having projects, we cannot distinguish the causes –the better teacher or having crowdfunding projects. Similarly, it is difficult to identify what cause school performance improvement, the mandatory disclosure of educators identities (researchers can observe) or student hard work (econometrician cannot observe). We therefore need to rely on some exogenous shocks that change the schools' access to DonorsChoose projects and teachers' exposure to mandatory disclosure requirement. The expansion of DonorsChoose into California in 2004 and several policy changes at DonorsChoose create natural experiment opportunities for identification, which allow us to examine the school performance changes before and after having projects and mandatorily disclosing information.

4.1 Variables and Descriptive Analysis

4.1.1 Dependent Variable

Following the study of Kane and Staiger (2002), we use API (Academic Performance Index) growth as the dependent variable to measure school performance improvement. API is annual measurement of test score performance, used to measure school progress for all schools in California. It is a single number on a scale of 200 to 1,000 that indicates how well students in a school or district performed during the previous academic year, with a score of at least 800 as the goal[38]. API score is calculated for a whole school and its numerically significant subgroups such as socioeconomically disadvantage students, English learners, and students with disabilities.

4.1.2 Control Variables

We draw on Koning and Model (2014) and Meer (2014) to include four groups of control variables shown in Table 3-3-2.

Table 3-3-2 Control Variables

Variable Groups	Explanations
Project characteristics	Information about projects such as year completed, donated amount, number of students reached, grade level, primary focus areas, primary subject areas, resource usage, and resource types.
School characteristics	School financial information (e.g., total expense and total revenue).
Teacher characteristics	Educator information such as gender, certificate for Teacher for America, and certificate for New York Teaching Fellow.
Local characteristics	Information about local school district such as total population, population between 5 and 17 years old, and number of households with 5 -17 years old in poverty.

Since many schools have multiple projects, we use proportional project characteristics in our study (e.g., if we have 3 projects in Maths, 2 projects in Science, and 1 projects in English for a school at a year, the Science projects will count for 3/7 in our model for a school that year). The basic statistics about the dataset are in Appendix E Table E-1.

[38] The details about API can be found from https://www.ed-data.k12.ca.us/Pages/UnderstandingTheAPI.aspx

4.2 The Impact of Crowdfunding on Education Improvement

4.2.1 Empirical Strategy and Model Specifications

We first examine school performance differences before and after schools have funded crowdfunding projects. Our hypothesis is that schools improve their performance after having crowdfunding projects. DonorsChoose entered California in August 2004. This creates a natural experiment setting, in which we can exploit exogenous variations as tools for identifying the effects of having project on education improvement by implementing a difference-in-difference (DID) strategy (Chan & Ghose, 2014; Greenwood & Agarwal, 2015; Sun & Zhu, 2013) .

4.2.2 DID with Propensity Score Matching

Although we can control many aspects that might affect our results such as characteristics of projects, schools, teachers, and locals, we cannot avoid biases caused by teacher self-selection. That is, because we do not assign the projects to schools randomly, we are unable to control for observed and unobserved variables that drive teachers to self-select themselves into the treatment group-creating funded DonorsChoose projects. It is also a reasonable assumption that some unobserved reasons might influence teachers to create projects and simultaneously improve student academic performance.

We follow the approach implemented by Aral et al. (2009), Oestreicher-Singer and Zalmanson (2013), Seamans and Zhu (2013), and Bapna and Umyarov (2015), and address this issue by using a propensity score matching approach (see Rosenbaum and Rubin (1985)). This technique will allow us to investigate heterogeneous treatment effects in non-experimental data using observed variables (Oestreicher-Singer & Zalmanson, 2013). We consider the group of schools that had successfully funded projects after August 2004 as the "treatment" group and the group of other schools that did not have any successfully funded project within our dataset as the "control" group.

We use data from September 2000 to September 2014, and divide them into 10 periods: each academic year each period. Within each period, we first use propensity scores to match schools in treatment group with schools in control group, based on school and local school district characteristics. We implement k-nearest neighbor algorithm for matching.

Then, we calculate the average treatment effects on treated group (ATET) and compare the differences.

4.2.3 Relative Time Model

Although previous tests provide a broad picture about the impact of having crowdfunding projects on school performance, it also leads to a well-known concern for this approach— the parallel path assumption. That is, there is no pre-treatment heterogeneity in the trends between treated and untreated groups (Angrist & Pischke, 2008). This concern arises because of the possibility that unobserved school or local factors, which exist in the school or local school district, may cause the heterogeneous trend in pretreatment API growth. If the dependent variable — API growth — shows heterogeneity over time, our defined untreated group cannot function as a valid control to reflect what would have happened in the absence of treatment. For example, it is possible that there were different trends in API growth between Ackerman Elementary and Adelanto Elementary school districts before DonorsChoose entry.

To address this issue, we implement a relative time approach as opposed to traditional DID estimation, a strategy widely used in previous studies (Chan & Ghose, 2014; Greenwood & Wattal, 2016; Greenwood & Agarwal, 2015). For each school at a certain time, in addition to the year control variable, this estimation will include an additional set of time dummies that measure the relative distances between this time and year when this school had the first project. We include 2 years of pre-having project dummies along with 4 years of after-having project dummies to capture potential intertemporal having projects effects. This model allows us not only to detect whether or not there is a heterogeneous pretreatment trend existing between schools with projects and schools without projects, but also to know how long after the first project the school can accumulate experience in improving student academic performance. Our proposed model is as follows:

$$Y_{st} = S_s + T_t + \theta \cdot PI_{st} + \delta \cdot SI_{st} + \vartheta \cdot LI_{st} + \sum_j \beta_j \cdot has\_project_s \cdot \emptyset + \varepsilon_{st} \quad (1)$$

where Y_{st} is API growth for school s at time t; S_s is a vector of schools fixed effects; T_t is a vector of time fixed effects; $PI_{st}$is a vector of project characteristics at time t; SI_{st} is a vector of school characteristics for school s at time t; LI_{st} is a vector of local school district characteristics for school s at time t; $has\_project_s$ measures whether or not school s had project after 2003 during the study; $has\_project_s$ =1 if the school s ever has at least one completed DonorsChoose project, otherwise 0; $\emptyset$ is a vector of relative time dummies that indicate the relative chronological distance between time t and the year at which school s had the first project; j is an indicator showing whether year t is the j^{th} year since school s had the first project; $and\ \varepsilon_{st}$ is an error time. The coefficients vector of β_j are the difference-in-difference estimates of the time distance effects from having first project on student performance. If $\beta > 0$, then accumulated experience from completed projects causes an improvement in school performance.

We include several fixed effects and additional control mechanisms in the above model to account for school, year, and location variances. The school fixed effects in above model specifications control for time-invariant differences cross all schools and the time fixed effects control for exogenous shocks through the years. We include these fixed effects to ensure that a school at a given year is comparable to other schools at the same year. In addition to school and year fixed effects, other factors may drive changes in API growth and thus we add several groups of controls in our model. Our first group of controls is project characteristics such as total donation, primary focus area, and resource type, which account for project differences. Teacher qualification, school financing, school district demographics and socioeconomic may also affect school performance and thus we control for these school and local district characteristics. For all model specifications, we employed a fixed effect estimator and clustered the error terms at the school level to account for the potential autocorrelation in the data (Bertrand, Duflo, & Mullainathan, 2002).

4.2.4 The Impact of Number of Donors

As we previously discussed, we expect that all else being equal, the number of donors has a larger impact on school performance. Since donation amount may affect our results, we condition on the donation amount and propose the following model:

$$Y_{st} = S_s + T_t + \theta \cdot PI_{st} + \delta \cdot SI_{st} + \vartheta \cdot LI_{st} + \beta \cdot Num\_Donors + \varepsilon_{st} \qquad (2)$$

where Num_Donors is the number of donors for school s at time t; and the rest of variables has the same meaning as those in previous model. The model specifications are the same as previous analysis.

4.3 Mandatory Disclosure on Educators' Behaviors

In the previous section, we study the impact of having DonorsChoose projects on school performance – API growth. In this section, w we test the magnifying effects of mandatory teacher information disclosure on the school performance impact of having Donorschoose projects. The mandatory disclosure of identities is a widely acceptable mechanism in finance and accounting for enforcing responsibilities (Sunstein, 2000) and is more effective when information is verifiable (Dranove & Jin, 2010), we focus on disclosed verifiable information.

4.3.1 Empirical Strategy and Model Specifications

Two policy changes at DonorsChoose platform that enact the mandatory disclosure of teachers' identities create natural experiment settings that allow the comparisons of the differences in school performance improvement before and after these changes. It can be argued that educators who had projects in both before and after policy change periods might differ from educators who only had projects before policy change (e.g., they might do not care about whether they have enough funding). If it is true, our results may suffer from a selection bias, as the assignment of educators to either treatment or control groups is not random. Similar to answering the impact of having projects on school performance, we employ the DID with propensity score matching strategy. The treatment group includes schools that had projects both before and after the policy change and other schools are in the control group.

4.3.2 Mandatory Disclosure of Gender

The first policy change happened on February 8, 2008 and project pages on DonorsChoose started including identifying information for educator gender. Prior to this date, DonorsChoose.org never published educators' gender. As a result donors for school projects can possibly use project information, such as grade level (only showing grade range, e.g., grade 2-5), subject, school name, location, to identify whom they are donating to – at least more likely to do so than before the policy was implemented. Educators who create the project also realize it. Meanwhile, it is indeed possible that identifying the gender may not be sufficient to uniquely identify a teacher, so this effect may be smaller compared to the next disclosure policy change (in section 4.3.3). We extract data between 2004 and

2014 for schools that had projects before this date and construct a panel dataset. Summary statistics are provided in Appendix E Table E-2.

We first estimate the propensity score of having projects after 2007 by using educators' school and local school district information. Then we match the schools in the control group to schools in treatment group by using k-nearest algorithm. Finally, we estimate following model:

$$Y_{it} = E_i + T_t + \theta \cdot PI_{it} + \delta \cdot SI_{it} + \vartheta \cdot LI_{it} + \beta \cdot Diclosure_{it} + \varepsilon_{it} \qquad (3)$$

where all variables have the same meanings as those in model (2) except of: $Diclosure_i t$=1 if the school i disclosed information at year t, otherwise 0; T_t is a time variable instead of fixed effects to avoid perfect collinearity. The coefficient of β is the difference-in-difference estimate of the impact of information disclosure. If $\beta >$0, then mandatory disclosure will have positive impact on school performance. We employed a fixed effect estimator and clustered the error terms at the school level to account for the potential autocorrelation in the data (Bertrand et al., 2002).

4.3.3 Mandatory Disclosure of Real Last Name

The second policy change took place in Fall 2011, which required educators to disclose the entire or the initial of their last names. Previously, educators displayed a pseudo name on project pages. Now donors can almost certainly know whom they are supporting. To address a similar concern to that in 4.3.1, we employ the same approach in section 4.3.2 to further test our hypothesis.

5. Results and Discussions

5.1 The Impact of Crowdfunding on Education Improvement

5.1.1 DID with Propensity Score Matching

Table 3-3-3 represents the main results of our DID model after propensity score matching. The second and third columns show the average API growth for treatment and control groups, respectively. The fourth column shows the differences in API growth between these groups. We see that schools that had successfully funded projects, in general, have larger API growth than schools without any completed projects. The largest difference was in 2005, the second year after DonorsChoose entered into California. Schools that had projects on average improve API about 8.13 points more than schools without any completed projects until then. The only exception is in 2008, schools in control group had a larger improvement in API than schools in treatment group. A possible explanation is the impact of the 2007-2009 recession, which resulted in less funded projects, and lower donated amount per project[39]. Schools that used to get funding from DonorsChoose had difficulty filling the funding gaps, but other schools with established stable alternative funding are less affected.

Table 3-3-3 API Growth

Years	Schools Have Projects	School Without Projects	Differences
2004	8.17	7.84	0.33**
3005	28.43	20.30	8.13***
2006	25.52	23.51	2.01**
2007	22.07	21.79	0.28***
2008	24.97	26.80	-1.84***
2009	29.83	23.40	6.43**
2010	25.26	24.90	0.35***
2011	10.87	8.43	2.44***
2012	8.63	5.51	3.12**
2013	-4.58	-6.15	1.57*

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Our initial insight confirms *Hypothesis IA* that schools having completed crowdfunding achieve better academic performance. That is, educators value their efforts and the opportunity of being funded, and in turn are willing to contribute to others' well-being. As a result they are motivated to achieve their goals — improving student performance (Ashton & Webb, 1986). If we examine further, since the API growth is built on previous academic years, the mostly positive signs for growth indicate continuous efforts from educators. This finding is consistent with previous studies: either previous efforts encourage educators to raise their goals and challenge themselves beyond the attainment of preset goals (Harmon-Jones et al., 2009) or reflecting on receiving funds motivates

[39] We calculated the average number of funded projects and average amount of donations per school cross years and present results in the Appendix F Table F-1

educators to continuously improve student performance (Carlson et al., 1988; Isen et al., 1976), or both.

5.1.2 Time Effectiveness of Having Crowdfunding Projects

Previous findings show us a broad picture that having crowdfunding projects positively affects student performance. The results in Table 3-3-4 from our relative time model, which provides more detailed accumulative effects of both efforts and reflecting of receiving funds, further confirm our initial finding. We observe that schools incrementally improve performance after their first projects (variables 1-4 years later). Within chorological distance defined in this study, 4 years after their first project, schools averagely can achieve 4 times more than their current growth in terms of academic improvement ((20.61+15.03+18.44+17.04-13/67)/13.67=4.3). When we consider the pre-treatment effects, the coefficients for both pre-treatment periods are insignificant and the data appear to support parallel assumption.

Other auxiliary results in this table are also interesting and suggest several areas where crowdfunding might be especially impactful. First, the effect of having crowdfunding projects is the largest in rural area. Rural areas in California have fewer resources, higher poverty rates, and worse school performance (Betts, Reuben, & Danenberg, 2000). Our finding suggests that when DonorsChoose has attracted over 73% of high poverty schools in US to its site, donations from educational crowdfunding may have a meaningful impact on those areas. Second, with respect to the subject areas and supplied resources, crowdfunding funding is more effective in mathematics, and technology resource helps the most in improving education performance. Third, we have mixed results about the importance of teacher certification. New York fellow recognition seems to present higher value than a certificate of Teacher for America.

Table 3-3-4 Results of Relative Time Model

Variables	Coefficients
4 years later	20.61***
	(2.409)
3 years later	15.03***
	(2.410)
2 years later	18.44***

	(2.420)
1 year later	17.04***
	(2.451)
Year of first project	13.67***
	(2.428)
1 year earlier	-0.362
	(2.457)
2 years earlier	-0.0917
	(2.528)
2005	3.193***
	(0.382)
2006	1.925***
	(0.465)
2007	2.829***
	(0.565)
2008	0.0433
	(1.090)
2009	1.846**
	(0.819)
2010	1.776*
	(0.947)
2011	3.629***
	(1.277)
rural	12.72***
	(0.532)
Suburban	0.000322***
	(1.32e-05)
Urban	2.471***
	(0.576)
High poverty	8.241
	(16.30)
Highest poverty	11.24
	(17.94)
Number of student	-0.0411***
	(0.00145)
Number of donors	0.00276
	(0.0123)
Science	-0.303
	(0.290)
Mathematics	0.514*
	(0.307)
English	0.196
	(0.245)
Books	-0.393
	(0.373)

Supplies	-0.158
	(0.335)
Technology	3.062***
	(0.991)
Trips	-0.293
	(0.341)
Visitors	2.572
	(2.118)
Mr.	2.247
	(8.323)
Mrs.	2.572
	(8.320)
Ms.	2.713
	(8.322)
Teacher for American	-2.507***
	(0.494)
New York teaching fellow	27.79*
	(15.64)
Total donation	-0.000272**
	(0.000126)
Constant	384.4***
	(24.28)
School financing information	Yes
Local school district information	Yes
Observations	65,597
Number of schools	5,023
R-squared	0.157

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

5.1.3 The Impact of Number of Donors

In addition to efforts and reflecting on receiving funds, we hypothesize that a social connection with visible donors may motivate educators to improve student performance. Table 3-3-5 shows the test results. The second columns show the coefficients of all variables and the coefficient of *number of donors* variable is statistically significant and positively related to school performance improvement with a value of 0.0373. It means that a one-donor increase will lift 0.0373 points in API growth. Though this increase is relatively small in magnitude, given the large number of donors, the total improvement will be much larger. Social connection in educational crowdfunding context indeed motivates educators to better use the funds, a positive finding consistent with previous studies

(Freedman & Jin, 2008; Lin et al., 2013; Liu et al., 2015). The *Hypothesis IB* therefore is also supported

Table 3-3-5 Results of Number of Donors

Variables	Coefficients
Number of donors	0.0373***
	(0.0132)
2005	51.84***
	(2.730)
2006	26.68***
	(2.667)
2007	19.86***
	(2.651)
2008	22.31***
	(2.623)
2009	20.55***
	(2.615)
2010	21.51***
	(2.583)
2011	7.468***
	(2.572)
2012	5.089**
	(2.565)
2013	-7.983***
	(2.567)
Rural	-2.043*
	(1.078)
Suburban	-0.543
	(0.571)
Urban	0.297
	(0.558)
Moderate poverty	1.727***
	(0.549)
High poverty	3.003***
	(0.507)
Highest poverty	3.323***
	(0.521)
Number of student	-0.968***
	(0.168)
Science	0.304
	(0.324)
Mathematics	0.625*
	(0.347)
English	0.167

	(0.276)
Books	0.254
	(0.430)
Supplies	0.428
	(0.388)
Technology	-0.248
	(0.390)
Trips	2.374**
	(1.138)
Visitors	-0.213
	(2.478)
Mr.	4.475
	(9.469)
Mrs.	2.058
	(9.467)
Ms.	3.660
	(9.468)
Teacher for American	0.453
	(0.499)
New York teaching fellow	32.60*
	(18.92)
Constant	2.214
	(9.896)
School financing information	Yes
Local school district information	Yes
Number of schools	5,023
Number of donation amount groups	2,715

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

All above results show that crowdfunding indeed has a positive impact on school performance. Both *Hypotheses IA and IB* are fully supported. Next we will discuss whether mandatory information disclosure in this context can reinforce such impact.

5.2 Mandatory Disclosure on Educators' Behaviors

Table 3-3-6 presents results about whether mandatory information disclosure may reinforce the positive impact of crowdfunding on school performance and encourage educators to take more responsibilities to improve student performance. Results partially support our hypothesis. The coefficient for disclosure of gender, though positively related to school performance improvement, is statistically insignificant. In other words, conditioning on being funded, disclosing gender may not further encourage educators to assume more responsibilities to improve teaching. In contrast, the disclosure of last name is positively and significantly related to better API improvement, lending support to *Hypothesis II*.

We attribute the discrepancy to the degree of identifiableness from such disclosure. As discussed previously, gender disclosure only partially provides the opportunity to identify a list of possible educators. It still may not uniquely identify the teacher who is requesting the funds. However, if patrons of the projects know the last name of an educator who is at a certain school and teaches certain grade levels, they are almost certain about whom they are supporting. Our results suggest that the more identifiable the disclosure is, the more likely that teachers will feel accountable, and the larger the effects the disclosure will have on student performance.

The higher value of identifiable information can be explained from economic perspectives. The disclosure directly links to educator's reputation. Lower class performance, an indicative of failure to make better usage of the funding, will certainly damage the reputation of the educator and generate higher reputation cost. In this sense the identifiable disclosure serves as an effective reputation mechanism (Dellarocas, 2005). In addition, our finding is consistent with economic studies about the positive impact of mandatory disclosure of school accountability on student outcomes (Jacob, 2005). The disclosure of educator names will motivate educators to assume more accountability and make more efforts in teaching.

The results fully support our hypothesis about the positive impact of crowdfunding on education improvement but partially support our hypothesis about mandatory disclosure. We will next address several concerns related to potential endogenous issues in this study.

Table 3-3-6 Results of Mandatory Information Disclosure

Variables	Gender Disclosure	Name Disclosure
Disclosure of Gender	5.043	
	(3.399)	
Disclosure of Name		15.18*
		(8.100)
2006	54.07*	33.22***
	(32.37)	(7.728)

2007	26.91***	22.04***
	(5.934)	(7.313)
2008	19.30***	16.49**
	(4.878)	(7.659)
2009	22.95***	17.07**
	(4.701)	(7.721)
2010	21.55***	15.17***
	(4.805)	(1.184)
2011	21.51***	0.757
	(4.206)	(1.159)
2012	7.678**	12.50***
	(3.570)	(1.179)
year	12.85***	27.68***
	(0.672)	(1.187)
Rural	12.09*	12.06**
	(2.118)	(10.05)
Suburban	-3.848	-2.773
	(3.166)	(4.572)
Urban	-3.422	-0.486
	(2.626)	(4.310)
Moderate poverty	2.104**	5.643*
	(0.944)	(3.064)
High poverty	3.567***	-4.711
	(1.316)	(3.555)
Highest poverty	3.370**	-3.501
	(1.365)	(3.455)
Number of student	-0.000328	-0.0101***
	(0.000987)	(0.00201)
Number of donors	0.00410	0.00450
	(0.0284)	(0.0126)
Mr.		3.567
		(5.208)
Mrs.		1.911
		(5.230)
Ms.		2.672
		(5.254)
Science	0.483	0.234
	(0.332)	(0.566)
Mathematics	0.655*	0.562
	(0.343)	(0.465)
English	0.0883	0.0254
	(0.272)	(0.389)
Books	0.465	0.496
	(0.492)	(0.495)
Supplies	0.571**	0.327
	(0.285)	(0.420)
Technology	0.0310	0.111
	(0.391)	(0.422)
Trips	2.232	3.241

	(3.879)	(3.548)
Visitors	1.493	2.145
	(6.552)	(2.207)
Total donations	-2.80e-05	-0.000252**
	(0.000131)	(0.000111)
Teacher for American	0.166	32.76***
	(2.152)	(5.391)
Constant	0.267	19.53*
	(6.136)	(10.29)
School financing information	Yes	Yes
Local school district information	Yes	Yes
Observations	56,915	56,915
Number of schools	5,023	5,023
R-squared	0.155	0.125

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

5.3 Endogeneity

Table 3-3-4 shows positive associations between having projects and school performance. However, the correlation could be spurious due to three major concerns. The first-order of concern is that some omitted variables may simultaneously make educators have crowdfunding projects and improve school performance. In addition to this concern, the endogeneity of having projects may be due to reverse causality. Good performance may encourage educators to seek funding for more school materials to reward students. The observed positive correlation between having projects and school performance improvement may be a result of fact that the students are excellent at the first place. That is, our finding may not necessarily reflect the causality path flow from having projects to improved performance. Last, the estimated correlation may also arise from selection biases among educators when they decide to request funds from crowdfunding sites. Although our model has addressed these issues to certain extent, we adopted two additional approaches to further validate our findings. First, we control for the endogeneity in the positive relation between having projects and API growth by using an instrumental variable approach. Second, we implemented the Heckman (1979) two-step procedure to address the selection among educators.

The first approach is using an exogenous shock in budget cuts as an instrumental variable. California passed budgets in 2009 with $15 billion in program cuts and spending reductions. K-12 schools took the largest hit in spending reductions, with cuts totaling $8.6 billion. These cuts to K-12 school funding lasted about 2 years and indeed forced educators to seek more funding from DonorsChoose as evidenced in Figure 3-3-1, in which the number of posted school projects is almost in perfect inverse relationship with California K-12 school budgets. However, research in the impact of school expenditure on student performance over years shows that variations in school expenditure does not systematically relate to the variations in school performance (Hanushek, 1989). Therefore, we use this exogenous shock in budget cut as our instrumental variable. We employ a school fixed effects with instrumental variable model and cluster the error at school level. Our second approach is to implement a Heckman 2-stage selection model.

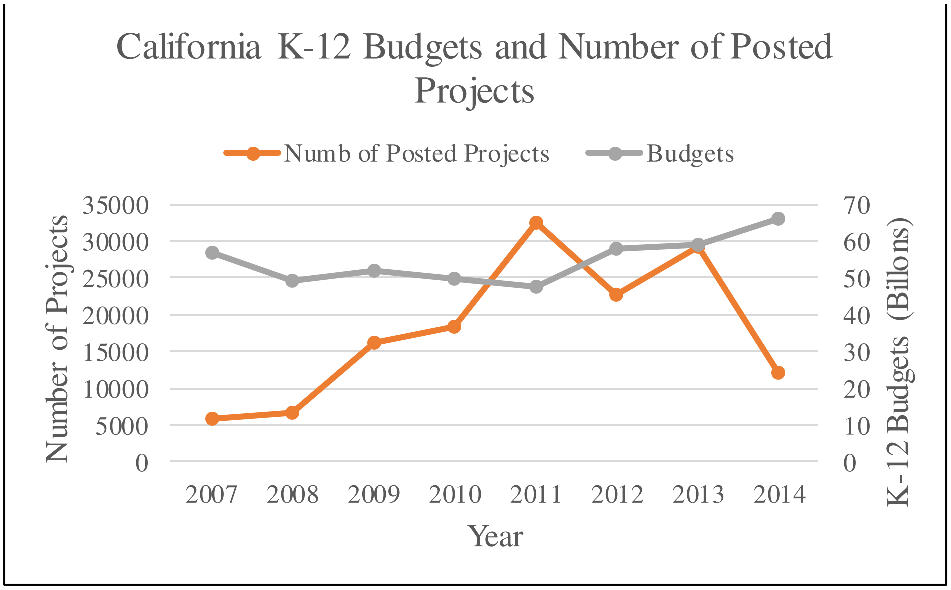


Figure 3-3-1 California K-12 School Budgets and Number of Posted Projects at DonorsChoose.org

Table 3-3-7 presents the results of our baseline (average improvement after having projects), instrumental variable, and Heckman selection models. The main variable of interest in all models is *Having projects*. All coefficients of *Having projects* are positive and significant at the 1% level but with different magnitudes, suggesting that after schools had completed crowdfunding projects during this period, the API of schools increased. These results therefore confirm our initial findings that school improves performance following completion of DonorsChoose projects. The coefficients for the rest of variables are largely consistent with our previous findings.

Table 3-3-7 Results from Instrumental Variable and Heckman Selection Models

Variables	Baseline Model	Instrumental Variables Model	Heckman Model
	Coefficients	***Coefficients***	***Coefficients***
Having projects	23.29115***		
	(2.320997)		
Having projects		12.72***	
		(0.348)	
Having projects			15.17***
			(0.345)
Suburban	-41.91***	-43.83***	-1.847
	(12.77)	(13.13)	(1.128)
Urban	17.63***	20.11***	10.173***
	(3.610)	(3.712)	(0.598)
High poverty	-4.202	-23.83	-0.167
	(16.46)	(16.93)	(0.595)
Highest poverty	-9.657	-27.76	-2.471***
	(18.12)	(18.63)	(0.576)
Number of student	-0.0400***	-0.0439***	-5.228***
	(0.00146)	(0.00150)	(0.532)
Number of donors	0.00815	0.0364***	0.00980
	(0.0124)	(0.0128)	(0.340)
Science	-0.212	-0.478	-0.202
	(0.293)	(0.301)	(0.364)
Mathematics	0.593*	0.129	0.0696
	(0.310)	(0.319)	(0.290)
English	0.277	-0.0443	-12.48
	(0.247)	(0.254)	(9.927)
Books	-0.418	0.184	10.12
	(0.377)	(0.388)	(9.926)
Supplies	-0.242	0.539	12.04
	(0.339)	(0.348)	(9.926)
Technology	0.259	0.0115	1.167***
	(0.345)	(0.354)	(0.450)
Trips	2.883***	6.861***	1.478***
	(0.999)	(1.025)	(0.406)
Visitors	1.996	7.477***	0.789*
	(2.137)	(2.196)	(0.406)
Mr.	4.106	5.094	10.18***
	(8.408)	(8.646)	(1.184)
Mrs.	4.384	5.484	10.47***

	(8.405)	(8.644)	(2.590)
Ms.	4.501	5.581	4.073***
	(8.407)	(8.645)	(0.521)
Teacher for American	-2.354***	-1.484***	-0.00980
	(0.498)	(0.512)	(0.340)
New York teaching fellow	24.51	24.05	29.40
	(15.80)	(16.24)	(19.84)
Total donation	-0.000261**	-0.000663***	-6.133***
	(0.000127)	(0.000131)	(0.835)
Constant	12,204***	503.1***	503.3***
	(168.8)	(23.18)	(21.11)
School financing information	Yes	Yes	Yes
Local school district information	Yes	Yes	Yes
Observations	65,597	65,597	65,597
Number of schools	5,023	5,023	5,023

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

5.4. Robustness

Given that the binary indicator of having projects may not adequately capture the granular effects of donation-supported projects on school performance, especially for schools with multiple projects. More projects possibly may further lead to larger improvement in school performance. To capture this number effect, we conduct two robustness tests. In the first test, we added an accumulated number of projects of a school to measure the effect of having one more project. In the second test, we include percentiles of number projects that a school can possibly have. We first conduct a descriptive analysis for the number of projects that a school has during our study period and Table 3-3-8 shows the results. Then, we include the percentile information as dummies in our model and Table 3-3-9 presents the results. Though the number of projects is positively correlated to API growth, the magnitude is small. One more project only increases API about 0.079 points. Having over 43 projects is less likely to improve school performance. These results further confirm our initial finding that having a project is more important.

Table 3-3-8 Distributions of Number of Projects for Schools

Percentiles	**Number of Projects**

25%	6
50%	18
75%	43
100%	86

Table 3-3-9 Results of Robustness Tests

Variable	One More Project	Number of Projects
	Coefficients	***Coefficients***
Number of projects	0.0719***	
	(0.0154)	
6-18 projects		0.901***
		(0.327)
18-43projects		1.132***
		(0.424)
43< projects		0.168
		(0.555)
2005	33.19***	41.15***
	(2.946)	(2.371)
2006	22.65***	22.18***
	(3.021)	(2.403)
2007	18.14***	16.58***
	(3.044)	(2.398)
2008	18.02***	20.07***
	(3.046)	(2.395)
2009	14.78***	18.67***
	(3.064)	(2.404)
2010	14.80***	15.12***
	(3.041)	(2.359)
2011	-0.238	1.173
	(3.038)	(2.353)
2012	-3.493	1.473
	(3.053)	(2.382)
2013	-15.26***	-10.39***
	(3.064)	(2.393)
Rural	13.48**	-
	(6.434)	
Suburban	-3.711	-42.12***
	(2.942)	(12.65)
Urban	-0.292	17.05***
	(2.822)	(3.576)
Moderate poverty	-4.836*	
	(2.628)	
High poverty	-3.878	-6.627
	(2.673)	(16.31)

Highest poverty	-2.688	-10.25
	(2.690)	(17.95)
Number of student	-0.00892***	-0.0407***
	(0.00104)	(0.00145)
Number of donors	0.00635	0.00252
	(0.0142)	(0.0124)
Science	0.241	0.291
	(0.392)	(0.290)
Mathematics	0.532	0.523*
	(0.397)	(0.307)
English	0.0405	0.185
	(0.312)	(0.245)
Books	0.554	-0.372
	(0.441)	(0.373)
Supplies	0.317	-0.160
	(0.398)	(0.335)
Technology	0.136	0.295
	(0.410)	(0.342)
Trips	3.083***	3.166***
	(1.107)	(0.991)
Visitors	2.122	2.668
	(2.445)	(2.120)
Teacher for American	-3.585	-2.516***
	(4.092	(0.494)
New York teaching fellow		27.54*
		(15.65)
Total donation	-0.000226*	-0.000270**
	(0.000131)	(0.000126)
Constant	23.41***	390.0***
	(4.891)	(24.19)
School financing information	Yes	Yes
Local school district information	Yes	Yes
Observations	56,877	65,597
Number of school	5,023	5,023

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In addition, given that crowdfunding projects are completed at different months of a year but school performance is measured at the year level, projects completed at different time of a school semester may have different effects on school performance. To address this issue, we construct two datasets by using projects completed at the first and second halves of an academic year, respectively. Then we apply propensity score matching. The results in Table 3-3-10 shows that projects from the second half of academic year are more effective.

Table 3-3-10 Results of Different Academic Semester

Data Year	Differences
First half of academic year	-.0843694
Second half of academic year	0.2934186**

Note:
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

6. Conclusions and Implications

Our paper examines the impact of crowdfunding on fundraisers' offline behaviors. By exploring the geographical expansion of a donation-based crowdfunding site for education, we empirically study the effects of having crowdfunding projects on school performance. We further investigate how mandatory disclosure reinforces these effects using the natural experimental opportunities provided by the policy changes at this site. Our empirical results show that crowdfunding can play a role for education, and this role is more than just dollar signs: It can positively affect education performance. Our findings further indicate that mandatory information disclosure encourages fundraisers to take more responsibilities but the effectiveness of this approach depends on types of disclosure.

Our study contributes to several streams of literature. First, our study fills the gap about the impact of crowdfunding on fundraisers' offline behaviors and supplements existing research that mostly concentrates on topics related to stakeholders' online behaviors (Burtch et al., 2014, 2015; Liu et al., 2015; Zhang & Liu, 2012). As crowdfunding research continues to grow in information systems, understanding the behaviors of both fund seekers and contributors has important implications.

Second, our research enriches the literature of mandatory disclosure. We examine the after-effects of mandatory information disclosure on individual behaviors, particularly in crowdfunding contexts, rather than corporate behaviors under legitimized governmental requirement (see (Dranove & Jin, 2010) for more details). In addition, we show that verifiable disclosure can serve as a motivation rather than a mitigating mechanism for adverse selection (Grossman, 1981).

Last, this paper contributes to literature in psychology and education. We provide evidence of future effects of efforts and extend the existing studies from Aronson and Mills (1959), Kruger et al. (2004), Norton et al. (2011), and many others, which mainly focus on the current impact of efforts. In addition, this study adds to existing discussion in education field about the effectiveness of information disclosure on school outcome (Hanushek & Raymond, 2004).

This study provides several important insights to practitioners. First, we show that requiring efforts in the fundraising process can in fact improve the marginal productivity of the funds. This has implications for how scarce financial resources in donations or other funding sources can be better allocated. Second, our results show that verifiable information disclosure motivates educators to take more responsibilities in improving student performance, so such disclosure improves welfare and should be encouraged. Website management can test and implement such mechanisms to enhance fundraisers' accountability. Policy makers can base on our results and design policies to encourage disclosure within the framework of law. Finally, the finding that crowdfunding affects school performance can provide evidence for education practitioners to utilize opportunities to both raise fund and stimulate the educators to improve their teaching. However, this should be used with caution and cannot be used as an excuses to replace funding from federal and local government.

Our paper has a few limitations; some of them can pave the way for future research. First, we do not have data at the student, class, and subjective levels. If we know which class and subject the fund is for and which students benefit from this funding, we will have finer understanding about the impact of crowdfunding. In addition, other crowdfunding sites for education purpose either started very recently or conceal their information. We cannot test the generalizability of our findings. Furthermore, we only use data from previous several years. The current effectiveness of crowdfunding may change in the future. However, our empirical framework and methods can be easily replicated with newer data, or even data from different platforms.

Chapter 4 Insights and Future Research

Three essays in this dissertation address two main issues in crowdfunding, information asymmetry and the impact of crowdfunding. Their findings and implemented approaches provide insights for possible future research.

4.1 Information Asymmetry

We propose using two unstructured data, namely, text and video provided by fundraisers, to mitigate asymmetric information. The results show that both can serve as effective mechanisms for addressing this issue. Building on these studies, we can include other aspects of text and video to further obtain a better understanding about their effectiveness in reducing information asymmetry.

First, we only use linguistic features of text written by the fundraisers and barely consider text contents. Herzenstein et al. (2011) and Michels (2012) all show that text content, which is a major informational channel through which fundraisers disclose information about themselves and campaigns they create, can signal fundraiser quality. In different crowdfunding contexts, text contents, including information about fundraisers such as their experience, education, and social status, may play various roles in reveal campaign quality. The performance comparisons of these potential mechanisms in future studies may provide more insights for both academic and practitioners.

Second, contents of videos that have not been studied in this dissertation can grant us more opportunities to address information asymmetry issue. Since campaigns in reward-based crowdfunding are diverse, contents of videos used in this dissertation only consider information that discloses general characteristics of campaigns and their creators. However, fundraisers in fact disclose all kinds of information, depending on types of crowdfunding projects. For example, they are likely to disclose more venture information in equity-based crowdfunding but more personal information in debt- and reward-based crowdfunding. Future studies can base on previous findings from offline contexts to examine how various disclosure functions as mitigating mechanisms in online environment. When stakeholders in this market are eagerly seeking mechanisms that both need less input and can maximize efficiency improvement, the results of these studies certainly provide valuable suggestions.

Last, other relevant questions are also interesting. For example, how fundraisers and investors in the market utilize the findings can provide opportunities for further studies and suggestions for practical improvement. If platforms implement the approaches and findings of this dissertation as screening mechanisms, we can probably observe fundraisers' gaming behaviors. That is, fundraisers are highly possible to manipulate these mechanisms in order to increase funding success rates, resulting in more serious "lemon market" problem (Akerlof, 1970). It is also highly possible that fundraisers with different characteristics, such as personality or quality, behave different. Therefore, how stakeholders in this market respond will allow us to improve or create new mitigating mechanisms for asymmetric information.

4.2 The Impact of Crowdfunding

One essay in this dissertation examines the impact of crowdfunding on school performance in donation-based crowdfunding. The results show that having crowdfunding projects has positive impacts on student performance improvement. This study, to the best of our knowledge, is one of first studies about the impact of crowdfunding on fundraisers' offline behaviors and provides insights for future research.

One possible reason is that the impact of crowdfunding on fundraiser behaviors, if any, may differ in various crowdfunding contexts. For example, the findings show that both giving higher valuations to efforts for raising funds and concerns about public images motivate fundraisers to take more responsibilities in donation-based crowdfunding (i.e., the context of third essay). But it may not be true in equity-based crowdfunding where fundraisers are more concerning about whether they raise enough funds for business and thus they might behave differently after funded. Hence, it is interesting to know how crowd funded companies, especially, startups, perform after receiving funds. In addition, the passing of JOB ACT allows small- and medium-sized companies to raise funds from crowdfunding sites and it is predicted that crowdfunding has and will be a major financing source for these companies (Agrawal et al., 2013). When these companies have become main drive for employment creation, understanding how entrepreneurs use funds raised from a large of number of strangers will be significantly meaningful.

Another possible reason is that the impact of crowdfunding on fundraiser individuals' behaviors may differ due to their own characteristics. The study in this dissertation only examines the average effects of having crowdfunding projects. Since individuals have various characteristics such as demographics, social status, and personality, the impact of crowdfunding on individuals may significantly differ case by case. Future research that examines the effects of participating online crowdfunding at individual level should be encouraged. By doing so, we can have a better understanding about the effective usage of limited crowdfunding resource.

4.3 Other Research Areas

In addition to issues of information asymmetry and the impact of crowdfunding, several areas related to crowdfunding are worth pursuing. One area is to examine the effects of mitigating mechanism in information asymmetry on social welfare. Mitigating mechanisms suggested in this dissertation can certainly help fundraisers increase success probabilities. As a result, the limited crowdfunding resources may be equally distributed to both "quality campaigns" (i.e., campaigns that have good quality) and "undesirable campaigns" (i.e., campaigns that pretend to be good quality). This will result in a decreased social welfare. If it is true, the implementation of such mechanisms should be thoroughly evaluated. Thus, it is a meaningful to understand the effects of such mechanisms on social welfare.

In addition, testing traditional offline theories in online context is an interesting area. For example, one possible topic could be: why do investors join in syndicates in equity-based crowdfunding? Theoretical explanations in offline context tend to suggest risk aversions as the main drive for syndication. However, transaction and communication via Internet have rendered the factors, which induce the risk aversion behaviors, less important. It is interesting to examine how other theories explain why investors still join in syndicates.

Furthermore, understanding how the emergence of crowdfunding affects different markets, particularly financial and labor markets, is also important. First, crowdfunding allows individuals and small- and medium-sized companies to obtain easier access to capital than offline context, directly affecting financial market. How the crowdfunding affects the landscape and underlying mechanisms of this market will be an interesting research area. Second, the crowdfunding benefit the establishment and growth of startups, creating new job positions. This will have unexpected impact on labor market, particularly on employment in high-tech industry. This is another interesting area worth studying. Last, crowdfunding itself as a new industry provides opportunities for exploring new market structure and operational mechanism.

Chapter 5 Conclusions, Implications, and Limitations

My dissertation, which includes three essays, targets two fundamental issues in crowdfunding: the information asymmetry and the understanding of the impact of crowdfunding; both are vital for the healthy development of this new emerging market. The first two essays in this dissertation examine the informational value of unstructured data, specifically text and video, in crowdfunding. The first essay shows that linguistic styles of loan descriptions written by the borrowers in debt-based crowdfunding can reveal the quality of borrowers, but their values have not been fully utilized. The findings in the second essay demonstrate that multi-dimensional video information possesses predictive power for crowdfunding campaign quality. Both essays suggest the usage of text and videos as effective mechanisms for mitigating asymmetric information in crowdfunding. The last essay examines the impact of educational crowdfunding on school performance. The findings show that crowdfunding plays a role far beyond that of a financial source and that it has important impacts on fundraisers' offline behaviors.

This dissertation contributes to both academics and practices. It contributes to several streams of literature. First, it contributes to the literature on crowdfunding in several aspects: (1) it attests to the values of text and videos in crowdfunding as mitigating mechanisms for information asymmetry and the feasibility of automating the extraction of linguistic features and video information; (2) it is the first study that treats video information as multi- instead of single-dimensional, supplementing the growing interests in videos from finance, IS, and marketing (Elliott et al., 2011; Ferran & Watts, 2008; Kumar & Tan, 2015; Xu et al., 2015); (3) it fills the information gap about the impact of crowdfunding on fundraisers' offline behaviors.

Second, it contributes to literature in education, finance, IS, linguistics, marketing, and psychology. It adds information to long-lasting debates about the effectiveness of voluntary information disclosure in education, finance, IS, and marketing (Dranove & Jin, 2010; Hanushek & Raymond, 2004; Lewis, 2011; Loewenstein et al., 2014) by proving its effectiveness in crowdfunding contexts. In addition, it provides suggestions for the informational value of the combination of multiple linguistic features of texts. Furthermore, it provides evidence of future effects of efforts in crowdfunding and extends the existing studies from Aronson and Mills (1959), Kruger et al. (2004), Norton et al. (2011), and many others, which mainly focus on the current impact of efforts.

This dissertation has direct managerial implications for crowdfunding practitioners, website management, and policy makers. Fundraisers can use our findings to increase their funding success while funders can utilize our results to improve their investment efficiency. In addition, website management can implement our approaches and findings to both better screen crowdfunding campaigns and design mechanisms to facilitate this process. Furthermore, policy makers can use our findings to better educate potential investors and to enact policies for improving market efficiency.

Though discussions in previous chapter point out many potential research opportunities, some limitations of this dissertation can also be fertile grounds for future research. First, we only use manual approaches to extract some information because there is no automatic approach available. With the rapid development of techniques, we expect the emergence of such technologies to improve our approach. Second, because of the limited data from crowdfunding, we could not verify some of the findings in this dissertation on other crowdfunding sites. Last, we only use data from the previous several years. The current effectiveness of our findings may change in the future. However, our empirical framework and methods can easily be replicated on newer data or even data from different platforms. Despite these limitations however, the comprehensive set of models and results in this dissertation fills an important gap in the crowdfunding literature, and provides a solid first step in understanding the crowdfunding.

Appendices

Appendix A: Borrower Panel Data Model on Funding Probability as a Function of the Presence of Texts

We estimate the following logit model with borrower fixed effects:

$\textit{Probability (Funded=1)}_{it} = \gamma_0 + \gamma_1 \times \textit{NoTexts}_{it} + \gamma_2 \times \textit{ControlVariables}_{it} + \eta_i + \varepsilon_{it}$

where *i* is the *i*th borrower, *t* is the t-th listing from that borrower, and *NoTexts* means that the borrower provides no texts on that listing (coded as "1"). The coefficient of *NoTexts* variable in our results is negative and shows that listings without texts are less likely to be funded. The odd ratio for loan requests without texts is 0.41, meaning that if a borrower switches from providing texts to not, their odds of being funded are reduced almost 60%.

Table A-1 Results of Borrower Panel Test

Variables	Coefficients
NoTexts	-0.775***
	-0.219
Credit grade A	3.597***
	-0.733
Credit grade B	2.720***
	-0.722
Credit grade C	2.120***
	-0.716
Credit grade D	1.375*
	-0.712
Credit grade E	0.483
	-0.71
Credit grade HR	-0.63
	-0.708
Debt-to-income-ratio	-0.160***
	-0.0227
Funding option	-0.748***
	-0.0398
Group member	1.991***
	-0.0551

Notes:

1. Robust standard errors in parentheses
2. * p<0.10 ** p<0.05 *** p<0.010)
3. For credit grades, AA is the baseline.

Appendix B: Machine Learning vs. Lexicon Approaches for Sentiment Analysis

Machine learning approach requires a pre-coded training dataset (derived from manual coding) that consists of texts and their labels. Researchers use models generated from this training set to accomplish certain tasks, such as classification, association rule mining, and clustering. This approach usually shows great accuracy in the domain in which the classifier is trained (Aue & Gamon, 2005). By contrast, the lexicon based approach can be faster, provided that an appropriate dictionary is available. Some well-known lexicons for sentiment analysis include the SentiWord Net (SWN) lexicon (Esuli & Sebastiani, 2006) and the Harvard-IV-4 dictionary (Tetlock, 2007; Tetlock, Saar-Tsechansky, & Macskassy, 2008).

Appendix C: Technical Details on Linguistic Feature Measurements

(1) Readability

Our spelling error corpuses are from two sources: Peter Norvig spelling errors list (Jurafsky & James, 2000), which includes Wikipedia misspelling list, and Birkbeck spelling error corpus (Mitton, 1987) gathered from Oxford Text Archive. The spelling error variable is defined as total spelling errors in a loan description. We randomly chose 300 loan descriptions and had one research assistant manually examined the calculated scores. We achieved precision 97% and recall about 91% for spelling error check.

Stanford statistical PCFG parser is a natural language parser that generates the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to produce the most likely grammatical structure of new sentences. However, this statistical parser still produces parsing score even the sentence is grammatically wrong. To maximize the probabilities of correctly parsed grammatical structures, we calculate the probability score by averaging the probability scores generated from top 5 most possible grammatical structures. The error probability is the difference between 1 and previous generated grammatical probability score. Our grammatical score is the log value of the error probability. The performance of this statistical parser is about 86.36% for F1 score (Klein & Manning, 2003).

(2) Positivity and Objectivity

To quantify positivity and objectivity, our first step is to prepare a manually coded sample of texts from our dataset of loan descriptions to help construct our classifiers. We used stratified sampling and extracted a 1% random sample of all loan request descriptions from each credit grade, and chunked each description into sentences to form our coding data set. To ensure accuracy and consistency, two research assistants coded sentences in this sample dataset. Each of them defined each sentence as negative, neutral, or positive in terms of positivity; and as objective or subjective in terms of objectivity. The agreement rate is 90% for 3,790 coded listings. Subsequently, we further divided the coded texts into training (70 % of total coded texts) and testing sets (30% of total coded texts). To build positivity classifier, we constructed several classifiers by combining different features and used SVMlight multiclass package (Crammer & Singer, 2002), to train and test our data sets with all parameters set to default values (Joachims, Finley, & Yu, 2009). As can be seen from the results in table below, classifier built on the combination of unigram and POS tag performed the best, achieving precision of 85.25% and recall of 98.73%[40].

Table C-1 Positivity Classification Results

Feature Sets	Precision	Recall
Unigram/POS tag	85.25%	98.73%
Unigram	84.73%	98.10%
Unigram/bigram/ tri-gram	84.70%	95.10%
Adjective	80.85%	78.57%

We built subjectivity classifier using features similar to those used in Barbosa and Feng (2010) but with extensions, including numbers of polarity words (negative, positive, and neutral words), strong or weak subjective words, modal words (such as "can" and "would"), numbers, unigrams, unigram/POS tag, bi-Char, adjectives, and adverbs. The polarity score and subjectivity clues are derived from the OpinionFinder lexicon (Wilson, Hoffmann, et al., 2005) and used in the study of Wilson, Wiebe, and Hoffmann (2005). This lexicon includes 8,000 subjectivity clues compiled from several sources, each clue was marked with either strong subjective (i.e., subjective in most contexts) or weak subjective (i.e., only have certain subjective usage), and also with polarity (i.e., positive, negative, or neutral). We constructed the objectivity classification model by using Joachims (1999) SVM package that implements the Support Vector Machine of Vapnik, Golowich, and Smola (1997) with all parameters set to default values. We achieve precision of 85.32% and recall of 87.56%.

(3) Deception Cues

Currently there are three approaches to detect deception (Masip, Sporer, Garrido, & Herrero, 2005). The first approach is based on the measurement, recording and analysis of the psychophysiological activity of the subject being examined. The most known example is polygraph test. The second approach focuses on the verbal content of subject's speech and the last approach focuses upon nonverbal aspect of deception. In online lending context,

[40] Precision is referred as true positive rate, the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved (see Jurafsky and James 2000)

we do not observe the physical activities of borrowers and listen to their speeches so our dimensions of deception cues are chosen based on the last approach-nonverbal cues. Nonverbal cues can be interpreted on the basis of linguistics-based cues such as self-reference (Zhou, Burgoon, Nunamaker, et al., 2004). Linguistic cues further fall into two general categories: nonstrategic linguistic cues, which reflect the psychological processes experienced by deceivers, and strategic cues that are strategically used by deceivers (Toma & Hancock, 2012).

We measure nonstrategic cue in three sub-dimensions: cognitive load, internal imagination, and negative emotion. Cognitive load is one of the most common non-strategic cue and measured by concreteness. The concreteness of each list description is calculated as the mean value of the concreteness of all content words in this text. The concreteness value of each content word is extracted from MRC Psycholinguistic Database described by Wilson (1988). One assistant manually examined calculated scores for randomly selected 100 loan descriptions and observed 98% accuracy.

We use temporal and spatial information to measure internal imagination based on the Reality Monitoring Theory (Johnson & Raye, 1981). We operationalized temporal information by using both Stanford SUTime parser (Chang & Manning, 2012) and LIWC time component (Pennebaker et al., 2001). The temporal score is the combination of results from both Stanford SUTime parser and LIWC time component. Stanford SUTime parser is a deterministic rule-based system designed for time expression extraction with accuracy of F1 score about 0.92 (Chang & Manning, 2012). LIWC time component is a list of words that have temporal semantic meaning. We achieved about 97% accuracy for finding correct temporal words in randomly selected samples. We measured spatial information also by implementing two approaches: Stanford name entity recognizer (Finkel et al., 2005) and LIWC space words (Pennebaker et al., 2001). Stanford name entity recognizer use Conditional Random Field (CRF) to extract spatial expression. The common accepted accuracy is 0.88 for F1 score (Finkel et al., 2005). Similarly, LIWC space component is a list of words that relate to spatial meaning. The spatial score is the combination of results from both Stanford name entity recognizer and LIWC space component.

Our last important nonstrategic cues is negative emotion. We quantify negation emotion from two sources: content negation word (Hancock et al., 2007) and functional negation word (Toma & Hancock, 2012). Content negation word are negating words such as “not” and “never”. Functional negation words are words that are semantically negative. Our functional 00negation words are from LIWC dictionary that categorizes words into different emotional states (Pennebaker et al., 2001). The emotion score is the combination of both results. Our accuracies for extracting both types of negation words are around 90% and 98 percent, respectively.

We only measure one dimension of strategic cue: dissociation. Deceivers tend to use more non-first person pronouns (e. g. “he”, “him”, or “her”) in their writings in order to dissociate themselves from their fabricated stories (Hancock et al., 2007; Newman et al., 2003; Toma & Hancock, 2012). A non-first person pronoun score is computed as the percentage of the number of non-first person pronouns to the total words of a text. We achieve 97% accuracy for identifying them.

(4) Variable Definition and Measurements

Table C-2 Variable Definition and Measurements

Variable Names		Definition and Measurements
Linguistic variables information		
Readability	Spelling errors	Spelling mistakes in loan descriptions. Total spelling errors in a loan description are used.
	Grammatical errors	The less likely probability of grammatical structure of a loan description. It is log value of the probability.
	Lexical Complexity (FOG)	The Gunning-Fog Index (FOG) score. It is used to measure the complexity of a sentence and calculated based on formula.
Positivity	Positivity	Average positivity score of a text.
Objectivity	Objectivity	Average objectivity score of a text.
Deception Cues	Concreteness	Log value of average concreteness score of a text.
	Spatial Information	Spatial words are from both Stanford name entity parser and LIWC space word list. Spatial score is the sum of these two values.
	Temporal Information	Temporal words are from both Stanford SUTime parser and LIWC time dictionary. Temporal score is the sum of these two values.

	Non-first-person pronouns	Percentage of first or second pronouns in a text
	Negative Emotion	Negative words are from both content and functional negation words. Negation score is the sum of these two values.
Hard Credit Information		
Credit Grade		Dummy variables indicating borrower's credit grade (letter AA, A, B,C,D,E, HH)
Debt-to-income-ratio		Borrowers' debt-to-income-ratio
Credit inquiries		Number of inquiries about credit report in the six months before listing.
Borrower rate		Given interest rate by borrowers
Is borrower home owner		Whether borrower owns a home
Amount delinquent		The amount the borrower failed to pay when he is requesting loan
BankcardUtilization		Number of bank card used
CurrentCreditLines		Number of credit lines
CurrentDelinquencies		Number of current delinquencies
DelinquenciesLast7Years		Number of delinquencies last 7 years
InquiresLast6Months		Number of credit scores inquired last 6 months
OpenCreditLines		Number of open credit lines
PublicRecordsLast10Years		Number of public record last 10 years
PublicRecordsLast12Months		Number of public record last year
RevolvingCreditBalance		Revolving credit balance
TotalCreditLines		The number of total credit lines
Auction characteristics		
Funding options		The funding options is a dummy with one of the following values: • Open for duration- The listing is open for its duration. • Close When Funded-The listing will close as soon as it is fully funded.
Loan amount		Loan amount requested by borrowers.
Category		A series of dummy variables specify loan category defined by borrower.
Text length		total length of a text.
Word length		Number of characters per word of a loan description
Social information		
Group membership		A dummy indicating whether borrower is a group member
Friend investments		A dummy indicating whether friends of this borrower have invested in this loan request.
Additional variables		

Monthly Fixed Effects	A series of dummy variables specify the months when the loans are originated.

(5) Summary statistics for Repayment Probability Models

Table C-3 Summary statistics for Repayment Probability Models

Variable Names	Obs	Mean	Std. Dev.	Min	Max
Spelling errors	33975	0.315	0.464	0	4
Grammatical errors	33975	-42.39	19.394	-108.7	0
Lexical complexity (FOG)	33975	6.747	5.662	0	22.489
Positivity	33975	1.528	0.450	0	3.039
Objectivity	33975	-3.396	3.156	-3.87	2.40
Concreteness	33975	5.04	4.08	0	6.1
Spatial information	33975	8.455	23.418	0	39
Temporal Iinformation	33975	8.811	9.444	0	76
Non-first-person pronouns	33975	1	1.3	0	15
Negative emotion	33975	1.207	1.735	0	37
Credit grade	33975	5.106	1.852	1	8
Debt-to-income-ratio	33975	0.275	0.722	0	10.01
Borrowers' bank card utilization	33975	2.113	0.017	0	124
Credit inquiries	33975	9.168	0.033	0	52
Borrower rate	33975	0.20	0.00045	0	0.36
Is borrower home owner	33975	0.498	0.002	0	1
Amount delinquent	33975	1064.4	34.02	0	223738
CurrentCreditLines	33975	9.22	0.03	0	56
CurrentDelinquencies	33975	0.84	0.013	0	83
DelinquenciesLast7Years	33975	4.53	0.06	0	99
InquiresLast6Months	33975	2.13	0.01	0	97
OpenCreditLines	33975	15.33	0.25	0	554
PublicRecordsLast10Years	33975	7.97	0.03	0	51
PublicRecordsLast12Months	33975	0.34	0.0040	0	30
RevolvingCreditBalance	33975	0.03	0.001	0	7
TotalCreditLines	33975	15238	190	0	1435667
Funding option	33975	0.880	0.324	0	1
Loan amount	33975	6,266.927	5,237.64	1,000	3,5000
Category	33975	2.490	3.195	0	20
Text length	33975	139.521	135.046	0	1141
Word length	33975	4.758	1.221	0	10.19

Group member	33975	0.197	0.398	0	1
Friend investments	33975	0.0385	0.192	0	1
Month-of-loan-origination	33975	8.454	3.358	1	12

(6) Multicollinearity Analysis

As shown in following table, there is no systematic correlation or multicolinearity among these variables because the VIF values are all less than 10 (O'brien, 2007). The condition number, which is 5.6281 and less than 10, also indicates the global stability of the regression coefficients.

Table C-4 Multicollinearity Test

Variables Nam	**VIF**	**Tolerance**	**R^2**	**Eigenval**	**Cond Index**
Spelling errors	1.23	0.8144	0.1856	0.9408 2.8314	2.3814
Grammatical errors	1.76	0.5683	0.4317	0.6749 3.3428	3.3428
Lexical Complexity (FOG)	1.75	0.5709	0.4291	0.5672 3.6465	3.6465
Positivity	3.79	0.2641	0.7359	0.4973 3.8944	3.8944
Objectivity	1.67	0.5983	0.4017	0.3298 4.7821	4.7823
Concreteness	2.73	0.3667	0.6333	0.1722 6.6185	1.9169
Spatial Information	1.16	0.8594	0.1406	0.1475 7.1511	2.2626
Temporal Information	2.47	0.4056	0.5944	0.0786 9.7971	5.6281
Negative Emotion	1.66	0.6007	0.3993	0.0373 14.2157	1.629
Non-first-person pronouns	1.1	0.9075	0.0925	0.0123 24.7141	1.7523
Mean VIF	1.93			Condition Number	5.6281

Appendix D: Summary Statistics of Variables (Loan Requests)

Table D-1 Summary Statistics of Variables (Loan Requests)

Variable Names	Obs	Mean	Std. Dev.	Min	Max
Spelling errors	331665	0.315	0.464	0	4
Grammatical errors	331665	-77.39	22.394	-138.7	0
Lexical Complexity (FOG)	331665	10.747	6.568	0	30.9437
Positivity	331665	1.377	0.56	0	3.039
Objectivity	331665	-2.9996	2.7285	-3.87	2.4
Concreteness	331665	4.82	3.7	0	6.11
Spatial Information	331665	7.01163	12.629	0	39
Temporal Information	331665	7.40687	8.63731	0	89
Non-first-person pronouns	331665	0.01081	0.01528	0	0.25
Negative Emotion	331665	1.01624	1.64895	0	37

Appendix E: Summary Statistics of Variables

Table E-1 Summary Statistics

Variables	Obs	Mean	Std	Min	Max
API Information					
Previous year	135,531	739.4	120.9	0	999
Current year	135,531	753.9	111.8	1	1,000
Project Information					
Number of students supported	65,642	81.38	143.2	0	12,143
Year completed	65,642	2,008	4.257	2,000	2,013
School metro	65,642	2.521	0.709	0	3
Grade level	65,642	1.102	1.070	0	3
Number of donations	65,642	621.1	750.1	0	211
Poverty level	65,642	3.043	0.789	1	4
Primary focus area	65,642	1.672	1.268	0	3
Second focus area	65,642	1.496	1.285	0	3
Resource usage	65,642	2.001	0.996	0	5
Charter school	65,642	0.123	0.329	0	1
Magnet school	65,642	0.118	0.323	0	1
Ready promise school	65,642	0.0220	0.147	0	1
Total donations	65,642	621.1	750.1	0	100,800
School Information					
Number of student	135,531	560.8	460.4	11	4,050
Total revenue	135,531	1.628e+09	3.128e+09	0	9.700e+09
Total expense	135,531	1.688e+09	3.252e+09	-2	1.050e+10
Local School District Information					
Number of age 5-17	135,531	178,574	302,975	6	888,621
Number of household in poverty with age 5_17	135,531	50,035	89,386	0	270,712
Total population	135,531	1.073e+06	1.781e+06	62	4.645e+06
Teacher Information					
Teacher gender	65,642	3.056	1.045	1	5
Teacher teach for American	65,642	0.0492	0.216	0	1
NY Teaching Fellow	65,642	3.05e-05	0.00552	0	1

Table E-2 Basic Statistics of School Projects

Variables	Obs	Mean	Std	Min	Max
API Information					
Previous year	56,906	736.7	121.9	0	999
Current year	56,906	751.5	112.7	1	1,000
Project Information					
Number of students supported	56,906	79.80	142.6	0	12,143
Year completed	56,915	2,011	1.771	2,004	2,013

School metro	56,915	2.530	0.698	0	3
Grade level	56,906	1.071	1.060	0	3
Number of donations	56,915	6.819	8.310	0	211
Poverty level	56,915	3.052	0.784	1	4
Primary focus area	56,909	1.666	1.267	0	3
Second focus area	56,906	1.494	1.285	0	3
Resource usage	56,915	1.995	0.994	0	5
Charter school	56,915	0.122	0.328	0	1
Magnet school	56,915	0.117	0.321	0	1
Ready promise school	56,915	0.0218	0.146	0	1
Total donations	56,915	620.9	782.2	0	100,800
School Information					
Number of student	56,906	555.1	455.5	11	4,050
Total revenue	56,906	1.560e+09	3.067e+09	-2	9.700e+09
Total expense	56,906	1.620e+09	3.195e+09	-2	1.050e+10
Local School District Information					
Number of age 5-17	56,906	172,894	299,387	6	888,621
Number of household in poverty with age 5_17	56,906	48,231	88,034	0	270,712
Total population	56,906	1.036e+06	1.755e+06	62	4.645e+06
Teacher Information					
Teacher gender	56,915	4.066	1.035	1	5
Teach for America	56,915	0.0510	0.220	0	1
NY Teaching Fellow	56,915	3.51e-05	0.00593	0	1

Appendix F: Average # of Projects and Donations Per School Cross Year

Table F-1 Average # of Projects and Donations Per School Cross Year

Year	Number of Schools	Total donation	Donation per project	Average # of funded projects
2004	79	49312	624.2025	1.61
2005	280	513984	1835.657	3.33
2006	583	1149796	1972.206	4.03
2007	857	1598403	1865.114	3.87
2008	1252	1764025	1408.966	2.20
2009	1470	2583658	1757.59	3.59
2010	2461	7259656	2949.881	5.01
2011	2974	8203494	2758.404	4.75
2012	3653	15670967	4289.89	6.79
2013	3698	12549426	3393.571	4.72
2014	4375	17756962	4058.734	4.75
2015	2922	7124748	2438.312	2.93

References

Abbasi, Ahmed, Albrecht, Conan, Vance, Anthony, & Hansen, James. (2012). Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly, 36*(4), 1293-1327.

Abbasi, Ahmed, & Chen, Hsinchun. (2008). CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly, 32*(4), 811-837.

Abbasi, Ahmed, Zhang, Zhu, Zimbra, David, Chen, Hsinchun, & Nunamaker Jr, Jay F. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly, 34*(3), 435-461.

Agrawal, Ajay K, Catalini, Christian, & Goldfarb, Avi. (2011). The Geography of Crowdfunding: NBER Working Paper, No. 16820.

Agrawal, Ajay K, Catalini, Christian, & Goldfarb, Avi. (2013). Some simple economics of crowdfunding: National Bureau of Economic Research.

Ahlers, Gerrit KC, Cumming, Douglas, Günther, Christina, & Schweizer, Denis. (2015). Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice, 39*(4), 955-980.

Akerlof, George A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 488-500.

Alessandri, Guido, Caprara, Gian Vittorio, Eisenberg, Nancy, & Steca, Patrizia. (2009). Reciprocal relations among self-efficacy beliefs and prosociality across time. *Journal of Personality, 77*(4), 1229-1259.

Alpizar, Francisco, Carlsson, Fredrik, & Johansson-Stenman, Olof. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics, 92*(5), 1047-1060.

Andreoni, James. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *The Journal of Political Economy, 97*(6), 1447-1458.

Andreoni, James. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100*(401), 464-477.

Andreoni, James. (2006). Philanthropy. *Handbook of the economics of giving, altruism and reciprocity, 2*, 1201-1269.

Andreoni, James, Brown, Eleanor, & Rischall, Isaac. (2003). Charitable Giving by Married Couples Who Decides and Why Does it Matter? *Journal of Human Resources, 38*(1), 111-133.

Angrist, Joshua D, & Pischke, Jörn-Steffen. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton university press.

Aral, Sinan, Muchnik, Lev, & Sundararajan, Arun. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences, 106*(51), 21544-21549.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science, 57*(8), 1485-1509. doi: 10.1287/mnsc.1110.1370

Archak, Nikolay, Ghose, Anindya, & Ipeirotis, Panagiotis G. (2007). *Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews.*

Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Aronson, Elliot, & Mills, Judson. (1959). The effect of severity of initiation on liking for a group. *The Journal of Abnormal and Social Psychology, 59*(2), 177.

Ashton, Patricia T, & Webb, Rodman B. (1986). *Making a difference: Teachers' sense of efficacy and student achievement*. Harlow, United Kingdom: Longman Publishing Group.

Aue, Anthony, & Gamon, Michael. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. Paper presented at the Proceedings of recent advances in natural language processing (RANLP).

Awad, Neveen F, & Ragowsky, Arik. (2008). Establishing trust in electronic commerce through online word of mouth: An examination across genders. *Journal of Management Information Systems, 24*(4), 101-121.

Axsom, Danny, & Cooper, Joel. (1985). Cognitive dissonance and psychotherapy: The role of effort justification in inducing weight loss. *Journal of Experimental Social Psychology, 21*(2), 149-160.

Baek, Tae Hyun, Yoon, Sukki, & Kim, Seeun. (2015). When environmental messages should be assertive: Examining the moderating role of effort investment. *International Journal of Advertising, 34*(1), 135-157.

Bandiera, Oriana, Barankay, Iwan, & Rasul, Imran. (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica, 77*(4), 1047-1094.

Bapna, Ravi, & Umyarov, Akhmed. (2015). Do your online friends make you pay? A randomized field experiment on peer influence in online social networks. *Management Science, 61*(8), 1902-1920.

Barbosa, Luciano, & Feng, Junlan. (2010). Robust Sentiment Detection on Twitter From Biased and Noisy Data. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

Bartlett, Monica Y, & DeSteno, David. (2006). Gratitude and prosocial behavior helping when it costs you. *Psychological Science, 17*(4), 319-325.

Basoglu, Kamile Asli, & Hess, Traci J. (2014). Online business reporting: A signaling theory perspective. *Journal of Information Systems, 28*(2), 67-101.

Baum, Joel AC, & Silverman, Brian S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing, 19*(3), 411-436.

Bekkers, René, & Wiepking, Pamala. (2010). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly, 50*(1).

Bekkers, René, & Wiepking, Pamala. (2011). Who gives? A literature review of predictors of charitable giving part one: religion, education, age and socialisation. *Voluntary Sector Review, 2*(3), 337-365.

Belleflamme, Paul, Lambert, Thomas, & Schwienbacher, Armin. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing, 29*(5), 585-609.

Bem, Daryl J. (1972). Self-perception theory. *Advances in Experimental Social Psychology, 6*, 1-62.

Bertrand, Marianne, Duflo, Esther, & Mullainathan, Sendhil. (2002). How much should we trust differences-in-differences estimates? : National Bureau of Economic Research.

Betts, Julian R, Reuben, Kim S, & Danenberg, Anne. (2000). *Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California*. San Francisco: ERIC.

Beyer, Anne, Cohen, Daniel A, Lys, Thomas Z, & Walther, Beverly R. (2010). The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics, 50*(2), 296-343.

Blei, David M, Ng, Andrew Y, & Jordan, Michael I. (2003). Latent Dirichlet Allocation. *the Journal of Machine Learning Research, 3*, 993-1022.

Boersma, Paul. (2002). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

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

Brown, Steven P, & Peterson, Robert A. (1994). The effect of effort on sales performance and job satisfaction. *The Journal of Marketing, 58*(2), 70-80.

Burgoon, Judee K, Buller, David B, Guerrero, Laura K, Afifi, Walid A, & Feldman, Clyde M. (1996). Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communications Monographs, 63*(1), 50-69.

Burtch, Gordon, Ghose, Anindya, & Wattal, Sunil. (2013). An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research, 24*(3), 499-519.

Burtch, Gordon, Ghose, Anindya, & Wattal, Sunil. (2014). Cultural differences and geography as determinants of online pro-social lending. *MIS Quarterly, 38*(3), 773-794.

Burtch, Gordon, Ghose, Anindya, & Wattal, Sunil. (2015). The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Science, 61*(5), 949-962. doi: doi:10.1287/mnsc.2014.2069

Campbell, Tim S, & Dietrich, J Kimball. (1983). The Determinants of Default on Insured Conventional Residential Mortgage Loans. *The Journal of Finance, 38*(5), 1569-1581.

Card, David. (1999). The causal effect of education on earnings. *Handbook of Labor Economics, 3*, 1801-1863.

Cardon, Melissa S, Wincent, Joakim, Singh, Jagdip, & Drnovsek, Mateja. (2009). The nature and experience of entrepreneurial passion. *Academy of Management Review, 34*(3), 511-532.

Carlson, Michael, Charlin, Ventura, & Miller, Norman. (1988). Positive mood and helping behavior: a test of six hypotheses. *Journal of Personality and Social Psychology, 55*(2), 211.

Chaiken, Shelly, & Eagly, Alice H. (1983). Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology, 45*(2), 241.

Chan, Jason, & Ghose, Anindya. (2014). Internet's Dirty Secret: Assessing the Impact of Online Intermediaries on HIV Transmission. *MIS Quarterly, 38*(4), 955-975.

Chang, Angel X, & Manning, Christopher D. (2012). *SUTime: A library for recognizing and normalizing time expressions.* Paper presented at the LREC.

Cheema, Amar, & Bagchi, Rajesh. (2011). The effect of goal visualization on goal pursuit: Implications for consumers and managers. *Journal of Marketing, 75*(2), 109-123.

Chen, Xiao-Ping, Yao, Xin, & Kotha, Suresh. (2009). Entrepreneur passion and preparedness in business plan presentations: A persuasion analysis of venture capitalists' funding decisions. *Academy of Management Journal, 52*(1), 199-214.

Chiu, Chao-Min, Hsu, Meng-Hsiang, & Wang, Eric TG. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems, 42*(3), 1872-1888.

Chow, Rosalind M, & Lowery, Brian S. (2010). Thanks, but no thanks: The role of personal responsibility in the experience of gratitude. *Journal of Experimental Social Psychology, 46*(3), 487-493.

Crammer, Koby, & Singer, Yoram. (2002). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *The Journal of Machine Learning Research, 2*, 265-292.

Cui, Geng, Wong, Man Leung, & Lui, Hon-Kwong. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science, 52*(4), 597-612.

Cunha Jr, Marcus, & Caldieraro, Fabio. (2009). Sunk-Cost Effects on Purely Behavioral Investments. *Cognitive Science, 33*(1), 105-113.

Cyr, Dianne. (2008). Modeling web site design across cultures: Relationships to trust, satisfaction, and e-loyalty. *Journal of Management Information Systems, 24*(4), 47-72.

Cyr, Dianne, Head, Milena, Larios, Hector, & Pan, Bing. (2009). Exploring human images in website design: a multi-method approach. *MIS Quarterly, 33*(3), 539-566.

Daugherty, Terry, Li, Hairong, & Biocca, Frank. (2008). Consumer learning and the effects of virtual experience relative to indirect and direct product experience. *Psychology & Marketing, 25*(7), 568-586.

Davis, Angela K, Piger, Jeremy M, & Sedor, Lisa M. (2006). Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Federal Reserve Bank of St. Louis Working Paper Series*(2006-005).

Dellarocas, Chrysanthos. (2005). Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research, 16*(2), 209-230.

Dimoka, Angelika, Hong, Yili, & Pavlou, Paul A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly, 36*(2), 359-426.

Dranove, David, & Jin, Ginger Zhe. (2010). Quality disclosure and certification: Theory and practice: National Bureau of Economic Research.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *Review of Financial Studies, 25*(8), 2455-2484. doi: 10.1093/rfs/hhs071

DuBay, William H. (2004). The Principles of Readability. *Impact Information*, 1-76.

Dushnitsky, Gary. (2010). Entrepreneurial optimism in the market for technological inventions. *Organization Science, 21*(1), 150-167.

Dye, Ronald A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research, 23*(1), 123-145.

Einolf, Christopher J. (2011). Gender differences in the correlates of volunteering and charitable giving. *Nonprofit and Voluntary Sector Quarterly, 40*(6), 1092-1112.

Eliashberg, Jehoshua, Hui, Sam K, & Zhang, Z John. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science, 53*(6), 881-893.

Elliott, W Brooke, Hodge, Frank D, & Sedor, Lisa M. (2011). Using online video to announce a restatement: Influences on investment decisions and the mediating role of trust. *The Accounting Review, 87*(2), 513-535.

Esuli, Andrea, & Sebastiani, Fabrizio. (2006). Sentiwordnet: A publicly Available Lexical Resource for Opinion Mining. Paper presented at the Proceedings of LREC.

Fawcett, Tom. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Fay, Scott, Hurst, Erik, & White, Michelle J. (2002). The household bankruptcy decision. *American Economic Review*, 706-718.

Ferran, Carlos, & Watts, Stephanie. (2008). Videoconferencing in the field: A heuristic processing model. *Management Science, 54*(9), 1565-1578.

Field, Andy. (2009). *Discovering Statistics Using SPSS*: London:Sage publications.

Finkel, Jenny Rose, Grenager, Trond, & Manning, Christopher. (2005). *Incorporating non-local information into information extraction systems by gibbs sampling.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Fisher, Jeffrey D, Nadler, Arie, & Whitcher-Alagna, Sheryle. (1982). Recipient reactions to aid. *Psychological Bulletin, 91*(1), 27.

Fishman, Michael J, & Hagerty, Kathleen M. (2003). Mandatory versus voluntary disclosure in markets with informed and uninformed customers. *Journal of Law, Economics, and Organization, 19*(1), 45-63.

Flint, Thomas A. (1997). Predicting student loan defaults. *Journal of Higher Education*, 322-354.

Flynn, Francis J, & Brockner, Joel. (2003). It's different to give than to receive: predictors of givers' and receivers' reactions to favor exchange. *Journal of Applied Psychology, 88*(6), 1034.

Forman, Chris, Ghose, Anindya, & Wiesenfeld, Batia. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*(3), 291-313.

Freedman, Seth, & Jin, Ginger Zhe. (2008). Do Social Networks Solve Information Problems for Peer-to-peer Lending? Evidence From Prosper. com: Working paper,University of Michigan, Ann Arbor.

Frydrych, Denis, Bock, Adam J, Kinder, Tony, & Koeck, Benjamin. (2014). Exploring entrepreneurial legitimacy in reward-based crowdfunding. *Venture Capital, 16*(3), 247-269.

Fung, Archon, Graham, Mary, & Weil, David. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge, United Kingdom: Cambridge University Press.

Gao, Pingyang. (2010). Disclosure quality, cost of capital, and investor welfare. *The Accounting Review, 85*(1), 1-29.

Gefen, David, Benbasat, Izak, & Pavlou, Paula. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems, 24*(4), 275-286.

Gefen, David, Karahanna, Elena, & Straub, Detmar W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly, 27*(1), 51-90.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science, 31*(3), 493-520. doi: 10.1287/mksc.1110.0700

Ghose, Anindya, & Ipeirotis, Panagiotis G. (2007). *Designing novel review ranking systems: predicting the usefulness and impact of reviews.* Paper presented at the Proceedings of the ninth international conference on Electronic commerce.

Ghose, Anindya, & Ipeirotis, Panagiotis G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on, 23*(10), 1498-1512.

Gilovich, Thomas, Medvec, Victoria Husted, & Savitsky, Kenneth. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology, 78*(2), 211.

Grant, Adam, & Dutton, Jane. (2012). Beneficiary or benefactor are people more prosocial when they reflect on receiving or giving? *Psychological Science, 23*(9), 1033-1039.

Grant, Adam M, & Gino, Francesca. (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology, 98*(6), 946.

Greenwood, B.N., & Wattal, S. (2016). Show Me the Way to Go Home: An Empirical Investigation of Ride Sharing and Alcohol Related Motor Vehicle Homicide. *MIS Quarterly., Forthcoming*.

Greenwood, Brad N, & Agarwal, Ritu. (2015). Matching Platforms and HIV Incidence: An Empirical Investigation of Race, Gender, and Socioeconomic Status. *Management Science, Forthcoming*.

Gregorio, Jose De, & Lee, Jong–Wha. (2002). Education and Income Inequality: New Evidence From Cross-country Data. *Review of Income and Wealth, 48*(3), 395-416.

Grossman, Sanford J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law & Economics, 24*(3), 461-483.

Gunning, Robert. (1969). The Fog Index After Twenty Years. *Journal of Business Communication, 6*(2), 3-13.

Hamermesh, Daniel S, & Biddle, Jeff E. (1993). Beauty and the labor market: National Bureau of Economic Research.

Hancock, Jeffrey T., Curry, Lauren E., Goorha, Saurabh, & Woodworth, Michael. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*(1), 1-23. doi: 10.1080/01638530701739181

Hanushek, Eric A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-62.

Hanushek, Eric A, & Raymond, Margaret E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*(2-3), 406-415.

Hargittai, Eszter. (2006). Hurdles to information seeking: Spelling and typographical mistakes during users' online behavior. *Journal of the Association for Information Systems, 7*(1), 1.

Harmon-Jones, Eddie, Amodio, David M, & Harmon-Jones, Cindy. (2009). Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. *Advances in Experimental Social Psychology, 41*, 119-166.

Heckman, James J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society, 49*(3), 153-161.

Herzenstein, Michal, Sonenshein, Scott, & Dholakia, Utpal M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research, 48*(SPL), S138-S149.

Hess, Traci, Fuller, Mark, & Campbell, Damon. (2009). Designing interfaces with social presence: Using vividness and extraversion to create social recommendation agents. *Journal of the Association for Information Systems, 10*(12), 889.

Hildebrand, Thomas, Puri, Manju, & Rocholl, Jörg. (2016). Adverse incentives in crowdfunding. *Management Science, Forthcoming*.

Hobson, Jessen L, Mayew, William J, & Venkatachalam, Mohan. (2012). Analyzing speech to detect financial misreporting. *Journal of Accounting Research, 50*(2), 349-392.

Isen, Alice M, Clark, Margaret, & Schwartz, Mark F. (1976). Duration of the effect of good mood on helping:" Footprints on the sands of time.". *Journal of Personality and Social Psychology, 34*(3), 385.

Iyer, Rajkamal, Khwaja, Asim Ijaz, Luttmer, Erzo FP, & Shue, Kelly. (2015). Screening peers softly: Inferring the quality of small borrowers. *Management Science, Forthcoming*.

Jacob, Brian. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of Public Economics, 89*(5-6), 761-796.

Jiang, Zhenhui, & Benbasat, Izak. (2007). Research note-investigating the influence of the functional mechanisms of online product presentations. *Information Systems Research, 18*(4), 454-470.

Jiang, Zhenhui, Heng, Cheng Suang, & Choi, Ben CF. (2013). Research note—privacy concerns and privacy-protective behavior in synchronous online social interactions. *Information Systems Research, 24*(3), 579-595.

Jin, Ginger Zhe. (2005). Competition and disclosure incentives: An empirical study of HMOs. *RAND Journal of Economics, 36*(1), 93-112.

Jin, Ginger Zhe, & Kato, Andrew. (2006). Price, quality, and reputation: Evidence from an online field experiment. *The RAND Journal of Economics, 37*(4), 983-1005.

Jin, Ginger Zhe, & Leslie, Phillip. (2003). The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards. *The Quarterly Journal of Economics, 118*(2), 409-451.

Joachims, Thorsten. (1999). Making Large Scale SVM Learning Practical *Advances in kernel methods - support vector learning, B. Schölkopf and C. Burges and A. Smola (ed.)* (pp. 44-56). Cambridge, US: MIT-Press.

Joachims, Thorsten, Finley, Thomas, & Yu, Chun-Nam John. (2009). Cutting-plane training of structural SVMs. *Machine Learning, 77*(1), 27-59.

Johnson, Marcia K, & Raye, Carol L. (1981). Reality Monitoring. *Psychological review, 88*(1), 67.

Judge, Timothy A, Higgins, Chad A, Thoresen, Carl J, & Barrick, Murray R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*(3), 621-652.

Jurafsky, Daniel, & James, H. (2000). *Speech and Language Processing An Introduction To Natural Language Processing, Computational Linguistics, and Speech* (2 ed.). Englewood Cliffs, NJ: Prentice-Hal.

Kane, Thomas J, & Staiger, Douglas O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives, 16*(4), 91-114.

Kaplan, Steven N, Klebanov, Mark M, & Sorensen, Morten. (2012). Which CEO characteristics and abilities matter? *The Journal of Finance, 67*(3), 973-1007.

Kaplan, Steven N, Sensoy, Berk A, & Strömberg, Per. (2009). Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies. *The Journal of Finance, 64*(1), 75-115.

Karhade, Prasanna, Shaw, Michael J, & Subramanyam, Ramanath. (2015). Patterns in information systems portfolio prioritization: Evidence from decision tree induction. *MIS Quarterly, 39*(2), 10929_RA_Karhade.

Karlan, Dean, & List, John A. (2006). Does price matter in charitable giving? Evidence from a large-scale natural field experiment: National Bureau of Economic Research.

Karlan, Dean S. (2007). Social connections and group banking. *The Economic Journal, 117*(517), F52-F84.

Kivetz, Ran. (2003). The effects of effort and intrinsic motivation on risky choice. *Marketing Science, 22*(4), 477-502.

Kivetz, Ran, & Simonson, Itamar. (2002). Earning the right to indulge: Effort as a determinant of customer preferences toward frequency program rewards. *Journal of Marketing Research, 39*(2), 155-170.

Klein, Dan, & Manning, Christopher D. (2003). *Accurate unlexicalized parsing*. Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Klein, Lisa R. (2003). Creating virtual product experiences: The role of telepresence. *Journal of Interactive Marketing, 17*(1), 41-55.

Koning, Rembrand Michael, & Model, Jacob. (2014). *Experimental study of crowdfunding cascades: When nothing is better than something*. Paper presented at the Academy of Management Proceedings.

Kruger, Justin, Wirtz, Derrick, Van Boven, Leaf, & Altermatt, T William. (2004). The effort heuristic. *Journal of Experimental Social Psychology, 40*(1), 91-98.

Kumar, Anuj, & Tan, Yinliang. (2015). The demand effects of joint product advertising in online videos. *Management Science, 61*(8), 1921-1937.

Lewis, Gregory. (2011). Asymmetric information, adverse selection and online disclosure: The case of eBay motors. *The American Economic Review, 101*(4), 1535-1546.

Li, Feng. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics, 45*(2), 221-247.

Li, Feng. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research, 48*(5), 1049-1102.

Li, Hairong, Daugherty, Terry, & Biocca, Frank. (2001). Characteristics of virtual experience in electronic commerce: A protocol analysis. *Journal of Interactive Marketing, 15*(3), 13-30.

Lim, Kai H, & Benbasat, Izak. (2000). The effect of multimedia on perceived equivocality and perceived usefulness of information systems. *MIS Quarterly, 24*(3), 449-471.

Lim, Kai H, Benbasat, Izak, & Ward, Lawrence M. (2000). The role of multimedia in changing first impression bias. *Information Systems Research, 11*(2), 115-136.

Lin, Mingfeng, Prabhala, Nagpurnanand R, & Viswanathan, Siva. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59*(1), 17-35.

Lin, Mingfeng, & Viswanathan, Siva. (2015). Home bias in online investments: An empirical study of an online crowdfunding market. *Management Science, 62*(5), 1393-1414.

Liu, Bing. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

Liu, De, Brass, Daniel, & Chen, Dongyu. (2015). Friendships in online peer-to-peer lending: Pipes, prisms, and relational herding. *MIS Quarterly, 39*(3), 729-742.

Lobo, Jorge M, Jiménez-Valverde, Alberto, & Real, Raimundo. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145-151.

Loewenstein, George, Sunstein, Cass R, & Golman, Russell. (2014). Disclosure: Psychology changes everything. *Annual Review of Economics, 6*(1), 391-419.

Long, William J, Griffith, John L, Selker, Harry P, & D'agostino, Ralph B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research, 26*(1), 74-97.

Loughran, Tim, & McDonald, Bill. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35-65.

Loughran, Tim, & McDonald, Bill. (2014). Measuring readability in financial disclosures. *The Journal of Finance, 69*(4), 1643-1671.

Ma, Yu-Fei, Hua, Xian-Sheng, Lu, Lie, & Zhang, Hong-Jiang. (2005). A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia, 7*(5), 907-919.

Masip, Jaume, Sporer, Siegfried L, Garrido, Eugenio, & Herrero, Carmen. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law, 11*(1), 99-122.

Mathwick, Charla, Malhotra, Naresh, & Rigdon, Edward. (2001). Experiential value: Conceptualization, measurement and application in the catalog and Internet shopping environment☆. *Journal of Retailing, 77*(1), 39-56.

Mayew, William J, & Venkatachalam, Mohan. (2012). The power of voice: Managerial affective states and future firm performance. *The Journal of Finance, 67*(1), 1-43.

Meer, Jonathan. (2014). Effects of the price of charitable giving: Evidence from an online crowdfunding platform. *Journal of Economic Behavior & Organization, 103*, 113-124.

Metzger, Miriam J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078-2091.

Michels, Jeremy. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review, 87*(4), 1385-1413.

Milgrom, Paul R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics, 12*(2), 380-391.

Mitra, Sabyasachi, & Ransbotham, Sam. (2015). Information disclosure and the diffusion of information security attacks. *Information Systems Research, 26*(3), 565-584.

Mitton, Roger. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing & Management, 23*(5), 495-505.

Mobius, Markus M, & Rosenblat, Tanya S. (2006). Why beauty matters. *The American Economic Review, 96*(1), 222-235.

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

Mollick, Ethan R. (2013). Swept away by the crowd? crowdfunding, venture capital, and the selection of entrepreneurs. *Venture Capital, and the Selection of Entrepreneurs (March 25, 2013).*

Mollick, Ethan R, & Kuppuswamy, Venkat. (2014). After the campaign: Outcomes of crowdfunding. *UNC Kenan-Flagler Research Paper*(2376997).

Morewedge, Carey K, Shu, Lisa L, Gilbert, Daniel T, & Wilson, Timothy D. (2009). Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. *Journal of Experimental Social Psychology, 45*(4), 947-951.

Munnell, Alicia H, Tootell, Geoffrey MB, Browne, Lynn E, & McEneaney, James. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 25-53.

Newman, Matthew L, Pennebaker, James W, Berry, Diane S, & Richards, Jane M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665-675.

Norton, Michael I, Mochon, Daniel, & Ariely, Dan. (2011). *The'IKEA effect': When labor leads to love*. Harvard Business School Marketing Unit Working Paper.

O'brien, Robert M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673-690.

Oestreicher-Singer, Gal, & Zalmanson, Lior. (2013). Content or community? A digital business strategy for content providers in the social age. *MIS Quarterly, 37*(2), 591-616.

Olivola, Christopher Y, & Shafir, Eldar. (2013). The martyrdom effect: When pain and effort increase prosocial contributions. *Journal of Behavioral Decision Making, 26*(1), 91-105.

Özdemir, Vural, Faris, Jack, & Srivastava, Sanjeeva. (2015). Crowdfunding 2.0: the next-generation philanthropy. *EMBO Reports, 16*(3), 267-271.

Pang, Bo, & Lee, Lillian. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pang, Bo, Lee, Lillian, & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment classification using machine learning techniques. Paper presented at the

Proceedings of the ACL-02 conference on Empirical methods in natural language processing.

Pennebaker, James W, Francis, Martha E, & Booth, Roger J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*, 2001.

Pennebaker, James W, Mehl, Matthias R, & Niederhoffer, Kate G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*(1), 547-577.

Pennington, Robin, Wilcox, H Dixon, & Grover, Varun. (2003). The role of system trust in business-to-consumer transactions. *Journal of Management Information Systems, 20*(3), 197-226.

Perols, Johan, Chari, Kaushal, & Agrawal, Manish. (2009). Information market-based decision fusion. *Management Science, 55*(5), 827-842.

Petersen, Mitchell A. (2004). Information: Hard and soft. Retrieve from http://www.kellogg.northwestern.edu/faculty/petersen/htm/papers/softhard.pdf (June 30, 2015): working paper, Northwestern University.

Pope, Devin G, & Sydnor, Justin R. (2011). What's in a picture? Evidence of discrimination from Prosper. com. *Journal of Human Resources, 46*(1), 53-92.

Porter, Stephen, & Yuille, John C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20*(4), 443.

Ransbotham, Sam, Mitra, Sabyaschi, & Ramsey, Jon. (2012). Are markets for vulnerabilities effective? *MIS Quarterly, 36*(1), 43-64.

Ravina, Enrichetta. (2012). *Love & Loans: the Effect of Beauty and Personal Characteristics in Credit Markets*. Working paper, Columbia University, New York.

Rennekamp, Kristina. (2012). Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research, 50*(5), 1319-1354.

Rosenbaum, Paul R, & Rubin, Donald B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Sargeant, Adrian. (1999). Charitable giving: Towards a model of donor behaviour. *Journal of Marketing Management, 15*(4), 215-238.

Saxton, Gregory D, & Wang, Lili. (2013). The social network effect: The determinants of giving through social media. *Nonprofit and Voluntary Sector Quarterly, 19*(1).

Schwartz, Eric M, Bradlow, Eric T, & Fader, Peter S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science, 33*(2), 188-205.

Schwartz, Shalom H, & Bardi, Anat. (2001). Value hierarchies across cultures taking a similarities perspective. *Journal of cross-cultural Psychology, 32*(3), 268-290.

Schwienbacher, A. , & Larralde, B. . (2012). Crowdfunding of small entrepreneurial ventures. In D. J. Cumming (Ed.), Crowdfunding of small entrepreneurial ventures. Oxford: Oxford University Press.

Seamans, Robert, & Zhu, Feng. (2013). Responses to entry in multi-sided markets: The impact of Craigslist on local newspapers. *Management Science, 60*(2), 476-493.

Shah, Avni M, Eisenkraft, Noah, Bettman, James R, & Chartrand, Tanya L. (2015). 'Paper or plastic?' How we pay influences post-transaction connection. *Journal of Consumer Research, Forthcoming*.

Shmueli, Galit. (2010). To explain or to predict? *Statistical science*, 289-310.

Shmueli, Galit, & Koppius, Otto R. (2011). Predictive analytics in information systems research. *MIS Quarterly, 35*(3), 553–572.

Short, John, Williams, Ederyn, & Christie, Bruce. (1976). *The social psychology of telecommunications*. Hoboken, NJ: John Wiley & Sons, Ltd.

Smith, Sarah, Windmeijer, Frank, & Wright, Edmund. (2015). Peer effects in charitable giving: Evidence from the (running) field. *The Economic Journal, 125*(585), 1053-1071.

Soetevent, Adriaan R. (2005). Anonymity in giving in a natural context—a field experiment in 30 churches. *Journal of Public Economics, 89*(11), 2301-2323.

Steuer, Jonathan. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication, 42*(4), 73-93.

Strahilevitz, Michal Ann, & Myers, John. (1998). Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell. *Journal of Consumer Research, 24*(4), 434.

Sun, Monic, & Zhu, Feng. (2013). Ad revenue and content commercialization: Evidence from blogs. *Management Science, 59*(10), 2314-2331.

Sunstein, Cass R. (2000). *Behavioral law and economics*. Cambridge , United Kingdom: Cambridge University Press.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. doi: 10.1177/0261927x09351676

Tetlock, Paul C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139-1168.

Tetlock, Paul C, Saar-Tsechansky, Maytal, & Macskassy, Sofus. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437-1467.

Thaler, Richard. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization, 1*(1), 39-60.

Toma, Catalina L., & Hancock, Jeffrey T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication, 62*(1), 78-97. doi: 10.1111/j.1460-2466.2011.01619.x

Unger, Jens M, Rauch, Andreas, Frese, Michael, & Rosenbusch, Nina. (2011). Human capital and entrepreneurial success: A meta-analytical review. *Journal of Business Venturing, 26*(3), 341-358.

Vapnik, Vladimir, Golowich, Steven E, & Smola, Alex. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281-287.

Verrecchia, Robert E. (2001). Essays on disclosure. *Journal of Accounting and Economics, 32*(1), 97-180.

Viscusi, W Kip. (1978). A note on" lemons" markets with quality certification. *The Bell Journal of Economics, 9*, 277-279.

Vrij, Aldert. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*: Chichester, UK:Wiley.

Vrij, Aldert. (2008). *Detecting lies and deceit: Pitfalls and opportunities*: John Wiley & Sons.

Vrij, Aldert, Fisher, Ronald, Mann, Samantha, & Leal, Sharon. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling, 5*(1-2), 39-43.

Wan, Chin-Sheng, & Chiou, Wen-Bin. (2010). Inducing attitude change toward online gaming among adolescent players based on dissonance theory: The role of threats and justification of effort. *Computers & Education, 54*(1), 162-168.

Wang, Tawei, Kannan, Karthik N, & Ulmer, Jackie Rees. (2013). The association between the disclosure and the realization of information security risk factors. *Information Systems Research, 24*(2), 201-218.

Watson-Manheim, Mary Beth, & Bélanger, France. (2007). Communication media repertoires: Dealing with the multiplicity of media choices. *MIS Quarterly, 31*(2), 267-293.

Wheat, Rachel E, Wang, Yiwei, Byrnes, Jarrett E, & Ranganathan, Jai. (2013). Raising money for scientific research through crowdfunding. *Trends in Ecology & Evolution, 28*(2), 71-72.

Wiepking, Pamala, & Bekkers, René. (2012). Who gives? A literature review of predictors of charitable giving. Part Two: Gender, family composition and income. *Voluntary Sector Review, 3*(2), 217-245.

Wiepking, Pamala, & Maas, Ineke. (2009). Resources that make you generous: Effects of social and human resources on charitable giving. *Social Forces, 87*(4), 1973-1995.

Wilson, Michael. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*(1), 6-10.

Wilson, Theresa, Hoffmann, Paul, Somasundaran, Swapna, Kessler, Jason, Wiebe, Janyce, Choi, Yejin, . . . Patwardhan, Siddharth. (2005). OpinionFinder: A system for subjectivity analysis. Paper presented at the Proceedings of HLT/EMNLP on Interactive Demonstrations.

Wilson, Theresa, Wiebe, Janyce, & Hoffmann, Paul. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Wooldridge, Jeffrey M. (2010). *Econometric analysis of cross section and panel data*: MIT press.

Xu, Pei, Chen, Liang, & Santhanam, Radhika. (2015). Will video be the next generation of e-commerce product reviews? Presentation format and the role of product type. *Decision Support Systems, 73*, 85-96.

Xu, Sean Xin, & Zhang, Xiaoquan. (2013). Impact of Wikipedia on market information environment: evidence on management disclosure and investor reaction. *MIS Quarterly, 37*(4), 1043-1068.

Zacharakis, Andrew L, & Meyer, G Dale. (2000). The potential of actuarial decision models: Can they improve the venture capital investment decision? *Journal of Business Venturing, 15*(4), 323-346.

Zadrozny, Bianca. (2004). *Learning and evaluating classifiers under sample selection bias.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Zhang, Juanjuan, & Liu, Peng. (2012). Rational herding in microloan markets. *Management Science, 58*(5), 892-912.

Zhou, Lina, Burgoon, Judee K, Nunamaker, Jay F, & Twitchell, Doug. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, Lina, Burgoon, Judee K, Twitchell, Douglas P, Qin, Tiantian, & Nunamaker Jr, Jay F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139-166.