NORTHWESTERN UNIVERSITY

Essays on Social Networks Analytics in Customer Relationship Management

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Panteleimon M. Loupos

EVANSTON, ILLINOIS

September 2018

Abstract

Essays on Social Networks Analytics in Customer Relationship Management

Panteleimon M. Loupos

The dissertation consists of three separate essays that lie at the interface of social network analytics and Customer Relationship Management (CRM). Essay 1 and Essay 2 cover completed research, while the research covered in Essay 3 is at a more preliminary stage.

Essay 1: Starting Cold: The Power of Social Networks in Predicting Non-Contractual Customer Behavior

In this work, we provide an integrated framework for marketing managers on how to appropriately measure and manage customer behavior in a non-contractual setting in the presence of social network data. Customer behavior is directly tied to customer lifetime value (CLV) and customer equity (CE). Predicting customer behavior and their spending patterns, and consequently CLV, in such settings is a very challenging problem due to the absence of a formal declaration of termination of the customer-firm relationship. This implies that inactivity does not necessarily signal the end of the relationship, as a user may temporarily become dormant, and return at a later point in time. Distinguishing between

dormant and churned consumers is a hurdle for marketeers who need to allocate their limited resources in a way that increases the overall value of a business's customer base. Another important implication of non-contractual relationships is evident in customer-based corporate valuations (CBCV). Performing a CBCV requires knowing ahead of time how long a customer will remain with the firm, which inevitably makes non-contractual businesses prone to misvaluations. Therefore, any improvement in the ability to predict behavior in non-contractual settings is highly valuable. In this work, we study the extent to which knowledge of a customer's social network can enhance the accuracy of forecasting their behavior in terms of future: (1) activity, (2) transaction levels and (3) membership to the group of best customers. We conduct a dynamic analysis on a sample of approximately one million users from the most popular peer-to-peer (P2P) payment application, Venmo. Our models produce high quality forecasts and demonstrate that social networks lead to a significant boost in predictive performance primarily during the first month of a customer's lifetime, thus providing a remedy to the "cold-start" problem. Finally, we characterize significant structural differences with regard to network centrality, density and connectivity between the top 10% and bottom 90% of users immediately after joining the service. We discuss how these structural dissimilarities provide a path towards proactive marketing and improved customer acquisition efforts.

Essay 2: Finding Strong Ties in a Facebook Haystack: A Multilayer Social Network Approach

In this work, we investigate the question of identifying the strong ties of an individual by just inspecting the person's underlying social network structure. Strong ties have been documented to play an influential role in people's decision making process across various settings. From our decision to donate goods to our decision to turn up and vote at the elections, strong ties are the ones who exert the greatest influence on us as they convey greater trust. The digital age has re-emphasized the importance and complexity of this task, as more and more companies have now access to online friendship data of their customers. We use and extend the "social bow tie" framework introduced in Mattie et al. (2017) and apply it to a unique dataset from Venmo, the most popular P2P mobile payment application, to expand our knowledge on tie strength prediction. Our dataset is unique because it combines two different but overlapping social networks. On the one hand, we have the Venmo social graph, which comprises of all friend relationships of users that signed up with Facebook (FB). On the other hand, we have the Venmo transactional graph which reflects offline transactions among users. By following the money trail, we are able to differentiate with whom a user is really closely connected to among his FB friends, and we study the extent to which knowledge of a customer's egocentric FB social network can enhance the accuracy of forecasting whether two individuals: (1) will transact at least once, (2) whether this transaction will be reciprocated and (3) their total number of transactions. Our models produce high quality forecasts for the tasks of predicting the formation of a financial relationship and its reciprocity, yielding final Accuracy scores in the range of 43%-90% and Area Under the Precision-Recall Curve (AUPRC) values in the range of 85%-98%, depending on the exact problem formulation. For the task of predicting the total number of transactions between a pair of users, we get a Mean Square Error (MSE) in the range of 7.38-25.48 and an R^2 in the range of 0.24-0.58. The most informative predictors are found to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends. These findings are consistent with

1) Granovetter's hypothesis: the stronger the tie between any two individuals, the higher the fraction of friends they share in common, and 2) Bott's hypothesis: the higher the degree of clustering in an individual's network the less likely to form a tie with somebody outside the group.

Essay 3: Venmo for Change: The Effect of Digital Donations on Customer Engagement

In today's competitive and connected environment, organizations are investing in corporate social responsibility (CSR) activities to differentiate themselves and create a meaningful engagement with their customers. Digital platforms have reemphasized this need by introducing new forms of donating mechanisms that use social cues to inform the users about a fundraising event. Research has documented the benefits of CSR activities to organizations in terms of enhanced consumer perceptions of the company, but there is little empirical evidence on the effect of digital platform donations on customer engagement as this is expressed by any potential interaction two existing users might have on the platform. In this work, we propose a setting to empirically explore this question. Specifically, we use data from charitable fundraising events in Venmo to investigate whether two users that have contributed to the same charity event and have not previously transacted up to that point in time are more likely to transact after the charity event. Our charitable events are created by exogenous random shocks (e.g., physical catastrophes), which allow us to causally identify the effect of donations on customer engagement. We seek to test whether donating to a common cause increases the likelihood of forming a relationship between two users and whether this likelihood is a decreasing function of the shortest path distance between the two users.

Acknowledgements

First, I have to express my sincerest gratitude to my advisor Eric T. Anderson for his guidance and advice throughout my PhD years. It was my childhood dream to come and study in the US, and Eric is the embodiment of the scientific virtues I was dreaming about; intellectual curiosity, idealism, and ethos. Eric gave me complete freedom to pursue my research interests, even when these diverged from his own, and he was always there to supply me with his feedback and sharp advice. Your most valuable lesson: that doubt and "I don't know" are not to be feared but welcomed and discussed.

I was fortunate to have an amazing group of co-authors and collaborators, which is yet another illustration of the power of social networks. First, I would like to thank my friend and co-author, Alexandros Nathan. Thank you for making my PhD journey much more interesting and fun, and for teaching me the virtue of perseverance. I consider myself lucky to have worked with one of the greatest minds in social network econometrics, Professor Marchel Fafchamps. Thank you for the opportunity you provided me; your patience, compassion, understanding and encouragement. Special thanks goes to my friend and co-author, Professor Can Ürgün. I hope to continue our mathematical and philosophical discussions in the future. Last, I want to thank Professor Moran Cerf for our collaboration, his witty sense of humor, and his perspective on life.

I would also like to thank my dissertation and proposal committee members: I am very grateful to Professors Jan Van Mieghem, Sunil Chopra and Chaithanya Bandi for their

advice and comments, and to Professor Florian Zettelmeyer for our rewarding interaction the last couple of months of my PhD journey and for all his valuable lessons.

Of course, life would not be the same without the Chicago Greek gang. Special thanks to Leonidas, George, Nicholas, and the boss Alex for all the fun and luben nights we had together! My deepest gratitude also goes to my first and best US friend, Michael Roytman, and his family for all the unconditional love and support they have provided me since my first day in the US.

Last but not least, I am grateful to my family for their invaluable presence in my life, my memories, my purpose: my grandparents Eirini-Michalis and Maritsa-Pantelis, my parents Elpida-Marios, my aunt Evangelia, my uncle Nikos, and my three cousins Eirini, Maritina and Natalia. You are my everything.

Dedication

Στην Μητϵρα

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

"From all the mysteries of the universe, people are the hardest for me to fathom."
- Albert Einstein

1.1. History and Motivation

Social networks and human behavior have long fascinated social scientists and lay people alike. Some of the ideas of network analysis date back to the ancient Greeks. Plato (“similarity begets friendship”) and Aristotle ("people love those who are like themselves”) were the first to write about homophily — our tendency to associate with similar people. But it was not until the early 1930s that the first systematic approach in social network analysis occurred. It started with Jacob Moreno and Helen Jennings who developed sociometry, and continued by bright scholars in the fields of psychology and anthropology (for an exciting overview see Freeman (2004)). Nevertheless, a rigorous mathematical approach began in 1951 by Solomonoff and Rapoport (1951), who introduced the notion of a random graph. Further advances followed by Erdos and Renyi (1959-1968), who are considered to be the fathers of modern graph theory and random graphs.

As with all models, random graphs are just an abstraction of reality, and as such represent humans as mere dots on a plane. This simplification posed several limitations in studying and analyzing real world social networks which exhibit a much richer structure.

Moreover, collecting data on real world networks was extremely challenging, and unavoidably social networks research reached a plateau. The advent of the World Wide Web in 1999, and all the social media applications and digital platforms that followed, came to change all this, and marked the beginning of a new era of research in the field. Access to larger networks datasets resulted in a rapid growth in networks research allowing us to study social networks much more quantitatively.

This explosion of big data occurred not only in the field of social networks, but across many industries. According to IBM[1] we create 2.5 quintillion bytes of data every day. More interestingly, 90% of all data generated and collected by 2013 was created between 2011 and 2013. But simply storing and crunching vast stores of data adds no particular value, and thus the field of data analytics (also termed "Data Science") emerged to make sense of this data. In the following section, I provide a brief data analytics framework and try to connect each type of data analytics with their corresponding stream of research in social networks. Hopefully, this will give the reader a more thorough understanding of the chapters of this dissertation, and, more specific, in which type of data analytics each chapter belongs to.

1.2. Data Analytics and Social Networks

Data analytics can broadly be categorized into three types: 1) Exploratory Analytics, 2) Predictive Analytics, and 3) Causal Analytics.

Exploratory Analytics: Exploratory analytics refers to the process of summarizing data through descriptive statistics, often over time and/or by segments of interest.

[1] https://www-01.ibm.com/software/sg/data/bigdata/

The first wave of research in the field of social networks immediately after the availability of large real world network data belonged to this type of analytics. Social scientists, especially in the fields of physics and computer science, empirically documented for the first time the properties of real word networks and made clear that graph theory models did not fit the observations. Therefore, they proposed new mechanisms that could explain the observed properties. However, the main criticism of this stream of research is that there could be more than one potential mechanisms that can give rise to the same observed behavior, and for this reason this stream of research is now considered obsolete.

Predictive Analytics: Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data.

The second major wave of research in the field of social networks falls into this category, and it is still very active today. Although predictive analytics do not provide a causal explanation why a predicted outcome will occur, they are still of great importance to academics and industry practitioners as they provide useful insights in their decision making process. In Chapter 2, we will see how social networks can help improve predictions about a customer's behavior in a non-contractual business setting. In Chapter 3, we will see how the structure of a person's egocentric network neighborhood on Facebook can be leveraged to identify the person's most important ties in the offline world.

Causal Analytics: Causal analytics is the use of observational data, or of newly created data through experimentation to determine whether and how an action of interest causes a change in the outcome you are predicting.

Causal inference in social networks is an emerging and very challenging line of research. The difficulty arises from the fact that the notion of causality is not clearly defined in social networks. A social network system is usually comprised of various subsystems that can vary across time and space. Some or all of these subsystems are by construction endogenous, but that does not mean that they are not causal (for an excellent overview of the topic see Fafchamps (2015)). Moreover, these subsystems are interconnected rather than independent, and thus, pinpointing a causal effect to a specific subsystem is extremely challenging or in most cases futile as it is the interaction of these subsystems that cause the outcome. In Chapter 4, we discuss an ongoing work that uses an event study regression method to infer causality from observational social network data. Specifically, we examine whether digital donations to fund raising events (e.g., due to physical catastrophes) posted in Venmo's public news feed drive up customer engagement.

1.3. Social Network Analytics and Customer Relationship Management

Customer Relationship Management (CRM) refers to processes, strategies and technologies that organizations use to manage and analyze customer data and interactions throughout the customer life-cycle with the goal of maximizing the customer value to the organization. It comprises of four dimensions, namely customer retention, acquisition, identification, and development (Ngai et al., 2009).

Marketing is one of the fields that has greatly benefited from the explosion of big data. This massive expansion in the breadth of individual-level ultra-fined customer data has allowed the implementation of a long existed marketing concept, that of customer centricity. Customer centricity can be applied in all dimensions of CRM, and it refers to the ability of understanding customers at a granular level, assessing their profitability, and aligning a firm's objectives with the needs of its best customers. The most important metrics of customer centricity are customer equity (CE) (Blattberg, 2001; Rust et al., 2015) and customer lifetime value (CLV) (Gupta et al., 2004). Predicting these two metrics alone allows firms to determine who are their valuable customers and to focus their resources on building a successful long term customer-firm relationship with them.

Although the aforementioned metrics have been proved to be very effective in settings where the decision to engage with a firm is purely individual, they don't capture the full story in settings where consumers exert influence and/or are being influenced to/by their social network. In such settings, it is imperative that the CLV of every individual must be complemented by her/his network or social lifetime value. For example, a customer might have a low CLV, but a high social value. Under the traditional customer centric model, this customer would not be worth investing resources on. This can in turn jeopardize the defection status of her/his social connections that are influenced by her/his decision to churn.

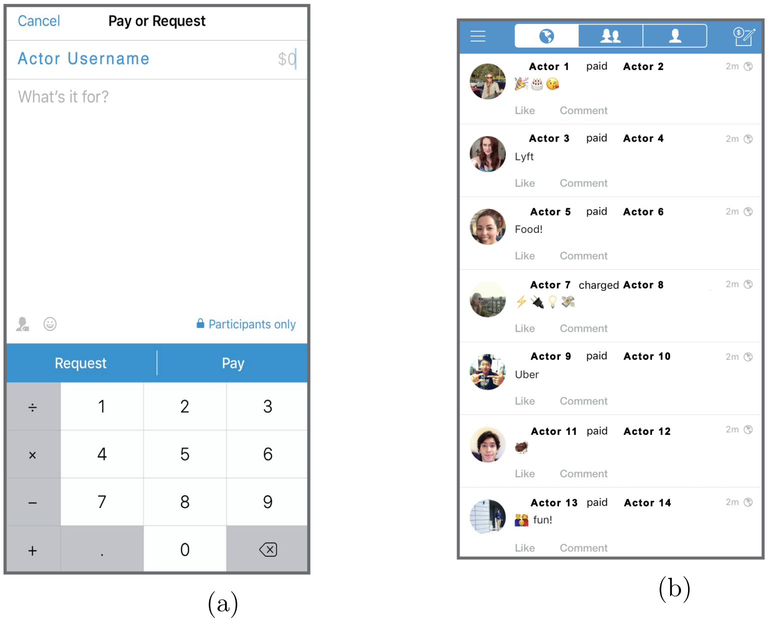
Most of the marketing applications of social network analysis have been focused on customer acquisition (Hill et al., 2006), but there is now an emerging body of research that investigates the effects of social connectivity on retention (Nitzan and Libai, 2011; Ascarza et al., 2017). The ultimate motivation of this thesis is to show that the incorporation of ultra fine-grained social network data can lead to a more integrated CRM framework.

1.4. Venmo Overview

All chapters of this dissertation use data from Venmo, a P2P mobile payment service owned by PayPal. In order to avoid repetition, a Venmo overview is provided here, along with its most interesting macroscopic statistical properties.

Venmo belongs to a broader set of P2P mobile payment applications (e.g., Square Cash, Zelle, Google Wallet, etc.) that are disrupting the traditional payment methods landscape. P2P payment services allow users to use their smart-phone to send and receive money instantaneously. What makes Venmo really unique though and has given it a competitive edge is its social nature, which has succeeded transforming financial transactions into a sharing experience. Upon logging into the application, users gain access to a Facebook-like news feed, which is composed of public transactions. The individual who initiates the transaction is required to accompany the post with a description of what the money was used for, while the dollar amount is left out for privacy reasons. Other users may ”like” or comment on the transactions that appear on their news feed. Although the public news feed is entirely open to any Venmo customer, users may opt to hide their transactions by adjusting their privacy settings. However, according to Dan Schulman, CEO of PayPal, "90% of transactions are shared"[2]. It is interesting to note that in contrast to other types of online social network ties (e.g., Facebook friendships), Venmo transactions reflect the activities of an offline network. Typical use cases for Venmo include splitting a restaurant bill (see “Actor 5 paid Actor 6” in Figure 1.1b), or paying rent or utilities (see “Actor 7 charged Actor 8” in Figure 1.1b). Venmo also offers a “charge” feature to its users, which merely serves as a reminder for payments.

[2] http://fortune.com/2017/11/17/dan-schulman-paypal-venmo/



Actor Username $0

What's it for?

(a)

Actor 1 paid Actor 2 2m

Actor 3 paid Actor 4 2m
Lyft

Actor 5 paid Actor 6 2m
Food!

Actor 7 charged Actor 8 2m

Actor 9 paid Actor 10 2m
Uber

Actor 11 paid Actor 12 2m

Actor 13 paid Actor 14 2m
fun!

(b)

Figure 1.1. Venmo's interface: Sub-figure (a) shows the tab for sending or requesting money, whereas sub-figure (b) shows the public news feed (user names have been anonymized for privacy reasons).

1.4.1. Data Collection and Pre-Processing

We used a snowball sampling approach[3] for data collection, which was implement in two stages. During the first stage, we programed a Python script to fetch public user profiles from Venmo's website (www.venmo.com), which displays public transactions from its news feed in real-time. Each publicly available transaction contains the following information: the unique user ID of the sender and the receiver, a timestamp, and the message associated with the transaction, as specified by the sender. Additionally, for each user we obtained the date they joined Venmo, as well as whether they signed up with Facebook. By the end

[3] While snowball or chain referral sampling is typically used to study a very small subset of the population, in our case it is used as a means to find as many users as possible in the Venmo community. Due to the sheer size of our final sample, issues regarding the representativeness of the sample or sampling bias do not affect our results.

of the first stage, we collected an initial sample of 1.2 million users. In the second stage, we programmed another Python script that requested a public user's full transaction history from the Venmo-Pay API (Kraft et al., 2014). After collecting the complete transaction history of all users in our initial sample, we identified all individuals they transacted with, thus augmenting our sample to over 2 million users. We repeated the second stage via Venmo's API for all newly identified users, in order to obtain their entire transaction history as well. All the aforementioned information comprises our raw dataset. For each chapter of this thesis, however, we have a different data pre-processing procedure in order to create the variables of interest. Therefore, in each chapter a more detailed description of our variables is provided. It is important to note that for all data pre-processing procedures we use Hadoop, MapReduce and Spark (Zaharia et al., 2010). This allows us to process and analyze our full dataset, and not to rely on traditional random sampling approaches.

1.4.2. Macroscopic Statistical Properties of Venmo

Venmo is the largest dataset of P2P financial activity ever to be analyzed. Its unique aspect of reflecting offline financial shared activities among friends make its properties documentation worth having. Therefore, a short overview of its key properties are summarized in Table 1.1 and explained below (we refer the interested reader to (Loupos and Nathan, 2018) for a complete overview of Venmo's properties).

Property	**Observation**
Degree Distribution	Not a Power Law
Clustering Coefficient	0.2
Median Degrees of Separation	6

Table 1.1. Venmo's statistical properties.

Degree Distribution: The degree distribution of a network is one of its most important properties. It is a power law if the number of nodes N_d of degree d is given by $N_d \propto d^{-\gamma}$ $(\gamma > 1)$ where γ is called the power law degree exponent.

Various studies have documented that power law distributions are common across many networks, such as citation networks (Redner, 1998), the Internet and the Web (Faloutsos et al., 1999; Kleinberg et al., 1999; Broder et al., 2000; Barabási and Albert, 1999; Barabási et al., 2000; Huberman and Adamic, 1999; Kumar et al., 1999), online social networks (Mislove et al., 2007) and phone call graphs (Abello et al., 1998). However, a recent study by Broido and Clauset (2018) investigated a thousand real world social networks and found that most of them do not follow a power law distribution. This is also the case with Venmo, demonstrating that real world networks exhibit a much richer structural diversity.

Clustering Coefficient: Clustering coefficient measures transitivity in social networks and takes values in $[0, 1]$ (Watts and Strogatz, 1998). In simple words, it measures the extent to which an individual's friends know each other, with a value of 1 meaning that all of an individual's friends know each other and a value of 0 meaning that no one knows each other. It has been documented that clustering coefficient in real networks is significantly higher than for random networks. Venmo's clustering coefficient reaches 0.2 in steady state, which is also common in online social networks. Its evolution, however, is different as it undergoes two distinct phases: a sharp increase, followed by a plateau around 0.2. This is in contrast with other networks, such as Google+, which shows three phases (decrease, increase and decrease again) (Gong et al., 2012).

Degrees of Separation: Degrees of Separation refers to the longstanding hypothesis that everyone in the world is connected by some short chain of acquaintances (also know as "small world" hypothesis). Venmo users are separated by a mean of 5.9 steps and a median of 6 steps. Travers and Milgram in their monumental experiments claimed that the degrees of separation across people are six (Milgram, 1967; Travers and Milgram, 1977). We should point out here that their results correspond to the "algorithmic" version of the small-world hypothesis which provides an upper bound on the average distance, whereas our results correspond to the "topological" version. Goel et al. (2009) corrected for the downward bias that existed in Milgram's experiment and estimated the median shortest path to be 7. In a recent study, Leskovec and Horvitz (2008) investigated the Microsoft Messenger instant-messaging system, a communication graph of 180 million nodes and 1.3 billion edges. They found that users were separated by a mean of 6.6 steps and a median of 7 steps. Later on, Backstrom et al. (2012) studied the Facebook social graph and found that the average degrees of separation are 4.5, claiming that the world is even smaller than we expected. Our results come to shed light to all these previous studies. On the one hand, we see that social networks tend to deflate the degrees by which people are separated. On the other hand, we find that the world is indeed smaller that we previously believed. In fact, it is 6 degrees separated.

CHAPTER 2

Starting Cold: The Power of Social Networks in Predicting Non-Contractual Customer Behavior

Joint work with Alexandros Nathan and Moran Cerf

2.1. Introduction

Over the last decade, there has been a surge of services that connect people and facilitate a wide range of interactions. Although this used to be primarily the case in telecommunications settings, social platforms have become commonplace in a variety of industries: online gaming (e.g., World of Warcraft), payment services (e.g., Google Wallet, Venmo), messaging applications (e.g., WhatsApp, SnapChat), and sharing economy services (e.g., Lyft or Uber, where riders split a fare). The majority of these services operate under a non-contractual business model, which increases the complexity of predicting future customer behavior, due to the inability of observing customer defection in real time, coupled with highly irregular spending patterns both in terms of inter-purchase time and amounts. The standard approach to making such predictions involves collecting data on a user's past behavior, and building statistical models to extrapolate a user's actions into the future. This framework relies on the assumption that past behavior is the best predictor of future behavior. However, this method fails in the case of newly acquired customers where past behavior is virtually non-existent. Drawing inferences about a user's future behavior in the absence of any historical data is known as the "cold-start" problem.

The “cold-start” problem poses challenges to marketing managers and financial professionals in a number of ways. First, businesses in today’s fast-paced environment need to allocate their limited marketing resources immediately after a new customer has been acquired. This inevitably implies that a large portion of the marketing budget may be wasted on one-time customers, who have no intention to engage with the firm in the long-term; it has been documented that 68% of newly acquired customers are non-profitable (McCarthy and Fader, 2017). Second, it is common practice to value businesses using a Customer-Based Corporate Valuation (CBCV) (McCarthy et al., 2017; McCarthy and Fader, 2017). Performing a CBCV requires knowing ahead of time how long a customer will remain with the firm, which consequently makes non-contractual businesses prone to mis-valuations. Therefore, any improvement in the ability to predict customer behavior in non-contractual settings is highly valuable.

In this work, we focus on utilizing social network information as a means to alleviating the “cold-start” problem in non-contractual settings. Specifically, we leverage a customer’s social network to investigate the following aspects of customer behavior: (1) distinguishing between active versus inactive customers, (2) predicting future transaction volumes at the individual level, and (3) identifying best future customers. To this end, we use data from Venmo, the most popular peer-to-peer (P2P) mobile payment application. We analyze over 100 million public transactions from approximately one million Venmo users, which to the best of our knowledge is the first large-scale analysis of P2P financial transactions.

Our investigation is organized in three parts. First, we predict short and long-term customer activity, where activity is defined as the act of using Venmo to send or receive money. In particular, we build and evaluate several competing models with the ultimate goal of quantifying the benefit of incorporating social network metrics in predicting future usage, as opposed to restricting the analysis to conventional user-based metrics. Our results confirm previous studies claiming that recency and transaction frequency play a key role in predicting customer activity (Coussement and De Bock, 2013). More importantly, however, we find that user and social network attributes are complementary to each other in the following manner: at the beginning of the customer lifecycle, social networks are most indicative of future activity, but later on, it is past customer behavior that takes over as the strongest predictor of future activity. Specifically, we find that social network metrics provide a significant boost in predictive performance early on in a customer's lifetime - relative increase for the Area Under the Receiver Operating Characteristic Curve (AUC) and top-decile Lift is in the range of 15%-28% and 28%-35% respectively, depending on the size of the prediction horizon. In the later stages of a customer's lifetime, user-based attributes yield a relative increase of 1%-4% and 1%-10% in AUC and top-decile Lift, respectively, compared to the network-based model.

Second, we address the tasks of predicting future transaction levels and identifying a firm's best customers. Due to the absence of dollar amounts for all transactions, a customer is deemed best based on their transaction volume. Similar to the activity problem, we find that social network attributes lead to a performance enhancement early on in a customer's lifetime; however, this effect persists into the later stages of the user lifecycle. During the first month of a customer's lifetime, models incorporating social network information yield a relative increase of 13.0% in Mean Square Error (MSE) in the case of future transaction levels. In the task of predicting future best customers the relative increase in AUC is in the range of 20.6% - 22.0% and in top-decile Lift it is in the range of 35.4% - 45.8% depending on the exact definition of a top customer.

Finally, we study the structural network differences between the top 10% and bottom 90% of customers. The motivation for examining these structural dissimilarities is that while transaction frequency and recency have consistently been identified as the best predictors of customer behavior, they are not purely proactive metrics; for example, the decline of a user's transaction frequency is the outcome and not the cause of their decision to stop using a service. We investigate the link between network centrality, connectivity, density and future tier immediately upon acquiring a customer. We find that top customers have a propensity to join dense and highly connected communities. This observation can lend itself to the improvement of customer acquisition initiatives.

Our main contribution is a framework for incorporating social network information to predict customer behavior. This approach is instrumental in overcoming the "cold-start" problem, which has significant managerial applications for many companies. The remainder of this chapter is organized as follows: the following section relates our work to the existing literature; then, we provide an overview of our data and methods; next, we outline the results of our analysis; and finally, we conclude with a summary of our findings and their managerial implications.

2.2. Literature Review

Customer-firm relationships can be divided into two categories: contractual and non-contractual (Schmittlein et al., 1987). One of the fundamental differences is that in contractual settings, customer defection can be observed in real-time as a customer opts to terminate a service by not paying the subscription fee. In non-contractual settings, however, customer inactivity does not necessarily imply defection. The inability to directly observe the end of the customer-firm relationship makes the task of calculating Customer Lifetime Value (CLV) or retention challenging. Over the years, researchers have developed two main approaches for modeling customer behavior in non-contractual settings: parametric probability models and machine learning models.

Probability models assume that customer behavior follows a specific parametric distribution. The assumed distribution is imposed on an individual-level, and is then aggregated across heterogeneous individuals to obtain the parameters of the joint distribution via maximum likelihood estimation. Some of the most popular parametric probability models include the Negative Binomial Distribution (NDB) model (Ehrenberg, 1959; Morrison and Schmittlein, 1988), the Pareto/NDB model (Morrison and Schmittlein, 1988; Schmittlein et al., 1987; Jerath et al., 2011) and its beta-geometric (BG)/NBD extension (Fader et al., 2005; Fader and Hardie, 2009). Although there exist successful applications of these models (Fader et al., 2010), they all impose strong distributional assumptions on a customer's behavior, which if not satisfied can create complications. For instance, it has been documented that the Pareto/NBD model may produce unrealistic customer lifetime estimates (Wübben and Wangenheim, 2008). Additionally, it is difficult to extend these models to non-stationary regimes, and to incorporate time-dependent covariates apart from recency and frequency. This stems from the difficulty of obtaining a closed-form to the likelihood function, which forces one to rely on simulations. Finally, it has been shown that simple, industry-specific heuristics for distinguishing between active and inactive customers often perform at least as well as complex models, thus not justifying the investment in implementing such solutions in managerial practice (Wübben and Wangenheim, 2008).

These limitations have given rise to the application of machine learning methods to modeling customer behavior, which can seamlessly incorporate a wide variety of covariates, and capture non-linear relationships among them. Some notable works showing the successful applicability of machine learning models in non-contractual settings include European financial services (Larivière and Van den Poel, 2004), Google AdWords (Yoon et al., 2010), e-commerce (Yu et al., 2011), Yahoo answers (Dror et al., 2012) and telecommunications (Tamaddoni et al., 2015; Ahn et al., 2006). A comparison of parametric probability models and machine learning methods in terms of churn predictions is provided in (Tamaddoni et al., 2015). In order for the comparison to be fair and meaningful, the authors use only the covariates (recency and frequency of purchases) that can be incorporated in the Pareto/NBD model. Even in that case, machine learning methods give superior results over the parametric probability models. In a recent paper, (Babkin and Goldberg, 2017) develop an extension of the BG/NBD model, which is able to utilize any kind of covariates, including time-dependent variables and monetary values from transactions. They show that their model is superior to the traditional BG/NBD model, but they do not compare it against any machine learning techniques. Such a task is beyond the scope of our paper.

The recent rise of information technology and online social networks has allowed marketing researchers and practitioners to investigate the effect of social influence on consumer preferences and behavior. Network-based marketing has proven to be a very effective tool in customer acquisition, especially in the area of new product adoption; for a concise review, see (Hill et al., 2006). Customer acquisition, however, is not the only practice to benefit from the abundance of social network data. As noted in (Hill et al., 2006; Bijmolt et al., 2010), customer churn can be contagious among groups of friends and, hence, researchers have proposed several strategies to incorporate social network information into their models.

The vast majority of this research has focused on contractual customer-firm relationships within the telecommunications industry. The underlying social network of customers is retrieved by analyzing phone call records and text messages, and the main finding of this line of work is that "word-of-mouth" effects, wherein individuals who churn (do not churn) can influence their friends to churn (not to churn), are present in the world of telecommunications. More specifically, (Nitzan and Libai, 2011) and (Verbeke et al., 2014) show that a customer is more likely to defect if their connections churn. (Haenlein, 2013) demonstrates that although social influence plays a key role in customer defection, this effect depends on the directionality of the communications between friends; individuals who primarily receive phone calls from churners are at a higher risk of defecting. In a related paper, (Haenlein, 2013; Ascarza et al., 2017) conduct a field experiment involving nearly 6,000 customers of a mobile telecommunications provider and find that customers are less likely to defect if their friends within the company continue to use the service. Another interesting result concerning customer defection is its relationship with "social embeddedness" (Haenlein, 2013; Ascarza et al., 2017; Benedek et al., 2014), wherein a high degree of connectivity within the provider's network is negatively correlated with churn. Finally, (Richter et al., 2010; Moldovan et al., 2017) show that customers typically leave in groups, and identify leaders of dense social groups whose departure from the service provider may lead to the entire group defecting.

Less work has been devoted to investigating social network effects in non-contractual settings. One such paper is that of (Dasgupta et al., 2008), which focuses once again on telecommunications networks but with an emphasis on pre-paid phones. The authors model social influence as a diffusion process, and show improved churn predictions. It is important to note that there exist some significant differences between social applications like Venmo, and telecommunications services. First, payment and telecommunication activities are quite different in nature, and therefore, it is likely that the respective social networks of customers exhibit different network structures. Second, due to the prevalent role of mobile phones in today's world, it is possible that churn in a telecommunications setting is equivalent to changing network providers. In Venmo, on the other hand, it is not clear whether a permanently inactive customer has joined a competitor, or has resorted to using other forms of payment (e.g., cash, credit cards). Finally, the notion of "social embeddedness" defined in (Benedek et al., 2014), where customers can place calls outside the network of their service provider, is unique to the telecommunications industry. Services such as Venmo, or WhatsApp allow transactions or communications only between users within the application, making it a more restrictive service compared to telecommunications.

Our work is also related to the network formation literature. Specifically, the occurrence of P2P transactions is essentially equivalent to the act of building and strengthening network ties. On one hand, predicting customer activity is a variation of the link prediction problem (Benedek et al., 2014; Liben-Nowell and Kleinberg, 2007). The main difference between the two problems is that in the link prediction setting the focus is not on forecasting whether an existing tie will be strengthened, but rather on whether a tie between every pair of nodes will be formed or not. On the other hand, our examination of the link between the architecture of Venmo's social network and customer behavior is closely connected with recent studies on how network structures influence behavior and outcomes (Fafchamps et al., 2010; Aral and Walker, 2014a); however, to the best of our knowledge, there has been no previous work that investigates the link between behavior and network structures in the context of customer activity and engagement.

Finally, the inclusion of social network information to solve the "cold-start" problem has been studied in several other settings. In the case of recommender systems, (Jamali and Ester, 2010, 2009; Bellaachia and Alathel, 2016; Liu et al., 2011) show that using social network information can significantly improve recommendation accuracy, especially when a user has just joined the service and has provided very little feedback about their preferences. In the setting of research activity, (Ductor et al., 2014) examine how the co-authorship network of an individual researcher can help predict their future research output. The authors' findings indicate that the predictive power of network information is strongest for young researchers, who typically have a smaller number of publications in comparison with more seasoned researchers.

The closest paper to our work is that of (Benoit and Van den Poel, 2012). The authors investigate the relative importance of network metrics versus user metrics in a retail banking setting, and find that social network attributes can indeed provide a boost in predictive performance. However, this work has two important limitations: first, their network data is limited to the relatives of each customer who are also members of the bank, and second, the authors use a static analysis to assess the importance of network-level metrics in predicting customer activity. As we show empirically in our work, the predictive power of network-based attributes is not fixed over time, and such metrics can be most valuable in the beginning of a customer's lifetime.

2.3. Data Overview and Methodology

2.3.1. Data

Our final sample exceeds 2 million users, but we restricted our analysis to 981,369 users for whom we had the complete transaction history over 12 consecutive months. An overview of our final dataset is shown in Table 2.1.

Dataset Overview	
Sample size	981,369
Time frame	January 2014 - June 2016
Lifetime analysis per customer	0 - 12 months
Calibration/Training sample size	686,952
Holdout/Testing sample size	294,408

Table 2.1. Descriptive statistics of final dataset.

2.3.2. Methodology

The unit of analysis in our work was an individual user. To evaluate the impact of social network attributes on predictive accuracy over time, we treat time as a sequence of discrete monthly intervals, and we use a design that groups customers by their "age" in Venmo (we call this age "lifetime" herein). A lifetime of 0 indicates that a user has just joined the service, and the termination point of our analysis is at lifetime 12 (months). We use this design because the effect of social network metrics on predictive performance may be different for customers with different lifetimes.

During the data pre-processing step, we create a set of 32 variables/features, which can be decomposed into two categories: user-specific (e.g., transaction frequency, early Venmo adopter) and social network attributes (e.g., number of friends, pagerank). At every discrete point in a user's lifetime we update all user and network-level attributes to reflect their most recent transactional behavior and make predictions about future usage; the time period over which the variables are re-evaluated is called the feature evaluation window. Some of our predictor variables remain static throughout a user's lifetime, but the vast majority of them evolve over time. Table A.1 in Appendix A.1 provides a detailed description of all 32 variables and indicates each variable's type. Depending on the task at hand (forecasting activity, number of transactions and best customers), the prediction window differs. Figure 2.1 demonstrates the prediction framework when the window is 30 days.

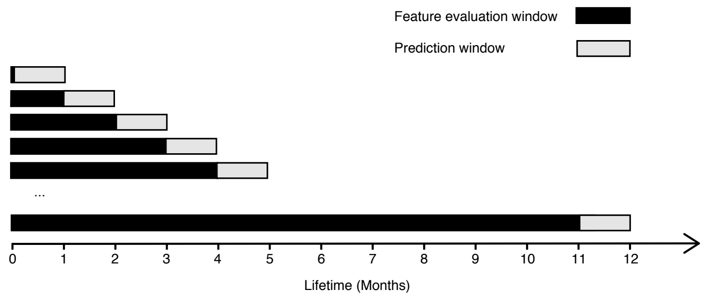


Figure 2.1. Illustration of the feature evaluation and prediction windows at each lifetime point of a user. In this example the prediction window is one month.

The main modeling framework that has been used in previous studies (Wübben and Wangenheim, 2008; Fader et al., 2005; Malthouse and Blattberg, 2005) is the following: given three distinct, evenly spaced time points t_0, t_1, and t_2, a statistical model is fit on the calibration period $[t_0, t_1]$ and the predictions are made on holdout data coming from the time interval $(t_1, t_2]$. There are two main challenges to this design compared to the one we follow in this chapter. First, it is not possible to examine the effect certain attributes have on predictive accuracy at all points of a customer's lifetime. Specifically, it is not possible to assess the importance of different predictors for any time point within the interval $[t_0, t_1]$. Second, it does not exploit the benefits of big data and new information technologies, such as the ability to process the continuous inflow of data in real time (Malthouse et al., 2013). That is, if new data has become available after t_1 it is wise to update any previous models to incorporate the latest information. This is particularly fitting in fast-paced environments, such as mobile applications, where 80% of customers churn within the first three months (Perro, 2016).

2.3.3. Modeling

We model three different but intertwined problems related to customer behavior in a non-contractual setting, namely predicting customer activity, predicting future transaction levels, and identifying the top 10% of customers at the end of the first year. Whereas the dependent variable will differ in each problem, they all share the same underlying modeling approach. In particular, in order to assess the relative importance of social network metrics compared to user metrics, we estimate three competing models. Benchmark Model 1 focuses on the predictive power of user-based variables $x_{i,t}$, i.e.

$$\text{Model 1: } y_{i,t+w} = F(x_{i,t}),$$

where w corresponds to the length of the prediction window. User variables include whether a user has signed up with Facebook, if they are an early adopter, the percent of their Venmo transactions taking place at night, the percent of their Venmo transactions taking place on weekends, and lastly, their transaction frequency. In Model 2, we examine the predictive power of network variables $z_{i,t}$, i.e.

$$\text{Model 2: } y_{i,t+w} = F(z_{i,t}).$$

Network variables include several variables that describe a user's network centrality (e.g., pagerank), density (e.g., triangle count) and connectivity (e.g., number of friends). Finally, in Model 3 we explore whether the combination of user and network variables can lead to higher quality forecasts compared to Models 1 and 2, i.e.

$$\text{Model 3: } y_{i,t+w} = F(x_{i,t}, z_{i,t}).$$

In all three models we include time dummies to account for the month and year that a user joined Venmo. We experiment with several functional forms of $F(\cdot)$. In particular, we fit and compare linear and logistic regression models (including regularization techniques) and random forests. In the case of linear and logistic regression models we also allow for the possibility of time-varying coefficients. As noted by (Bijmolt et al., 2010; Ascarza et al., 2018), while most models assume that customer behavior remains stable over a user's lifetime, a more realistic approach should allow for time-varying coefficients, which implies that the coefficient or weight of a covariate can vary over the lifetime of a customer. Table 2.2 summarizes the prediction problems and functional forms we test. For more details on the definition of each dependent variable see the corresponding subsection under Empirical Results. Note that the model evaluation procedure outlined below is applicable to all three predictive tasks we undertake.

Predictive Task	**Dependent Variable Type**	**Functional Forms**
Customer Activity	Dichotomous	Logistic, Lasso, Random Forests
Future Transaction Volume	Continuous	Linear Regression, Random Forests
Top 10% and 20% Customers	Dichotomous	Logistic, Lasso, Random Forests

Table 2.2. Description of predictive tasks, dependent variable types, and functional forms.

2.3.4. Model evaluation

To evaluate our models, we split our data into two random, non-overlapping sets: the calibration or training sample (686,952 users - 70%), and the holdout or testing sample (294,408 users - 30%). Given that we make multiple predictions per customer at different lifetime points, in order to obtain fair, out-of-sample estimates of predictive accuracy, we ensure that all lifetime points of a customer belong to one of the two sets. For each classification model we report the AUC score, which is a popular performance metric in the machine learning literature, and top-decile Lift, which is a common metric in marketing. For our linear regression model we use the evaluation scheme outlined in (Fader et al., 2005), which focuses on the expected number of transactions, conditional on the number of observed transactions. We also report the Mean Square Error (MSE) for the linear regression model.

2.4. Empirical Results

2.4.1. Predicting Customer Activity

The goal of this section is to address one of the fundamental questions in customer relationship management: which customers will be active in the future? To answer this question one needs to first define a measure of activity. Over the years researchers have proposed a variety of metrics to capture customer activity, the simplest one being recency (how long has it been since a customer last used a service?), which we also employ here. To ensure the robustness of our results, we experiment with four thresholds for defining customer activity: 30, 60, 90 and 120 days. To formalize this process, let $y_{i,t} \in \{0, 1\}$ be a dichotomous variable equal to 1 if individual i has not transacted at least once during the activity window that ends at time t, and 0 otherwise. For example, if the activity window is 30 days, c means that user i has not been active during the 9th month of their lifetime. Note that we require the activity window to be the same as the predictive window. In other words, if the activity window is equal to 30 days, all predictions at time t will be for 30 days into the future, i.e. time $t + 1$.

We report the results of our three models in Figure 2.2. Note that across all these models, all types of functional forms (variations of logistic regression and random forests) achieve approximately the same predictive performance - see Table A.2 and Table A.3 in Appendix A.2. For this reason, we present the results obtained using a lasso logistic regression model, which performs a variable selection procedure and sheds light into the importance of the different features. We find that at lifetime 0, Model 1 performs almost as poorly as a random guess (AUC of 53%, top-decile Lift of 1.13). Model 2 outperforms

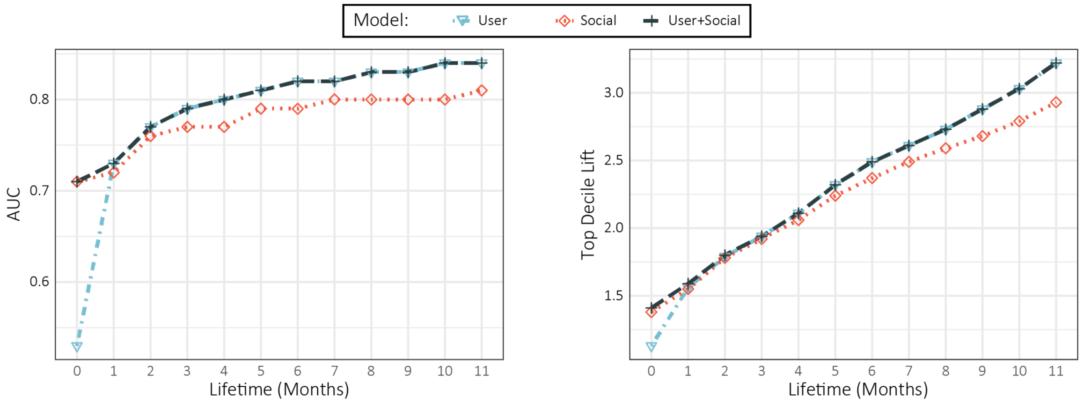


Figure 2.2. Classification results for predicting future customer activity by lifetime; the predictive window is set to 30 days. The figure on the left depicts the AUC score at each distinct time point, whereas in the figure on the right the lifetime points are plotted against top-decile Lift.

Model 1 (AUC of 71%, top-decile Lift of 1.38), while Model 3 yields the best overall performance (AUC of 71%, top-decile Lift of 1.41): a relative increase of 33.9% in AUC and 24.8% in top-decile Lift compared to Model 1.

Our results indicate that social network metrics provide a significant boost in predictive performance early on in a customer's lifetime, and can play an instrumental role in predicting customer activity in the absence of a user's historical data. Immediately after lifetime point 0, however, the effect of the social networks metrics tails off, and it is a user's past behavior that becomes most indicative of future activity. Consistent with prior work, we find that transaction frequency and recency are the strongest predictors of future customer activity (Coussement and De Bock, 2013). This finding is confirmed via the lasso model, which in later stages of a user's lifetime drops all variables but transaction frequency and recency. As far as the time-varying coefficient approach is concerned, we only observe a small improvement at lifetime 0 (relative increase in AUC of 6%, 6%

and 7% for Models 1, 2 and 3 respectively), which goes away immediately after. Our findings remain unchanged regardless of the size of the activity and prediction windows (see Table A.4 and Table A.5 in Appendix A.2).

2.4.2. Predicting future transaction levels

While being able to accurately predict which customers will be active in the future is valuable for a firm, it does not provide deep insights into customer profitability. For example, a customer who transacts once per month and a customer who transacts ten times per month are identical from an activity standpoint but are very different in terms of profitability. To distinguish between such customers in a non-contractual setting, one needs to model repeat purchases and generate forecasts of future transaction levels. In this section we explore the extent to which social network attributes can yield improved forecasts when predicting the total number of transactions completed by lifetime 12. That is, the dependent variable for user i is

$$y_{i,12} = \text{ transactionCount}_{12}.$$

We again build three competing models and we follow the approach of (Fader et al., 2005) to compare their performance. Simply put, we examine individual-level predictions on the test set conditional on the number of observed transactions at a particular time point. Figure 2.3 summarizes our results at four distinct lifetime points, namely 0, 3, 6 and 9.

All models provide excellent predictions of the expected number of transactions at all lifetime points. Clearly, the closer the predictions get to lifetime 12, the more accurate

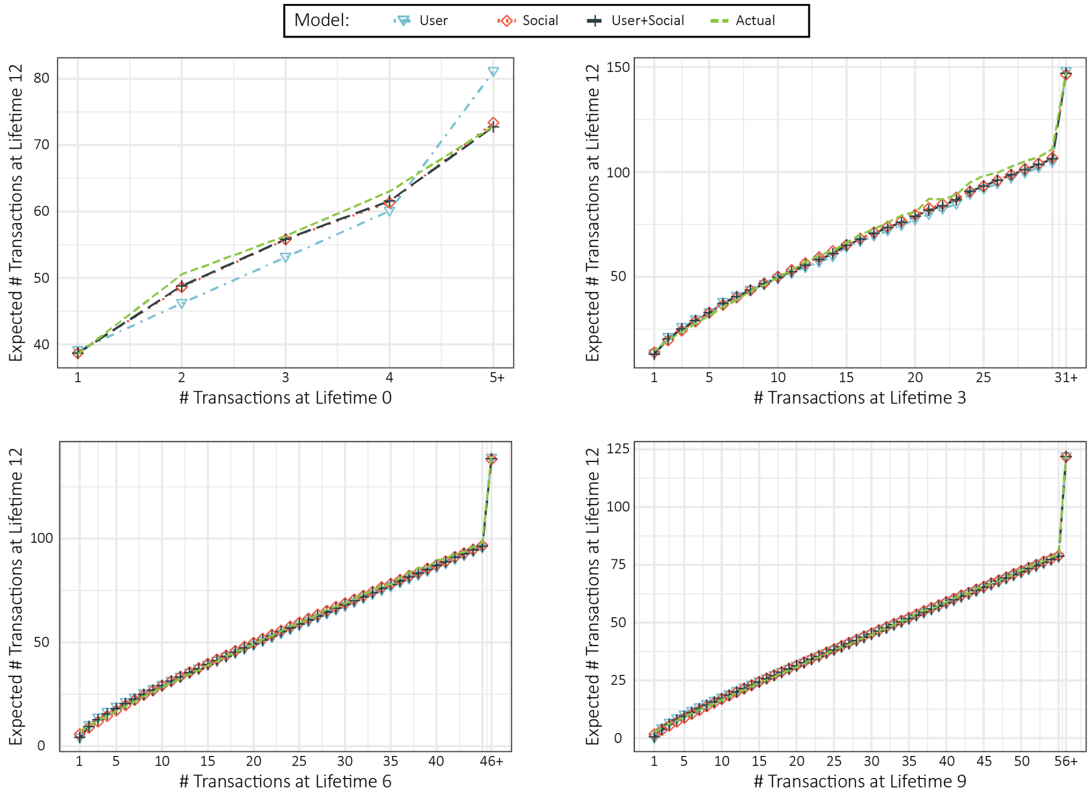


Figure 2.3. Results on predicting the expected transaction levels at lifetime 12, conditional on past number of transactions at lifetime points 0, 3, 6 and 9.

they become. We should note that Model 2 and Model 3 have a slight edge at all lifetime points, and particularly at lifetime 0 (relative decrease of 13.0% in MSE for both models). Table 2.3 summarizes the MSE of each model.

2.4.3. Identifying future best customers

In a truly customer centric setting, a firm should be able to identify the most profitable customers, and invest disproportionate resources to keep them loyal and develop them. The notion that not all customers should be treated equally relies on the fact that typically

Lifetime	MSE		
	Model 1	Model 2	Model 3
0	1699	1503	1503
1	1274	1197	1197
2	1040	995	995
3	863	815	815
4	708	662	662
5	558	520	520
6	423	395	395
7	311	291	291
8	218	203	203
9	141	131	131
10	75	69	69
11	27	65	65

Table 2.3. Mean Square Error (MSE) when predicting future transaction levels via a linear regression model.

a small percentage of customers is responsible for generating the largest portion of revenues and profits for a firm (Mulhern, 1999). For a customer to belong to the top 10% in our dataset, they need to complete at least 90 transactions by lifetime 12. That is, the dependent variable for user i is defined in the following manner:

$$y_i = \begin{cases} 1 & \text{if } transactionCount \geq 90 \\ 0 & \text{o.w.} \end{cases}$$

We test our model when the best group is the top 20% of customers (at least 63 transactions) for robustness purposes (see Table A.6 in Appendix A.2). Figure 2.4 summarizes the effect of social metrics in determining Venmo's best customers. The previously observed theme of a performance improvement at lifetime 0 is seen here once again for the top 10% (20%) of customers: the relative increase in AUC is 22.0% (20.6%), and for top-decile Lift it is 45.8% (35.4%). Nevertheless, in contrast with the activity prediction problem, Model 2 can essentially achieve the same predictive power as Model 1, or even outperform it, in all lifetime points after 0. In other words, it is possible to forecast a customer's future tier by taking under consideration only the past behavior of their friends. Note that while it is possible to predict the most elite group of customers using the the linear regression model from the previous section, the logistic regression model produces more accurate estimates (see Table A.7 in Appendix A.3). For this reason, we only present the results obtained from the lasso logistic regression model.

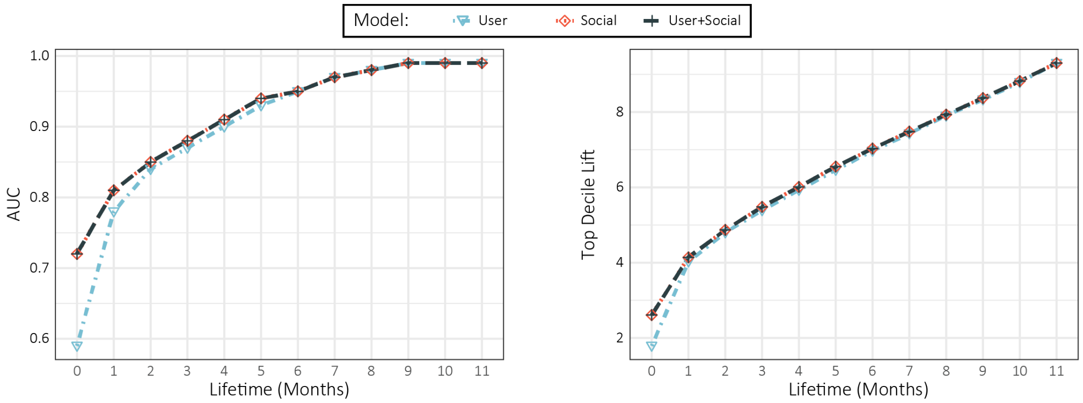


Figure 2.4. Classification results for top 10% of customers by lifetime. The figure on the left depicts the AUC score at each distinct time point, whereas in the figure on the right the lifetime points are plotted against top-decile Lift.

2.4.4. Investigating network structural properties of best customers

Evaluating the structural network predictors of Venmo's best customers is an intricate task. Customer activity in the context of Venmo is precisely the act of building and strengthening network ties. This implies that judging solely by the total number of completed transactions, a customer who belongs to the top 10% is much more likely to have higher degree, triangle count and pagerank compared to a user who belongs to the bottom 90%, especially in later lifetime points.To alleviate this issue, we focus our attention on lifetime 0, when the number of transactions of the top 10% and the bottom 90% of customers is identical. Specifically, at lifetime 0, both user cohorts have a median of 1 transaction. In this way, the network differences we identify between the two groups of customers will provide insights about the communities that best customers typically join.

2.4.5. Overview of structural network characteristics

The structural network characteristics of interest (see Table 2.4) can be further subdivided into local and global metrics. This distinction emphasizes the fact that local metrics are obtained by considering only a user's first and second degree connections, whereas global metrics reflect an individual's position in the greater network, and take into account the overall organization of ties. Since some of the structural metrics at hand have not been studied in such a content before, they warrant a more detailed description which we provide in Appendix A.4.

2.4.6. Top 10% versus the rest

To understand which are the most important predictors of the top customer tier, we present the coefficients of a logistic regression model containing only the seven structural network covariates in Table 2.5. While all variables turn out to be significant, due to the large size of our data set, we are cautious with the use of statistical significance tests. As explained in (Coe, 2002), the p-values associated with statistical significance tests depend on two quantities: the magnitude of the effect and the size of the sample. Consequently,

Measure	Type	Description
Triangle count	Local	Measures connectedness and density in a community. Provides an absolute count of triadic relationships, in contrast with cohesion, which is normalized.
Cohesion	Local	Measures the extent to which a user's friends know each other. Cohesion has been linked in the literature to brokerage (Stovel and Shaw, 2012).
Mutual Friends of Friends	Local	Measures the extent to which a user's friends share mutual friends, even if they are not connected directly to each other.
Number of friends	Local	Number of first degree connections, irrespective of direction of the link (i.e. does not account for who initiated the transaction).
Outgoing transaction percentage	Local	The percentage of outgoing transactions captures the directionality of money transfers.
Friends average number of friends	Local	Average number of friends of all first degree connections.
Friends of friends average number of friends	Local	Average number of friends of all second degree connections.
Giant component	Global	Indicator of whether a user belongs to the giant component. Network theory suggests that most social networks contain one large, connected component comprised by a significant fraction of all nodes.
Pagerank	Global	Measures centrality in a directed network, and can be interpreted as a metric of importance based on a node's connections.

Table 2.4. Description of structural network variables.

it is possible that a result is deemed significant entirely because of the large size of the sample. To avoid such issues, we use Cohen's effect size to quantify the mean difference between the top 10% and bottom 90% of users. Cohen's effect size is known to emphasize the size of the difference rather than conflating it with the size of the data sample (see Appendix A.5 for details).

Metric	**Coefficient**
Triangle count	0.04 *** (0.003)
Cohesion	0.76 *** (0.013)
Mutual Friends of Friends	1.32 *** (0.036)
Number of friends	0.01 ** (0.006)
Outgoing transaction percentage	0.11 *** (0.008)
Friends average number of friends	0.003 *** (0.0003)
Friends of friends average number of friends	0.020 *** (0.0005)
Giant component	0.19 *** (0.013)

** p<.01
*** p<.001

Table 2.5. Coefficient for logistic regression model predicting the top 10% of customers using only structural network variables.

We present our results in Table 2.6, which shows Cohen's effect size for the mean difference at lifetime 0 - the table also includes the mean and standard deviation of all metrics for each group. We observe significant structural dissimilarities between the two cohorts. In particular, it is clear that density (triangle count, cohesion, mutual friends of friends) and centrality (pagerank, number of friends) measures are positively correlated with the propensity of users to become top customers. Additionally, membership in the giant component has a small but noticeable effect size, which can be attributed to the general concept of connectivity. Giant component is a term that describes a frequently observed property of social networks: the vast majority of nodes belong to a single, connected component, i.e. there is a path connecting any two nodes that belong to the component (Easley and Kleinberg, 2010). The metric that particularly stands out in our data is mutual friends of friends (MFF), since it hints that a customer's tier can be predicted by individuals who are two degrees apart.

Metric	**Lifetime = 0**		
	Top 10%	**Bottom 90%**	**Cohen's**
Triangle count	0.75 (4.22)	0.18 (0.97)	0.35**
Cohesion	0.24 (0.39)	0.10 (0.28)	0.49**
Mutual Friends of Friends	0.06 (0.14)	0.02 (0.09)	0.43**
Number of friends	1.28 (1.11)	1.13 (0.51)	0.25**
Outgoing transaction percentage	0.52 (0.48)	0.52 (0.49)	0
Friends average number of friends	9.36 (10.62)	7.87 (10.38)	0.14*
Friends of friends average number of friends	10.06 (7.98)	8.19 (7.69)	0.24**
Giant component	0.87 (0.33)	0.80 (0.40)	0.18*

Very small: *
Small: **

Table 2.6. Cohen's effect size on group mean differences at lifetime 0. According to (Cohen, 2013), a value of 0.01(*) is considered very small, a value of 0.2(**) is regarded small, and a value of 0.5(***) or higher is medium.

2.5. Discussion

2.5.1. Summary of findings

Identifying a firm's active/inactive customers and evaluating their profitability (in terms of repeat transactions) at the individual level is a critical task in non-contractual settings. Our work focuses on harnessing the power of social networks to improve the predictive performance in all aforementioned tasks. We first evaluate the effect of social network attributes when predicting future customer activity. Our results indicate that social network attributes lead to a significant boost in performance early on in a customer's lifecycle: relative increase in AUC is 24%-33.9% and for top-decile Lift is 23.9%-37.8%. This enhancement in predictive power fades away in the subsequent months, and the models containing user-based variables produce the best results. Furthermore, we reaffirm findings from previous studies that show recency and transaction frequency to be the best predictors of future customer activity.

In the second part, we focus on modeling repeat transactions and identifying Venmo's best customers, as expressed by the total number of completed transactions. Given that typically a small number of customers are responsible for the largest proportion of revenue and profits, it is imperative to determine which customers have the potential to reach the best customer tier, and allocate disproportionate marketing resources towards them. We find that our approach can accurately predict the future number of transactions of all customers, and also identify the top 10% (20%) of customers. Once again, social network attributes lead to improved forecasts at lifetime 0, especially in determining the future best customers: relative increase in AUC is 22.0% (20.6%), and for top-decile Lift it is 45.8% (35.4%) for top 10% (20%) of customers. Furthermore, contrary to the previous part, models containing social network information either outperform or match the predictive power of user-centered ones beyond lifetime 0.

In the final section, we characterize the differences of the top 10% and bottom 90% of customers from a structural network perspective. We focus our attention on lifetime 0, when the median number of transactions for each cohort is identical. We find evidence of a strong, positive correlation between future customer tier and various individual-level measures of centrality and density. The best customers tend to join communities that exhibit high density and connectivity, and the most important metrics turn out to be cohesion and MFF.

2.5.2. Managerial Implications

In relation to the incorporation of social network attributes into our models, our thesis is that while they improve the overall quality of the predictions, their power is most evident in the beginning of a customer's lifetime. Social networks can help overcome the "cold-start" problem, where there is little to no information about a customer's past behavior.

It has not escaped our notice that as a result, firms can save valuable time in the model lifecycle process; they can start making reliable forecasts immediately upon acquiring a customer instead of wasting time collecting sufficient amounts of data before undertaking the task of predictive modeling. The solution we provide to the "cold-start" problem significantly improves one's ability to differentiate between one-time and best customers from day one. This observation has two main implications. As far as allocating marketing resources is concerned, firms can start investing into their customers immediately upon acquiring them, while at the same time being more confident that they are not wasting their resources on unprofitable individuals. Moreover, a company's main source of revenue, and therefore its overall valuation, typically relies on a relatively small number of highly engaged customers (Fader, 2012). Our findings can enhance the ability of performing CBCV for firms in two scenarios: when experiencing rapid growth of their user base, and when customer lifecycles are very short.

Finally, our investigation of the structural properties of Venmo's best customers reveals that the future most engaged users tend to join communities that exhibit high density and connectedness. This clear correlation pattern can potentially be leveraged to increase the effectiveness of customer acquisition efforts in at least two ways. First, a firm may identify customers who belong to these communities and incentivize them to invite their friends into the service. Second, if the firm has access to a user's network of friends who have not yet joined the service, it can determine whether it is worth targeting these friends directly.

2.6. Conclusions

In this chapter, we investigate customer behavior using a comprehensive dataset from Venmo, consisting of the entire transaction history of approximately one million users. Our work introduces a framework for incorporating social network information when predicting customer behavior, and we demonstrate its ability to lead to improved forecasts, especially during the beginning of a customer's lifecycle.

There are three limitations to this research. First, our focus is on predicting customer behavior, rather than providing a causal interpretation of the mechanisms behind various types of behavior. Second, due to the social nature of applications like Venmo, where customers can only use the service when their friends are using it as well, it is not clear whether these findings will generalize to other non-contractual settings with different formats (e.g., hybrid setting where customer can transact both directly with firm and with friends). Finally, we determine the top 10% of customers on the basis of the number of completed transactions, rather than the monetary value of the transactions. A minor issue that we should also acknowledge is that we only analyze publicly available Venmo transactions. While our analysis excludes private transactions, our results should not be affected by this as, according to the CEO of PayPal, 90% of transactions are shared publicly.[1]

Although our work is by no means exhaustive, it highlights several future research directions. It would be of great interest to examine the generalizability of our findings to other non-contractual contexts, either in hybrid settings, where customers can use the product or service either directly on their own or with their peers, or in traditional non-contractual settings, when external social network data (e.g., Facebook friends and interactions) is available. Finally, further research could shed light into the relationship between best customers and their propensity to join highly connected communities. It is unclear if these individuals become top customers as a result of their community's activity, or whether they serve as catalysts of high engagement.

[1] http://fortune.com/2017/11/17/dan-schulman-paypal-venmo/

CHAPTER 3

Finding Strong Ties in a Facebook Haystack: A Multilayer Social Network Approach

3.1. Introduction

As Aristotle famously noted almost twenty four hundred years ago, man is by nature a social animal. Human sociability includes building and maintaining close relationships with a few individuals (Mac Carron et al., 2016), as well as less intimate interactions with a larger group of people (Dunbar, 1993). This spectrum of tie strength is crucial for several decisions in our lives. Among others, strong ties have been shown to be responsible for adopting a new product or service, while weaker ties foster the diffusion of information (Granovetter, 1973; Centola and Macy, 2007; Hansen, 1999). Moreover, a recent study by (Nathan et al., 2018) shows the existence of a strong positive correlation between tie strength and customer engagement in social digital platforms. Therefore, it is clear that predicting tie strength has several important applications in the fields of marketing and targeted advertisement. This is especially true for marketeers in the digital age, as more and more people use the "social login" feature (e.g., Facebook, Twitter, Gmail, Google+, etc.) to sign up for a service or product. According to a survey report from Gigya [1], 88% of U.S. consumers have used social login, mainly because of the convenience it offers to the user of not having to fill in another registration form or remember another new username

[1] https://marketingland.com/gigya-survey-88-of-u-s-consumers-say-they-have-used-social-logins-134933

and password. Facebook (FB) is the major player in the social authentication space; it accounted for 62% of the overall market and an astounding 80% on mobile applications in 2015[2]. Social login allows companies to collect data not only on individual users, but also on their online social networks by granting access to their friends list. This enables marketeers to engage in a new form of advertising, namely social advertising, via social media or in-house, personalized targeting.

Social advertising uses cues, such as likes from your peers, to influence your decision to engage with a product or service. The social media ads industry accounted for $9.5 billion in revenue just in the first half of 2017, growing at an incredible rate of 37% from 2016[3]. This heavy investment by companies is greatly justified by their effectiveness, which several studies have documented (Tucker, 2016; Bakshy et al., 2012; Aral and Walker, 2014b). More important, this influence increases with tie strength. In other words, it is more likely that a user will open a digital ad, if one of their strong ties have liked the ad rather than a weak tie. However, measuring tie strength is a difficult and time-consuming task. Therefore, social network scientists and companies alike would love to have access to an easy to implement method through which they could simply inspect the underlying network structure and be able to determine the strong ties of a user in advance.

In this work, we employ the "social bow tie" framework introduced in (Mattie et al., 2017) and apply it to a unique dataset from Venmo, the most popular peer-to-peer (P2P) mobile payment application, to expand our knowledge on tie strength prediction. Our dataset is unique because it combines two different but overlapping social networks. On

[2] https://www.gigya.com/blog/the-landscape-of-customer-identity-facebook-slides-again/

[3] https://www.iab.com/wp-content/uploads/2017/12/IAB-Internet-Ad-Revenue-Report-Half-Year-2017-REPORT.pdf

the one hand, we have the Venmo social graph, which comprises of all friend relationships of users that signed up with FB. On the other hand, we have the Venmo transaction graph which reflects offline transactional activities among the users. By following the money trail, we are able to determine to whom a person is really closely connected to and we study the extent to which knowledge of a customer's egocentric social network can enhance the accuracy of forecasting whether two individuals: (1) will transact at least once, (2) if they do transact, whether this transaction will be reciprocated and (3) their total number of transactions.

We break down our investigation into two parts. In the first part, we study the above questions at the time when a new user has just signed up in a service through FB and has given access to his friend list. The only information we use to predict his strong ties is the structure of his online egocentric social network. In the second part, we investigate whether the addition of a user's transactional egocentric social network improves our forecasting accuracy. Our models produce high quality forecasts for the tasks of predicting the formation of a financial relationship and its reciprocity, yielding final Accuracy scores in the range of 43%-90% and Area Under the Precision-Recall Curve (AUPRC) values in the range of 85%-98%, depending on the exact problem formulation. For the task of predicting the total number of transactions between a pair of users, we get a Mean Square Error (MSE) in the range of 7.38-25.48 and an R^2 in the range of 0.24-0.58.

The main contributions of our work are the following. First, we apply the social bow tie framework on a multilayer social network and expand it to include several structural network metrics that have not been investigated previously. Next, we exploit the above multilayer nature to empirically document FB's predictive power in inferring a user's strong ties. We find the most informative predictors to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends. These findings provide correlational evidence on the applicability of Granovetter's and Bott's hypotheses on online social networks. Finally, we show how the information of the underlying transactional social network can be used in conjunction with the online social network information to improve the forecasting accuracy of tie strength prediction.

3.2. Related Literature

Granovetter was the first to make theoretical predictions regarding tie strength and the factors that influence it in his seminal paper (Granovetter, 1977). However, due to the absence of real world social network data, these predictions remained untested for a long period of time. Onnela et al. (2007) conducted the first comprehensive study by examining a who-talks-to-whom network generated by mobile phone users. They found that strong ties are associated with densely connected network neighbourhoods, while weak ties provide global connectivity at the network level; providing empirical evidence in support of Granovetter's hypothesis. Another more recent study with a who-talks-to whom network showed that the use of transactional information, as expressed by frequency of communication or phone call duration, can indeed be indicative of tie strength (Wiese et al., 2014).

Online social networks gave rise to another stream of literature, where researchers examined whether online social media increase the size of people's personal networks and whether they affect the structure of our social networks. In a blog-post by FB's data scientist team, Marlow et al. (2009) analyzed the friendship links and communication network of a random sample of users over the course of 30 days. Their main finding was that FB enables people to keep passive engagement with their friends. Strong ties maintain regular communication, while weak ties simply keep up with their network. In a related paper, Huberman et al. (2008) investigated what proportion of a user's followees are actually close ties in Twitter. They define a user's strong tie to be a person with whom the user has posted at least two messages on his Twitter page, and find that only a few followees are indeed close ties. More recently, Backstrom and Kleinberg (2014) used FB data to investigate a particular category of strong ties, those of romantic partners. They introduced a new structural network measure, dispersion, which measures the extent to which two people's mutual friends are not themselves well-connected, and showed it can achieve highly accurate results in predicting this particular type of strong tie.

In a slightly different vein, Gilbert (2012); Pappalardo et al. (2012) explored whether a tie strength model developed for one social medium adapts to another. Gilbert (2012) focused on mapping the respective relational features of FB to Twitter, and found that his FB tie strength model can generalize to Twitter. Pappalardo et al. (2012) collected friendship links across the same 7500 individuals, in three online social networks (Foursquare, Twitter and FB), and proposed a multidimensional definition of tie strength which captures the existence of multiple online social links between two individuals. They found that the more links two individuals share across the different social media, the stronger their tie.

Our work is also related to the link prediction literature (Benedek et al., 2014; Liben-Nowell and Kleinberg, 2007; Lü and Zhou, 2011). However, our setting is different in the sense that our focus is not on whether a tie between every pair of nodes will be formed or not, but rather on whether an existing online social tie will also be formed in an offline setting. More so, what will that tie's strength be? De Sá and Prudêncio (2011) used the weights of current links of a co-authorship network to predict the links that will appear in the future and found the predictive performance to be better than that without weights. Kahanda and Neville (2009) used the transactional information among users, defined as communication in FB messenger and file transfers, from the public Purdue FB network to predict tie strength. The authors investigate the relative predictive importance of attribute-based, topological, and transactional features and find the latter ones to be the most informative.

The closest papers to our work are Lewis et al. (2008); Gilbert and Karahalios (2009); Mattie et al. (2017). Lewis et al. (2008) collected FB data from college students when FB was still at its infancy, and tried to infer how many of their FB friends are actually close friends in real life. To proxy who is socially close, they used the photographs that users posted on their FB page. If one user tags another in a photograph, they are considered to be close ties. They found that users had on average 6.6 close friends out of their approximately 145 online connections. Gilbert and Karahalios (2009) recruited 35 students in the lab and asked them to rate the strength of their FB friendships, as well as answer a different set of questions yielding more than 70 variables. Their predictive model performed quite well in distinguishing between strong and weak ties. However, both of the aforementioned works have important limitations on their definition of close ties. On the one hand, the picture method may represent an overly strict criterion for defining a close tie as people can be close friends but not post pictures together. On the other hand, survey created metrics are usually biased due to cognitive constraints. Most important, both of these approaches require time and effort to be collected. Last, Mattie et al. (2017) introduced the social bow tie framework to study tie strength. They tested their framework in two distinct offline social networks: a collection of 75 rural villages in India and a who-talks-to-whom network generated by a European mobile service provider. They find that bow tie metrics are highly predictive of tie strength, and that the more the social circles of two individuals overlap, the stronger their tie. In this work, we enrich the social bow tie framework with several other structural network metrics and apply it to an online-offline setting.

3.3. Data and Methods

3.3.1. Data Description

We collected and analyzed data from Venmo, the most popular P2P mobile payment service. For the purposes of this work, we consider Venmo to be a multilayer network. The first layer consists of Venmo's social graph. When a new user signs up in Venmo, he is given the option to allow access to his friend list either through his FB account or through his mobile contact list. We used a two-step snowball sampling approach to collect this data (Goodman, 1961). During the first step, we collected the friend lists of 127,218 users that signed up through FB in the period of January 2014 to May 2015. In the second step, we collected the friend lists of the friends of the aforementioned users. This resulted in a social graph of 1,655,348 users. The second layer consists of Venmo's transactional graph. Venmo allows users to easily transfer money electronically with their friends for offline shared social activities. We collected the full transactional history of the aforementioned 127,218 users. This allows us to infer with whom of their online friends they have transacted with, and at what intensity. A Venmo user that signs up with FB has on average 94.5 friends that are also users of Venmo, and transacts on average with 11.2 users in a period of one year. Figure 3.1 shows the egocentric social network of a sampled Facebook user. The picture in the left shows the set of all declared friendships in this user's profile that are also using Venmo, while the picture on the right shows with whom the user had a financial transaction. An overview of our final dataset is shown in Table 3.1.

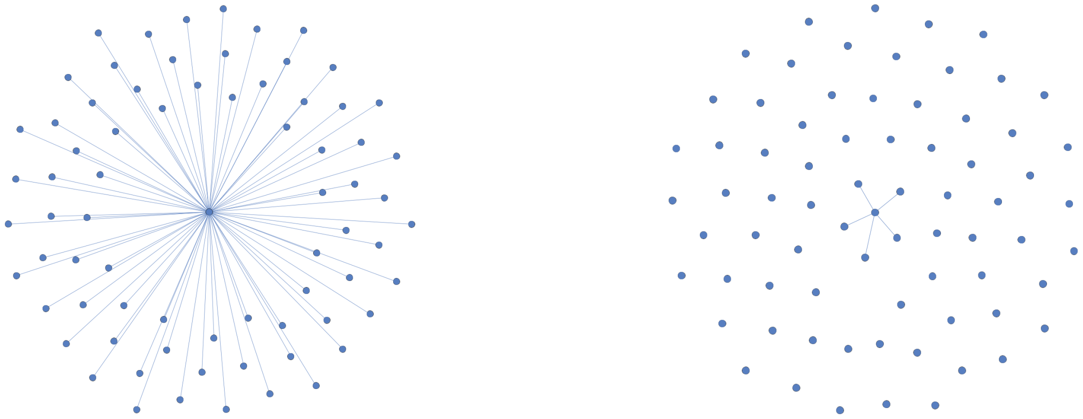


Figure 3.1. Egocentric social network of a sampled Venmo user that signed up through Facebook. The figure on the left depicts the set of all declared friendships in this user's profile that are also using Venmo, whereas the figure on the right depicts with whom the user had a financial transaction.

Dataset overview	
Time Frame (in months)	12
Total Number of Users that Signed Up with FB	127,218
Average Number of Friends of a User who Signed Up with FB	94.5 (87.2)
Average Number of Distinct people a user has transacted with	11.2 (9.2)

Table 3.1. Summary statistics. Standard deviation is shown in parentheses.

We should mention here some caveats of our data set. First, we have no access to geolocation. Since Venmo reflects offline social interactions, most of which take place in the same geographic venue, it is likely that access to geolocation would increase the predictive accuracy of our results, as previous research has documented (Kylasa et al., 2015). However, geolocation is a node attribute and not a structural one, and therefore, it is out of the scope of our investigation. Second, we might lose some of the social graph information, because some of a user's friends have not signed up through FB or have not given access to their contact list. Another possibility is that they have decided to make their profile private. This missing information can bias the computation of the structural network metrics of a user's egocentric network. However, this is not a limitation in the way we collected the data, rather it is the status-quo of what every service with social login feature is facing.

3.3.2. Methodology

We define the tie strength prediction problem as follows. Let G be a multilayer network graph, consisting of V nodes and M layers, each one representing a different type of relation. The structure of G can be fully described by the set of adjacency matrices

$$G \equiv \mathbf{A} = \{A^{[1]}, ..., A^{[M]}\},$$

where $A^{[k]} = \{a_{ij}^{[k]}\}$, with $a_{ij}^{[k]} = 1$ if there is a link between i and j, and 0 otherwise. When the links among nodes are weighted, G can be described by a set of weighted adjacency matrices

$$\mathbf{W} = \{W^{[1]}, ..., W^{[M]}\},$$

where $W^{[k]} = \{w_{ij}^{[k]}\}$ and $w_{ij}^{[k]}$ is the weight of the link between i and j.

In our case, we have $M = 2$ layers. The first layer represents the online friendships between pairs of users, and it is undirected and unweighted. The second layer represents the financial transactions between pairs of users, and it is directed and weighted. The link weights correspond to the total number of transactions two users have shared in one year's time.

Our first goal is to learn a predictive model that given the adjacency matrix $A^{[1]}$ will predict $A^{[2]}$ and $W^{[2]}$. Our second goal is to investigate whether snapshots of the transactional network at different time points, $A_t^{[2]}$ and $W_t^{[2]}$, can increase the predictive accuracy of $A^{[2]}$ and $W^{[2]}$, respectively. In words, we address a set of three different but interrelated research questions: (1) Will two individuals transact at least once? (2) Will that transaction going to be reciprocated? and (3) What is the total number of transactions these two individuals will share in one year's period. The first two are classification problems, while the third is a regression one.

For our second goal, we examine two snapshots. The first snapshot is taken at month 1, where 28.4% of the transactional relationships have been formed. The second is taken at month 5, where almost 50% of the transactional relationships have been formed. Figure B.1 in Appendix B.2 depicts the cumulative distribution function of forming a new financial relationship with a distinct user.

To answer all the above questions, we use the social bow tie framework introduced in Mattie et al. (2017). For each pair of individuals, say i and j, social bow tie framework partitions their network of friends into three disjoint sets, namely i's friends, j's friends, and the friends they share in common. The structure of social bow tie captures two hypotheses: 1) Granovetter's: the stronger the tie between any two individuals, the higher the fraction of friends they share in common (Granovetter, 1973), and 2) Bott's: the higher the degree of clustering in an individual's network the less likely to form a tie with somebody outside the group (Bott and Spillius, 2014). We enrich the bow tie framework by incorporating several other structural metrics that have not been studied before. Table B.1 in Appendix B.1 provides a detailed description of all the structural network variables.

3.3.3. Measuring Tie Strength

Granovetter suggested that tie strength between a pair of individuals should be measured along four dimensions: 1) amount of time spent together, 2) the level of intimacy, 3) emotional intensity and 4) reciprocity of interactions. Further research has expanded these dimensions to include social distance (Lin et al., 1981), emotional support (Wellman and Wortley, 1990), and network structural factors (Burt, 2009).

Measuring tie strength in practice, though, is a challenging task. Most studies have used proxies that focused on the frequency of interactions (Gilbert et al., 2008; Bond et al., 2012; Mattie et al., 2017). Other proxies include the social context of the relationship of two individuals (e.g. whether they attended the same college, share common hometown or institutional affiliations) and their common interests (e.g. FB pages likes) (Aral and Walker, 2014a). Most of these proxies, however, focused on data from online social networks. Last, some studies have used surveys to ask individuals about their relationships category and their perceived intimacy (Marsden and Campbell, 1984; Brown and Reingen, 1987; Frenzen and Davis, 1990; Gilbert and Karahalios, 2009). These survey proxies though cannot scale up to large systems and are subject to perception bias.

In this work, we use the financial transactions in Venmo to define tie strength. Our measurement is by construction very robust, as it reflects offline shared social experiences. In other words, a pair of individuals need to be in the same physical location and share the same experience to exchange money in Venmo. Therefore, looking at the transaction history between two individuals captures the amount of time they spent together and the level of their intimacy.

3.3.4. Predictive Framework

To evaluate our models, we split our data into two random, non-overlapping sets: the training set (89,053 users - 70%), and the testing set (38,165 users - 30%). As mentioned before, our dataset is highly imbalanced. To account for this imbalance, we try several different methods, namely Undersampling, Oversampling, SMOTE and ROSE. Furthermore, we perform 10-fold cross validation for all models to establish their predictive accuracy. For each classification problem, we report the Accuracy and AUPRC, which is a popular performance metric in the machine learning literature for imbalanced problems (Saito and Rehmsmeier, 2015). For our regression problem, we report the Mean Square Error (MSE).

3.4. Results

3.4.1. Forming a Transactional Relationship

We start off our investigation by examining the one time formation of transactional links. That is given a user that singed up through FB and has given access to his friend list, with whom of his online friends will transact at least once within one year in Venmo. We report our results in Figure 3.2.

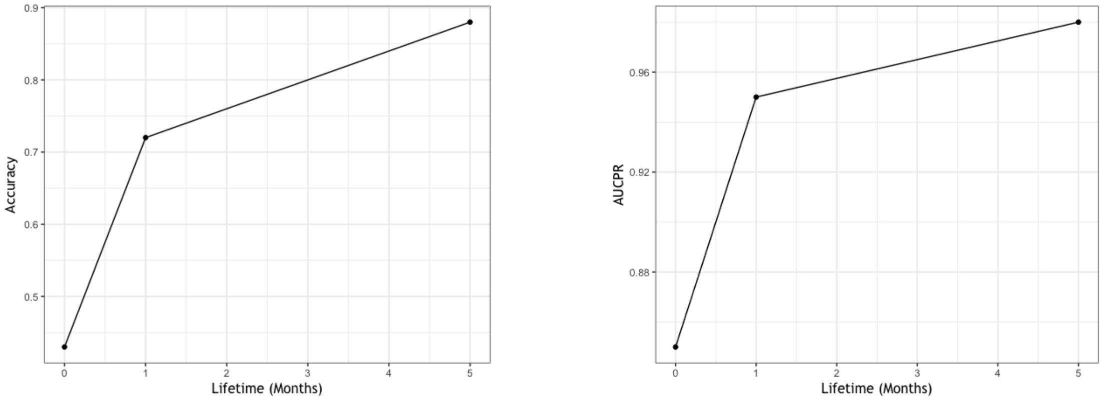


Figure 3.2. Classification results for predicting future formation of transactional relationships among pair of users. The figure on the left depicts the Accuracy score, whereas the figure on the right depicts the AUCPR score.

Note that all types of machine learning models (variations of logistic regression and random forests) and sampling methodologies achieve approximately the same predictive performance - see Table B.2 and Table B.3 in Appendix B.3. Therefore, we present the results obtained using an oversampling technique and a lasso logistic regression model, which performs a variable selection procedure and sheds light into the importance of the different features.

We find that even at lifetime 0, with the only available information being the online social network of a user, our machine learning models achieve an Accuracy of 43% and an AUCPR score of 85%. The most informative predictors are found to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends. Moreover, at lifetimes 1 and 5, when we take into account the information from the underlying transactional graph this predictive accuracy is increased significantly. More specific, at lifetime 1, Accuracy is increased to 72% and AUCPR to 95%, while at lifetime 5, Accuracy goes up to to 88% and AUCPR to 98%. These results indicate that knowing part of your transactional relationships can lead to a significant improvement in predicting the rest of your transactional relationships. Again, the most informative predictors are found to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends, but this time these metrics correspond to the transactional graph.

3.4.2. Reciprocity of Transactional Relationships

We continue our investigation by examining the act of reciprocating a transactional relationship. As explained earlier, reciprocity is an important dimension of tie strength. Dyadic relationships that are reciprocated are signals of trust and social capital. We report our results in Figure 3.3.

Surprisingly, our predictive performance is better than the one for forming a transactional relationship. At lifetime 0, our machine learning models achieve an Accuracy of 45% and an AUCPR score of 86%. Moreover, at lifetime 1, Accuracy is increased to 76% and AUCPR to 97%, while at lifetime 5, Accuracy goes up to 90% and AUCPR to 98%. At lifetime 0, the most informative predictors are found to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends. However, at lifetime 1 and 5, the most informative predictor is the overlap of friends between two individuals. One possible explanation for this observation, is that the more common friends a dyadic relationship has transacted with, the more likely that they will all go out as a group and financial trust among the group's people is established.

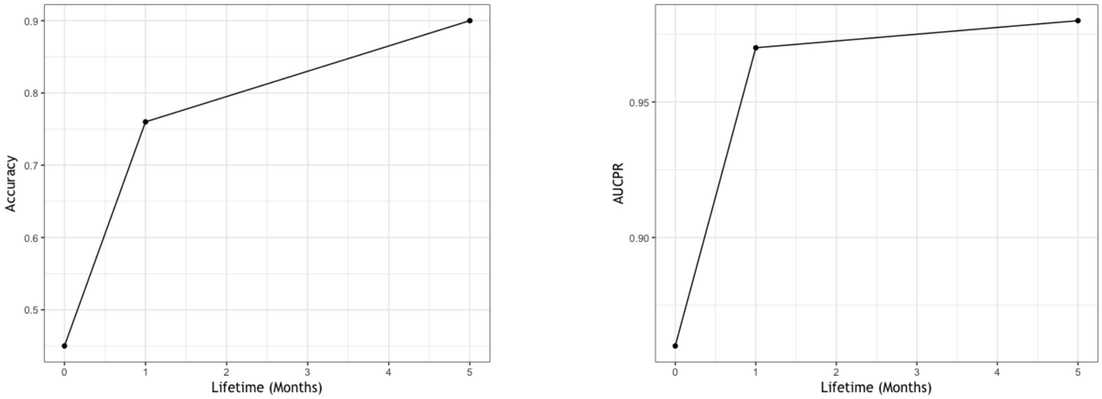


Figure 3.3. Classification results for predicting reciprocity of transactional relationships among pairs of users. The figure on the left depicts the Accuracy score, whereas the figure on the right depicts the AUCPR score.

3.4.3. Intensity of Financial Interactions

A final aspect we investigate is the intensity of financial interactions between two individuals, which is the main dimension of a strong tie in a P2P financial setting. We report our results in Figure 3.4.

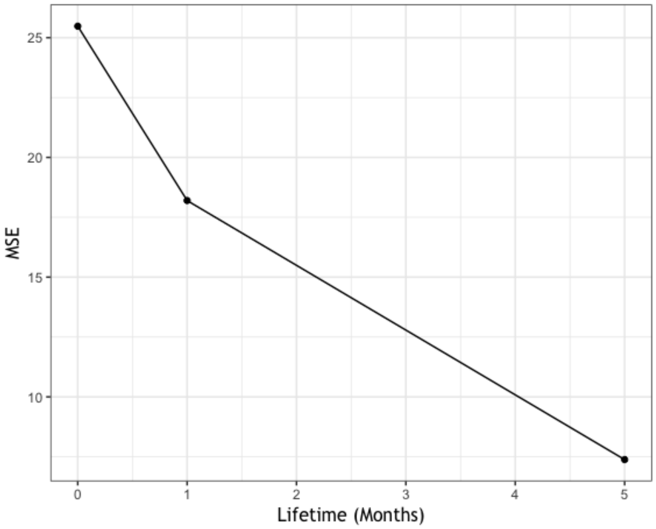


Figure 3.4. Results on predicting the expected number of financial interactions between two individuals at lifetime points 0, 1, and 5.

Our predictive results are quite impressive given the difficult nature of this question. At lifetime 0, we get an MSE of 25.48 and an R^2 of 0.24. At lifetime 1, the MSE and R^2 are improved to 18.20 and 0.33, respectively. Finally, at lifetime 5, we get an MSE of 7.38 and an R^2 of 0.58. Again, the most informative predictors in all lifetimes are found to be the overlap of friends between two individuals, and the clustering coefficient of their non-overlapping friends.

3.5. Discussion

Strong ties have been documented to play an influential role in people's decision making across various settings. From our decision to donate goods (Carman, 2003; Leider et al., 2009) to our decision to turn up and vote at the elections (Huckfeldt, 1984; Nickerson, 2008), strong ties are the ones who exert the greatest influence on us as they convey greater trust (Coleman, 1988).

The digital age has re-emphasized the importance and complexity of predicting an individual's strong ties, as more and more companies now have access to online friendship data of their customers. There has been a great deal of debate recently about the role of FB in influencing decision making in people's lives. According to a FB case study, Republican governor, Rick Scott, used FB ads campaigns to create a "22% increase in Hispanic support", "a deciding factor" in his re-election [4]. However, FB does not make its raw data public, and even if it did, answering questions such as, "Did FB's social digital ads swing the US elections?" is an extremely hard, if not impossible task. In this work, we exploit the multilayer nature of Venmo data to offer the first comprehensive answer regarding FB's power in predicting the strong ties of an individual. Our results provide evidence that such a task is feasible under the right methodology. We find that one is able to predict tie strength by just inspecting the underlying structure of a person's egocentric online network. Moreover, this predictive accuracy can be enhanced if one uses information from the underlying transactional network. In our case, this transactional network is expressed by financial transactions, but we conjecture that similar results will hold true in other transactional networks, such as call-logs and instant messaging.

We should point out that our study focused entirely on predictive analytics, and although we provide correlational evidence on the applicability of Granovetter's and Bott's hypotheses on online social networks, this by no means constitutes a causal explanation of why this is the case. As such, this study opens up new avenues for further research. First,

[4] https://www.facebook.com/business/success/rick-scott-for-florida

it would be of great interest to perform a matching strategy not only on the observed user characteristics, but also on the calculated network characteristics. For example, we could create two "artificial" control and treatment groups, where all network characteristics are similar but cohesion. This way we can treat cohesion as a treatment and check if we can provide causal evidence for Granovetter's hypothesis. Finally, further research in the form of field experiments is required to determine the effectiveness of our predictions in real-world settings. These types of experiments will enable governments and corporations to better understand the power of online social networks in propagating real-world behaviors and economic outcomes.

CHAPTER 4

Venmo for Change: The Effect of Digital Donations on Customer Engagement

Joint work with Marcel Fafchamps

4.1. Introduction

In today's competitive and connected environment, organizations are investing in corporate social responsibility (CSR) activities to differentiate themselves and create a meaningful engagement with their customers (Morrison and Crane, 2007; Devinney et al., 2012; Muller, 2006). CSR initiatives include but are not limited to environmental responsibility awareness campaigns and donations to philanthropic causes. With regards to the latter initiatives, organizations donate every year millions of dollars to causes that align with their core values and brand image. For example, Walmart donates food and fresh produce to the anti-hunger charity Feeding America, while Wells Fargo donated 25 million dollars in 2015 to the nonprofit NeighborWorks to support financial education and down payments on homes [1].

On top of the traditional donating methods, digital platforms have significantly decreased the barriers of making a donation by lowering the required effort and reducing transactional costs (Huck and Rasul, 2010). A Facebook (FB) user, for example, can now view the donations his/her favorite companies are making, increasing this way visibility

[1] http://fortune.com/2016/06/22/fortune-500-most-charitable-companies/

and brand image awareness. What is more, digital platforms have introduced a new form of a donating mechanism that uses social cues to inform the users about a fundraising event. In other words, a user can view that his/her connections have donated to a particular cause and potentially be influenced to engage with the cause, as well. This new form of donating is especially evident on FB, which allows its users to raise money for 501c3 nonprofits and gives them the option to create fundraisers for personal needs since 2016 [2].

Research has documented the benefits of CSR activities to organizations in terms of enhanced consumer perceptions of the company (Drumwright, 1996; Brown and Dacin, 1997; Sen and Bhattacharya, 2001), but there is little empirical evidence on the effect of digital platform donations on user engagement. For the purposes of this study, we define customer engagement on social digital platforms to be any interaction two existing users might have on the platform. Going back to the previous FB example, it might be the case that a user who donated to a cause observed one of his high school connections, who is not yet friends on FB with him, to have also donated to the same cause. Immediately after the donation, they become friends on FB and they start interacting though instant messaging, increasing this way their engagement with the FB platform. In this work, we propose a setting to empirically explore this question. Specifically, we use data from charitable fundraising events in Venmo to investigate whether two users that have contributed to the same charity event and have not previously transacted up to that point in time are more likely to transact after the charity event. Our charitable events are created by exogenous random shocks (e.g., physical catastrophes), which allow us to causally identify the effect of donations on customer engagement.

[2] https://techcrunch.com/2017/11/28/the-gates-foundation-is-matching-2m-in-donations-on-facebook-for-givingtuesday/

We seek to test the following hypotheses:

H1 Donating to a common cause increases the likelihood of forming a relationship between two users.

H2 This likelihood is a decreasing function of the shortest path distance between the two users.

4.2. Related Literature

Prior studies have documented that customers tend to empathize with organizations with which they share common traits (Bhattacharya et al., 1995; Ashforth, 1998; Elsbach, 1998; Sen and Bhattacharya, 2001). Organizational commitment to social issues enhances customer perceptions, which in turn increases customer loyalty to the organization (O'Brien et al., 2015; Brown and Dacin, 1997; Aguinis and Glavas, 2012; Becker-Olsen et al., 2006; Maignan and Ferrell, 2004; Arora and Henderson, 2007). As a result, organizations are now treating CSR as a strategic marketing tool to maximize their customer value (Knox and Maklan, 2004).

Particular importance is placed in using CSR activities on the development of customer engagement. This is greatly justified by the ever-increasing emergence of digital platforms which have introduced new mechanisms for organizations to facilitate: dialogue and community engagement (Guo and Saxton, 2014; Curtis et al., 2010; Auger, 2013), advocacy purposes (Bortree and Seltzer, 2009; Briones et al., 2011) and financial mobilization (Greenberg and MacAulay, 2009). Waters (2010); Kanter and Fine (2010); LaCasse et al. (2010) provide qualitative evidence that nonprofit organizations use social media to streamline their management functions and engage with current and potential stakeholders, clients and donors through the sharing of real-time information. In one of the few studies that have empirically examined donations in social networking sites, Saxton and Wang (2014) used data from FB causes to investigate the nature and determinants of charitable giving in those settings. They find that online donations are not driven by the same factors as in offline settings, and that a social network effect model takes over traditional economic explanations.

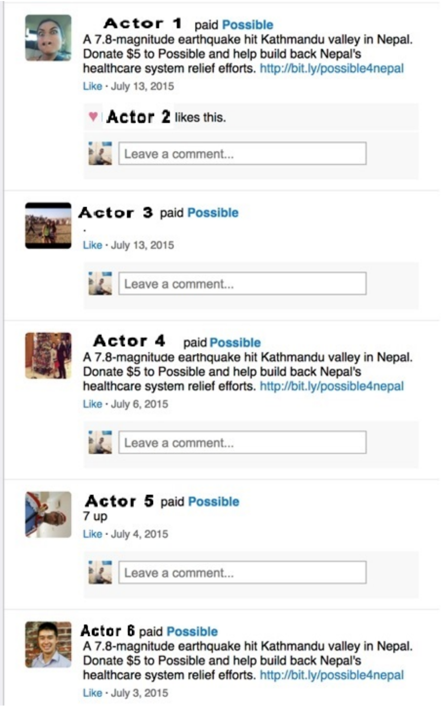
There is also some notable work that use field experiments to investigate the various aspects motivating donations. In offline settings, Karlan and List (2007) conducted a natural field experiment to test the effectiveness of matching funds strategy on charitable giving, and find it to increase both the revenue per solicitation and the response rate. In a similar vein, Huck and Rasul (2010) find that straight linear matching schemes raise the total donations received including the match value, but partially crowd out the actual donations given excluding the match. Meier (2007) adds to these observations that a matching strategy increases contributions to a public good. However, in the long run, the willingness to contribute may be undermined. In online settings, Castillo et al. (2014) investigated the costs and benefits of peer-to-peer fundraising through online social networks. They find that asking friends to donate generates new donations, and that charities should target incentives to those who have the smallest cost to fund raise online. Smith et al. (2015) examined the underlying peer effects in charitable giving and find that an increase of past donations increases future giving.

The closest papers to this work are (O'Brien et al., 2015; Lichtenstein et al., 2004). O'Brien et al. (2015) empirically studied the role of CSR activities on customer loyalty. They find that CSR efforts can indeed increase customer loyalty, and they are more effective when the charity causes are aligned to the core business values. Lichtenstein et al. (2004) find that CSR activities have both corporate benefits (e.g., more favorable corporate evaluations, increased purchase behavior, etc.) and social benefits in the form of consumer donations to corporate-supported non-profits. Moreover, CSR initiatives increase consumers' identification with the corporation.

4.3. Research Setting and Data

In this section, we present the research setting and data for testing our hypotheses.

Our research setting is Venmo, the most popular Peer-to-Peer (P2P) mobile payment application among millenials. Venmo allows users to easily transfer money to one another for shared social activities like paying for food or utility bills. Venmo users have the ability to view their connections' financial transactions on a FB-like news feed. Although Venmo was designed for P2P transfers among individuals, its ease of use and popularity makes it ideal for fundraising charity events. Therefore, many non-profit organizations (e.g., Planned Parenthood, the Red Cross) created a Venmo account to raise money for various causes. We focus our attention to charity events that are created by exogenous random shocks, e.g., physical catastrophes. For example, Possible Health is a non-governmental organization that delivers health care through public-private partnership agreements with the Government of Nepal. During the Nepal Earthquake in April 2015 (also knows as Gorkha earthquake), Possible Health collected donations from thousands of Venmo users (see figure 4.1 for illustration). We should note here that while a Venmo account corresponds to a single non-profit organization, the same account can collect donations for multiple charity events. We have collected data for a total of 28 such events.



Actor 1 paid Possible
A 7.8-magnitude earthquake hit Kathmandu valley in Nepal. Donate $5 to Possible and help build back Nepal's healthcare system relief efforts. http://bit.ly/possible4nepal
Like · July 13, 2015

♥ Actor 2 likes this.

Actor 3 paid Possible
.
Like · July 13, 2015

Actor 4 paid Possible
A 7.8-magnitude earthquake hit Kathmandu valley in Nepal. Donate $5 to Possible and help build back Nepal's healthcare system relief efforts. http://bit.ly/possible4nepal
Like · July 6, 2015

Actor 5 paid Possible
7 up
Like · July 4, 2015

Actor 6 paid Possible
A 7.8-magnitude earthquake hit Kathmandu valley in Nepal. Donate $5 to Possible and help build back Nepal's healthcare system relief efforts. http://bit.ly/possible4nepal
Like · July 3, 2015

Figure 4.1. Donations collected by Possible Health for the Gorkha earthquake in 2015.

4.4. Proposed Research Design

4.4.1. Unit of Analysis

The unit of analysis is a pair of users. We want to test whether user j is more likely to transact with user i in period $t+1$, if j and i had not transacted up to period t, and either j or i or both contributed to a fundraising event at time t.

4.4.2. Identification Strategy and Econometric Specification

We employ an event study methodology based on regression estimation for our identification. We provide here a brief overview of the general setup (for a concise overview of the traditional event study analysis see MacKinlay (1997); Binder (1985); Fama et al. (1969)). The first step is to define the event of interest and the event window over which the dependent variable of interest will be examined. In our case, the events of interest are fundraising events and we set the time window to be one month (a Venmo user transacts on average twice a month, so we believe that this is a reasonable time window. However, we will also test other values for robustness purposes). The next step is to define our dependent variable of interest. In our case, it is $Y_{i,j,t+w}$. It is equal to 1, when a user i has made a financial transaction with user j at time $t+w$; otherwise, it is equal to 0. We denote by t the time of the event and by w the event window that we allow the users to start transacting because of the event. The last step is to select the set of pairs to be included to the study. Although our social network is comprised of around 1 million nodes, we exploit Apache Spark and pooled regression to run it over the whole network. This is crucial for our analysis as we avoid biases caused by sampling techniques (we discuss later these potential biases). We can write our regression as:

$$
\begin{aligned}
Y_{i,j,t+w} &= \sum_{n=1}^{N} \beta_{\beta_n} 1\{w = -n\} + \sum_{n=1}^{N} \beta_{\alpha_n} 1\{w = +n\} \\
&+ \sum_{n=1}^{N} \beta_{\beta\beta_n} 1\{w = -n\} Event_{i,j} + \sum_{n=1}^{N} \beta_{\alpha\alpha_n} 1\{w = +n\} Event_{i,j} \\
&+ \sum_{n=1}^{4+} \beta_{1n} 1\{Distance_{i,j,t+w} = n\} + \sum_{n=1}^{4+} \beta_{11n} 1\{Distance_{i,j,t+w} = n\} Event_{i,j} \qquad (4.1) \\
&+ \beta_2 NooneContributed_{i,j,t} \\
&+ \beta_3 BothContributed_{i,j,t} + \beta_4 OneContributed_{i,j,t} \\
&+ \beta_5 FriendsBeforeEvent_{i,j,t^-} + \beta_{55} FriendsBeforeEvent_{i,j,t^-} Event_{i,j}
\end{aligned}
$$

This is an event regression, so we need N observations before and N observations after the event. The coefficients of the 2N time dummies (one for each period before and each period after the event) give the reduced-form effect of the event on the average Y. Our regressors of interest are:

1) $NooneContributed_{i,j,t}$: This dummy will be 1, if neither i nor j contributed to the event.

2) $BothContributed_{i,j,t}$: This dummy will be 1, if both i and j contributed to the event.

3) $OneContributed_{i,j,t}$: This dummy will be 1, if either i or j contributed to the event.

By definition of an event, all these variables will always be equal to 0 at times from $t - w$ up to $t - 1$. Moreover, we are only interested in their lagged effect on our dependent variable. Our explanatory variables are:

1) $Distance_{i,j,t+w}$: the shortest path distance between nodes i and j at time $t+w$. Since it is computationally expensive to exactly determine the shortest path in large graphs, we make an approximation and compute only the following distances: 1, 2, 3 and greater or equal to 4. Note here that we take the indicator function inside our regression equation. This in turn makes all of our variables binary, which allows us to speed up the computations considerably.

2) $FriendsBeforeEvent_{i,j,t^-}$: This dummy will be 1, if i and j have transacted at least once before the event.

The dependent variable might have a slight upwards trend, i.e. conditional on not having transacted before, two arbitrary individuals have a positive probability of transacting. This is different from standard event studies where the dependent variable can either go up or down, e.g., financial stocks. We, therefore, need to net out this trend from our estimation. To do this, we need a way of estimating what this "background" $Y_{i,j,t+w}$ would be without the event. There are two types of observations that can be used as controls: (1) individuals who do not contribute to an event; and (2) individuals without an event. To model this, we use the event dummy $Event_{i,j}$, where $Event_{i,j} = 1$, if the i, j observations correspond to an event, and 0 if the i, j observations correspond to a non-event, i.e., are a sequence of 2N+1 weeks without any event. The purpose of using these observations is to estimate "background" coefficients for all relevant regressors, so as to ensure that we estimate our coefficients of interest net of this background correlation.

4.5. Concluding Remarks

Digital platforms are reshaping the way organizations and individuals donate to philanthropic causes. In this work, we ask how this increased visibility of fundraising charity events affects customer engagement. To answer this question, we propose the use of an event study design to causally identify the effect of individual donations on customer engagement. We hypothesize that donating to a common cause will increase the likelihood of forming a relationship between two users. Furthermore, we conjecture that this likelihood is a decreasing function of the shortest path distance between the two users. Our findings will have significant managerial implications with regards to customer engagement marketing. For example, FB partnered with the Bill & Melinda Gates Foundation and matched $2M of donations to fundraisers held on FB by nonprofits. If our hypotheses are true, and donations do indeed increase customer engagement, then other companies might want to follow FB's paradigm.

Bibliography

Abello, James, Adam L Buchsbaum, Jeffery R Westbrook. 1998. A functional approach to external graph algorithms. *European Symposium on Algorithms*. Springer, 332–343.

Aguinis, Herman, Ante Glavas. 2012. What we know and don't know about corporate social responsibility: A review and research agenda. *Journal of management* **38**(4) 932–968.

Ahn, Jae-Hyeon, Sang-Pil Han, Yung-Seop Lee. 2006. Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry. *Telecomm. Policy* **30**(10-11) 552–568.

Aral, Sinan, Dylan Walker. 2014a. Tie strength, embeddedness, and social influence: A Large-Scale networked experiment. *Manage. Sci.* **60**(6) 1352–1370.

Aral, Sinan, Dylan Walker. 2014b. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science* **60**(6) 1352–1370.

Arora, Neeraj, Ty Henderson. 2007. Embedded premium promotion: Why it works and how to make it more effective. *Marketing Science* **26**(4) 514–531.

Ascarza, Eva, Peter Ebbes, Oded Netzer, Matthew Danielson. 2017. Beyond the target customer: Social effects of customer relationship management campaigns. *J. Mark. Res.* **54**(3) 347–363.

Ascarza, Eva, Scott A Neslin, Oded Netzer, Zachery Anderson, Peter S Fader, Sunil Gupta, Bruce G S Hardie, Aurélie Lemmens, Barak Libai, David Neal, Foster Provost, Rom Schrift. 2018. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions* **5**(1-2) 65–81.

Ashforth, Blake E. 1998. Becoming: How does the process of identification unfold. *Identity in organizations: Building theory through conversations* 213–222.

Auger, Giselle A. 2013. Fostering democracy through social media: Evaluating diametrically opposed nonprofit advocacy organizations' use of facebook, twitter, and youtube. *Public Relations Review* **39**(4) 369–376.

Babkin, Andrey, Ina Goldberg. 2017. Incorporating Time-Dependent covariates into BG-NBD model for churn prediction in Non-Contractual settings. *SSRN Electronic Journal* .

Backstrom, Lars, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna. 2012. Four degrees of separation. *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 33–42.

Backstrom, Lars, Jon Kleinberg. 2014. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 831–841.

Bakshy, Eytan, Dean Eckles, Rong Yan, Itamar Rosenn. 2012. Social influence in social advertising: evidence from field experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 146–161.

Barabási, Albert-László, Réka Albert. 1999. Emergence of scaling in random networks. *science* **286**(5439) 509–512.

Barabási, Albert-László, Réka Albert, Hawoong Jeong. 2000. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: statistical mechanics and its applications* **281**(1) 69–77.

Becker-Olsen, Karen L, B Andrew Cudmore, Ronald Paul Hill. 2006. The impact of perceived corporate social responsibility on consumer behavior. *Journal of business research* **59**(1) 46–53.

Bellaachia, Abdelghani, Deema Alathel. 2016. Improving the recommendation accuracy for cold start users in Trust-Based recommender systems. *International Journal of Computer and Communication Engineering* **5**(3) 206–214.

Benedek, Gábor, Ágnes Lublóy, Gyula Vastag. 2014. The importance of social embeddedness: Churn models at mobile providers. *Decision Sciences* **45**(1) 175–201.

Benoit, Dries F, Dirk Van den Poel. 2012. Improving customer retention in financial services using kinship network information. *Expert Syst. Appl.* **39**(13) 11435–11442.

Bhattacharya, Chitrabhan B, Hayagreeva Rao, Mary Ann Glynn. 1995. Understanding the bond of identification: An investigation of its correlates among art museum members. *The Journal of Marketing* 46–57.

Bijmolt, Tammo H A, Peter S H Leeflang, Frank Block, Maik Eisenbeiss, Bruce G S Hardie, Aurélie Lemmens, Peter Saffert. 2010. Analytics for customer engagement. *J. Serv. Res.* **13**(3) 341–356.

Binder, John J. 1985. On the use of the multivariate regression model in event studies. *Journal of Accounting Research* 370–383.

Blattberg, Robert. 2001. Customer equity: Building and managing relationships as valuable assets (:) .

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**(7415) 295.

Bortree, Denise Sevick, Trent Seltzer. 2009. Dialogic strategies and outcomes: An analysis of environmental advocacy groups' facebook profiles. *Public relations review* **35**(3) 317–319.

Bott, Elizabeth, Elizabeth Bott Spillius. 2014. *Family and social network: Roles, norms and external relationships in ordinary urban families*. Routledge.

Briones, Rowena L, Beth Kuch, Brooke Fisher Liu, Yan Jin. 2011. Keeping up with the digital age: How the american red cross uses social media to build relationships. *Public* *relations review* **37**(1) 37–43.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, S Raghavan, P Rajagopalan, R Stata, A Tomkins, J Wiener. 2000. Graph structure inthe web: Experiments and models. *9th World Wide Web Conference*.

Broido, Anna D, Aaron Clauset. 2018. Scale-free networks are rare. *arXiv preprint arXiv:1801.03400* .

Brown, Jacqueline Johnson, Peter H Reingen. 1987. Social ties and word-of-mouth referral behavior. *Journal of Consumer research* **14**(3) 350–362.

Brown, Tom J, Peter A Dacin. 1997. The company and the product: Corporate associations and consumer product responses. *The Journal of Marketing* 68–84.

Burt, Ronald S. 2009. *Structural holes: The social structure of competition*. Harvard university press.

Carman, Katherine Grace. 2003. Social influences and the private provision of public goods: Evidence from charitable contributions in the workplace. *Manuscript, Stanford University* 1–48.

Castillo, Marco, Ragan Petrie, Clarence Wardell. 2014. Fundraising through online social networks: A field experiment on peer-to-peer solicitation. *Journal of public economics* **114** 29–35.

Centola, Damon, Michael Macy. 2007. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**(3) 702–734.

Coe, Robert. 2002. It's the effect size, stupid: What effect size is and why it is important. *Annual Conference of the British Educational Research Association*.

Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Coleman, James S. 1988. Free riders and zealots: The role of social networks. *Sociological Theory* **6**(1) 52–57.

Coussement, Kristof, Koen W De Bock. 2013. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *J. Bus. Res.* **66**(9) 1629–1636.

Curtis, Lindley, Carrie Edwards, Kristen L Fraser, Sheryl Gudelsky, Jenny Holmquist, Kristin Thornton, Kaye D Sweetser. 2010. Adoption of social media for public relations by nonprofit organizations. *Public Relations Review* **36**(1) 90–92.

Dasgupta, Koustuv, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, Anupam Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*.

De Sá, Hially Rodrigues, Ricardo BC Prudêncio. 2011. Supervised link prediction in weighted networks. *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2281–2288.

Devinney, Timothy M, Pat Auger, Giana Eckhardt. 2012. Can the socially responsible consumer be mainstream? .

Dror, Gideon, Dan Pelleg, Oleg Rokhlenko, Idan Szpektor. 2012. Churn prediction in new users of yahoo! answers. *Proceedings of the 21st international conference companion on World Wide Web*.

Drumwright, Minette E. 1996. Company advertising with a social dimension: The role of noneconomic criteria. *The Journal of Marketing* 71–87.

Ductor, Lorenzo, Marcel Fafchamps, Sanjeev Goyal, Marco J van der Leij. 2014. Social networks and research output. *Rev. Econ. Stat.* **96**(5) 936–948.

Dunbar, Robin IM. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences* **16**(4) 681–694.

Easley, David, Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Ehrenberg, Andrew S C. 1959. The pattern of consumer purchases. *Applied Statistics* **8**(1) 26–41.

Elsbach, Kimberly D. 1998. The process of social identification: With what do we identify. *Identity in organizations: Building theory through conversations* **232** 237.

Fader, Peter S. 2012. *Customer Centricity: Focus on the Right Customers for Strategic Advantage*. Wharton Digital Press.

Fader, Peter S, Bruce G S Hardie. 2009. Probability models for Customer-Base analysis. *Journal of Interactive Marketing* **23**(1) 61–69.

Fader, Peter S, Bruce G S Hardie, Ka Lok Lee. 2005. "counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science* **24**(2) 275–284.

Fader, Peter S, Bruce G S Hardie, Jen Shang. 2010. Customer-Base analysis in a Discrete-Time noncontractual setting. *Marketing Science* **29**(6) 1086–1108.

Fafchamps, Marcel. 2015. Causal effects in social networks. *Revue économique* **66**(4) 657–686.

Fafchamps, Marcel, Marco J van der Leij, Sanjeev Goyal. 2010. Matching and network effects. *J. Eur. Econ. Assoc.* **8**(1) 203–231.

Faloutsos, Michalis, Petros Faloutsos, Christos Faloutsos. 1999. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, vol. 29. ACM, 251–262.

Fama, Eugene F, Lawrence Fisher, Michael C Jensen, Richard Roll. 1969. The adjustment of stock prices to new information. *International economic review* **10**(1) 1–21.

Freeman, Linton. 2004. The development of social network analysis. *A Study in the Sociology of Science* **1**.

Frenzen, Jonathan K, Harry L Davis. 1990. Purchasing behavior in embedded markets. *Journal of Consumer Research* **17**(1) 1–12.

Gilbert, Eric. 2012. Predicting tie strength in a new medium. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1047–1056.

Gilbert, Eric, Karrie Karahalios. 2009. Predicting tie strength with social media. *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 211–220.

Gilbert, Eric, Karrie Karahalios, Christian Sandvig. 2008. The network in the garden: an empirical analysis of social media in rural life. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

Goel, Sharad, Roby Muhamad, Duncan Watts. 2009. Social search in small-world experiments. *Proceedings of the 18th international conference on World wide web*. ACM, 701–710.

Gong, Neil Zhenqiang, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, Dawn Song. 2012. Evolution of social-attribute networks: measurements, modeling, and implications using google+. *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM, 131–144.

Goodman, Leo A. 1961. Snowball sampling. *Ann. Math. Stat.* **32**(1) 148–170.

Granovetter, Mark S. 1973. The strength of weak ties. *American Journal of Sociology* **78**(6) 1360–1380. doi:10.1086/225469. URL `https://doi.org/10.1086/225469`.

Granovetter, Mark S. 1977. The strength of weak ties. *Social networks*. Elsevier, 347–367.

Greenberg, Josh, Maggie MacAulay. 2009. Npo 2.0? exploring the web presence of environmental nonprofit organizations in canada. *Global Media Journal: Canadian Edition* **2**(1).

Guo, Chao, Gregory D Saxton. 2014. Tweeting social change: How social media are changing nonprofit advocacy. *Nonprofit and voluntary sector quarterly* **43**(1) 57–79.

Gupta, Sunil, Donald R Lehmann, Jennifer Ames Stuart. 2004. Valuing customers. *Journal of marketing research* **41**(1) 7–18.

Haenlein, Michael. 2013. Social interactions in customer churn decisions: The impact of relationship directionality. *International Journal of Research in Marketing* **30**(3) 236–248.

Hansen, Morten T. 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative science quarterly* **44**(1) 82–111.

Hill, Shawndra, Foster Provost, Chris Volinsky. 2006. Network-Based marketing: Identifying likely adopters via consumer networks. *Stat. Sci.* **21**(2) 256–276.

Huberman, Bernardo A, Lada A Adamic. 1999. Internet: growth dynamics of the world-wide web. *Nature* **401**(6749) 131.

Huberman, Bernardo A, Daniel M Romero, Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045* .

Huck, Steffen, Imran Rasul. 2010. Transactions costs in charitable giving: evidence from two field experiments. *The BE Journal of Economic Analysis & Policy* **10**(1).

Huckfeldt, R Robert. 1984. Political loyalties and social class ties: the mechanisms of contextual influence. *American Journal of Political Science* 399–417.

Jamali, Mohsen, Martin Ester. 2009. Using a trust network to improve top-n recommendation. *Proceedings of the third ACM conference on Recommender systems*.

Jamali, Mohsen, Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. *Proceedings of the fourth ACM conference on Recommender systems*.

Jerath, Kinshuk, Peter S Fader, Bruce G S Hardie. 2011. New perspectives on customer "death" using a generalization of the Pareto/NBD model. *Marketing Science* **30**(5) 866–880.

Kahanda, Indika, Jennifer Neville. 2009. Using transactional information to predict link strength in online social networks. *ICWSM* **9** 74–81.

Kanter, Beth, Allison Fine. 2010. *The networked nonprofit: Connecting with social media to drive change*. John Wiley & Sons.

Karlan, Dean, John A List. 2007. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review* **97**(5) 1774–1793.

Kleinberg, Jon M, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew S Tomkins. 1999. The web as a graph: Measurements, models, and methods. *International Computing and Combinatorics Conference*. Springer, 1–17.

Knox, Simon, Stan Maklan. 2004. Corporate social responsibility:: Moving beyond investment towards measuring outcomes. *European Management Journal* **22**(5) 508–516.

Kraft, Ben, Eric Mannes, Jordan Moldow. 2014. Security research of a social payment app.

Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins. 1999. Trawling the web for emerging cyber-communities. *Computer networks* **31**(11-16) 1481–1493.

Kylasa, S. B., G. Kollias, A. Grama. 2015. Social ties and checkin sites: Connections and latent structures in location based social networks. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 194–201. doi:10.1145/2808797.2809308.

LaCasse, Kaitlin, LS Quinn, Chris Bernard. 2010. Using social media to meet nonprofit goals: The results of a survey. *Portland, ME: Idealware* .

Larivière, Bart, Dirk Van den Poel. 2004. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Syst. Appl.* **27**(2) 277–285.

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

Leskovec, Jure, Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th international conference on World Wide Web*. ACM, 915–924.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, Nicholas Christakis. 2008. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks* **30**(4) 330–342.

Liben-Nowell, David, Jon Kleinberg. 2007. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58**(7) 1019–1031.

Lichtenstein, Donald R, Minette E Drumwright, Bridgette M Braig. 2004. The effect of corporate social responsibility on customer donations to corporate-supported nonprofits. *Journal of marketing* **68**(4) 16–32.

Lin, Nan, Walter M Ensel, John C Vaughn. 1981. Social resources and strength of ties: Structural factors in occupational status attainment. *American sociological review* 393–405.

Liu, Nathan N, Xiangrui Meng, Chao Liu, Qiang Yang. 2011. Wisdom of the better few. *Proceedings of the fifth ACM conference on Recommender systems*.

Loupos, P., A. Nathan. 2018. The Structure and Evolution of an Offline Peer-to-Peer Financial Network. *ArXiv e-prints* .

Lü, Linyuan, Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* **390**(6) 1150–1170.

Mac Carron, Pádraig, Kimmo Kaski, Robin Dunbar. 2016. Calling dunbar's numbers. *Social Networks* **47** 151–155.

MacKinlay, A Craig. 1997. Event studies in economics and finance. *Journal of economic literature* **35**(1) 13–39.

Maignan, Isabelle, OC Ferrell. 2004. Corporate social responsibility and marketing: An integrative framework. *Journal of the Academy of Marketing science* **32**(1) 3–19.

Malthouse, Edward C, Robert C Blattberg. 2005. Can we predict customer lifetime value? *Journal of Interactive Marketing* **19**(1) 2–16.

Malthouse, Edward C, Michael Haenlein, Bernd Skiera, Egbert Wege, Michael Zhang. 2013. Managing customer relationships in the social media era: Introducing the social CRM house. *Journal of Interactive Marketing* **27**(4) 270–280.

Marlow, Cameron, L Byron, T Lento, I Rosenn. 2009. Maintained relationships on facebook. *Retrieved February* **15**(8).

Marsden, Peter V, Karen E Campbell. 1984. Measuring tie strength. *Social forces* **63**(2) 482–501.

Mattie, Heather, Kenth Engø-Monsen, Rich Ling, Jukka-Pekka Onnela. 2017. The social bow tie. *CoRR* **abs/1710.04177**. URL `http://arxiv.org/abs/1710.04177`.

McCarthy, Daniel, Peter S Fader. 2017. Customer-Based corporate valuation for publicly traded Non-Contractual firms. *SSRN Electronic Journal* .

McCarthy, Daniel, Peter S Fader, Bruce G S Hardie. 2017. Valuing subscription-based businesses using publicly disclosed customer data. *J. Mark.* **81**(1) 17–35.

Meier, Stephan. 2007. Do subsidies increase charitable giving in the long run? matching donations in a field experiment. *Journal of the European Economic Association* **5**(6) 1203–1222.

Milgram, Stanley. 1967. The small-world problem. *Psychology Today* **1**(1).

Mislove, Alan, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 29–42.

Moldovan, Sarit, Eitan Muller, Yossi Richter, Elad Yom-Tov. 2017. Opinion leadership in small groups. *International Journal of Research in Marketing* **34**(2) 536–552.

Morrison, Donald G, David C Schmittlein. 1988. Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *J. Bus. Econ. Stat.*

6(2) 145–159.

Morrison, Sharon, Frederick G Crane. 2007. Building the service brand by creating and managing an emotional brand experience. *Journal of Brand Management* **14**(5) 410–421.

Mulhern, Francis J. 1999. Customer profitability analysis: Measurement, concentration, and research directions. *Journal of Interactive Marketing* **13**(1) 25–40.

Muller, Alan. 2006. Global versus local csr strategies. *European Management Journal* **24**(2-3) 189–198.

Nathan, Alexandros, Noshir Contractor, Pantelis Loupos, Moran Cerf. 2018. Not all adoptions are equal: Predicting structural virality in venmo.

Ngai, Eric WT, Li Xiu, Dorothy CK Chau. 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications* **36**(2) 2592–2602.

Nickerson, David W. 2008. Is voting contagious? evidence from two field experiments. *American Political Science Review* **102**(1) 49–57.

Nitzan, Irit, Barak Libai. 2011. Social effects on customer retention. *J. Mark.* **75**(6) 24–38.

O'Brien, Ingrid M, Wade Jarvis, Geoffrey N Soutar. 2015. Integrating social issues and customer engagement to drive loyalty in a service organisation. *Journal of Services Marketing* **29**(6/7) 547–559.

Onnela, J-P, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences* **104**(18) 7332–7336.

Pappalardo, Luca, Giulio Rossetti, Dino Pedreschi. 2012. How well do we know each other? detecting tie strength in multidimensional social networks. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 1040–1045.

Perro, Justina. 2016. Mobile apps: What's a good retention rate? `http://Info.Localytics.Com/Blog/Mobile-Apps-Whats-A-Good-RetentionRate`.

Redner, Sidney. 1998. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* **4**(2) 131–134.

Richter, Yossi, Elad Yom-Tov, Noam Slonim. 2010. Predicting customer churn in mobile networks through analysis of social groups. *Proceedings of the 2010 SIAM International Conference on Data Mining*. 732–741.

Rust, Roland T, James Kim, Yue Dong, Tom J Kim, Seoungwoo Lee. 2015. Drivers of customer equity. *Handbook of Research on Customer Equity in Marketing* 17–43.

Saito, Takaya, Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**(3) e0118432.

Saxton, Gregory D., Lili Wang. 2014. The social network effect: The determinants of giving through social media. *Nonprofit and Voluntary Sector Quarterly* **43**(5) 850–868. doi:10.1177/0899764013485159. URL `https://doi.org/10.1177/0899764013485159`.

Schmittlein, David C, Donald G Morrison, Richard Colombo. 1987. Counting your customers: Who-Are they and what will they do next? *Manage. Sci.* **33**(1) 1–24.

Sen, Sankar, Chitra Bhanu Bhattacharya. 2001. Does doing good always lead to doing better? consumer reactions to corporate social responsibility. *Journal of marketing Research* **38**(2) 225–243.

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

Solomonoff, Ray, Anatol Rapoport. 1951. Connectivity of random nets. *The bulletin of mathematical biophysics* **13**(2) 107–117.

Stovel, Katherine, Lynette Shaw. 2012. Brokerage. *Annu. Rev. Sociol.* **38**(1) 139–158.

Tamaddoni, Ali, Stanislav Stakhovych, Michael Ewing. 2015. Comparing churn prediction techniques and assessing their performance. *J. Serv. Res.* **19**(2) 123–141.

Travers, Jeffrey, Stanley Milgram. 1977. An experimental study of the small world problem. *Social Networks*. Elsevier, 179–197.

Tucker, Catherine. 2016. Social advertising: How advertising that explicitly promotes social influence can backfire .

Verbeke, Wouter, David Martens, Bart Baesens. 2014. Social network analysis for customer churn prediction. *Appl. Soft Comput.* **14** 431–446.

Waters, Richard D. 2010. The use of social media by nonprofit organizations: An examination from the diffusion of innovations perspective. *Social computing: Concepts, methodologies, tools, and applications*. IGI Global, 1420–1432.

Watts, Duncan J, Steven H Strogatz. 1998. Collective dynamics of'small-world'networks. *nature* **393**(6684) 440.

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

Wiese, Jason, Jun-Ki Min, Jason I Hong, John Zimmerman. 2014. Assessing call and sms logs as an indication of tie strength .

Wübben, Markus, Florian v Wangenheim. 2008. Instant customer base analysis: Managerial heuristics often "get it right". *J. Mark.* **72**(3) 82–93.

Yoon, Sangho, Jim Koehler, Adam Ghobarah. 2010. Prediction of advertiser churn for google AdWords. *JSM Proceedings, American Statistical Association.*

Yu, Xiaobing, Shunsheng Guo, Jun Guo, Xiaorong Huang. 2011. An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Syst. Appl.* **38**(3) 1425–1430.

Zaharia, Matei, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, Ion Stoica. 2010. Spark: cluster computing with working sets. *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. 10–10.

APPENDIX A

Chapter 2 Supporting Material

A.1. Variable Description

Variable Type	**Dynamic**	**Variable Name**	**Description**
User	**No**	facebookSignIn	Binary variable indicating is user signed up through Facebook.
		earlyAdopter	Binary variable indicating early adoption of Venmo.
		month	Month that a user created Venmo account.
		year	Year that a user created Venmo account.
	Yes	daysFromLastTransaction	Days since last transaction. This feature is used to create the labels/dependent variable.
		nightAverage	Percent of transactions taking place at night (6:00pm - 4:00am)
		weekendAverage	Percent of transactions taking place on weekends.
		transactionFrequency	Number of transactions divided by the number of days in specified time interval.
Social Network	**Yes**	chargeCount	Number of "charge" transactions.
		payCount	Number of "pay" transactions.
		numberOfFriends	Number of distinct people a user has transacted with.
		isEmoji	Percent of transactions that contain at least one emoji in their description.
		numberOfPeopleBrought	Number of users who transacted with this customer for the first time.
		outgoingTransactionPctg	Percentage of outgoing transactions, i.e. user of interest was the initiator of transaction.
		pageRank	User's pagerank in the network.
		isGiant	Binary variable indicating membership to the giant component.
		triangleCount	Number of triangles in user's local network.
		cohesion	Measure of the degree to which a user's friends know each other.
		mutualFriendsOfFriends	Measure of the degree to which a user's friends have mutual friends.
		friendAvgTransactionFreq	Average transaction frequency of a user's friends.
		friendSDTransactionFreq	Standard deviation of transaction frequency of a user's friends.
		friendAvgTransFreqLagged	One month lagged average transaction frequency of a user's friends.
		friendAvgIsEmoji	Average percent of a user's friends transactions containing at least one emoji in description.
		friendSDIsEmoji	Standard deviation of the percent of a user's friends transactions containing at least one emoji in description.
		friendAvgNumFriends	Average number of individuals who have transacted with a user's friends.
		friendSDNumFriends	Standard deviation of the number of individuals who have transacted with a user's friends.
		friendOfFriendAvgTransFreq	Average transaction frequency of user's friends of friends.
		friendOfFriendSDTransFreq	Standard deviation of transaction frequency of a user's friends of friends.
		friendOfFriendAvgTransFreqLagged	One month lagged average transaction frequency of a user's friends of friends.
		friendOfFriendAvgIsEmoji	Average percent of a user's friends of friends transactions containing at least one emoji in description.
		friendOfFriendSDIsEmoji	Standard deviation of the percent of a user's friends of friends transactions containing at least one emoji in description.
		friendOfFriendAvgNumFriends	Average number of individuals who have transacted with a user's friends of friends.
		friendOfFriendSDNumFriends	Standard deviation of the number of individuals who have transacted with a user's friends of friends.

Table A.1. Variable Description

A.2. Robustness Checks

A.2.1. Comparison of alternative functional forms for predicting customer activity

Lifetime	Model 1			Model 2			Model 3		
	Logistic	**Lasso**	**RF**	**Logistic**	**Lasso**	**RF**	**Logistic**	**Lasso**	**RF**
0	0.53	0.53	0.51	0.71	0.71	0.72	0.71	0.71	0.72
1	0.73	0.73	0.72	0.72	0.72	0.72	0.73	0.73	0.72
2	0.77	0.77	0.76	0.76	0.76	0.73	0.77	0.77	0.76
3	0.79	0.79	0.78	0.77	0.77	0.75	0.79	0.79	0.78
4	0.80	0.80	0.79	0.77	0.77	0.75	0.80	0.80	0.79
5	0.81	0.81	0.80	0.79	0.79	0.76	0.81	0.81	0.80
6	0.82	0.82	0.80	0.79	0.79	0.76	0.82	0.82	0.80
7	0.82	0.82	0.80	0.80	0.80	0.79	0.82	0.82	0.80
8	0.82	0.83	0.80	0.80	0.80	0.79	0.82	0.83	0.80
9	0.82	0.83	0.80	0.80	0.80	0.79	0.83	0.83	0.80
10	0.83	0.84	0.80	0.80	0.80	0.79	0.83	0.84	0.80
11	0.83	0.84	0.81	0.81	0.81	0.79	0.83	0.84	0.81

Table A.2. AUC Score for alternative functional forms. Note that since our analysis time frame ends after one year, it is not possible to make forecasts when the current lifetime point plus the predictive window extends beyond 12 months. (RF stands for Random Forests).

Lifetime	Model 1			Model 2			Model 3		
	Logistic	Lasso	RF	Logistic	Lasso	RF	Logistic	Lasso	RF
0	1.12	1.12	1.10	1.38	1.38	1.54	1.40	1.41	1.54
1	1.57	1.56	1.55	1.55	1.55	1.55	1.59	1.59	1.57
2	1.79	1.79	1.79	1.77	1.78	1.80	1.80	1.80	1.80
3	1.94	1.94	1.92	1.91	1.92	1.93	1.94	1.94	1.93
4	2.10	2.11	2.09	2.05	2.06	2.09	2.10	2.11	2.11
5	2.32	2.32	2.29	2.23	2.24	2.28	2.32	2.32	2.31
6	2.49	2.49	2.47	2.37	2.37	2.40	2.49	2.49	2.47
7	2.61	2.61	2.59	2.49	2.49	2.50	2.61	2.61	2.59
8	2.73	2.73	2.71	2.59	2.59	2.59	2.74	2.73	2.72
9	2.89	2.88	2.86	2.68	2.68	2.70	2.89	2.88	2.86
10	3.03	3.03	3.00	2.79	2.79	2.83	3.03	3.03	3.00
11	3.22	3.22	3.19	2.93	2.93	2.96	3.22	3.22	3.19

Table A.3. Top 10% decile Lift for alternative functional forms. Note that since our analysis time frame ends after one year, it is not possible to make forecasts when the current lifetime point plus the predictive window extends beyond 12 months. (RF stands for Random Forests).

A.2.2. Alternative predictive windows for customer activity

We experiment with predictive windows of size 60, 90 and 120 days. We report the AUC and top decile Lift below. Our results remain unchanged.

Lifetime	60 days			90 days			120 days		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
0	0.53	0.68	0.68	0.53	0.67	0.68	0.54	0.67	0.67
1	0.73	0.72	0.73	0.73	0.72	0.73	0.73	0.73	0.74
2	0.77	0.76	0.78	0.78	0.76	0.78	0.78	0.77	0.79
3	0.79	0.77	0.79	0.80	0.78	0.80	0.81	0.79	0.81
4	0.81	0.79	0.81	0.82	0.80	0.82	0.83	0.81	0.83
5	0.82	0.80	0.83	0.83	0.81	0.84	0.84	0.82	0.84
6	0.84	0.81	0.84	0.85	0.82	0.85	0.85	0.83	0.86
7	0.84	0.82	0.84	0.85	0.83	0.86	0.86	0.83	0.86
8	0.85	0.82	0.85	0.86	0.83	0.86	0.87	0.84	0.87
9	0.85	0.83	0.85	0.87	0.84	0.87	-	-	-
10	0.86	0.83	0.86	-	-	-	-	-	-

Table A.4. AUC Score for alternative predictive windows of length 60, 90 and 120 days. Note that since our analysis time frame ends after one year, it is not possible to make forecasts when the current lifetime point plus the predictive window extends beyond 12 months.

Lifetime	60 days			90 days			120 days		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
0	1.15	1.45	1.50	1.17	1.54	1.60	1.19	1.62	1.66
1	1.76	1.78	1.83	1.90	1.94	2.00	2.04	2.12	2.14
2	2.11	2.13	2.15	2.35	2.40	2.43	2.54	2.63	2.70
3	2.40	2.38	2.42	2.76	2.73	2.81	3.03	3.00	3.12
4	2.71	2.63	2.72	3.16	3.04	3.18	3.48	3.36	3.51
5	3.04	2.89	3.04	3.51	3.37	3.53	3.89	3.73	3.91
6	3.27	3.11	3.28	3.83	3.63	3.84	4.27	4.03	4.28
7	3.48	3.27	3.48	4.13	3.85	4.13	4.64	4.27	4.64
8	3.72	3.44	3.73	4.41	4.04	4.42	4.93	4.47	4.94
9	3.93	3.58	3.93	4.67	4.19	4.67	-	-	-
10	4.19	3.77	4.19	-	-	-	-	-	-

Table A.5. Top 10% decile Lift for alternative predictive windows of length 60, 90 and 120 days. Note that since our analysis time frame ends after one year, it is not possible to make forecasts when the current lifetime point plus the predictive window extends beyond 12 months.

A.2.3. Predicting the future top 20% of customers

In addition to predicting the top 10% of customers, we also test our models in the task of predicting the future 20% of customers. We report the AUC and top decile Lift below. Once again, our results remain unchanged.

Lifetime	AUC			Lift		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
0	0.58	0.70	0.70	1.61	2.18	2.18
1	0.76	0.79	0.79	3.11	3.19	3.19
2	0.82	0.83	0.83	3.59	3.66	3.66
3	0.85	0.87	0.87	3.93	3.99	3.99
4	0.88	0.90	0.90	4.22	4.26	4.26
5	0.91	0.92	0.92	4.46	4.50	4.50
6	0.94	0.94	0.94	4.69	4.70	4.70
7	0.96	0.96	0.96	4.84	4.85	4.85
8	0.97	0.97	0.97	4.94	4.94	4.94
9	0.98	0.98	0.98	4.97	4.97	4.97
10	0.99	0.99	0.99	4.97	4.97	4.97
11	0.99	0.99	0.99	4.97	4.97	4.97

Table A.6. Classification results for top 20% of customers by lifetime. The left three columns show the AUC score for the three competing models, whereas the rightmost columns illustrate top decile lift for the same three models.

A.3. Comparison of logistic and linear regression in predicting best customers

We provide a comparison between a logistic regression model and a linear regression when predicting the future top 10% customers of a firm. A customer will belong to the top 10% if he completes at least 90 transactions by the end of their first year with Venmo. Given that the dependent variable for each model is of different nature (dichotomous versus continuous), we select an evaluation metric that can be computed in both cases, namely the F1-score. The F1-score is the harmonic mean of precision and recall, and it is widely used in the machine learning literature. Our results indicate that the logistic regression model performs best, and especially from lifetime 0 up until lifetime 5.

Lifetime	**Logistic Regression F1-Score**			**Linear Regression F1-Score**		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
0	0.20	0.29	0.29	0.01	0.03	0.03
1	0.41	0.42	0.42	0.25	0.29	0.28
2	0.48	0.49	0.49	0.39	0.42	0.42
3	0.54	0.55	0.55	0.49	0.50	0.50
4	0.60	0.60	0.60	0.56	0.57	0.57
5	0.65	0.66	0.66	0.63	0.64	0.64
6	0.70	0.71	0.71	0.69	0.70	0.69
7	0.75	0.75	0.75	0.74	0.75	0.75
8	0.79	0.80	0.80	0.79	0.79	0.79
9	0.84	0.84	0.84	0.83	0.84	0.84
10	0.88	0.89	0.89	0.88	0.88	0.88
11	0.94	0.94	0.94	0.93	0.93	0.93

Table A.7. F1-score when predicting the future top 10% of customers. On the left we provide the results of the logistic regression model, and on the right the results of the linear regression model.

A.4. Cohesion, triangle count and mutual friends of friends

We begin with a more rigorous definition of cohesion, and explain in more detail its relationship to triangle counts. Let $G_t(V_t, E_t)$ be a dynamic social network, where V_t is the set of nodes, and E_t is the set of edges at time t. For any actor $u \in V_t$ at time t, let $F_{u,t}$ denote the set of all his connections (friends), i.e. all v for which $(u, v) \in E_t$. Further, let $FF_{u,t}$ denote the set of all the connections of actor's u friends at time t, i.e. $FF_{u,t} = \bigcup_{v \in F_{u,t}} F_{v,t} \backslash \{u\}$. It is important to exclude actor u from all $FF_{v,t}$, where $v \in F_{u,t}$, since u will always be in that set by default.

Cohesion The cohesion metric of user u at time t is given by

$$cohesion_{u,t} = \frac{|FF_{u,t} \cap F_{u,t}|}{|F_{u,t}|}.$$

Cohesion can take values between 0 and 1, with 1 implying that user u and his friends form a rather dense graph; however, this does not mean that the graph must be complete. A graph is complete if there is an edge connecting any pair of vertices. Consider the network in Figure A.1c. This graph would be complete with the addition of edges (E, C) and (B, D), and yet, cohesion is equal to 1. Cohesion is closely related to triangle count, but it is normalized by a user's number of friends. In the network on the left, the cohesion of user A is 0, since none of her friends is connected to each other. In Figure A.1c, user A has cohesion equal to 1 and triangle count equal to 4, but this is an absolute count and does not adjust for the node's degree.

Another metric we elaborate on is the Mutual Friends of Friends (MFF). One limitation of the cohesion metric is that it does not capture the extent to which a user's friends have mutual connections, even if they are not connected to each other. To account for such ties, we need to look beyond the first degree connections. Given our network notation, MFF is defined as follows:

Mutual Friends of Friends (MFF) The MFF metric of user u at time t is given by

$$mutualFriendsOfFriends_{u,t} = \frac{\left| \bigcup_{\substack{v,w \in F_{u,t} \\ v \neq w}} (F_{v,t} \cap F_{w,t}) \backslash \{u\} \right|}{|FF_{u,t}|}.$$

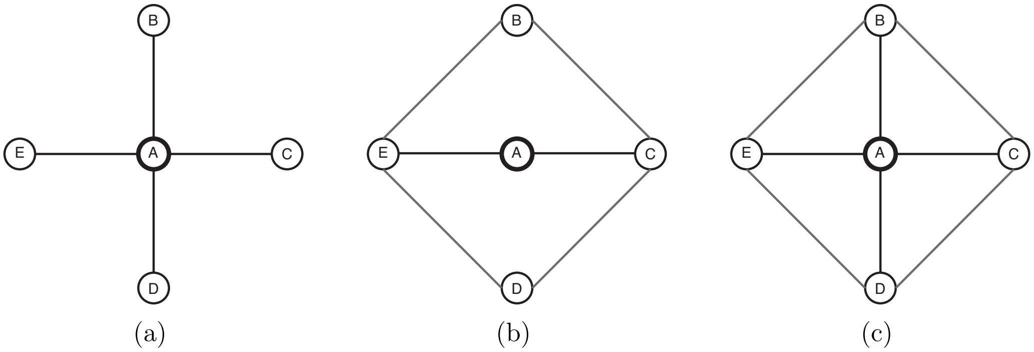


Figure A.1. Similarities and differences between cohesion, triangle count and MFF. (a) exhibits a star formation, and all three metrics for actor A are equal to 0. (b) actor's A cohesion and triangle count is equal to 0, but MFF metric is equal to 1: although C and E do not directly know each other, they share 2 mutual friends. (c) actor's A cohesion is equal to 1, triangle count is equal to 4, and MFF remains 1.

A.5. Cohen‘s effect size

We provide here the formula for calculating Cohen's effect size. Given two groups of size n_1 and n_2, let M_1, M_2 and SD_1, SD_2 denote their respective means and standard deviations. Further, let the pooled standard deviation of the two groups be denoted by:

$$SD_p = \sqrt{\frac{(n_{1-1})SD_1^2 + (n_2 - 1)SD_2^2}{n_1 - n_2 - 2}}$$

Then, Cohen's effect size is given by:

$$C = \frac{M_1 - M_2}{SD_p}$$

APPENDIX B

Chapter 3 Supporting Material

B.1. Variable Description

Variable Name	Graph	Description
Number of Online Friends of Actor i	S	Number of distinct friends user i had when he signed up through FB.
Number of Online Friends of Actor j	S	Number of distinct friends user j had when he signed up through FB.
Number of Common Friends of i and j	S&T	Number of distinct friends users i and j have in common.
Unweighted Edge Overlap between i and j	S&T	The proportion of friends shared between i and j.
Common Clustering Coefficient	S&T	Measure of the degree to which the common friends of i and j know each other.
Non-common Clustering Coefficient	S&T	Measure of the degree to which the non-common friends of i and j know each other.
Common Friends' Average Number of Friends.	S&T	Average number of friends which the common friends of i and j have.
Non-common Friends' Average Number of Friends.	S&T	Average number of friends which the non-common friends of i and j have.
Number of Friends of Actor i	T	Number of distinct friends user i has transacted with.
Number of Friends of Actor j	T	Number of distinct friends user j has transacted with.
Online-Offline Edge Overlap between of i and j	T	The proportion of online common friends users i and j has transacted with so far.
IFB	S	Dummy variable indicating whether both i and j signed up through FB.

Table B.1. Variable Description. "S" stands for Social and "T" for Transactional.

B.2. Number of distinct friends a user has transacted with over his lifetime.

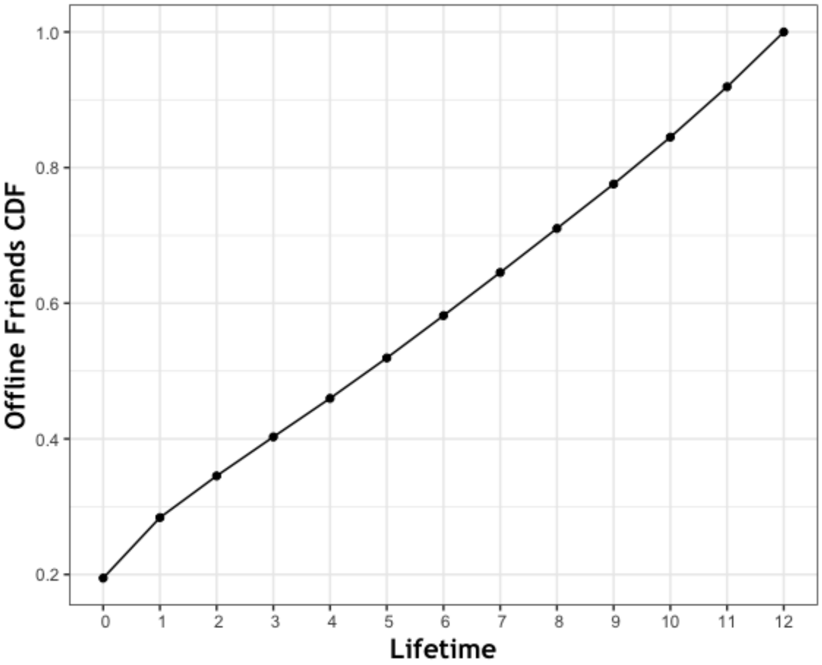


Figure B.1. Cumulative distribution function of making new financial distinct friendships. Lifetime is measured in months.

B.3. Robustness Checks

B.3.1. Comparison of alternative machine learning models and re-sampling techniques

Forming a Transactional Relationship

Lifetime = 0	**Machine Learning Model**		
Re-sampling Technique	**RF**	**Logistic**	**Lasso**
Undersampling	0.43	0.43	0.43
Oversampling	0.42	0.43	0.43
SMOTE	0.41	0.41	0.42
ROSE	0.41	0.41	0.41

Table B.2. Accuracy Score for alternative machine learning models and re-sampling techniques.

Forming a Transactional Relationship

Lifetime = 0	**Machine Learning Model**		
Re-sampling Technique	**RF**	**Logistic**	**Lasso**
Undersampling	0.85	0.85	0.85
Oversampling	0.84	0.84	0.84
SMOTE	0.84	0.84	0.84
ROSE	0.83	0.83	0.84

Table B.3. AUCPR Score for alternative machine learning models and re-sampling techniques.

Reciprocity

Lifetime = 0	**Machine Learning Model**		
Re-sampling Technique	**RF**	**Logistic**	**Lasso**
Undersampling	0.45	0.44	0.45
Oversampling	0.44	0.44	0.45
SMOTE	0.45	0.44	0.44
ROSE	0.44	0.44	0.45

Table B.4. Accuracy Score for alternative machine learning models and re-sampling techniques.

Reciprocity

Lifetime = 0	**Machine Learning Model**		
Re-sampling Technique	**RF**	**Logistic**	**Lasso**
Undersampling	0.86	0.86	0.87
Oversampling	0.85	0.84	0.86
SMOTE	0.85	0.84	0.86
ROSE	0.84	0.84	0.86

Table B.5. AUCPR Score for alternative machine learning models and re-sampling techniques.