ABSTRACT

Title of Dissertation: BEYOND SOCIAL STRUCTURE: PREDICTING INFLUENCE IN ONLINE REVIEW COMMUNITIES

Vanecia L. Fluelling, Doctor of Philosophy
December 2019

Dissertation Chair: Sanjay Bapna, Ph.D.
Department of Information Science and Systems

The intelligence derived from this dissertation provides businesses and online review communities (ORCs) with ways to exploit opportunities to predict influence for these communities. To my knowledge, there is little research in the prediction of influence. Moreover, few studies of influence use ORCs as a basis of investigation. This dissertation seeks to fill the gap by investigating the prediction of influence in an ORC and generating predictive models for this type of online social network (OSN), specifically Yelp.

Emphasizing the differences between an ORC and a traditional OSN, past research and studies, the susceptible infective removed model, threshold model, cascade model, expectation's theory, social influence and status characteristics theory, self-categorization theory and concepts of homophily and reciprocity, are used to ground a discussion of modifications to traditional measurements of influence for ORCs. Then a

comparison is conducted into different measurements. The results are used to create predictive models using supervised machine learning algorithms.

Two separate investigations are conducted to study the prediction of influence. First, an investigation is conducted into determinants of predicting influence, based on the differences associated with ORCs. Influence for the Yelp dataset is operationalized as a change in friends and a change in votes from one time period to another. Second, an investigation is conducted to generate models that predict influence. In comparing the determinants, three different levels are examined: the member, the member's network, and the member along with the member's network. Based on the results from the first investigation, the member level as well as the member along with the member's network level are examined in the second investigation. In both investigations, members of the Yelp community are separated into four groups, based on their time since joining the site, to see if there are differences between the groups. The receiver operating characteristic (ROC) graph with its accompanying area under the curve (AUC) data is chosen for analysis in both investigations. This metric is used to assess the performance of classifiers.

In investigating the best determinant of future influence, I compare the AUC values between the proposed determinants (friends count, votes count, and review count) and the dependent variable future influence. The results indicate that the review count is the best predictor of influence, whether using a change in friends or a change in votes as the dependent variable. In investigating the levels of the determinants, I compare the AUC data between the levels of the determinants and the dependent variable future influence. For a change in friends, the member along with the member's network is the

best level of measurement for all groups, no matter the predictor: friends count, votes count, and review count. In the case of change in votes, the member along with the member's network is the best level of measurement (for the predictors votes count and review count) for members that have been a part of Yelp's community for one month. The member is the best level of measurement (for the predictors votes count and review count) for groups in which the member has been a part of Yelp's community for more than one month. In the case of friends count, the predictor does not have a member level. Thus, the member's network is the same as the member along with the member's network, so the two levels have the same results.

Logistic regression, naive Bayes, neural networks (NN), and support vector machine (SVM) algorithms, are used in the second investigation. Logistic regression is introduced to generate models. Eight models are created and through statistical tests of the individual predictors (Wald chi-square statistic), goodness-of-fit statistics (Cox and Snell R^2 and Nagalkerke R^2), validation of predicted probabilities (classification table with its derived sensitivity, specificity, false positive, false negative, and overall percentage rates), a comparison of the AUC for the models probability and the AUC best predictors, NN algorithms, naive Bayes algorithms, and SVM algorithms, there is support for the robustness of these models.

BEYOND SOCIAL STRUCTURE: PREDICTING INFLUENCE IN ONLINE REVIEW COMMUNITIES

by

Vanecia L. Fluelling

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

MORGAN STATE UNIVERSITY

December 2019

BEYOND SOCIAL STRUCTURE: PREDICTING INFLUENCE

IN ONLINE REVIEW COMMUNITIES

by

Vanecia L. Fluelling

has been approved

October, 2019

DISSERTATION COMMITTEE APPROVAL:

_____, Chair
Sanjay Bapna, Ph.D.

_____
Shimi Ahmad, Ph.D.

_____
Gregory Ramsey, Ph.D.

_____
Xingxing Zu, Ph.D.

DEDICATION

This dissertation is dedicated to my family. They have shown me unconditional love and support throughout this process. They have given me strength when I thought I had none left. They picked me up when I fell. They believed in me when I was not sure I believed in myself. They protected me. Most importantly, they taught me to fight for what I believe in and to fight for all I want to achieve.

ACKNOWLEDGEMENTS

I want to thank my dissertation chair Dr. Sanjay Bapna for supporting me through this process. Also, for imparting knowledge and guidance that will be engrained in me forever. I would also like to thank my dissertation committee members Dr. Gregory Ramsey, Dr. Xingxing Zu, and Dr. Shimi Ahmad. Finally, I want to thank each one of my friends for their continued support. It really does take a village.

TABLE OF CONTENTS

List of Tables vii

List of Figures viii

List of Abbreviations ix

Chapter 1: Introduction 1
1.1 Importance of the Research 2
1.2 Contribution to the Domain 5
1.3 Structure of the Dissertation 7

Chapter 2: Background and Literature Review 9
2.1 Big Data 9
2.2 Online Social Networks 11
2.2.1. Social network analysis 13
2.2.2. Online review communities 13
2.3 Influence 15
2.4 Influencers 18
2.4.1 Centrality measures and modifications 19
2.4.2. Classification and topic level analysis 25

Chapter 3: Framework 28
3.1 Existing Models 33
3.2 Operationalization of Constructs 36
3.2.1 Measures of future influence 36
3.2.2 Measure of social structure 40
3.2.3 Measure of member engagement 40
3.2.4 Measure of overall review quality 41
3.2.5 Measure of elite 42
3.2.6 Action processes and network structure 43
3.3 Scope of Research and Study Development 46
3.3.1 Member's importance 47
3.3.2 Network importance 52

Chapter 4: Methods 55
4.1 Centrality Measures 55
4.2 Network Measures 57
4.2.1 Computation 57
4.3 Levels of Measurement 61
4.4 Data Mining Techniques 64
4.5 Dataset 65
4.5.1 User file 65

4.5.2 Experiment set-up 67

Chapter 5: Results and Analysis 75
5.1 Study 1: Determinants of Future Influence 75
5.1.1. Change in friends 76
5.1.2 Change in votes 80
5.2 Study 2: Models of future influence 84
5.2.1 Change in friends 87
5.2.2 Change in votes 95

Chapter 6: Discussion 102
6.1 Change in friends 102
6.2 Change in votes 105

Chapter 7: Challenges, Limitations, Future Research Recommendations, Practical Implications, and Conclusion 108
7.1 Challenges, Limitations, and Future Research Recommendations 108
7.2 Practical Implications 110
7.2.1 Firms 110
7.2.2 Online review communities 111
7.2.3 Yelp 111
7.2.4 Seekers of influence 113
7.3 Conclusion 113

References 121

Appendices 133

List of Tables

Table 1: Business Strategies: Comparison of Social Networks 30
Table 2: Theoretical Framework 47
Table 3: TIS (Traditional Influence Score—Degree Centrality) 59
Table 4: VIS (Value Influence Score—A Modified Degree Centrality) 60
Table 5: EIS (Engagement Influence Score—A Modified Degree Centrality) 60
Table 6: Data for Scenario 1 64
Table 7: Data for Scenario 2 (Removal of M4) 64
Table 8: User File Statistics (July 2014) 66
Table 9: File Descriptive Statistics 68
Table 10: Elite Descriptive Statistics (Totals for 1-year time frame) 68
Table 11: Description of Data Set for Logistic Regression—CIV 71
Table 12: Description of Data Set for Logistic Regression—CIF 72
Table 13: AUC Values for Change in Friends by Predictor and Level 77
Table 14: AUC Values for Change in Votes by Predictor and Level 81
Table 15: Change in Friends Models 88
Table 16: Change in Friends Logistic Regression Analysis by SPSS 90
Table 17: Classification Table Derived Statistics—CIF 91
Table 18: AUC Comparison—CIF 93
Table 19: AUC Model Comparison—CIF 94
Table 20: Change in Votes Models 96
Table 21: Change in Votes Logistic Regression Analysis by SPSS 97
Table 22: Classification Table Derived Statistics—CIV 98
Table 23: AUC Comparison—CIV 100
Table 24: AUC Model Comparison—CIV 101

List of Figures

Figure 1: Reciprocal Networks vs. Non-Reciprocal Networks 29
Figure 2: Yelpers 2 step Action Process 44
Figure 3: Characteristics of Influence Network Structure 45
Figure 4: Node Graph 59
Figure 5: An ORC Example 63

List of Abbreviations

AUC	Area under the Curve
CIF	Change in Friends
CIV	Change in Votes
COSN	Content Oriented Social Networks
eWOM	Electronic Word of Mouth
EIS	Engagement Influence Score
ENS	Engagement Network Score
GFP	Generalized Friendship Paradox
GPS	Global Positioning System
IT/IS	Information Technology and Information Systems
LSA	Latent Semantic Analytics
MIV	Model Influence Value
NN	Neural Network
ORC	Online Review Community
OSN	Online Social Network
ROC	Receiver Operating Characteristic
SAN	Social Activity Network
SIR	Susceptible Infective Removed
SNA	Social Networking Analysis
SNS	Social Networking Site
SVM	Support Vector Machine
TIS	Traditional Influence Score
VIS	Value Influence Score
VNS	Value Network Score
UGC	User Generated Content
WOM	Word of Mouth

Chapter 1: Introduction

As Greek philosopher Heraclitus asserted, change is the only constant (as cited in Luftman, Lewis, & Oldach, 1993). With it comes challenges and opportunities. What a firm does to exploit the opportunities and reduce or prevent the challenges is what enables one to stay competitive in an ever-changing environment. An area of constant change and growth that has brought rise to all types of opportunities and challenges is that of big data. Data has been increasing exponentially every day. This is due to changes in the structure of which data can be stored and analyzed and the speed of which data can be created and disseminated to others. Big data is traditionally defined by its three characteristics (volume, variety, and velocity), and is discussed in greater detail in Chapter 2.

A segment of big data that firms are actively trying to uncover opportunities from, and that plays a vital part in this study, is the online social network data associated with customers. Social Networking Sites (SNSs), also known as social media, are an advancement in technology allowing for an abundance of big data. These websites are mediums for connections and communication among users. The communities of users that emerge on these websites have been termed online social networks (OSNs).

Members of online communities or OSNs are afforded the ability to create, share, and view content in a user-friendly, informal environment. Members are also able to create networks for which to communicate, view, and share content (Livingstone, 2008).

The number of sites and users of these sites is increasing rapidly. Businesses want to collect, store, and analyze as much of this data as possible—the content the customers produce, information regarding those who create the content, information in

regards to those who consume the content, the relationships between the creators and the consumers, and the relationships between the people and the content.

This study aims to investigate these OSNs, the content, the users, and the relationship dynamics derived from them. Specifically, of interest in this study are the dynamics in online review communities (ORCs), a segment of OSNs. ORCs are understudied in research. This study will utilize data from Yelp to uncover patterns that predict influence—identifying determinants, comparing their effect, and generating predictive models. The goal is to analyze the prospective influence of members of ORCs.

1.1 Importance of the Research

To measure the market's acceptance, firms have looked to recommendations from consumers since before reviews existed on online mediums. Since the advent of the Internet; however, barriers to time and space have lifted (Fruth & Neacsu, 2014) and now there is an even greater benefit of that content coming in different structures from different mediums, signaling opportunities and a need for research of this big data segment.

Not only can value be derived from this content in terms of measuring acceptance rates, value can also be derived in terms of understanding the usefulness that potential buyers are gaining from the information and the influence that current consumers have regarding prospective consumers.

Today consumers can easily share their opinions and rapidly spread messages on to others (Fruth & Neacsu, 2014). Thirty years ago, it may have taken a customer, hours, days, weeks, and maybe even longer to spread a message about an experience with a business to a network of friends and associates. In today's highly technologically

interconnected society, spreading a message to tens, hundreds, thousands, or even millions can happen in seconds. It can be done through countless avenues.

Research indicates that consumers tend to be influenced more by word of mouth (WOM) messages of their peers than by company marketing campaigns (Muruganantham & Ghandi, 2015). Survey results show that consumers trust online WOM and are increasingly looking to this medium to make purchase decisions (Comscore & The Kelsey Group, 2007; Dimensional Research, 2013; Podium, 2017). This occurrence is based on the greater trust of the opinions of peers in comparison to the company's marketing efforts. A survey deployed in fifty-six countries to more than twenty-eight thousand Internet users shows that online reviews are the second most trusted source of business information (Aral, 2014).

The study of influence is rooted in psychology and sociology (AlFalahi, Atif, & Abraham, 2014); however, as we can see from the above paragraph, this topic is relevant to other fields, such as marketing (Kiss & Bichler, 2008; Muruganantham & Ghandi, 2015).

In the field of marketing, influence is commonly examined in relation to networks or communities of people within a market. Researchers observe how information and perceptions travel across a market and influence others. WOM lets businesses know how well a good or service is being accepted by its consumers (Fruth & Neacsu, 2014). If companies see a majority of positive comments, they can continue on their current product or service strategy. If a company is seeing a majority of negative comments, they should identify the factors causing the negative WOM and strategize to fix the issue(s). By instituting monitoring processes and strategizing actions based on

monitoring activities, companies can influence their customers (Fruth & Neacsu, 2014). Businesses must also monitor more than just the message. They must also focus on those who are creating the message.

If businesses know who the creators of the content are that significantly influence others; they can use these crowd-pleasing content creators in a way that gets the firm's message across to a maximum amount of people with little effort from the firm (Muruganantham & Ghandi, 2015). Dens, De Pelsmacker, and Purnawirawan (2015) note that firms can also springboard off the WOM message from these influencing members to either uphold the positive feeling from positive WOM (thanking the customer for their patronage and comments) or change perceptions and influence in the case of negative WOM (by apologizing for any issues and offering to make it up to the customer). Negative WOM can severely hurt a business (Dens et al., 2015).

This strategy of identifying influential users is seen as an imperative for success (Kiss & Bichler, 2008; Muruganantham & Ghandi, 2015). A study by Danaher and Rust (1996) finds that satisfaction has a positive effect on WOM and the positive effect on WOM has a positive effect on sales and market share. These results illustrate why it is important to investigate influence and identify influential users, making sure they are kept satisfied so that they provide positive WOM (satisfaction has a positive effect on WOM).

Kiss and Bichler (2008) call the influencing members of the OSNs, influencers. The researchers define an influencer as an individual who has the ability to move a message swiftly and consistently on to others. An important characteristic of these

individuals is that they tend to have ties to groups with large numbers (Kiss & Bichler, 2008).

The aim of this study is not to identify or study influencers. It is instead to understand influence—investigating different ways to determine influence, analyzing the patterns that surround influence, and using this information to predict influence. With that said, it is important to note, findings from this study can be used by businesses and ORCs to distinguish influencers.

The measures and models derived can be used to convert ORC data to business intelligence, expanding current influence research and providing businesses with a way to stay competitive by capitalizing on the abundance of data provided online.

Specifically, the results from this study provide 1) different determinants to predict influence, both in its traditional form (social structure) and modern form (social structure and actions) and 2) models that can be used to predict the future influence of users based on their ORC tenure.

1.2 Contribution to the Domain

Researchers in the marketing field are interested in this area for the purpose of understanding the market. From the information systems (IS) perspective; analyzing the data using scientific tools to provide intelligence is of importance to further the study of influence.

The plentiful data from marketing efforts (Wood, 2000), advancements in technology (Kiss & Bichler, 2008), and the growing popularity of certain websites (Wood, 2000) demands the need for tools designed by the information technology and

information systems (IT/IS) fields, as well as, analytics conducted by the IT/IS fields (Kiss & Bichler, 2008; Lu, Li, & Liao, 2012; Wood, 2000).

Areas of concentration within IT/IS such as business intelligence and knowledge management denote the importance of understanding data, information and intelligence, and the systems for creating, organizing, and managing it. Also illustrating the importance are areas such as data warehousing that focus on techniques such as data mining for classification and prediction, or information architecture and the area of security, discussing the building blocks of communication networks and safeguarding information.

Research papers such as "Identification of Influencers—Measuring Influence in Customer Networks" (Kiss & Bichler, 2008) and "A Graph-based Action Network Framework to Identify Prestigious Members Through Member's Prestige Evolution" (Lu et al., 2012), published in Decision Systems Support, demonstrate IT/IS's role. These papers also highlight how the study of influence can be done from a multidisciplinary (a topic that spans across disciplines that can be worked on separately by the different disciplines, little to no necessity for work or research between disciplinary units) perspective or an interdisciplinary (a topic that spans across disciplines and between disciplines, where work is done between units in the different disciplines; a breaking down of the boundaries between the fields and working on the topic in an integrative manner) perspective.

My research interest lies in this interdisciplinary area, the marketing and IT/IS crossover, of investigating OSNs and IT/IS's part in investigating influence. The goal is to use various scientific tools and algorithms to expand research in determinants of future

influence and predict influence for different groups. To my knowledge, there is little research looking into this while acknowledging the differences specific to ORCs. This dissertation will delve into this and prediction of influence, another understudied area. The aim is to take advantage of the plethora of OSN user data and exploit this along with the social structure to predict influence. This work will also contribute to present research by utilizing data from an understudied network in influence, Yelp, which has some significant differences from traditional OSNs, Facebook and Instagram.

1.3 Structure of the Dissertation

In the next chapter, Chapter 2, background and a review of the literature is provided. An overview of big data is offered and provides context for how technological advancements set the stage for a new avenue and medium to be studied, OSNs. The literature review facilitates distinguishing between traditional SNSs and ORCs, noting ORCs as the environment for this study. Next, a succinct review of the influence literature is provided to aid in understanding the historical context of influence in sociology literature. Last, but not least, a detailed review of the literature for identifying and/or predicting influence(rs) in OSNs is described in this section. The review of these different aspects provides evidence for the need to expand the current literature by means of this research.

In Chapter 3, a discussion of the different types of networks and aligning business strategies, as well as the existing business models, are introduced to provide the framework for this study. The chapter will then operationalize constructs and develop the basis for the subsequent analysis.

Chapter 4 will explore the different methods used in social network analysis, indicating the measurements that will be used in this investigation. It will also present data from the Yelp dataset being examined for this study.

Chapter 5 provides data analysis and results. This section is separated into two studies, determinants of future influence and models of future influence.

In Chapter 6, there is a discussion of the results. Chapter 7 provides the limitations of the study, implications to practice, future research recommendations, and conclusion.

Chapter 2: Background and Literature Review

Decades ago, data was hard to come by and companies had to spend much of their resources to collect and store it. Today, the amount of accessible data is abundant, and businesses have a different problem, effectively managing it (Economist Intelligence Unit, 2011).

2.1 Big Data

What has changed about data that has given rise to this abundance and this accessibility? One change is volume. The amount of data that now flood the marketplace has increased exponentially over the last few years and the volume is only continuing to increase over time (Economist Intelligence Unit, 2011; Gandomi & Haider, 2015; Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, & Byers, 2011). Businesses are storing more detailed transaction information, personal information, and consumer behavior information on a more frequent basis (Manyika et al., 2011). In the year 2007, we had the first occurrence where the volume of digital data created in one year exceeded the storage capacity of the world (Manyika et al., 2011). Brown, Chui, and Manyika (2011) report that on average, U.S. businesses with over one thousand employees store over 235 terabytes of data. Globally in 2010, businesses stored more than 7 exabytes of data on disk drives (Manyika et al., 2011). In 2014, the forecast of stored data for the world for the end of the year was 2.16 zettabytes. The prediction for 2016 was 3.77 zettabytes (Mearian, 2014). A zettabyte equals a little more than 1,000 exabytes. Current predictions put the sum of the world's data at 175 zettabytes by 2025 (Patrizio, 2018).

Contributing to the rise in the volume of data and its accessibility is another important item, velocity. With today's technologies, we can get real-time data relatively

cheap (Gandomi & Haider, 2015; Manyika et al., 2011; McAfee & Brynjolfsson, 2012). While in the past, the issue may have been not being able to collect enough data, real-time data has caused some companies to have the problem of excessive data. The healthcare industry discards 90% of their data. Most of that data is their real-time surgical video feeds (Manyika et al., 2011). The Global Positioning System (GPS) is an advancement in technology that enables real-time location data (Manyika et al., 2011). It is being used in a wide variety of industries such as construction, banking, and package delivery. GPS is used in various devices such as ATMs and cell phones (GPS.gov, 2013). This type of data becomes rapidly multiplied and excessive because firms have quick and easy access to it. We can see from these examples that velocity contributes to an increase in the volume of data. Velocity also operates alongside volume as a separate dimension to describe an aspect of data which did not before exist, data in real-time.

In working to create and transmit this real-time data, we are introduced to an additional element. This element is variety. It is another reason for the continued growth in the volume of data in today's era and a contributor to the increase in the velocity of data (Gandomi & Haider, 2015; McAfee & Brynjolfsson, 2012). Variety refers to the different sources and types of data available today (Lohr, 2012). Video feeds, GPS, and sensors discussed earlier are examples of the variety dimension (Manyika et al., 2011). These types of data are called unstructured to denote the absence of organization (McAfee & Brynjolfsson, 2012). Unstructured data can take the form of text from email, webpages, customer comments, internal documents, service personnel documents, (Kuechler, 2007), videos (Kuechler, 2007; McAfee & Brynjolfsson, 2012), images, sensor data, data from GPS signals, and social media (McAfee & Brynjolfsson, 2012).

Unstructured data is not usually in a form favorable for traditional databases; however, it is used in many of today's new data sources (Lohr, 2012).

These three dimensions: volume, velocity, and variety make up what has been termed big data. Firms that are able to effectively manage their big data can set themselves apart from others and gain substantial rewards (Economist Intelligence Unit, 2011).

Social Media also known as OSNs or Social Networking Sites (SNSs) are a repository for big data that researchers and practitioners are very interested in. This medium offers data that touches every dimension of big data. As was just discussed, the avenue allows for an unstructured format; therefore, it touches the variety dimension. The information that is created can be shared in real-time which touches on the velocity dimension. Practically anyone anywhere can create content as long as they have access to a machine with an operating system and the internet, so it also touches on the volume dimension.

OSNs with its rich data is what is of interest in this study.

2.2 Online Social Networks

SNSs have become one of the most popular applications on the Web (Acquisti, 2005; Gross & Kumar, Novak, & Tomkins, 2006; Kuss & Griffiths, 2017; Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007). They house OSNs. The purpose of SNSs is to connect people with other people (Kuss & Griffiths, 2017; Mislove et al., 2007) and to foster communication among its members (Livingstone, 2008).

Users of SNSs can do more than just connect with others. Members have the ability to create and share content such as photos, videos, and diaries. A user can invite

others into the community and their specific network, make their profile available for all to view or only for specific members to view, block or add other members, and a plethora of other activities (Livingstone, 2008).

Depending on the site housing the OSN, the connections created between the members can be undirected or directed. The terms directed and undirected denote both the route in which a user is connected to another user and the information flows from one user to another user (Fu, Liu, & Wang, 2007; Pal, Kundu, & Murthy, 2014).

An undirected network is one in which there is a bilateral agreement to the connection and sharing of information flow. In a directed network, there is a clear distinction between an incoming connection and an outgoing connection and the route in which information flows (Fu et al., 2007; Pal et al., 2014). An undirected network's strategy is to maximize information flow. A directed network's strategy is to maximize useful information for each person.

Popular SNSs include Myspace (Kumar et al., 2006; Mislove et al., 2007), LinkedIn (Mislove et al., 2007), Instagram, Twitter, and so on. All of these sites have directed networks.

The variety of SNSs is so vast that they have been grouped into categories of interests, friends, pets, photos, dating, face-to-face facilitation, and business (Gross & Acquisti, 2005).

While different sites may provide different options to users, they all share certain commonalities (Gross & Acquisti, 2005). A study of OSNs, within two social networks, Flickr and Yahoo! 360, show that the interconnections follow a certain pattern in which the associations grow quickly at the start, decline, and then taper off with slow and steady

growth (Kumar et al., 2006). In addition to looking at the properties of social networks, the researchers classify the OSN users. Singletons are the users who join a network but have no connections. Giants are those that make connections, direct or indirect, with others that enable the group as a whole to link to a large segment of the network population. The middle group is the group that makes connections and form small communities within the larger network (Kumar et al., 2006).

2.2.1. Social network analysis

Social Network Analysis (SNA), also known as Social Network Theory, looks at the relationships and exchanges between the individuals that are the make-up of the network. The underlying premise is that there are patterns to the social connections (Scott, 1988). The analyst's goal is to use methods to uncover the patterns and circumstances of which arose the social structure (Scott, 1988).

2.2.2. Online review communities

This study analyzes the effect that happens between actors in a specific OSN segment; a segment where the goal is maximizing information flow and many connections are made as a result of content versus familiarity with others. This segment I call online review communities (ORCs) to denote the site's emphasis, reviews.

Earlier it was noted that users of SNSs generate content such as photos, videos, and diaries (Livingstone, 2008). Another important type of user-generated content (UGC) are reviews (Ngo-Ye & Sinha, 2014). Reviews are also called word of mouth (Dens et al., 2015; Fruth & Neacsu, 2014). Word of mouth (WOM) has been identified as a leading source of information and influence (Kiss & Bichler, 2008).

Examples of sites where users generate review content are tripadvisor.com and Yelp.com (Ngo-Ye & Sinha, 2014). Reviews make up the majority of UGC. This content makes users aware of the opinions of consumers (Ngo-Ye & Sinha, 2014). Reviews are information that users provide regarding personal experiences in dealing with a product or service (Ngo-Ye & Sinha, 2014).

Users view this content as a trustworthy source of information (Ngo-Ye & Sinha, 2014). Comscore and the Kelsey group's 2007 report shows that approximately twenty-five percent of Internet users read online reviews before making an offline purchase. Seventy-five percent of those users who read online reviews indicate that reviews make an impact on their purchase decisions (Comscore & The Kelsey Group, 2007). Dimensional Research's (2013) report six years later shows a rise in the numbers. Sixty-six percent of U.S. Internet users read online reviews and eighty-eight percent say it impacts their purchase decisions (Dimensional Research, 2013). Podium's (2017) report just two years ago indicates that the numbers have risen yet again. Fifty-eight percent of Internet users read online reviews. Ninety-three percent say that online reviews impact their purchase decisions. In a new finding, sixty-three percent of consumers reveal that they are willing to pay a higher price if they believe they will have a positive experience (Podium, 2017). Kumar et al. (2006) proclaim that many businesses owe their success to the Internet and communities that share review content.

This influence effect that users have on other users in these ORCs is the basis for study in this work.

2.3 Influence

Research on influence started in the psychology and sociology fields. One of the earlier papers on influence (French, Raven, & Cartwright, 1959) led to the topic of influence by examining social power. In this early paper, influence is recognized as the basis of power and it is defined as a psychological change (French et al., 1959). "Psychological change is defined as any alteration of the state of system a over time" (French et al., 1959, p. 151). Examples given are behavior, opinions, or values, but the authors note that psychological aspects include any aspects of an individual's psychological field. A change in any item can be prompted by different forces, such as a driving force like another person or a restraining force like a group's opinion. In the same year, Jahoda (1959) defines social influence, a slightly different term used than solely the term 'influence' in French et al.'s paper (1959), as conforming to the majority (Jahoda, 1959).

In 1961, Kelman delves deeper into this notion of change and conformity and discusses the categories of social influence: normative and informational. Normative social influence refers to the changing of one's opinions or beliefs in order to comply with others' expectations (Kassarjian, 1971; Kelman, 1961; Kuan, Zhong, & Chau, 2014). Informational social influence refers to the changing of one's opinions or beliefs based on the evaluation of evidence (Kassarjian, 1971; Kelman, 1961; Kuan et al., 2014).

Approximately thirteen years later, Kelman (1961, 1974) continues with influence research and offers three bases of influence: compliance, influence centered around reward and punishment; internalization, influence based on whether one is accepting of

another's beliefs; and identification, influence based on one's positive or negative feelings toward another person.

Rashotte (2007) identifies five categories of social influence research. Minority influence research examines the attempts of minority groups to change the behavior or opinion of majority groups. Persuasion research concentrates on the verbal or written messages and the change that occurs based on the information. Dynamic social impact theory research recognizes that the impact of a message is based on the closeness of the source of the message, the number of people that make up that source, and the strength of that group of people and then use this information to explain and predict the circulation of messages through social systems (Rashotte, 2007). Structural approach research highlights the make-up of the people within a group and the process by which individuals weigh the opinions of those in the group and assimilate their beliefs (Rashotte, 2007). Expectation states theory research posits that individuals of equal status within a group will set expectations of the ability level of each person, including themselves, based on a future task. This expectation of ability level corresponds with the level of a member's influence within the group, the higher the perceived ability towards a task, the higher the level of influence within the group (Berger, Rosenholtz, & Zelditch, 1980; Rashotte, 2007). Oldmeadow, Platow, Foddy, and Anderson (2003) discuss expectation states theory in their article but refer to it as social influence and status characteristics theory. Participation rates of members within a group and the influence they exert are a function of the expected ability level of the members within the group to perform a specific task. They also discuss social influence in terms of Turner, Hogg, Oakes, Reicher, and Wetherell's (1987) self-categorization theory. As individuals develop their social

identity, they gravitate towards certain groups they share values in common with and they tend to adopt the beliefs and behaviors of those within the group that are perceived to be ideal members. These beliefs and behaviors are seen as items that should be held collectively across the group (Oldmeadow et al., 2003).

Anagnostopoulos, Kumar, and Mahdian (2008) describe social influence as the use of an individual's actions to motivate others to behave in a similar fashion. Kuan et al. (2014) define social influence as an individual modifying their actions based on the actions of another. Muruganantham and Gandhi (2015) define social influence as the behavioral change an individual makes in response to the actions of others in their network. AlFalahi et al. (2014) and Tang, Sun, Wang, and Yang (2009) note that social influence involves an individual changing his or her behavior after an interaction with others in their social network. Godinho de Matos, Ferreira, and Krackhardt (2014) use the term peer influence in their paper. Their definition is similar to those used for social influence. Peer influence is defined as the process of which an individual adopts behaviors, beliefs, or attitudes based on that of others in the individual's social system.

It appears that though influence research has advanced, and its span has increased, the definition French et al. (1959) established from psychology and sociology remain applicable to other streams and in newer environments. Influence is a psychological change. A psychological change is a difference in a state of a system that happens over time. The definition is general and therefore can encompass both normative and informational influence and can be applied to different environments across different disciplines.

In this work, I do not seek to investigate or track the change that happens within a person's psychological state due to influence. I seek to modify established measures and methods to identify the best predictors of influence emanating from the actions of users in ORCs. In an ORC, the influencing members are those who can move a message swiftly and consistently to others. This is done by means of writing reviews that are then voted on by others. Based on the findings, I then seek to generate predictive models of influence that will identify, in ORCs, those who will likely evoke psychological change in others.

2.4 Influencers

The study on the topic of influence and related topics in OSNs has exploded in the last several years. Articles examine the influence of WOM messages on purchase decisions and or revenue (Comscore & The Kelsey Group, 2007; Dimensional Research, 2013; Goh, Heng, & Lin, 2013; Kumar et al., 2006; Liu & Zhang, 2010; Lu, Ba, Huang, & Feng, 2013), influence of WOM messages on users to join product specific groups (Coulter & Roggeveen, 2012), factors that influence users to create content (Li & Yang, 2014), estimation of (positive or negative) influence for a user creating a certain type (positive, negative or neutral) of message (Li, He, Wang, & Zhang, 2014), on the influence of businesses' response to negative reviews on the perceptions of other users (Dens et al., 2015), influence of reviewer status on ratings (Zhu, Yin, & He, 2014), approaches to influence OSNs (Faletra, Palmer, & Marshall, 2014), influence of number of people who friend or follow someone on other's opinion formation (Battiston & Stanca, 2015), and influence of reciprocal links in comparison to non-reciprocal links in networks (Zhu, Zhang, Sun, Tang, Zhou, & Zhang, 2014).

While these are worthwhile research areas; my work is specifically concerned with the influence of members in a network on others. I have broken down the current literature in this realm based on how influence has been researched and influencers identified, centrality measures (and modifications) and classification and topic level analysis. I have provided a review below.

2.4.1 Centrality measures and modifications

Dubois and Gaffney (2014) investigate influencers amongst Twitter data in regard to two major Canadian political parties. Five measures of influence were employed: indegree, eigenvector, clustering coefficient, knowledge, and interaction. The first two are centrality measures, the third is a measure of grouping, and the last two are content or semantic-based measures. The study finds that there are different influencers identified based on the measure used by the researcher. Influencers, considered traditional elites, such as journalists were considered influencers under indegree (number of users wanting a relationship) and eigenvector (measure of importance based on how highly connected one is with others who are highly connected). Measures in content analysis, such as looking at content-specific terms, identified influencers that were more local experts—bloggers or commentators (Dubois & Gaffney, 2014).

Xu, Sang, Blasiola, and Park's (2014) study uses both content analysis of user's tweets, and network analysis (betweenness centrality), to identify the characteristics of opinion leaders on twitter. Tweets are read, coded by coders, and classified into action, information, commentary/community. These categories represent the involvement characteristic. By identifying the characteristics, the researchers can predict user performance. The results show that 1) users with higher connectivity and issue

involvement are the influencers in the network and 2) tweets from organizations have a higher influence than tweets from individuals.

Kim and Han (2009) identify influencers of OSNs by examining both the user's social structure in a network and the user's activity history. They first find the candidate group by looking at degree centrality and then look at the activity of the users to find the actual group of influencers.

Lu et al. (2012) work both tries to identify and predict future influencers, called prestigious members, by examining actions of users on Flickr. By examining both the structure of the network (measure for degree, indegree-outdegree correlation, and assertive mixing pattern) and the actions of the users in the network, the researchers identify current influencers. Then the researchers use sociological theory factors to identify key factors for action intentions and use this information to identify potential influencers.

Zhang and Li (2014) use agent-based modeling to simulate OSNs to understand the superiority of measurements when working with different types of network structures and different types of diffusion models. The centrality measures they study are out-degree, sender rank, weighted out-degree, and betweenness. The network structures they examine are scale-free networks, small-world networks, and random networks. The diffusion methods they investigate are informational and normative under the susceptible-infective-removed model. Zhang and Li (2014) find that centrality measures show a different pattern under each type of network, therefore the structure of a network, infection probability, and infection mode in small-world networks all have a significant

impact on the superiority of the measurements. They also find that the weighted out-degree centrality measure performed at the top in every experiment condition.

Pal et al. (2014) are primarily interested in devising efficient algorithms for computing influence in graph networks that also examine the influence of neighbors. They look at the influence maximization issue, trying to maximize influence by selecting the users that influence the most users in a network. They create two measures for identifying influencers. The researchers argue that the measurements of influence, degree and betweenness centralities, only look at how close a user is to the action. To include also the neighbors' contributions, they come up with a diffusion degree and maximum influence degree. The first considers the initial user or node and the contributions of his or her neighbors (diffusion degree). The second looks at the maximum possible influence of a user (maximum influence degree). They conduct their experiment using Twitter, Amazon, Slashdot, and two university friendship networks. Findings indicate that the computations for their two new measures are significantly quicker than other algorithms and that taking into consideration the contributions of neighbors is significant (Pal et al., 2014).

Liu, Jiang, Lin, Ding, Duan, and Xu (2015) develop an approach based on the Epinions data set for identifying influencers. By looking at user trust networks, variance over time, and domain, the researchers categorize users as holding influencers, vanishing influencers, and emerging influencers. The trust network is derived from direct information from trust lists and trusted lists. The domain and time are extracted from the review data. Also extracted are the review ratings. Centrality measures indicating how

many incoming relationships and outgoing relationships (indegree and outdegree) are then used to determine influence.

Li, Lai, and Chen (2011) examine networks of users, content (subject, length, and time it stays available), and activity to come up with an influence score and identify the most influential bloggers. Combining these three aspects leads to what the researchers call model influence value (MIV). Using an artificial neural network for the experiment, the researchers find that their model performs better than out-degree and betweenness centrality network measures and popular author and review rating content-based measures.

Momeni and Rabbat (2016) examine twitter data, specifically looking at quality through a user's attributes: their activities (total number of user's tweets, including re-tweets and total number of user's unique tweets, restricted to only the ones that were original and initiated by them) and their influence (number of times a user's tweets were retweeted, number of tweets that were retweeted at least once, the average number of retweets that a user receives for an original tweet, and normalized version of the number of tweets that were retweeted at least once). Once the researchers assess the quality, they use a neighbor superiority measure to get to their goal of examining the generalized friendship paradox (GFP). The GFP states that each user will show less action on average than those of their friends. Measures of neighbor superiority used were based on the fraction of median/mean and follower/followee for all the attributes. Momeni and Rabbat (2016) find that 1) three quarters of the users in the data set have no influence over others, indicating they are simply observers and not producers of content and 2)

OSNs have a hierarchical structure rather than a star configuration, indicating that people follow those with similar attributes or higher attributes (Momeni & Rabbat, 2016).

Nguyen and Zheng (2014) use Klout and PeerIndex services to determine user influence scores for a data set of Twitter users. They examine the structure of the Twitter community and find that whether a user retweets a message is based on who of their friends was the first to tweet the message. Due to this, the researchers propose a model of diffusion that takes the friend who first tweeted the message into consideration.

Subbian and Melville (2011) propose combining centrality measurement methods (in degree, out-degree, and page rank) to create a different measure of influence. They argue that in using a single measurement, one can miss out on potential influencers due to different measurements having different results. They look to aggregating aspects of the methods to combine the centrality measurements and predict the retweeting of tweets on Twitter.

Zhang, Li, and Wang's (2013) research examines the role of tie strength in identifying influencers. Tie strength is a characteristic of relationships in social networks. The researchers assert that while centrality measures have been used to identify influencers on the quantitative level, a measure of tie strength captures the qualitative level. Strong ties are shown through consistent communication and weak ties are shown through inconsistent or occasional communication. The results of their agent-based modeling simulation indicate that when the ratio of strong ties to weak ties is high, or when strong ties are at a high percentage throughout a network, the measurement of tie strength is a valid option in identifying influencers.

Vidmer, Medo, and Zhang (2015) argue that popular items, such as trending topics or images, in the data can skew measurements of influence. By measuring the probability someone is going to take an action based on a friend's behavior, they work to identify influence strength of users without the bias of popular items. They compare how many times a person takes an action based on a friend's behavior (count between pairs) to how many times it would happen in a null model. Vidmer et al. (2015) measure the spread of an action or behavior by comparing a count within the original network to a count with the subnetwork. The subnetwork only consists of those who are found to have taken a specific action. The researchers use the member's network and how many friends are likely to take an action, how many times each friend will take the action, and how the action will spread through networks. This research follows the networks of users and looks at the actions of the users to see who is influencing whom and how much (Vidmer et al., 2015).

Zhao, Li, Xie, Wu, Xu, Ma, and Lui (2016) propose to look at the influence maximization problem from a social activity network (SAN) perspective. While the traditional way of looking at the problem is to look at the network, the researchers assert that looking at the online activities, such as giving ratings to products, needs to be added into the equation. Zhao et al. (2016) reformulate the influence maximization problem to account for the user's social network and online activities. They create a measure using a random walk approach on a hypergraph to find the influencers for Ciao, Yelp, and Flixster (Zhao et al., 2016).

Budalakoti, DeAngelis, and Barber (2011) seek to identify the most trustworthy people in user-generated content (UGC) online networks, knowledge sharing oriented

OSNs, or content-oriented social networks (COSN). They define trustworthy as those individuals who add the most value by being loyal and productive in the community. Since the site's purpose is content-oriented, the social or networking aspect can bring rise to items such as nepotism or reciprocity, identifying trustworthy members is not a simple task. The researchers therefore create an algorithm that takes into consideration both similarity-based motivation (occurs when a user follows a content creator because of a preference for content by that content creator) and social influence-based motivation (occurs when a user endorses content from a creator because of a social relationship with that creator) to more accurately identify the more highly loyal and productive users in ORC Digg (Budalakoti et al., 2011).

2.4.2. Classification and topic level analysis

Alarcón-del-Amo, Lorenzo-Romero, and Gómez-Borja (2011) classify users based on experience in SNSs, sociodemographic variables, number of times users perform certain activities, and interaction patterns. They use latent segmentation methodology and cluster analysis. Their classification, expert-communicator, is considered the influencer group. The expert-communicator performs more activities more times than others and they perform more marketing-related activities, such as gathering information on brands and or commenting on advertisements.

Fang, Sang, Xu, and Rui (2014) use Flickr data to identify topic level influential users and images. Fang et al. (2014) employ a hypergraph approach to graphically view the users and images (the vertices) and the relationships between users and images and the image-content relationships among images (the edges). Their topic level mining

algorithm first learns the topic for each user and then comes up with an influence score based on the topic for each user.

By building a single topic comments network, Huang, Yu, and Karimi (2014) work to identify the most influential comments within a SNS. Then, from the results of the single topic network, they map the comments to users and identify the most influential members. The researchers use Weibo, a SNS for news in China for their investigation. Huang et al. (2014) examine emotions and take into consideration the time when a user adds a topic and comments are made. A topic that has more positive comments is considered influential. The longer it takes for replies to a topic, the less influential the topic is considered. A random walk algorithm they call Dynamic Opinion Rank is employed for the study. Their work is inspired by the PageRank algorithm. A finding from the study is that opinion leaders vary with time.

Li and Du (2011) propose to identify influencers through first identifying hot blogs. They look at the blog content, the author and reader expertise and blog preference, and the homophily and social tie relationships between the author and reader. After identifying hot blogs based on selected topics from the factors just named, Li and Du (2011) identify who are the opinion leaders.

Jindal (2015) conducts a study, for a master's thesis, aimed at using Yelp data to find local experts for different business categories. The data used includes business categories, business locations, user reviews, user review rating, user start date, and user friends. Using this data, Jindal (2015) creates an algorithm that ascertains both topical authority (expertise on a specific topic) and local authority (association with a locale).

Through a classifier system, this algorithm identifies the experts for a specific topic for a specific geographic location.

Based on examining content (topic, similarities in posts), behavior (viewing, replying, posting, and forwarding) and time, Li, Ma, Zhang, and Huang (2013) use a mixed approach to identify opinion leaders in online learning communities. These aspects assess expertise, novelty influence, and activity. The researchers rank leaders based on their influence, expertise, activity, and novelty. They investigate the centrality of leaders (uses the number of posts for a user on a topic and number of replies for that user on a topic) and longevity (based on behaviors).

Chapter 3: Framework

More recent influence research has focused mainly on Twitter's OSN (Dubois and Gaffney, 2014; Momeni & Rabbat, 2016; Nguyen & Zheng, 2014; Pal et al., 2014; Subbian & Melville, 2011; Xu et al., 2014); however, as noted in Chapter 2, there are different types of networks. It is imperative that research reflects examinations of different networks, specifically when the relationships formed follow a different structure that may, therefore, offer different results (Zhang & Li, 2014). The business strategy chosen by the SNS defines the relationship structure and information focus of the OSN. Members of the site can be connected to each other through two dominant means (reciprocal or non-reciprocal), thus resulting in two dominant network structures (directed or undirected respectively). In Figure 1 below, I have illustrated the difference between the two networks.

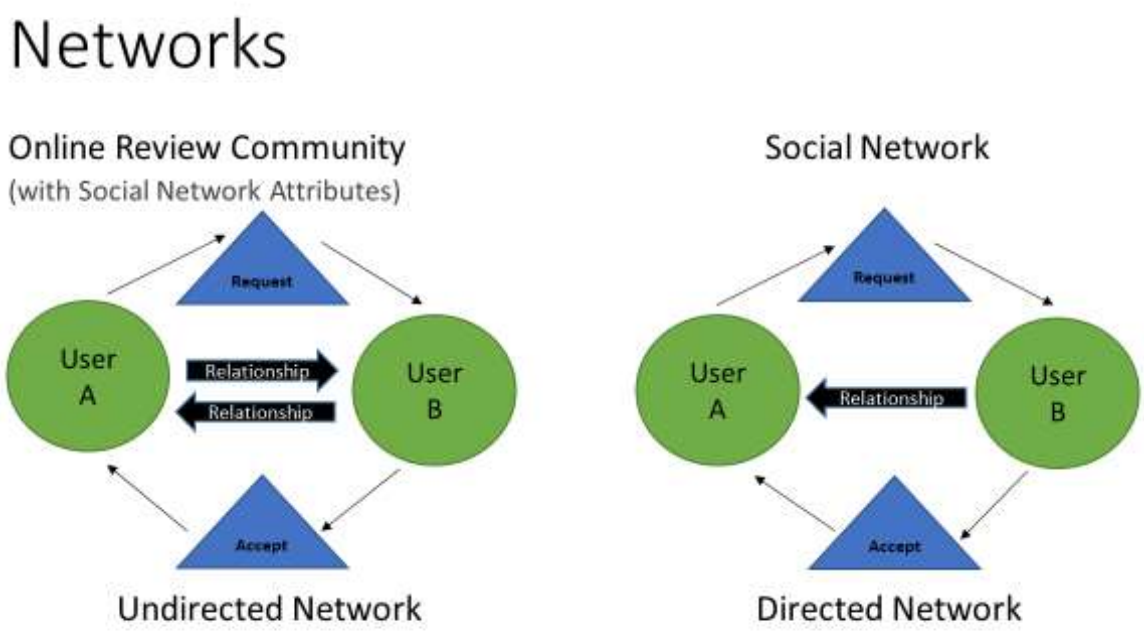


Figure 1: Reciprocal Networks vs. Non-Reciprocal Networks

In a reciprocal business strategy for an OSN, illustrated on the left side of the figure, an automatic bi-directional connection occurs when one party accepts a friendship request of an initiating party. Information flows from one to the other, each is following and a follower. In a non-reciprocal business strategy, illustrated on the right side of the figure, a one-way connection is created between two parties once a party accepts a friendship request from an initiating party. There is no automatic reciprocation of the accepting party in also following the initiator. Information flows from one to the other but not in reverse.

Twitter, Facebook, and Instagram are non-reciprocal OSNs. The connections and information flow between users are specific to the way in which the relationship was initiated between the users. Yelp and Digg are examples of reciprocal networks.

Table 1 below provides a comparison of the structure, focus, and information flow of the two business strategies.

Table 1: Business Strategies: Comparison of Social Networks

Business Strategy	Network/Relationship Structure	Information Focus	SNSs
Reciprocal	Undirected Network Bilateral Relationship	Maximizing information flow of messages • Information flows between users in both directions	Yelp and Digg
Non-reciprocal	Directed network Unilateral Relationship	Maximizing useful information for each individual • Information flows between users in 1 direction (direction depends on who initiates the friend request)	Twitter, Facebook, and Instagram

An example to illustrate the advantages and disadvantages of each of the business strategies and their focus as emphasized in Table 1 follows.

As illustrated in Figure 1, in a non-reciprocal network, agent A has 1,000 friend requests. Agent A accepts these requests and now has 1,000 followers; however, agent A only chooses to follow 25 of these individuals. Agent A has made the choice to provide information to 1,000 others while only receiving information from 25 of these individuals. Agent A will only see information from those he or she deems as useful. From a business standpoint, agent A is missing out on 975 opportunities to be marketed

to; however, agent A is receiving messages from a small group. This means he or she is less likely to be resilient to the messages and more likely to be susceptible to the messages. For a reciprocal network, if agent A has 1,000 followers, he or she is automatically following 1,000 people. This is great for maximizing the number of messages being received by agent A and being promoted by agent A. The problem with this situation is that we know from the non-reciprocal network scenario, agent A only wants to follow 25 people. By now following 1,000, he or she is getting bombarded with information from 975 people he or she does not prefer. This may cause agent A to be too inundated with messages. Agent A can either become immune to messages or he or she can miss messages from those individuals that he or she actually deems important.

Vidmer et al. (2015) point out that a user can become resilient to messages when they receive too many of them. When a user has many friends, it takes more messages from friends to break through the user's resistance. The other issue that may happen is that when agent A's friend, agent B, decides to join the social network, agent A would like to befriend agent B. Because agent A has hit the maximum number of friends for his or her network, he or she must either choose not to befriend agent B or to dissolve one of his or her current relationships.

It is not uncommon in social networks for reciprocity to exist (Budalakoti et al., 2011). By taking away the choice of whether to reciprocate and limit the number of relationships, users will reach their maximum quicker than that of a non-reciprocal network.

This issue may not arise for most users. There tends to only be a finite group of users who make a lot of direct connections. That finite group; however, is of interest to businesses because of their potential ability to influence, directly or indirectly, many others.

Undirected networks tend to be understudied in comparison to directed networks. This study will extend the research of undirected networks as well as highlight other aspects pertinent to predicting influence that are specific to ORCs, such as Yelp. Yelp's goal is to connect users with great local businesses (Stoppelman, 2013). The method of connection is reviews (WOM). As of June 2019, Yelp had 192 million reviews on the site ("An Introduction to Yelp Metrics", 2019). The reviews have not only worked to connect people with businesses but have also worked as a vehicle to connect users with other users. The common interest of users lies in the act of reviewing, reading reviews, or in a shared topic or interest. The reviews have brought people together.

Restricting the viewing capacity of users to only those in one's network would hinder the forming of new relationships and the mission of Yelp. Thus, Yelp enables members and non-members to view member reviews.

A few studies discussed in Chapter 2 attempt to use Yelp data, (Jindal, 2015; Vidmer et al., 2015; Zhao et al., 2016) to investigate both social structure and review activity for the purpose of identifying influence(rs). Jindal's (2015) study uses Yelp data to find local experts for different business categories. Zhao et al. (2016) utilize both member's social network and online activities to define a random walk approach to find influencers (Zhao et al., 2016). Vidmer et al. (2015) measure influence by examining the probability of someone taking an action based on the behavior of a friend.

This study seeks not to categorize users by their expertise or use a significantly different methodology than classic measures, or predict someone being influenced. This study seeks to modify the calculation of a classic measure to incorporate measures of actions, examine measures of potential influence, and predict influence in ORCs.

3.1 Existing Models

Pal et al. (2014) and Zhao et al. (2016) frame their studies of influence under the influence maximization problem. Using this framework, one seeks to maximize the number of people being influenced by a message by identifying the best group of people to start sending a message (Pal et al., 2014; Zhao et al., 2016). In marketing, the goal is to identify individuals to promote and manage WOM that increases brand awareness (Lu et al., 2012). These marketing campaigns are usually called viral marketing, denoting the cascade effect that happens in the diffusion process (Lu et al., 2012; Vidmer et al. 2015).

Many models have been proposed of which to examine the influence maximization problem. Models include the susceptible infective removed model, threshold model, and cascade model.

Most of the models are based on the concept of a virus. A virus spreads through what is called an infection. The infection spreads through networks passing from one agent to another agent. In the case of a bacterial infection, a virus may spread through touch. In the case of a computer virus, a hacker may spread infection through a coded message in an email.

In the case of marketing, the infection is spread through WOM (Pal et al., 2014). The goal of the marketing team is to infect individuals that will then infect many other

individuals (Pal et al., 2014; Vidmer et al., 2015). Infection is analogous to influence in this respect (Vidmer et al., 2015). Some of these models are described next.

The susceptible infective removed (SIR) model follows the diffusion process of a virus. It adds that an agent that becomes infected may be cured of the affliction and become immune to its effects in the future. This immune status causes the agent to be removed from the network because they become resilient to more messages (Zhang & Li, 2014).

The threshold model indicates there is a level of infection needed before which an individual will become infected with a condition. In other words, an individual will become infected once a certain proportion of his or her neighbors is infected with it (Vidmer et al., 2015).

The cascade model assigns a probability to which a non-infected agent will be infected by an infected agent (Pal et al., 2014). Each time an agent's neighbor is infected, it increases the probability of infection for the non-infected agent (Vidmer et al., 2015). The probability of infection is termed propagation probability (Pal et al., 2014).

Kempe, Kleinberg, and Tardos (2003) find that the threshold and cascade models, in their general state, are equivalent. Other attempts at modeling have included the representation of a network as multi-levels in which there are relationships between users and between users and items (Kim & Han, 2009; Vidmer et al., 2015; Zhao et al., 2016).

This dissertation follows the social structure and action analysis theme of past research (Jindal, 2015; Kim & Han, 2009; Lu et al., 2012; Vidmer et al., 2015; Zhao et al., 2016) and frames the influence maximization problem under the multi-level network virus model. Both actions and the social structure network is used to investigate who will

influence and what items play a part in that influence. In seeking to predict who will influence, I work towards the influence maximation problem by differentiating between the groups of users a company should and should not target. The differentiation is based on the prediction of a user to influence. This is valuable especially when a person is new to an OSN and may not have much to measure in terms of current influence. As Momeni and Rabbat (2016) report, seventy-five percent of the users in their study had no influence over others. It is beneficial to predict which members will fall into what group.

Research into social structure (friends) and actions (number of posts, votes, or shares) that users take in OSNs has proven to be a valuable area of study for influence research (Jindal, 2015; Kim & Han, 2009; Lu et al., 2012; Vidmer et al., 2015; Zhao et al., 2016). This study extends current research by also investigating aspects of an action (in ORC case, the review itself) and comparing different actions of users (reviews—action of a user, and votes—response action to a user) at the individual and network levels.

Lu et al. (2012) combine social structure, action network creation, and data mining for their study, but the key differences are that 1) they conduct their study on a directed network; therefore, using measurements key to that social structure cannot be used in this study of an undirected network, 2) their network looks at the actions that take place between users and weighs that connection, while my network variables place a value on a user by the user's action and the actions of those in the network (to be discussed more in Chapter 5), and 3) they use data mining and sociological theory factors for prediction while I use strictly data mining techniques. Also, of note, this study

investigates determinants and prediction models based on the length of time a user has been a member of the ORC.

Lu et al. (2012) use data mining to identify factors (homophily, triadic interaction rule, continuity and recency sociological theory) that signal the intention of a user to take an action along with measures of a user's current structure to predict influencers. Subbian and Melville (2011) use an aggregate of different measurement methods to predict influencers through the prediction of retweets. These studies have a commonality which makes them different from this study. They use directed network data and employ measures that quantify the incoming and outgoing relationships of a user. These items are not relevant to an undirected network.

3.2 Operationalization of Constructs

Five constructs are utilized in this SNA: influence (specifically defined below as future influence), social structure, user engagement, overall review quality, and elite status.

3.2.1 Measures of future influence

There are two measures of future influence examined in this dissertation, a measure of the change in the social structure (friends) and a measure of the change in votes (CIV).

3.2.1.1 Change in social structure

Social structure is an established construct in influence research (Li et al., 2011; Liu et al., 2015; Pal et al., 2014; Xu et al., 2014; Zhang & Li, 2014). Centrality measures, discussed in Chapter 4, are the method of measurement for this construct. These measures are specifically designed to quantify the influence of users by their social

structure (Li et al., 2011; Liu et al., 2015; Pal et al., 2014; Xu et al., 2014; Zhang & Li, 2014). The idea is that the position a user has in a network indicates the breadth or potential reach of a user.

Traditionally, researchers have used the measure of influence for identifying current influence. In this study, influence is measured as it pertains to future influence. The centrality of members at two time periods is calculated and used to quantify influence at both time points. The difference (change) in the social structure over the two time periods establishes whether influence occurred within the time frame. This change is used to train the prediction models in the supervised machine learning methods used and discussed in Chapters 4 and 5. Noteworthy here is that the influence being studied in this dissertation is a positive influence or a positive change.

3.2.1.2 Change in votes

In the more recent research discussed in the review of influence literature above, the actions of users have become an area of study in examining influence (Jindal, 2015; Kim & Han, 2009; Lu et al., 2012; Vidmer et al., 2015; Zhao et al., 2016).

In addition, the SIR, threshold, and cascade models reviewed in the previous section highlight the importance of looking for ways, other than the traditional format, to more accurately examine influence.

The SIR, threshold, and cascade models emphasize the need to look beyond the overall potential of a user to influence others. The models highlight the importance of attempting to identify how many people are actually susceptible to the message (and not those who are immune) as well as what efforts and or how much effort (threshold and cascade models) is necessary to spread the message (Pal et al., 2014; Vidmer et al., 2015;

Zhang & Li, 2014). The models indicate that while the network informs us about the people in one's network who could potentially be affected by a message, we do not know if they are infected. In the case of Yelp, an undirected network, the bilateral relationship is imposed on an individual if they want others to directly see their message(s). This bi-lateral relationship can cause issues of immunity to messages.

With the results of past studies and the viewpoint of the three models mentioned, a look into the actions of members is conducted to identify how many users are known to have been infected by a message. This aspect is particularly important in undirected networks where social structure is reciprocal but may not be preferred by a member.

Votes have been used by sites such as Amazon and Yelp to indicate the quality of a review; however, just as Lu et al. (2012) indicate, the action can be used to track interaction, specifically indicating influence. One of influence's earliest definitions states simply that influence is "*any alteration* of the state of system a over time" (French et al., 1959, p. 151). This includes opinions but is not limited to it. The change can also be in behavior. The act of voting itself is a change in the behavior of an individual over the duration of the time it takes for the user to read the review. Users come to Yelp to read reviews on businesses. A review, or those characteristics inherent in it, may prompt a user to take an action, such as voting, which has altered the user from simply reading a review to taking an action. That action indicates that the review made an impact on the reader. The action of voting embodies the interaction between users, which indicates an impact was made by one user on another user. Thus, members who write reviews that have positive votes can influence other users.

Specific to communities such as Yelp, votes capture the influence of members within the network and influence that may be garnered outside of the network. In Yelp's case, all reviews are public; all members and non-members can view it. A vote of support can be cast by a Yelp member's friends, members who are not friends, and even individuals who use Yelp but choose not to become a member of Yelp's community. On sites such as Instagram, Facebook, and Flickr, members are limited to only seeing and liking posts from individuals who are public or are part of the member's network (those listed as friends). One must be a member of the site's community to like or vote on a post. It is noteworthy to mention that non-members can share a photo or download a photo from Flickr and tracking this action could fit closer with the action of tracking votes; however, there is still the issue that any members that mark photos private cannot be seen by others and therefore cannot be favorited, downloaded or shared.

Yelp's mission is to connect users with businesses. Allowing all users, regardless of their commitment to building Yelp's community, to view and vote on reviews, enables Yelp to most effectively achieve their mission. In the case of Instagram or Facebook, where the mission is centered around building community, it makes more sense to make a requirement of becoming a member. Total votes for a member, therefore, encompass an impact factor that cannot be examined when solely surveying social structure. Yelp's mission and its strategy, of public reviews and voting, is yet another reason why Yelp is indeed a different entity than many other OSNs and is an appropriate basis for a separate study. For these reasons, a positive CIV is used as another indication of future influence.

3.2.2 Measure of social structure

Previously discussed, social structure, or what I have termed for this study, traditional influence score (TIS) or friends count, is an already established construct. The social structure extracted in the first time period of the study is used as the independent variable.

3.2.3 Measure of member engagement

In Chapter 2, it is discussed that the primary goal of ORC sites is to provide content to users. That content comes in the form of reviews. In order to more accurately predict influence in ORCs, the actions of members, their reviews, are taken into consideration. The inclusion of actions of users is a newer measure in influence research; however, many researchers have emphasized the need and benefit of its use (Jindal, 2015; Kim & Han, 2009; Lu et al., 2012; Vidmer et al., 2015; Zhao et al., 2016). Much of the past research pertains to blog sites, but it can be applied to ORCs as actions are taken on these sites as well.

Budalakoti et al. (2011) assert that trust is an important factor in ORCs. Trust is based on a user's reputation. It takes repeated exposure to a stimulus to increase one's reputation and build trust (Budalakoti et al., 2011). The most trustworthy individuals are also seen as the most knowledgeable (Budalakoti et al., 2011). Attempts at measuring knowledge come in the form of counting the number of times a user writes on a certain topic (Dubois & Gaffney, 2014; Jindal, 2015; Li & Du, 2013).

If a user reviews only once, the likelihood of him or her gaining one's trust, displaying enough knowledge, or building a good reputation is low and therefore the user has a low likelihood of engaging and gaining followers. The more exposure a user

receives, the higher the chances are of gaining one's trust, displaying knowledge or expertise, and building a good reputation. These all work to increase one's followership.

For this reason, each members' review counts (action) is examined as well as newly created scores, engagement network score (ENS) and engagement influence score (EIS). ENS is based on the review counts of those in a member's network, excluding the review count of the member. EIS is based on each member's review count and the review counts of those in their network. A more thorough explanation of these measures is discussed in Chapter 4.2.

3.2.4 Measure of overall review quality

Amazon was one of the first companies to use a quality rating mechanism (Wan & Nakayama, 2014). As the number of reviews for a product increases, reading through all the reviews to find the ones of value becomes increasingly time-consuming. Providing users a voting button for each review allows readers a way to easily indicate whether the user believes the review to hold any value. The sum of the votes indicated by each review provides a quick quantitative way for readers to identify whether a review is worth his or her time to read. The votes a review receives is an indicator of the quality of the review (Wan & Nakayama, 2014).

A study by Otterbacher (2011) investigates the construct of quality as measured by votes. Otterbacher (2011) studies votes in investigating communication tactics and prominence for three online communities. In the study, Otterbacher (2011) investigates votes, termed overall review quality to specify any and all votes assigned to a review. For sites that offer only one voting option, only this number is used and for sites that offer multiple voting options, the aggregate is used.

In following with Otterbacher's (2011) research, this study uses any and all votes assigned to a review, aggregating the number of votes for all voting options, as the indicator of review quality. The construct is called overall review quality to denote the allowance of more than one type of vote for the measurement. In addition to examining total votes of members (response action), newly created value network score (VNS) and value influence score (VIS) are examined. VNS is based on the vote counts of those in a member's network, excluding the review count of the member. VIS is based on each member's vote count and the vote counts of those in their network. A more thorough explanation of these measures is discussed in Chapter 4.2.

3.2.5 Measure of elite

Yelp, the online review community used for this study, has created its own user importance measure, called elite. The elite designation is chosen by Yelp through a manual process. Members must apply online, filling out a form, identifying whether they are nominating themselves or someone else, the user profile URL of the nominee, the city, and information on why he or she deserves the status. Yelp then reviews each application and decides which members will receive the designation. The Yelp site states that their council manually goes through applications and profiles to identify the next year's elite (www.yelp.com). The program is yearly, and the designation extends only until the end of the calendar year. Members must apply each year for the status. Yelp does not provide specifics on exactly what it takes to become an elite member, but they do note that overall it has to do with being active and quality. They also note that being an elite member is about trust, so one must provide their real name and photo on their

personal profile to be considered for the status. Lastly, Yelp has a requirement that in order to receive elite status, one must be of legal drinking age.

While the elite construct is specific to Yelp and therefore not used to build any prediction model, the elite measure is evaluated in comparison to the other measures identified previously. The goal is to see how well Yelp's manual measure of importance compares to the current and modified algorithmic measures.

3.2.6 Action processes and network structure

As described in the previous section, social structure and actions are the items being investigated in this study. Figure 2 below illustrates the two-step action process, for the two actions of interest, for a Yelp member, reviews and votes. Figure 3 below illustrates how these actions are related to a member and how a member has a relationship with another member, building a member's network (also known as social structure).

Reviewing is the action most vital to Yelp's purpose of matching users with great businesses. Votes are the response from users to a review.

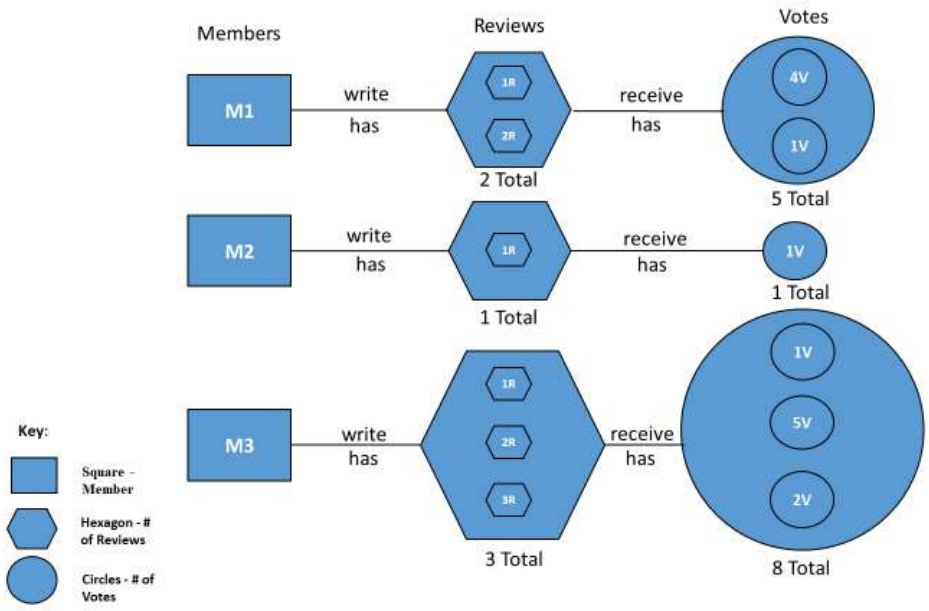


Figure 2: Yelpers 2 step Action Process

Figure 2 shows the actions of three Yelpers and the response actions or reactions given to those three Yelpers by other members. The squares denote the members, the hexagons denote the number of reviews written by each member, the circles denote the number of votes each member has received for their reviews. The response actions are based on the initial actions of members, writing a review. The action word on top of the relationship line in the figure indicates the fundamental process—members write reviews and reviews receive votes. Below that, the relationship specific to each of the three members is shown. In Figure 2, member one has written two total reviews and those two reviews have received five total votes; member two has only one review and that review has one vote; member three has three total reviews and those three reviews have eight total votes. Based on the data provided by Yelp, while the number of votes each review

has received is available, I use the totals for reviews and the total votes for all reviews as the measure of user engagement and user overall review quality respectively. The data provided does not indicate which members voted on a specific review. This distinction makes this model a value network rather than the action network that Lu et al. (2012) propose for identifying prestigious members.

An example graphical illustration of the network in terms of the three characteristics of a member's influence is shown in Figure 3 below.

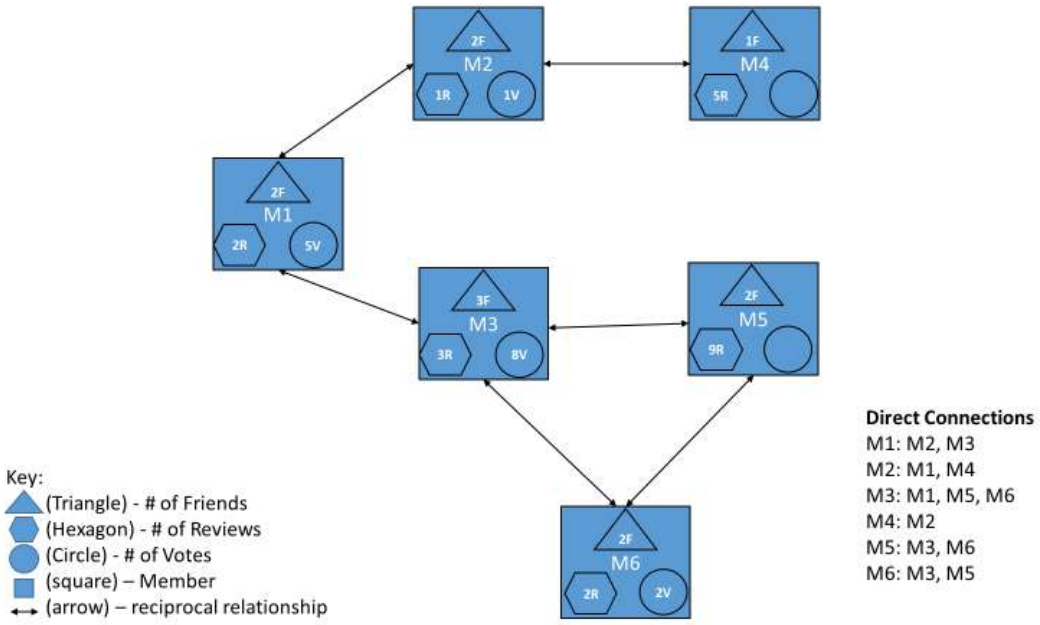


Figure 3: Characteristics of Influence Network Structure

Figure 3 shows a network of six members. The squares denote the members, the triangles within the squares denote the number of direct relationships the member has (also known as degree centrality), the hexagons within the squares denote the number of

total reviews written by each member, the circles within the squares denote the number of votes each member has received for their reviews, and the two-sided arrows denote the reciprocal relationships between the members.

3.3 Scope of Research and Study Development

The previous section explains the measures of the constructs and why they are used in the investigation of predicting influence. This section provides an understanding of the levels of investigation of the independent variables and their relationship with the dependent variables, a change in friends and a change in votes. Table 2 below provides the theories and concepts used as the bases of the investigation.

Table 2: Theoretical Framework

Variable(s)	Concept/Theories	Reference	Change in
Reviews, Friends	Social influence	Anagnostopoulos, Kumar, and Mahdian, 2008; Muruganantham and Gandhi, 2015	Friends, Votes
Reviews, Votes, Friends	Social influence and status characteristics	Oldmeadow, Platow, Foddy, & Anderson, 2003	Friends, Votes
Reviews, Votes, Friends	Expectation states	Berger, Cohen, & Zelditch, 1966	Friends, Votes
Reviews, Votes, Friends	Self-Categorization	Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S., 1987	Friends, Votes
Reviews, Votes, Friends	Homophily	Li and Du, 2013; Lu, Li, & Liao, 2012	Friends, Votes
Friends, Votes	Exchange	Emerson, 1976	Votes

3.3.1 Member's importance

While traditionally, the connections made in a SNS allow researchers to study influence in terms of interactions, studying the member's characteristics and how they lead to an individual having influence may also provide valuable insight. Table 2 indicates the theories used in this dissertation that recognize the individual (the member) as the driving force of influence.

3.3.1.1 Change in friends

An individual's contribution is the initial driving force behind influence. As discussed by Anagnostopoulos et al. (2008), social influence is the use of an individual's actions to motivate behavior in others; specifically, to motivate a similar behavior. In an ORC, the actions used to influence are writing reviews and befriending others. By performing these actions, a member hopes to increase their relationships and therefore their influence.

Social influence and status characteristics theory indicate that there are certain characteristics of an individual that others use to evaluate the individual's ability at a task (Oldmeadow et al., 2003). The number of reviews a member writes (engagement), the number of friends a member has (social structure), and the number of votes a member's reviews received (overall review quality) are characteristics one can use to determine ability level.

Expectation states theory conceives that ability level corresponds with influence. The higher one's perceived ability toward performing a task, the higher the level of influence within the group (Berger et al., 2007). Characteristics in an ORC are a member's number of friends, number of reviews, and number of votes. According to the expectation states theory, these aspects would be employed to examine the ability level of a member to write reviews. If any or all these characteristics are perceived as high, members will look to this member as knowledgeable. Knowledgeable members are followed because of their perceived expertise (Levi & Mokryn, 2014; Li & Du, 2013; Xu et al., 2014). This perceived ability level thereby increases a member's friends count.

As one's ability level increases, the member's relationships grow as well. Members will seek out other members they deem to be of similar status to that of themselves. This concept is termed homophily and discussed in the literature (Li & Du, 2013; Lu et al., 2012). Homophily is the tendency of people to befriend other people that are similar to themselves. Members with a higher number of reviews (friends or votes) will likely be friends with other members who have a higher number of reviews (friends or votes). Members with a lower number of reviews (friends or votes) will tend to be friends with members with a lower number of reviews (friends or votes). Based on homophily, those members who have characteristics perceived as high within the community are likely to seek out other members that have characteristics perceived as high; thereby increasing their friends.

Self-categorization theory posits that as individuals develop their identity, they move toward groups that they share common interests with. While in these groups, the individuals identify ideal members and adopt their beliefs and behaviors. Based on self-categorization, members will move to groups based on their shared interests. As members join an ORC, they will increase their friends count as they find the group(s) they share a common interest with. Also, as members identify the ideal member in the group, they are likely to find that ideal member who has many friends. Members will then seek to adopt the behavior of befriending others.

Social influence, social influence and status characteristics theory, expectation states theory, homophily, self-categorization theory, and past studies lead to the investigation of a member's characteristics as a predictor of change in friends (CIF).

3.3.1.2 Change in votes

Expectation states theory and social influence and status characteristics theory emphasize the characteristics of an individual as an indicator of ability level (Oldmeadow et al., 2003). In the case of an ORC, there are a few characteristics that users may perceive as an indicator of ability. The ability level may be assessed by the number of reviews, votes, and friends of a member. These characteristics drive votes.

Driving votes is the result of the ability to write reviews. The absence of reviews equals the absence of votes. Members that write more reviews increase their exposure and therefore their ability to receive votes. Bakhshi, Kanuparthy, and Shamma (2015) find that reviews that receive feedback, whether in the form of votes or comments, drive writers of those reviews to write more reviews. Bakhshi et al. (2015) also find that lengthier reviews tend to be more popular in Yelp's ORC. They propose this to be the case because more active and dedicated users are the ones writing these reviews and therefore are likely to provide more information about the business, leading to longer reviews which are considered better quality. In other words, they find that review popularity (tested in terms of each of the different types of votes) is a function of review quality. Review quality is a function of activity level. The more activity (increased friendships and reviewing), the more votes are likely to accrue to the writer of the reviews.

Levi and Mokryn (2014) assert that members that write more consistently become better known in ORCs. They declare that familiarity is a requirement of trust (Levi & Mokryn, 2014). Trust comes from the belief that the information being provided is

accurate and truthful. Trust is expressed in the form of voting (Levi & Mokryn, 2014). Writing consistently begets familiarity; familiarity begets trust; trust begets votes.

Exchange theory affirms that an action will only continue if there is a valued return. Psychologists term this reinforcement. Economists term this exchange (Emerson, 1976). In an ORC, an increase in friends and or votes may be seen by reviewers as a valued return.

Studies from Cameron and Trivedi (2013) and Lampe and Johnston (2005) indicate that a higher level of influence (as measured by friends) creates a drive to increase activity. In other words, as members see an increase in their social network and the impact (as measured by feedback) of their influence (reviews in this case), the more the member feels a need to produce more reviews for their network. The more a member produces, the more an opportunity to receive votes. Along similar lines, Zhu et al. (2014) investigate hotel reviews on Yelp and find that influence, defined as reviewer expertise and attractiveness, drives votes. Zhu et al. (2014) find that member attractiveness, measured by the number of friends, impacts votes. Friends drive votes.

Along similar lines, votes drive votes. Feedback, such as votes (Cameron & Trivedi, 2013; Lampe & Johnston, 2005) and or comments (Cameron & Trivedi, 2013), results in increased participative behavior. A study of Slashdot users shows that users who post a comment and receive a number rating or a reply, return more quickly (than others who did not receive any feedback) to post another comment (Lampe & Johnston, 2005). Other experimentation shows that when a member receives messages from other users, the member’s contributions increase. This phenomenon can be explained by people's need for social approval (Cameron & Trivedi, 2013) as well as exchange theory.

As the participative behavior of the member increases, so too does the opportunity to receive feedback.

3.3.2 Network importance

The importance of the network as a measure of influence is well established in literature; however, this section discusses the psychology and theories that explain the importance of examining the network's actions in predicting influence. Self-categorization theory, homophily, and exchange theory shown in Table 2 support using the network's review count, votes count, and friends count in predicting influence.

3.3.2.1 Change in friends

Self-categorization theory conceives that new members of a community gravitate towards groups that they share an interest with and evaluate different people within the group(s), comparing and looking for the ideal person to emulate. This theory is consistent with the concept of homophily. People have relationships with those that are like themselves. Both self-categorization theory and homophily explain the finding of Rosenquist, Murabito, Fowler, and Christakis (2010) and Walther, Van Der Heide, Kim, Westerman, and Tong (2008) that members within a group have shared behaviors. They find that the company you keep is an indicator of the behaviors you exhibit. There are a few sayings depicting these findings, you are the friends you keep and birds of a feather flock together. This means that members of OSNs are judged not only by their abilities but by the abilities of those in their member network.

Because others judge you by the company you keep (Rosenquist et al., 2010; Walther et al., 2008), I investigate the network characteristics of the members in terms of the effect on CIF.

3.3.2.2 Change in votes

Review sites, such as Yelp, rely on two methods to spread messages, the reviews being open to the public and the member created relationships. When a Yelp member logs into their Yelp profile, the first link in the navigation bar, under his or her own profile link, is the friends link. By clicking on the friends link, members are quickly able to access a list of friends and view all the reviews written; therefore, information is effortlessly and quickly spread through a member's network. The more connections a member has, the more the likelihood that the reviews will be seen by others and voted on by others.

A person who has many votes and or reviews and or friends is likely to be more engaged on the site (Bakhshi et al., 2015). A person who starts relationships with members who have many friends, votes, and or reviews may see an upsurge in his or her votes since the member(s) he or she is connected to are engaged and active members. These members who have many friends, votes, and or reviews may be more likely to read the person's reviews and vote.

Based on their study, Bakhshi et al. (2015) suggest that as members of networks get more familiar with the community, they learn from the community. Individuals that have connections to high characteristic members may learn from those members (by reading the reviews) how to write better reviews. This learning from those in the network could lead to an increase in votes. This theme falls in line with the self-categorization theory. As a person tries to imitate their ideal member of a group, the individual will perform the actions of the ideal member.

Exchange theory and the concept of reciprocity may also play a part in the upsurge of votes of a member based on the member's reviews. A member may feel that if they vote on the reviews of their friends, the friends will reciprocate with votes. Connections of a member to members with many votes and or reviews and or friends, would therefore, impact votes.

Chapter 4: Methods

Chapter 3 provides a framework for the contribution of this dissertation. The contribution is not to create and test hypotheses about the relationships discussed above. The relationships between social structure and actions on influence have been well established in previous studies. This dissertation reflects on the relationships in order to frame the investigation as an algorithm to predict influence in ORCs. The computations and methods of which to test, evaluate, and predict influence, specifically for ORCs is the contribution.

4.1 Centrality Measures

In network theory, the users in a network are called nodes and the relationships nodes have with each other are called edges (Lu et al., 2012). Centrality indices are measurements of influence or relative importance of nodes (Pal et al., 2014). Measures of centrality provide a ranking of nodes (Subbian & Melville, 2011).

Degree is the count of nodes a user is directly linked to in a network (Zhang & Li, 2014). Indegree is the count of the number of nodes that provide information that flows into the user (Zhang & Li, 2014). Graphically it is the number of edges that point inwards (Dubois & Gaffney, 2014). Outdegree is the count of the number of nodes that the user provides information out to (Zhang & Li, 2014). Graphically this is the number of edges that point outwards. The difference in the two types of degrees is in the direction of the flow of information (Zhang & Li, 2014). Even simpler, the number of followers of the user is the item to count for outdegree and those the user is following is the item to count for indegree (Momeni & Rabbat, 2015). In a reciprocal network, indegree and outdegree are equivalent.

Closeness measures the proximity of nodes to a user. The measure looks for the shortest path between two pairs of nodes. Nodes in central locations can be productive in communicating messages as there is less of a distance for the message to have to travel. (Kiss & Bichler, 2008).

Betweenness measures how well a node is positioned to facilitate information flow between nodes. It measures the regularity of which a user lies in the shortest path connecting other nodes (Pal et al., 2014; Xu et al., 2014; Zhang & Li, 2014). Lower betweenness means a lower level of connectivity; a high betweenness is desired (Xu et al., 2014). High betweenness means a user lies between many other nodes in the network and therefore the agent controls the interactions between non-adjacent nodes (Kiss & Bichler, 2008).

Eigenvector centrality measures the importance of a user by summing up the centrality of other nodes the user is connected to (Dubois & Gaffney 2015; Kiss & Bichler, 2008). A high score represents a node that has connections with other highly connected nodes (Dubois & Gaffney 2015; Kiss & Bichler, 2008). Simply put, eigenvector centrality takes into consideration all of the nodes connected to a user in order to derive an importance (centrality) value for the user.

Lu et al. (2012) proclaim that the centrality measures used in SNA must match the network flow.

Degree centrality with its focus on direct connections is an appropriate choice for this study. Users in ORCs do not repost messages such as what is allowable in OSNs like Twitter or Facebook. Reviews are often viewed either by a user looking for a business or by a member entering their own profile page and viewing the recent reviews of friends.

For the purpose of this study, degree centrality (friend's network), which may be used interchangeably with the terms traditional influence score (TIS) or friends count, is used for deriving the relative importance of a member. It is used as the basis for creating the modified measures discussed in Chapter 3: VNS, VIS, ENS, and EIS.

4.2 Network Measures

In the modified measures, VNS, VIS, ENS, and EIS, the premise of degree centrality is used to provide scores that weight connections. This is similar to the premise of eigenvector centrality; however, where a user's importance is based on weighted connections in eigenvector, actions (number of reviews written) of the user or response actions given to a user (votes) give weight to the connections for ENS, EIS, VNS, and VIS.

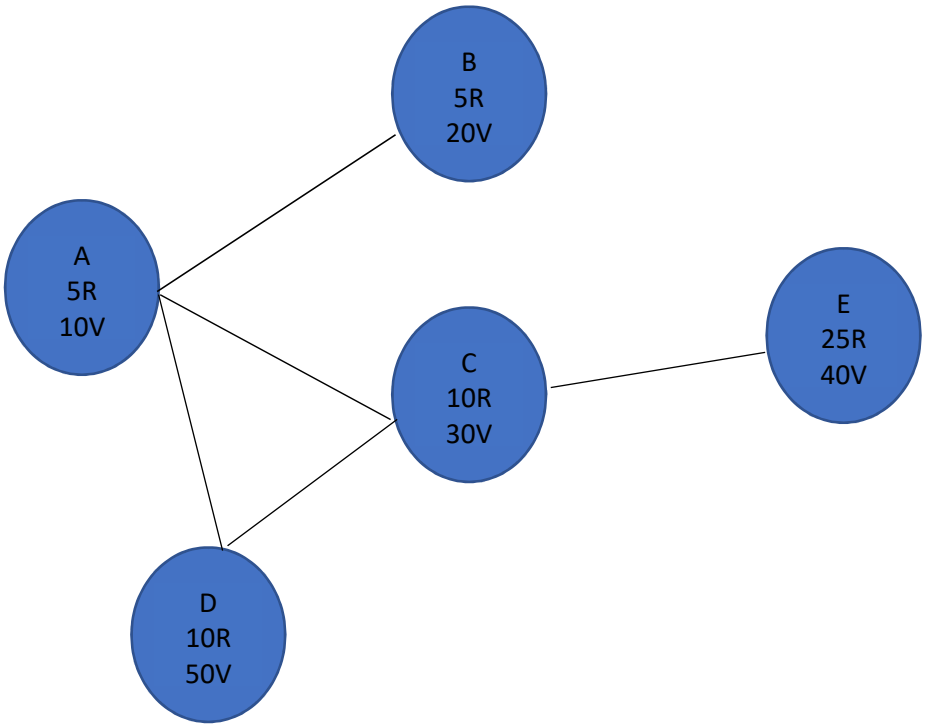
While degree centrality has the benefit of simplicity, both in its premise and its computational aspect, a criticism of the measure is that all vertices are set to be equal (Zafarani, Abbasi, & Liu, 2014). The modified measures preserve the advantages of simplicity of degree centrality for computational purposes while combatting the criticism. These measures do this by weighing vertices using the more modern concept of incorporating actions into examining influence.

4.2.1 Computation

In the TIS (degree centrality) calculation, each node is given equal weight (weight of one). These ones are added up thus producing TIS. TIS is essentially the number of friends of each node. In other words, TIS is the friends count of a node. Because the degree is counting the connections a member has, it is a network measure. The modifications (ENS, EIS, VNS, and VIS) use the premise of looking at the direct

network, but substitutes the count of friends, for a sum of a variable attributed to each friend. The ENS measure sums the review count (number of reviews by a member) for each direct friend of a member. This becomes the member's influence score. The EIS measure sums the review count (number of reviews by a member) for each direct friend of a member and adds it to that member's review count. The VNS measure sums the votes received for all the reviews written by a member for each direct friend of a member. This becomes the member's influence score. VIS sums the votes received for all the reviews written by a member for each direct friend of a member and adds it that member's votes count.

Below is an example of a node (network) graph. The graph is accompanied by examples of the TIS (Table 3), VIS (Table 4), and EIS (Table 5) matrices, scores, and ranks. Figure 4 and the accompanying tables are illustrations of the existing and modified influence measure calculations.



*Key: R = Review and V = Votes

Figure 4: Node Graph

Table 3: TIS (Traditional Influence Score—Degree Centrality)

Degree Matrix					
	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	0	0
C	1	0	0	1	1
D	1	0	1	0	0
E	0	0	1	0	0

Node	Degree (# of Friends)	Rank
A	3	1
B	1	3
C	3	1
D	2	2
E	1	3

Table 4: VIS (Value Influence Score—A Modified Degree Centrality)

Matrix without division by average votes								
	A	B	C	D	E		Sum	Rank
A	10	20	30	50	0		110	2
B	10	20	0	0	0		30	5
C	10	0	30	50	40	=	130	1
D	10	0	30	50	0		90	3
E	0	0	30	0	40		70	4

Table 5: EIS (Engagement Influence Score—A Modified Degree Centrality)

Matrix without division by average reviews								
	A	B	C	D	E		Sum	Rank
A	($M_{1,1}$) 5	5	10	10	0		30	3
B	5	5	0	0	0		10	5
C	5	0	10	10	25	=	50	1
D	5	0	10	10	0		25	4
E	0	0	10	0	25		35	2

The calculation of average votes (or reviews) can be used to deconstruct the modified calculation of TIS. The traditional format of giving each node equal weight of one is equivalent to giving each node a weight based on the average votes (or average reviews). Instead of giving each connection equal weight, each connection can be given a weight that is proportional to the votes (or reviews) of the friend the node is connected to.

Dividing each node's votes (or reviews) by the average weight of the votes (or reviews) provides the relative importance of that node. Thus, the starting node can also be weighted by dividing its votes (or reviews) by the average value. Aggregating the weight of the node and the weights of the friends he or she is connected to becomes the influence score. This illustrates that some nodes should have a higher score/rank because

their connections to certain other nodes are more valuable and they themselves have a value that should be taken into consideration and weighed.

Formally, the formula/calculation for the modification to the TIS is as follows, $Rank_i \mid \sum_j (M_{i,j}/c) \mid = 1/c\ Rank_i \mid \sum_j (M_{i,j}) \mid = Rank_i \mid \sum_j (M_{i,j}) \mid$, where $M_{i,j}$ is a member's number of reviews for EIS and number of votes for VIS; c is average number of reviews for EIS and average number of votes for VIS.

Of note, as well, is the difference between the EIS and ENS (VIS and VNS) calculations. In ENS and VNS, the member's counts are not included in the calculation, only the member's network counts are included. This allows for the separation between a member's contribution and the contribution of the member and the network. Using only the member's network counts could give a different rank than using the member along with the member's network. This is discussed in the next section.

4.3 Levels of Measurement

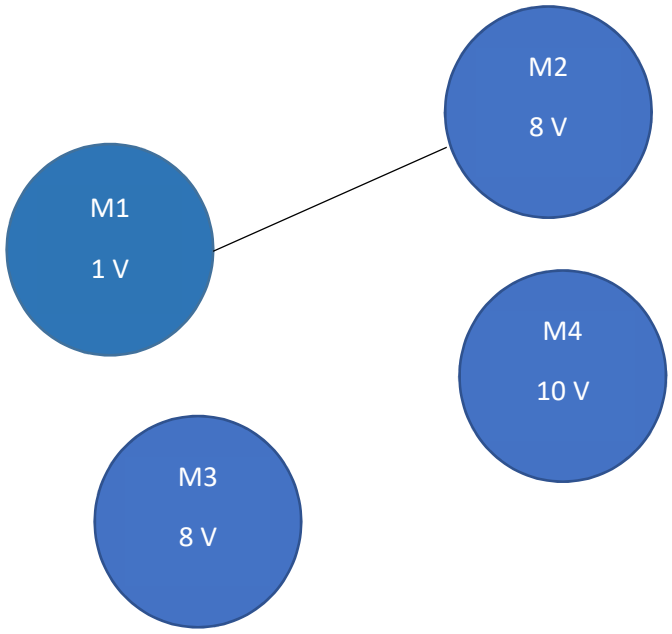
This dissertation investigates predictors of influence by investigating the Yelp member (the individual), their network (the individual's network), and the Yelp member plus the members' network. Investigating and measuring effect in this way, a comparison of measurement methods can be conducted for this type of OSN. This allows an examination of the best predictors of influence. By testing all levels, one can analyze the effects of the individual, the effects of having the connections, and the effects of the combination of the individual and the connections of the individual.

The following two scenarios described next provide a practical understanding of why it is important to investigate measures at the different levels: member, the member's network, and the member plus the member's network. In scenario 1 shown in Figure 5

below, there are four nodes. Two nodes are connected to each other and two nodes have no connections. The nodes that have no connections are called dangling nodes. These dangling nodes have votes. The production of votes tells us that though the nodes do not have any connections, the nodes have influenced others. The production of votes from a member that does not have a network (no connections) indicates the importance of examining measures outside of just a member's social structure.

The network in terms of friends (degree centrality) does not measure the influence a member has on members outside of their specific network. Since degree centrality only considers connections, there is no way to gauge this. The network in terms of votes (network score using votes) provides a way to look beyond connections; however, there is still no measurement of the influence the original member (or starting node) has on members outside of their network. Table 6 below, based on the illustration in Figure 5, indicates that using just the member's score results in the opposite ranking than the member along with the member's network score. The member's score does not recognize that though a member's review may be seen and voted on by anyone, there is a higher likelihood of their review being seen and voted on by those in their network.

In scenario 2 (Table 7 below), member 4 is removed from the community. A comparison of the ranks of the nodes indicates that the rank for the degree centrality is equivalent to the rank of the member along with the member's network. The other measures provide different ranks. We see that the two measures were not equivalent in the ranks in scenario 1 (Table 6). Different scenarios can result in certain measures being equivalent; however, with other scenarios, this may not be the case. Thus, there is a need to compare measurements to see the best fit for the network.



Key: M indicates Member. V indicates Votes.

Figure 5: An ORC Example

Scenario 1: Four nodes have joined the community but only two have started a relationship

Table 6: Data for Scenario 1

	Measurement				Rank			
	Degree (friends)	Member score using votes	Network score using votes	Member score of votes + Network score of votes	Degree (friends)	Member score using votes	Network score using votes	Member score of votes + Network score of votes
Member 1	1	1	8	(1+8) = 9	1	3	1	2
Member 2	1	8	1	(8+1) = 9	1	2	2	2
Member 3	0	8	0	(8+0) = 8	2	2	3	3
Member 4	0	10	0	(10+0) = 10	2	1	3	1

Table 7: Data for Scenario 2 (Removal of M4)

	Measurement				Rank			
	Degree (friends)	Member score using votes	Network score using votes	Member score of votes + Network score of votes	Degree (friends)	Member score using votes	Network score using votes	Member score of votes + Network score of votes
Member 1	1	1	8	(1+8) = 9	1	2	1	1
Member 2	1	8	1	(8+1) = 9	1	1	2	1
Member 3	0	8	0	(8+0) = 8	2	1	3	2

4.4 Data Mining Techniques

Classification techniques are utilized for the examination into of the prediction of influence. Classification is a common procedure used in data mining. Characteristics of an object are examined, and objects are grouped into classes based on these characteristics. Classes are defined with pre-existing examples to train the model. The goal is to build a method for which to classify unclassified objects (Berry & Linoff,

2004). The dichotomous dependent variables change in friends and change in votes are the predicted behavior sought after in this study. Logistic (also known as logit) models are utilized for predictions. In this study, the significant variables found in the logistic models are used to train neural networks, SVM, naive Bayes classifiers and are utilized for prediction.

4.5 Dataset

Data from Yelp's dataset challenge has been obtained for this study. Data from a time cut of July 2014 and December 2015 were pulled from the data challenge. The data from both times consist of five json files: user, business, review, check-in, and tips. The file relevant to this study is the user file. The December 2015 cut enables the study of changes to members' outcomes from July 2014.

4.5.1 User file

The user file contains the date members created an account on Yelp, the total number of votes (broken into three types: useful, funny, and cool) a user has received for their reviews, the total number of reviews written, their name, user identification, list of friends (by user id), total number of fans, average star rating given to businesses, number of compliments (funny, cool, writer, photos, plain, useful, hot, note, more, profile, cute, and list) given by other members, and the years the status of elite was held by each user. The file holds records for 252,898 members. Based on the variables germane to this study, below are the results of the review for the July 2014 time cut of the user file.

Table 8: User File Statistics (July 2014)

Data Variable	Range		Zero Statistics	
	Min	Max	# of 0s	% of 0s
Yelping Since	Oct–04	Jul–14	N/A	N/A
Review Count	0	8,062	1	0.0004%
Fans*	0	1,259	185,912	73.52%
Friends	0	5,000	129,530	51.22%
Votes—Useful (U)	0	31,982	41,860	16.55%
Votes—Funny (F)	0	31,236	99,760	39.45%
Votes—Cool (C)	0	30,996	92,941	36.75%
Total Votes (U + F + C)	0	87,296	34,880	13.79%

As Table 8 shows, users have been yelping since October 2004. The data was cut for the challenge on July 2014. The total number of reviews written by each user ranges from 0 to 8,062, with only one user showing zero reviews written by the cut-off date. The number of fans a user has ranges from 0 to 1,259 with 73.52% of the users showing no fans. The number of friends a user has ranges from 0 to 5,000 with 51.22% of users showing no friends. Useful votes range from 0 to 31,982 with 16.55% of the users receiving no useful votes. Funny votes range from 0 to 31,236 with 39.45% of the users showing no funny votes. Cool votes range from 0 to 30,996 with 36.75% of users receiving no cool votes. The total number of votes for users ranges from 0 to 87,296 with 13.79% of the users receiving no votes.

The difference between fans and friends is that fans can follow a user without having to initiate a request. In order to become a friend, a user must initiate a request and

then the relationship becomes automatically bi-directional. Because the dataset does not provide information on the user id of fans and three-quarters of the users have no fans, the number of friends is used for this investigation.

4.5.2 Experiment set-up

MongoDB, a free distributed database program that stores data in JavaScript Object Notation (JSON) format ("The Most Popular Database", n.d.), is used to convert the JSON files from Yelp to comma-separated value (CSV) files. Microsoft Excel and R are then used to manipulate the data, converting a member's friends from a list of members to lists of friends counts, vote counts, and review counts. The different counts for each member are separately examined as individual predictors of influence. The counts are also used, as discussed in Chapter 4, to create network predictors of influence.

Change in friends (CIF) and change in votes (CIV) are derived from subtracting data of interest at time point A (July 2014) from time point B (December 2015). The variables CIF and CIV are then coded to dichotomous variables. Members with a positive change are coded as a one and members with no change or a negative change are coded as a zero.

The data is then broken into four groups: members with a time frame of a year on Yelp, members with six months on Yelp, members with two months on Yelp and members with one month on Yelp, i.e., months since joining Yelp. The choice to divide the data into groups is based on the findings from Kumar et al. (2006) study of two OSNs. The study finds that member associations have heavy growth in the beginning, suffer a decline, and then show slow and steady growth. With this in mind, this study

investigates different durations of time of Yelp members to see if there are differences between the groups.

Table 9 below provides descriptive statistics for the four groups used in the study. Table 10 below provides descriptive statistics for the Yelp defined elite measure.

Table 9: File Descriptive Statistics

	Time Frame	Dates	# of Members	# of Elite	# of Members with CIF*	% of Members with CIF*	# of Members with CIV*	% of Members with CIV*
Members Time Yelping	1 Year	July 2013–July 2014	27,436	197	9,115	33%	20,024	73%
	6 Months	Feb 2014–July 2014	9,558	35	3,278	34%	7,133	75%
	2 Months	June 2014–July 2014	1,755	1	587	33%	1,364	78%
	1 Month	June 2014	668	0	176	26%	489	73%

*Key: CIF = Positive Change in Friends and CIV = Positive Change in Votes

Table 10: Elite Descriptive Statistics (Totals for 1-year time frame)

Total	CIF*	%	CIV*	%	Members with 0 friends	%	Members with 0 Votes	Members with 0 Reviews
197	197	100%	188	95%	9	5%	0	0

*Key: CIF = Positive Change in Friends and CIV = Positive Change in Votes

Each member's friends count (or TIS), votes count, review count, elite status, ENS, VNS, EIS, VIS, CIF, and CIV are then imported into SPSS, a statistical analysis software. The predictors, friends count (or TIS), votes count, review count, elite status, ENS, EIS, VNS, and VIS are plotted against the dichotomous variables CIF and CIV.

This is done to obtain the area under the curve (AUC) statistic and the receiver operating characteristic (ROC) curve.

The ROC curve illustrates the performance of a variable's probability in prediction at different thresholds. The AUC statistic provides a value for the ROC curve and denotes how well the predictor is doing in distinguishing between the different outcomes, change or no change. When assessing the area under the curve data, a higher number is best. AUC for each predictor is used and compared with others to assess each measure's ability to predict the proper outcome.

SPSS is used to perform logistic regression to examine combinations of variables as predictors (different models) of influence. Logistic regression allows modeling the probability of an event taking place. The goal is to predict a discrete outcome based on a combination of variables.

The outcome variables in this study are CIV and CIF. As noted earlier, these outcomes are dichotomous, either there is a positive change (one) or there is no change or a negative change (zero). Furthermore, this study is focused on using the observed data to directly approximate and to interpret and explain the contributions of each variable, making this a suitable method. Logistic regression is more appropriate than the use of other modeling techniques such as neural networks (NN), naive Bayes, and support vector machine (SVM) algorithms, because these latter methods estimate assumptions and distributions, working within a black box in which the variable contributions are unknown. With that said, NN, naive Bayes, and SVM are however utilized to see the improvement that these classifiers can provide over logistics regression.

The elite measure is removed as a predictor for logistic modeling. While elite is a suitable measure to compare predictors, it is not suitable for modeling. The elite measure is employed by Yelp and the specifics on its creation are unknown; therefore, it is not reproducible or generalizable.

The value influence factor (VIF) is utilized to test for multicollinearity of the independent variables. The VIF is no greater than 3.126 for any independent variable. These low scores indicate multicollinearity is not an issue. Standard rules dictate a cutoff of five or ten (Craney & Surles, 2002). The VIF scores for the independent variables can be found in the collinearity tables in Appendix A. Tables 11 and 12 below provide the descriptive statistics of sample size, means, and standard deviation for each group based on the outcome variable and the predictors.

Of note in the tables and discussion below as well as in the collinearity tables in Appendix A is the R in front of VIS and EIS, where R refers to the respective ranks. VIS and EIS observations are highly skewed, with most values towards the lower end with a few extremely high values. Due to the extreme nonlinearity of predictors VIS and EIS (see Appendix B) resulting in the inability of the variables to be transformed to a parametric function for modeling, the variables are transformed into ranks using SPSS's rank cases feature. VIS and EIS could not be transformed into a log scale due to the prevalence of a large number of values that have zeros in the data set.

For new observations, VIS and EIS values may be mapped into the historical ranks of VIS and EIS. Periodically, the mapping table would need to be updated. The transformation is indicated by an R in front of the variable names for logistics regression output.

Table 11: Description of Data Set for Logistic Regression—CIV

Group A: July 2013–July 2014 (13 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Votes?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	20,024	73%	10.66	62.54	8.85	28.29	1.94	12.00	12470.69	7646.02	12218.94	7490.62
No	7,412	27%	1.72	12.84	2.08	2.34	0.76	3.10	17089.54	7318.49	17769.66	7344.85
Summary	27,436	100%	8.25	53.99	7.02	24.38	1.62	10.39	13718.50	7832.12	13718.50	7848.47

Group B: Feb 2014–July 2014 (6 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Votes?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	7,133	75%	6.21	39.06	6.08	12.08	1.48	6.22	4420.81	2668.28	4310.09	2623.60
No	2,425	25%	1.00	2.11	1.82	1.49	0.63	2.39	5834.57	2498.25	6160.24	2540.64
Summary	9,558	100%	4.89	33.83	5.00	10.63	1.26	5.52	4779.50	2697.15	4779.50	2724.35

Group C: June 2014–June 2014 (2 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Votes?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	1,364	78%	4.68	43.89	4.99	11.15	0.87	3.65	830.01	484.54	798.75	485.94
No	391	22%	0.60	1.23	1.77	1.33	0.36	1.42	1045.41	412.70	1154.48	438.74
Summary	1,755	100%	3.77	38.74	4.28	9.94	0.76	3.29	878.00	477.86	878.00	798.22

Group D: July 2014 (1 month)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Votes?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	489	73%	4.20	57.12	3.65	5.79	0.67	4.29	318.54	176.83	299.13	183.97
No	179	27%	0.54	1.25	1.50	1.03	0.04	0.35	378.09	143.31	431.12	135.59
Summary	668	100%	3.22	48.89	3.07	5.07	0.50	3.69	334.50	170.46	334.50	181.91

Key:
M = Mean
SD = Standard Deviation

Table 12: Description of Data Set for Logistic Regression—CIF

Group A: July 2013–July 2014 (13 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Friends?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	9,115	33%	16.32	82.94	11.22	38.36	4.29	16.77	9204.33	7670.42	9121.87	7566.77
No	18,321	67%	4.23	29.92	4.93	12.04	0.30	4.07	15964.37	6885.46	16005.40	6929.59
Summary	27,436	100%	8.25	53.99	7.02	24.38	1.62	10.39	13718.50	7832.12	13718.50	7848.47

Group B: Feb 2014–July 2014 (6 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Friends?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	3,278	34%	9.42	56.43	7.45	15.38	3.27	9.01	3370.64	2738.42	3285.16	2676.23
No	6,280	66%	2.53	7.99	3.72	6.61	0.21	0.88	5514.89	2362.55	5559.51	2405.12
Summary	9,558	100%	4.89	33.83	5.00	10.63	1.26	5.52	4779.50	2697.15	4779.50	2724.35

Group C: June 2014–June 2014 (2 months)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Friends?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	587	33%	7.82	65.06	6.34	15.78	1.99	5.37	640.42	508.52	618.30	503.74
No	1168	67%	1.74	10.82	3.24	4.51	0.14	0.78	997.40	413.16	1008.52	441.23
Summary	1,755	100%	3.77	38.74	4.28	9.94	0.76	3.29	878.00	477.86	878.00	798.22

Group D: July 2014 (1 month)			Votes count		Review count		Friends count		Rank of VIS (RVIS)		Rank of EIS (REIS)	
Change in Friends?	Total Sample (N)	Sample %	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	176	26%	9.77	95.05	4.34	7.73	1.81	7.02	267.21	196.11	254.23	196.42
No	492	74%	0.88	2.43	2.62	3.59	0.04	0.26	358.57	153.49	363.22	167.53
Summary	668	100%	3.22	48.89	3.07	5.07	0.50	3.69	334.50	170.46	334.50	181.91

Key:
M = Mean
SD = Standard Deviation

Group A (one year membership since joining the ORC) consists of 27,436 members, group B (six months membership since joining the ORC) consists of 9,558 members, group C (two months membership since joining the ORC) consists of 1,755 members, and group D (one month membership since joining the ORC) consists of 668 members. The predictors used are continuous variables: votes count, review count, friends count (individual member characteristics), value influence score (member's individual votes plus their network's votes), and engagement influence score (member's reviews plus their network's reviews). Value influence score and engagement influence score are transformed into ranks as described earlier. As noted above, the outcome variables are binary variables, CIF (1 = yes and 0 = no) and CIV (1 = yes and 0 = no).

The model summary, classification tables, and variables in the equation tables, resulting from the logistic regression procedure in SPSS, are used to define and assess the models. The Beta coefficients, P-values, and Wald chi-square statistic from the variables in the equation table are used to define the model. Wald chi-square and P-values are used to test for the significance of the classifiers. The sign for the Beta coefficient is used to determine whether the classifier is negatively or positively related to the outcome variable. The Beta coefficient for each classifier is used to ascertain the log of the odds of a change in the outcome variable. The Cox and Snell R^2 and Ngagelkerke R^2 from the model summary along with the sensitivity, specificity, false positive, false negative, and overall percentage correct statistics derived from the classification tables are used to assess the fit and robustness of the logistic model. The probabilities of the logistic models are extracted and tested against the outcome variables using AUC-ROC. A comparison of the AUC for the model probability of the outcome variables and the AUC

best predictor of the outcome variables is used as an additional test for the robustness of the model. Higher numbers for AUC model probability indicate the model is better at predicting the outcomes. Finally, a comparison of probabilities of AUC for neural networks (NN), AUC for naive Bayes, and AUC for support vector machine (SVM) is done using SPSS and R and compared to the AUC logistic regression model probability. Higher numbers for AUC logistic model probability indicate the model is better at predicting the outcomes.

Chapter 5: Results and Analysis

Two separate investigations are conducted to study the prediction of influence. First, an investigation is conducted into determinants of predicting influence, based on the differences associated with ORCs. Influence for the Yelp dataset is operationalized as a change in friends and a change in votes from one time period to another. Second, an investigation is conducted to generate models that predict influence. In comparing the determinants, three different levels are examined: the member, the member's network, and the member along with the member's network. Based on the results from the first investigation, the member and the member along with the member's network levels are examined in the second investigation. In both investigations, members of the Yelp community are separated into four groups, based on their time since joining the site, to see if there are differences between the groups. The receiver operating characteristic (ROC) graph with its accompanying area under the curve (AUC) data is chosen for analysis in both investigations. This metric is used to assess the performance of classifiers.

5.1 Study 1: Determinants of Future Influence

The ROC curve shows the false positive rate against the true positive rate (Provost & Fawcett, 2013). A point that is more northwest of another point on the graph is considered better. Points near to the x-axis have low true positive rates and low false-positive rates. Points on the upper right side have high false-positive rates and high true positive rates (Provost & Fawcett, 2013). AUC is a summary statistic that provides a single number that can be used for the analysis of the classifiers. The higher the AUC, the better the variable is at predicting (Provost & Fawcett, 2013).

The AUC ROC results for the study were obtained from SPSS version 23. The analysis of the results is below. The individual plots and data tables can be found in Appendix C and D.

5.1.1. Change in friends

Table 13 below provides the AUC values for CIF by predictor and level. Friends count, votes, and reviews are values from the July 2014 cut of data. AUC values are determined based on a positive change in friends from the difference between the December 2015 cut of data and the July 2014 cut of data. The largest AUC value is shaded grey in the table.

Table 13: AUC Values for Change in Friends by Predictor and Level

Change in Friends				
		Friends Count	Votes	Reviews
Group A: 1 year	Member	N/A	0.616	0.608
	Member's Network	0.743	0.732	0.736
	Member + Network	0.743	0.746	0.751

Change in Friends				
		Friends Count	Votes	Reviews
Group B: 6 months	Member	N/A	0.603	0.604
	Member's Network	0.721	0.709	0.716
	Member + Network	0.721	0.724	0.738

Change in Friends				
		Friends Count	Votes	Reviews
Group C: 2 months	Member	N/A	0.605	0.608
	Member's Network	0.685	0.678	0.683
	Member + Network	0.685	0.703	0.722

Change in Friends				
		Friends Count	Votes	Reviews
Group D: 1 month	Member	N/A	0.570	0.599
	Member's Network	0.625	0.614	0.624
	Member + Network	0.625	0.637	0.663

Key: Light grey highlights indicate the highest AUC.

In investigating the levels of the determinants, the AUC data between levels of the determinants and the dependent variable future influence are compared. The absence of a measure at the member level for friends count is denoted with an N/A located at the member level of each group. Friends count is a measure of connections and therefore represents a network measure. The absence of a score at the member level thus results in the same AUC value for a member's network and a member along with a member's network.

In examining CIF as a measure of future influence, the member's network is the same as the member along with the member's network. For CIF, the member along with the member's network is the best level of measurement for all groups, no matter the predictor: friends count, votes, and reviews.

In investigating the best determinant of future influence, the AUC ROC data between the proposed determinants and the dependent variable future influence are compared. The best predictor of influence, measured as CIF, for all groups is the reviews of the member along with the reviews of the member's network.

When examining the CIF for Yelp members, EIS (review count of member along with review counts of member's network) is the best predictor and VIS (votes count of member along with vote counts of member's network) is the second-best predictor of CIF for all groups of members, members who joined Yelp one year, six months, two months, and one month prior.

Based on the results, evaluating the network's contributions in conjunction with a member's contribution is the best predictor of CIF. In all instances, the number of reviews of a member along with the number of reviews of a member's network is the best

predictor of influence. In line with the definition of social influence, the individual is the driving force for predicting influence. A member must first take an action (befriending another and or writing a review) in order to influence others. From these actions originate relationships. As expectation states theory, social influence and status characteristics theory, and the concept of homophily reveal, people in groups observe and make assumptions about a person based on certain characteristics and perceived abilities. As a member's characteristics are attractive to a person based on their abilities (reviewing and befriending others) being perceived as comparable or higher, others will want to befriend the member. In addition, self-categorization theory indicates that not only are members observed by their abilities, but members are also observed by the abilities of those around them. Members evaluate themselves, those within groups within the community, and the community. As a member's identity is formed, members will evaluate groups and move toward groups that coincide with their identity. Thus, the member, their actions, and the actions of those in the member's network work to increase a member's relationships.

The elite AUC value has been left out of Table 13 due to its inability to be fit to a level; however, a table with this data can be found in Appendix E. The AUC value for the elite measure ranged from 0.500–0.510, depending on the group. These values indicate that the elite measure has little to no predictive power in determining CIF.

5.1.2 Change in votes

Table 14 below provides the AUC values for CIV by predictor and level. Friends count, votes, and reviews are values from the July 2014 cut of data. AUC values are determined based on a positive change in friends from the difference between the December 2015 cut of data and the July 2014 cut of data. The largest AUC value is highlighted in the table.

Table 14: AUC Values for Change in Votes by Predictor and Level

Change in Votes				
		Friends Count	Votes	Reviews
Group A: 1 year	Member	N/A	0.705	0.754
	Member's Network	0.558	0.561	0.562
	Member + Network	0.558	0.668	0.702

Change in Votes				
		Friends Count	Votes	Reviews
Group B: 6 months	Member	N/A	0.675	0.740
	Member's Network	0.550	0.551	0.553
	Member + Network	0.550	0.648	0.694

Change in Votes				
		Friends Count	Votes	Reviews
Group C: 2 months	Member	N/A	0.626	0.723
	Member's Network	0.549	0.550	0.551
	Member + Network	0.549	0.623	0.703

Change in Votes				
		Friends Count	Votes	Reviews
Group D: 1 month	Member	N/A	0.570	0.684
	Member's Network	0.547	0.542	0.547
	Member + Network	0.547	0.589	0.698

Key: Light grey highlights indicate the highest AUC.

In investigating the levels of the determinants, the AUC data between levels of the determinants and the dependent variable future influence are compared. The absence of a measure at the member level for friends count is denoted with an N/A located at the member level of each group. Friends count is a measure of connections and therefore represents a network measure. The absence of a score at the member level thus results in the same AUC value for the member's network and the member along with the member's network. In the case of reviews and votes for groups A–C, the member is the best level predictor of CIV. In the case of reviews and votes for group D, the member along with the member's network is the best level predictor for CIV. For group D, the member along with the member's network is the best level predictor for both the CIF and CIV.

In investigating the best determinant of future influence, the AUC ROC data between the proposed determinants and the dependent variable future influence are compared. The review count of a member is the best predictor of influence for groups in which the member has been on Yelp for more than a month. In the case of the group that has only been on Yelp for a month, reviews of the member along with the reviews of the member's network is the best predictor of CIV.

When examining the CIV for Yelp members, review count is the best predictor of CIV for members who have been Yelping for at most one year, six months, and two months; however, when looking at members who started one month prior, EIS (review count of member along with review counts of member's network) is the best predictor. Number of votes is the second-best predictor of CIV for members who have been Yelping for at most one year, six months, and two months. For members who started one

month prior, VIS (votes count of the member along with the votes count of the member's network) is the second-best predictor.

Based on the results, examining the member's contribution in terms of an action and response action is the best and second-best single predictor of CIV for groups A–C. Evaluating the network's contributions in conjunction with a member's contribution is the best and second-best single predictor of CIV for group D. No matter the group, examining an action or response action of a member yields better AUC values than examining just the network in terms of friends. As stated in the previous section, a member's ability or abilities drive whether others will be influenced by a member. A member must first take the action of writing a review before the member can receive votes. As the number of reviews for a member increases, the more reviews are available to receive votes. Also, as the number of friends, reviews, and votes of a member increases, a member's ability is perceived as high. Other members and non-members will believe the member to be engaged and knowledgeable and therefore, others are more likely to read the member's reviews. The more people reading the reviews, the more opportunity there is for the member to receive votes. Additionally, exchange theory posits that actions are based on the belief of receiving a valued return. As a member receives feedback in terms of increased relationships and votes (a valued return), the member will continue their actions. In other words, the cycle of reviewing and receiving votes will continue if there is a valued return (increase in friends and increase in votes).

An interesting finding is that the member's network is not predictive of change in votes (maximum AUC value 0.56). This lends credence to the dominance of exchange

theory. This was not the case with the prediction of change in friends, where the network of friends also drives change in friends.

As previously noted, the elite AUC value has been left out of the table due to its inability to be fit to a level. A table with this data can be found in Appendix E. The AUC value for the elite measure ranged from 0.500–0.505, depending on the group. These values indicate that the elite measure has little to no predictive power to determine CIV.

A more comprehensive discussion of the results in Tables 13 and 14 can be found in Chapter 6.

5.2 Study 2: Models of future influence

Logistic models are fitted to the data to create models that can predict two types of influence, CIF and CIV, at different points in time as a Yelp member. The foundation of the logistic model is the logit or the natural logarithm of an odds ratio (Peng, Lee, & Ingersoll, 2002).

The formula for the logistic model is shown below:

$$\text{Logit}(Y) = \text{natural log (odds)} = \ln(\pi / (1 - \pi)) = \alpha + \beta X .$$

Logit is the regression coefficient β. After parsing out the antilog, a new equation is derived that provides the probability of the event happening or not happening (Peng et al., 2002).

$$\Pi = \text{Probability} = e^{(\alpha + \beta X)} / (1 + e^{\alpha + \beta X})$$

The probability of the event is π. The regression coefficient is β. The natural logarithm base is e. The Y-intercept is the α. X is the predictor variable, continuous or categorical. Y is the outcome variable.

The logistic regression analysis is carried out by the Logistic procedure in SPSS version 23 in the Windows 2019 environment. The forward conditional method is used for the logistic procedure. Forward conditional is a process in which a model is fitted through an automatic procedure where variables are tested for statistical significance based on a criterion, "…at a given step, [forward conditional] selects the independent variable that maximizes the squared partial correlation efficient with the dependent variable, given the variables already selected" (Bendel & Afifi, 1977, p. 46).

Each group in the study is randomly split into a training set and a test set; seventy percent of each group is used to train the model while thirty percent of each group is used to test the predictive capability of the model.

The Wald chi-square statistic is used to evaluate the statistical significance of the regression coefficients.

Cox and Snell R^2 and Nagelkerke R^2 are descriptive statistics used to assess the fit of the logistic model. While R^2 in linear regression is clearly defined, no equivalent variance explanation exists in logistic regression; however, Cox and Snell R^2 and Nagelkerke R^2 are indices that can be used as supplements to other measurements of the model, such as the statistical tests of individual predictors discussed above (Peng et al., 2002).

SPSS provides a classification table that substantiates the predicted probabilities of the logistic regression. The rows indicate the observed outcomes and are broken down to show each step of the forward conditional logistic regression process. The columns indicate the predicted outcomes, based on the cutoff point of 0.5 set by SPSS. The columns are broken into two types of cases: selected cases (the training set) and the

unselected cases (the test set). The results from the classification tables along with the model summaries and variables in the equation data can be found in Appendix F and G. Based on similar percentages of outcomes for the training and test sets, the models are deemed to be robust.

A summary of the sensitivity, specificity, false positive, false negative, and overall percentage correct statistics derived from the classification tables for the models validate the predicted probabilities. Sensitivity measures the proportion of members correctly classified as having a change. Specificity measures the proportion of members correctly classified as having no change. The false-positive rate measures the proportion of misclassifications of a CIF/CIV. The false-negative rate measures the proportion of misclassifications of no CIF/CIV. The cut off was set by SPSS at 0.5 for all models.

In addition, the probabilities generated in the logistic regression step are used to generate AUC statistics. This new AUC data provides insight into the predictive power of a combination of variables which contrasts with the AUC data provided in on earlier where each predictor was examined individually. The purpose of comparing the two sets of AUC data is to see how well the model is doing in comparison to the best individual predictor. ROC curves and the AUC results from SPSS can be found in Appendix F and G.

Lastly, neural networks, naive Bayes, and SVM model probabilities are generated and used to produce ROC curves and corresponding AUC values for each machine learning method. The AUCs of these methods along with the AUCs of the logistic model probability are compared and utilized to see the improvement that these classifiers can provide over logistics regression.

The results of the logistic models for CIF and CIV are shown and discussed below.

5.2.1 Change in friends

According to Table 15 below, the CIF model for group A, the log of the odds of a member having a CIF is positively related to review count ($p < .05$) and friends count ($p < .05$) and review count ($p < .05$) and negatively related to RVIS ($p < .05$) and REIS ($p < .05$). RVIS and REIS, which are ranks, are inversely related to the VIS and EIS values respectively. The higher the review count, the more likely the member will have a higher CIF. The higher the friends count, the more likely the member will have a higher CIF. The lower the number for RVIS (or higher VIS rank), the more likely the member will have a higher CIF. The lower the number for REIS (or higher EIS rank), the more likely the member will have a higher CIF. According to the CIF model for group B, the log of the odds of a member having a CIF is positively related to review count ($p < .05$) and friends count ($p < .05$) and review count ($p < .05$) and negatively related to RVIS ($p < .05$) and REIS ($p < .05$). The higher the review count, the more likely the member will have a higher CIF. The higher the friends count, the more likely the member will have a higher CIF. The lower the number for RVIS, the more likely the member will have a higher CIF. The lower the number for REIS, the more likely the member will have a higher CIF. According to the CIF model for group C, the log of the odds of a member having a CIF is positively related to friends count ($p < .05$) and negatively related to REIS ($p < .05$). The higher the friends count, the more likely the member will have a higher CIF. The lower the number for REIS, the more likely the member will have a higher CIF. According to the CIF model for group D, the log of the odds of a member

having a higher CIF is positively related to friends count ($p < .05$) and negatively related to REIS ($p < .05$). The higher the friends count, the more likely the member will have a higher CIF. The lower the number for REIS, the more likely the member will have a higher CIF.

Table 15: Change in Friends Models

Group	**Change in Friends Model**
A: July 2013–July 2014	Predicted logit of (Change in Friends) =.01525177309 + (.2603271016)*FriendsCount + (-.0000382090)*RVIS + (-.0000531895)*REIS
B: Feb 2014–July 2014	Predicted logit of (Change in Friends) = -.3823552653 + (.0097098204)* ReviewCount + (.5334898008)*FriendsCount + (-.0000529210)*RVIS + (-.0000960457)*REIS
C: June–July 2014	Predicted logit of (Change in Friends) = .0615672203 + (.5009021471)*FriendsCount + (-.0012498318)*REIS
D: July 2014	Predicted logit of (Change in Friends) = -.6895643248 + (.9251009198)*FriendsCount + (-.0018286586)*REIS

For all groups, the number of friends of a member and the number of reviews written by members (REIS or rank of EIS) are predictors of influence. These results substantiate the use of actions in combination with social structure as essential in best predicting a change in friends. Theories of social influence, social influence and status characteristics, and expectation states emphasize the importance of action as a precondition to influence. In addition to these theories, the concept of homophily and theory of self-categorization indicate that a member's actions (number of reviews, number of friends befriended, and number of votes) are assessed by others in the ORC. The greater the perceived ability of a member based on his or her actions, the more attractive members become in terms of relationships. Thus, the result of actions is an

increase in friendships. A more comprehensive discussion of the findings from Table 15 can be found in Chapter 6.

The Wald chi-square statistic in Table 16 below is used to evaluate the statistical significance of the regression coefficients (βs) for CIF.

Table 16: Change in Friends Logistic Regression Analysis by SPSS

Group A: July 2013–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	0.1525177309	0.048	10.094	1	0.001	1.16476311
Friends Count	0.2603271016	0.013	386.927	1	0.000	1.29735438
RVIS	-0.0000382090	0.000	69.301	1	0.000	0.99996618
REIS	-0.0000531895	0.000	168.249	1	0.000	0.99994681
Note. Cox and Snell R2 = .205. Ngagelkerke R2 (Max rescaled) R2 = .285						
Group B: Feb 2014–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	-0.3823552653	0.108	12.436	1	0.000	0.68225263
Review Count	0.0097098204	0.005	4.371	1	0.037	1.00975711
Friends Count	0.5334898008	0.038	202.372	1	0.000	1.70487160
RVIS	-0.0000529210	0.000	8.915	1	0.003	0.99994708
REIS	-0.0000960457	0.000	25.19	1	0.000	0.99994553
Note. Cox and Snell R2 = .210. Ngagelkerke R2 (Max rescaled) R2 = .290						
Group C: June 2014–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	0.0615672203	0.168	0.134	1	0.714	1.06350198
Friends Count	0.5009021471	0.900	31.223	1	0.000	1.65020933
REIS	-0.0012498318	0.000	53.680	1	0.000	0.99875095
Note. Cox and Snell R2 = .200. Ngagelkerke R2 (Max rescaled) R2 = .279						
Group D: July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	-0.6895643248	0.249	7.7	1	0.006	0.50179464
Friends Count	0.9251009198	0.234	15.564	1	0.000	2.52212278
REIS	0.0018286586	0.001	7.388	1	0.007	0.99817301
Note. Cox and Snell R2 = .141. Ngagelkerke R2 (Max rescaled) R2 = .207						

For group A, friends count, RVIS, and REIS are significant predictors of CIF ($p < .05$). The test of the intercept in this data set ($p > .05$) suggests that an alternative model without the intercept could be applied to the data set. Group B shows review count, friends count, RVIS, and REIS are significant predictors of CIF ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set. Group C shows both friends count and REIS are significant predictors of CIF ($p < .05$). The test of the intercept in this data set ($p > .05$) suggests that an alternative model without the intercept could be applied to the data set. Group D shows both friends count and REIS are significant predictors of CIF ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set.

Table 17 below provides a summary of the sensitivity, specificity, false positive, false negative, and overall percentage correct statistics derived from the classification tables for the models.

Table 17: Classification Table Derived Statistics—CIF

	Group	Sensitivity	Specificity	False Positive	False Negative	Overall Percentage
Time Yelping	A: 1 Year	46.0%	92.7%	24.1%	22.5%	77.2%
	B: 6 Months	42.6%	93.0%	24.0%	24.4%	75.7%
	C: 2 Months	38.2%	94.3%	22.8%	24.8%	75.6%
	D: 1 Month	25.0%	97.6%	21.4%	21.6%	78.4%

Based on Table 17, with the cut off of the predicted probability set at 0.5, the prediction for no change in friends (CIF) for groups A, B, C, and D is more accurate than for those who did have a change in friends. This was supported by the sensitivity and specificity of all groups. For group A, the sensitivity is 46% in comparison to that of

specificity at 92.7%. The false-positive rate is 24.1% while the false-negative rate is 22.5%. The overall prediction rate is 77.2%. For group B, the sensitivity is 42.6% in comparison to the specificity at 93%. The false-positive and negative rates are just above 24%. The overall prediction rate is 77.2%. For group C the sensitivity is 38.2% compared to that of specificity at 94.3%. The false-positive rate is 22.8% while the false-negative rate is 24.8%. The overall prediction rate is 75.6%. For group D, the sensitivity is 25% in comparison to the specificity at 97.6%. The false-positive and negative rates are just above 21%. The overall prediction rate is 78.4%.

While the results indicate that the prediction for no change in friends (specificity) for all groups is more accurate than for those who did have a change in friends (sensitivity), the results also indicate an increase over time in the accuracy of identifying cases of a change in friends (sensitivity). At one month, the sensitivity is 25.0%. At two months, the sensitivity is 38.2%. At six months, the sensitivity is 42.6%. At one year, the sensitivity is 46.0%. The longer the duration of time, the higher the sensitivity. This increase in sensitivity implies that more factors play a role in accurately classifying cases of change in friends. Also, of note is the downward trend of the specificity over time. This is as expected, sensitivity and specificity have an inverse relationship. As sensitivity goes up, specificity goes down (and vice versa).

Table 18 below compares the model's predictive power to that of the best individual predictor for CIF.

Table 18: AUC Comparison—CIF

Group	AUC (Model Probability of CIF)	AUC (Best Individual Predictor of CIF)
A: July 2013–July 2014	0.765	0.751 (EIS)
B: Feb 2014–July 2014	0.751	0.738 (EIS)
C: June–July 2014	0.731	0.722 (EIS)
D: July 2014	0.671	0.663 (EIS)

Based on the results in Table 18, the probabilities from the logistic model as the predictor of CIF is slightly better at predicting CIF for all the groups A, B, C, and D in comparison to that of the best predictors of CIF using the original classifiers. ROC curves and the AUC data from SPSS can be found in Appendix F.

While EIS (reviews of a member along with reviews of a member's friends) can solely predict a change in friends for each of the groups, the models, which also utilize EIS, generate a slightly more accurate prediction. The findings of EIS as a significant predictor in all of the models (Tables 15 and 16) coincides with the findings of EIS as being the significant individual predictor for all of the groups (Tables 13 and 18).

Table 19 below compares the AUC probabilities of the different machine learning algorithms used to generate models from the predictors of CIF as found from the logit model. The best model AUC values are depicted shaded in the table. Note that RVIS and REIS (R indicating ranks) are used in generating the logit models, while VIS and EIS (the scores, not the ranks) are used in generating the NN, naive Bayes, and SVM models. The inability of the variables VIS and EIS to be transformed to a parametric function for modeling caused the necessity to transform the variables to ranks for logistic modeling. However, the NN, naive Bayes, and SVM algorithms can handle variables that are non-linear.

Table 19: AUC Model Comparison—CIF

Group	AUC (LOGIT Model Probability vs CIF)	AUC (NN Model Probability vs CIF	AUC (naive Bayes Model Probability vs CIF)	AUC (SVM Model Probability vs CIF)
A: July 2013–July 2014	0.765	0.765	0.758	0.762
B: Feb 2014–July 2014	0.751	0.751	0.737	0.720
C: June–July 2014	0.731	0.731	0.725	0.729
D: July 2014	0.671	0.678	0.673	0.675

For groups A, B, and C, the AUC for the logistic model probability and neural networks model probability has the same value. These two models are the highest of the models in groups A, B, and C. For group D, the AUC for the neural networks model probability produces the highest AUC.

While the best model probability is NN for group D, the number does not have a wide distribution from the logit model; therefore, confirming the robustness of the

logistic model used to generate models that predict influence. Non-linear machine learning classifiers, neural networks, naive Bayes, and SVM do not significantly outperform the linear logit model.

5.2.2 Change in votes

According to Table 20 below, the CIV model for group A, the log of the odds of a member having a CIV is positively related to votes ($p < .05$) and review count ($p < .05$) and negatively related to RVIS ($p < .05$) and REIS ($p < .05$). RVIS and REIS, which are ranks, are inversely related to the VIS and EIS values respectively. In other words, the higher the votes, the more likely the member will have a higher CIV. The higher the review count, the more likely the member will have a higher CIV. The lower the number for RVIS (or higher VIS rank), the more likely the member will have a higher CIV. The lower the number for REIS (or higher EIS rank), the more likely the member will have a higher CIV. According to the CIV model for group B, the log of the odds of a member having a CIV is positively related to votes ($p < .05$) and review count ($p < .05$) and negatively related to REIS ($p < .05$). The higher the votes, the more likely the member will have a higher CIV. The higher the review count, the more likely the member will have a higher CIV. The lower the number for REIS, the more likely the member will have a higher CIV. According to the CIV model for group C, the log of the odds of a member having a CIV is positively related to review count ($p < .05$) and negatively related to REIS ($p < .05$). The higher the review count, the more likely the member will have a higher CIV. The lower the number for REIS, the more likely the member will have a higher CIV. According to the CIV model for group D, the log of the odds of a

member having a CIV is negatively related to REIS ($p < .05$). The lower the number for REIS, the more likely the member will have a higher CIV.

Table 20: Change in Votes Models

Group	**Change in Votes Model**
A: July 2013–July 2014	Predicted logit of (Change in Votes) =.3602113295 + (.0348592645)*Votes + (.2948504568)*ReviewCount + (-.0000154248)*RVIS + (-.0000108280)*REIS
B: Feb 2014–July 2014	Predicted logit of (Change in Votes) =.3567740047 + (.0832226264)*Votes + (.3454507659)*ReviewCount + (-.0000619633)*REIS
C: June–July 2014	Predicted logit of (Change in Votes) = .9596715970 + (.3410057297)*ReviewCount + (-.0005574559)*REIS
D: July 2014	Predicted logit of (Change in Votes) = 3.0990953153 + (-.0055873036)*REIS

For all groups, the number of reviews written by a member (review count and or REIS) is a predictor of change in votes. There can be no votes without reviews. The more reviews that are written by a member, the more exposure of a member; therefore, the more opportunity for a member to receive votes. Writing reviews is the pre-condition for predicting influence (as measured here by a change in votes). The models from Table 15 and Table 20 show that whether the future influence is measured as a change in friends or a change in votes, the action of reviewing is a vital component. This finding supports the use of actions in measuring influence and coincides with the definition of social influence. A more comprehensive discussion of the findings from Table 20 can be found in Chapter 6.

The Wald chi-square statistic in Table 21 below is used to evaluate the statistical significance of the regression coefficients (βs) for CIV.

Table 21: Change in Votes Logistic Regression Analysis by SPSS

Group A: July 2013–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	0.3602113295	0.069	27.068	1	0.000	1.43363235
Votes	0.0348592645	0.008	19.320	1	0.000	1.03547397
Review Count	0.2948504568	0.014	455.613	1	0.000	1.34292552
RVIS	-0.0000154248	0.000	9.295	1	0.002	0.99998458
REIS	-0.0000108280	0.000	4.467	1	0.035	0.99998917
Note. Cox and Snell R2 = .160. Ngagelkerke R2 (Max rescaled) R2 = .232						
Group B: Feb 2014–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	0.3567740047	0.123	8.453	1	0.004	1.42871295
Votes	0.0832226264	0.016	27.979	1	0.000	1.08678373
Review Count	0.3454507659	0.025	185.556	1	0.000	1.41262654
REIS	-0.0000619633	0.000	18.637	1	0.000	0.99993804
Note. Cox and Snell R2 = .146. Ngagelkerke R2 (Max rescaled) R2 = .216						
Group C: June 2014–July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	0.9596715970	0.363	6.980	1	0.008	2.61083893
Review Count	0.3410057297	0.070	23.913	1	0.000	1.40636130
REIS	-0.0005574559	0.000	6.114	1	0.013	0.99944270
Note. Cox and Snell R2 = .117. Ngagelkerke R2 (Max rescaled) R2 = .179						
Group D: July 2014						
Predictor	β	SE	Wald's X^2	df	P	e^{β} (odds ratio)
Constant	3.0990953153	0.331	87.897	1	0.000	22.17787821
REIS	-0.0055873036	0.001	55.907	1	0.000	0.99442828
Note. Cox and Snell R2 = .141. Ngagelkerke R2 (Max rescaled) R2 = .206						

For group A, votes, review count, RVIS, and REIS are significant predictors of CIV ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set. Group B shows votes, review count, and REIS are significant predictors of CIV ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set. Group C shows both review count and REIS are significant predictors of CIV ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set. Group D shows REIS is a significant predictor of CIV ($p < .05$). The test of the intercept in this data set ($p < .05$) indicates that the intercept should be included in the data set.

Table 22 below provides a summary of the sensitivity, specificity, false positive, false negative, and overall percentage correct statistics derived from the classification tables for the models.

Table 22: Classification Table Derived Statistics—CIV

	Group	Sensitivity	Specificity	False Positive	False Negative	Overall Percentage
Time Yelping	A: 1 Year	100.0%	0.0%	27.0%	100.0%	73.0%
	B: 6 Months	100.0%	0.0%	25.4%	0%	74.6%
	C: 2 Months	100.0%	0.0%	22.3%	0%	77.7%
	D: 1 Month	100.0%	0.0%	26.8%	0%	73.2%

Based on Table 22, with the cut off of the predicted probability set at 0.5, the prediction for change in votes (CIV) for groups A, B, C, and D is more accurate than for those who did not have a change in votes. This was supported by the sensitivity (100%) compared to that of specificity (0%). If the cut off was set a different value, the

sensitivity, specificity, false positive, false negative, and overall percentage rates would change. This is captured in the AUC analysis described later. For group A, the false positive rate is 27% while the false-negative rate is 100%. The overall prediction rate is 73%. For group B, the false positive rate is 25.4% while the false-negative rate is 0%. The overall prediction rate is 74.6%. For group C, the false positive rate is 22.3% while the false-negative rate is 0%. The overall prediction rate is 77.6%. For group D, the false positive rate is 26.8% while the false-negative rate is 100%. The overall prediction rate is 73.2%.

The results show that for all time periods, the proportion of members correctly classified as having a change in votes is 100% (sensitivity). For all time periods, the proportion of members correctly classified as having no change in votes is 0% (specificity). This indicates that for all time periods, the algorithm predicts every member to have a change in votes. If the predicted probability (0.5) were set to a different value, the specificity and sensitivity results would vary. The ROC curve and the corresponding AUC values, however, are generated by using continuously varying probabilities.

Table 23 below compares the model's predictive power to that of the best individual predictor for CIV. ROC curves and the AUC data from SPSS can be found in Appendix G.

Table 23: AUC Comparison—CIV

Group	AUC (Model Probability of CIV)	AUC (Best Individual Predictor of CIV)
A: July 2013–July 2014	0.763	0.754 (Reviews)
B: Feb 2014–July 2014	0.752	0.740 (Reviews)
C: June–July 2014	0.727	0.723 (Reviews)
D: July 2014	0.663	0.698 (EIS)

Based on the results in Table 23, the probabilities from the logistic model as the predictor of CIV is slightly better at predicting CIV for groups A, B, and C in comparison to that of the best predictors of CIV using the original classifiers. For group D, the individual determinant classifier, EIS, has a higher AUC (0.698) in comparison to the logistic model's probability's predictor (0.663).

The findings of number of reviews for a member as a significant predictor in the models for groups A, B, and C (Tables 20 and 21), coincides with the findings of number of reviews as being the significant individual predictor for change in votes for groups A, B, and C. Also, the finding of EIS as the predictor for the model for group D (Tables 20 and 21) coincides with EIS as the best individual predictor of change in votes for group D. The finding of the AUC best individual predictor of change in votes for group D (0.698) producing a higher AUC than the probabilities from the logistic model (0.663) is an unexpected outcome. This result is likely due to the transformation of EIS to REIS for logistic regression. The use of ranks rather than the true value may have caused a loss in information. Due to the severe nonlinearity and prevalence of zeros, EIS could not be used for logistic modeling. EIS had to be transformed into a rank, REIS.

Table 24 below compares the AUC probabilities of the different machine learning algorithms used to generate models from the predictors of CIV as found from the logit model. The best model AUC values are depicted shaded in the table. As noted earlier with the comparison for CIF, RVIS and REIS (ranks) are used in generating the logit models, while VIS and EIS (scores) are used in generating the other models.

Table 24: AUC Model Comparison—CIV

Group	AUC (LOGIT Model Probability vs CIV)	AUC (NN Model Probability vs CIV)	AUC (naive Bayes Model Probability vs CIV)	AUC (SVM Model Probability vs CIV)
A: July 2013–July 2014	0.763	0.763	0.726	0.722
B: Feb 2014–July 2014	0.752	0.752	0.722	0.567
C: June–July 2014	0.727	0.729	0.713	0.709
D: July 2014	0.663	0.703	0.684	0.694

For groups A and B, the AUC for the logistic model probability and neural networks model probability has the same value. These two models are the highest of the models in group A and group B. For groups C and D, the AUC for neural networks model probability produces the highest AUC.

While the best model probability is NN for groups C and D, the numbers do not have a wide distribution from the logit model; therefore, confirming the robustness of the logistic model. Non-linear machine learning classifiers, neural networks, naive Bayes, and SVM do not significantly outperform the linear logit model, which is a similar finding for the change in friends. Noteworthy is the 0.567 AUC value for SVM group B. SVM is outlier dependent, causing the issue of a very low AUC value for group B.

Chapter 6: Discussion

This chapter provides a discussion of the results in Chapter 5, providing insight into both the findings for the best individual predictors as well as the best combination of predictors for CIFs and CIVs. The discussion is divided into two sections based on the dependent variable, future influence. A discussion of the results for CIV follows a discussion of the results of CIF.

6.1 Change in friends

The AUC results in the last row and last column of Table 13 (shaded in grey) above indicate that the number of reviews written is the most important factor for predicting influence in ORCs.

For all groups, the modified measure EIS (number of reviews of the member along with the number of reviews of the member's network) outperforms all other measures, including the traditional measure (degree or TIS) and Yelp's elite measure. In line with previous research, social structure and actions both play a part in influence. The results of this study indicate that while a member's current number of friends can predict the likelihood on making friends, examining actions and the actions of those in the friend group increases the accuracy of predicting an increase in friends.

Social influence theory with its focus on the individual's action as the driver of influence explains why the individual makes an impact on predicting influence. Expectation states theory, as well as social influence and status characteristics theory, explain why the number of reviews of a member makes an impact on whether one will increase their number of friends. These theories emphasize the characteristics of an individual as an indicator of ability level. The perception of one's ability to write many

reviews corresponds to one's ability to be engaged in the community and or knowledgeable on topics. As one's perception of being engaged and or knowledgeable goes up, so too do the reviewer's followership.

The concepts of homophily and self-categorization explain why the number of reviews of those in a member's network increases the predictability of CIF. Homophily is the tendency of people to befriend other people like themselves. The self-categorization theory specifies that as people are figuring out their identity, they look for ideal people in a group to emulate. Homophily and self-categorization theory signal that members of ORCs evaluate different aspects of a member (including who they are friends with and the number of review contributions of their network) when deciding whether to have a relationship. Members of OSNs are judged not only by their abilities but of the abilities of those in their member network.

Logistic regression is utilized to provide insight into what combination of variables work together to produce the best accuracy in predicting a change in friends (Tables 15 and 16). The AUC values in Table 18 indicate the models become more predictive when a member has been in the network for a longer time. At one month, the AUC value is 0.671. At two months, the AUC value is 0.731. At 6 months, the AUC value is 0.751. At one year, the AUC value is 0.765. These gains in predictability; however, are not very significant after a member has been in the network for two months. Even when a member has been in the network for one month, the model has an AUC value of 0.671

Tables 15 and 16 show that two essential predictors of change in friends are a member's friends count and the number of reviews for the member plus that of their

friends (REIS or rank of EIS). The friends count and REIS are significant predictors in each of the four models. For groups, C and D, friends count and REIS are the only two predictors. For groups A and B, friends count and REIS are two amongst other predictors. The AUC values for the model provided in Table 18 indicate that the predictability of the models increases over time.

When a member is new to an ORC, the member has likely not been very productive in terms of writing reviews. The member is attempting to get familiar and acclimated to the environment. In this process, the member will assess others in the community and go through self-categorization. The member will identify the best group fit(s) for he or she based on the groups within the ORC. Then the member will move toward the group(s) with the best fit, thereby increasing their friends count. This process indicates a desire to participate in the community and the ability to make friends. The action of making a friend or friends shows the likelihood to do so in the future as found empirically and tabulated in Table16. With this said, it must be kept in mind that an ORC is built on the premise of reviews. A new member may have few reviews at such an early stage. The ability to at least show some productivity of writing a review (or reviews) and to befriend those who write reviews (REIS) provides an indication of future action as found empirically and tabulated in Table 16. People are often judged by the company kept; therefore, a member is likely judged in terms of new relationships based on what they have done and what their friends have done.

Tables 15 and 16 show an increase in the number of predictors of the model over time. By the time members get to about six months in the community, their likelihood to make friends is dependent on the same factors as those members of one or two months.

In addition to these factors are the member's review count and the combination of votes count and votes counts of friends. The increase in variables over time is primarily due to more data being available.

6.2 Change in votes

The AUC results in the last column of Table 14 (shaded in grey) above reveal again that the number of reviews is the best predictor. For group D, the modified measure EIS (number of reviews of the member along with the number of reviews of the member's network) outperforms all other measures. For groups A–C, solely the number of reviews for a member outperforms all other measures.

The results indicate that when a member has been part of an ORC for two months or more, their votes are predicted more accurately based solely on their actions (reviews and votes). Social influence is in effect. When someone is new to joining a community (one month), the best way to gauge the person is by their actions primarily and to a much lesser extent, the actions of their friends.

When a member has been part of an ORC for two months or more, the member's actions speak for themselves. Enough time has passed to see whether there is productivity (based on the number of reviews). When a member has more recently joined the community (one month), the member has not yet had much time to be productive. In line with Bakhshi et al. (2015), as members get more acclimated, they produce more and of better quality. When a member is new to an ORC, they have likely written few reviews. In this case, an understanding of the likelihood of the member to increase the number of current reviews is best sought after by looking at the number they have produced and the numbers of those in the member's network. In this case, the

company that one keeps is an indicator of likelihood as long as the member has also shown some initiative to produce. The member's network (or relationships created) on its own may show that a member can be productive in gaining friends, but not that the member takes the action needed to obtain votes.

Noteworthy here, the use of votes to predict influence is a more modern way of studying the topic. The results show that the number of reviews a member writes is not only an optimal way to predict CIF, it is also an optimal way to predict CIV. The level of the measurement is however different based on how influence is measured and the duration of time a member is part of an ORC.

The logistic regression models for change in votes follow a similar, but not exact, pattern to that of change in friends. The AUC values for the models increase over time.

The AUC results of the model in Table 23 and the models generated in Tables 20 and 21 support the findings of the AUC results for the best individual predictor (Wald's X^2) of change in votes. When a member has been a part of an ORC for as short as one month, the number of reviews of a member plus the number of reviews of the member's network (REIS) is the best predictor for change in votes. As time goes on, a member's review count (aside from that of their network) becomes a significant predictor of change in votes.

As was true with change in friends, the logistic procedure generates more variables in the model over time (Tables 20 and 21). For members who have been a part of the community for one month, the combination of the number of reviews of a member and that of their friends (REIS) is needed for the best prediction. For members who have been a part of the community for two months, the number of reviews for the member is

needed as well as the combination of their review count and that of their friends (REIS). For members who have been a part of the community for six months, the review count and votes are needed as well as the combination of the number of reviews of the member and the total number of reviews from friends (REIS). For members who have been a part of the community for one year, the review count and votes are needed, the combination of the number of reviews of a member and the total number of reviews from friends (REIS), and the combination of the number of votes of a member and the total number of votes from friends (RVIS). The increase in variables over time is primarily due to more data being available.

Chapter 7: Challenges, Limitations, Future Research Recommendations, Practical Implications, and Conclusion

Chapter 7 provides a discussion of the challenges and limitations of this dissertation, providing recommendations for future researchers. An examination into the implications for firms, members of ORCs, the ORC SNS, and the environment of study Yelp, is also provided in this chapter. Lastly, this chapter provides a summary, concluding the dissertation

7.1 Challenges, Limitations, and Future Research Recommendations

Although this study was able to examine individual variables and generate models for prediction, it did so with challenges. Issues arose in trying to generate models with the independent variables. Severe nonlinearity for the variables EIS and VIS and missing values necessitated a transformation of the variables to ranks. Utilizing ranks however has a couple of drawbacks: loss of information for highly skewed values and the requirement to periodically update the models due to structural changes.

I urge future researchers to examine the review text using latent semantic analysis to provide an understanding of the patterns of the text used to influence. This would be a valuable contribution. For text to be utilized, Yelp would need to provide clean textual data. The review counts indicated in the user file do not match with the number of reviews provided by Yelp in the review file. The discrepancy varies but is found to be most severe in cases where a member has a high number of reviews (indicated by the user file). The user file may show a member as having written thirty reviews, but the review file may only have one review for that member. Yelp is aware of the discrepancy.

I used degree centrality as the measure to compare and modify determinants in the dissertation; however, I encourage future researchers to examine the use of other centrality measures. A suggestion would be to use eigenvector centrality. I also encourage the comparison of degree centrality and its modification with other centrality measures and the modification provided in this dissertation.

This work emphasizes the difference between traditional OSNs and ORCs but focused the examination on ORCs. Future researchers could compare findings from a study examining both types of communities.

The examinations were conducted using binary categorical variables. A further research direction is to examine multinomial logistics models. Suggestions are categorizing by different ranges of scores or categorizing positive change, no change, and negative change.

This research used an aggregate of votes as a predictor of influence; however, future researchers may want to examine each type of vote (funny, useful, and cool) as a predictor of influence.

This study found that the number of reviews written by a member is a highly significant variable in predicting friends and votes. Future researchers may want to examine the prediction of change in reviews and use this as a factor to examine future influence, CIF and CIV.

Finally, the data provided by Yelp did not allow for tracking of votes of a member back to the castor of the vote. This led to the inability of examining influence through the connections or interactions. Future studies that can obtain this data, tracking of votes of a

member back to the castor of the vote, are urged to separately investigate (current and future) influence within and from within to outside of the ORC.

7.2 Practical Implications

The intelligence derived from this dissertation can be used by firms, ORCs, Yelp, and seekers of influence on ORCs.

7.2.1 Firms

The models produced in this dissertation provide decision-makers in marketing with algorithms to predict which members of ORCs will yield influence. Firms can target these individuals with campaigns or efforts to spread messages. Specifically, decision-makers have a clear endorsement of what measure has the best ability to predict influence, as well as, which combinations of measures have the best ability to predict influence. This dissertation endorses actions as well as the social structure in investigating influence. Using the models produced from this dissertation, companies can increase awareness of their firm and their products and services. Companies may also seek to respond or rectify negative impressions from a member who will yield influence, based on the knowledge that this member's message will infect others.

By providing decision-makers in marketing with algorithms based on the duration of time a member has been registered with the ORC, firms can start efforts toward individuals as early as one month when a user becomes a member.

Finally, the computations used to produce scores fed into the prediction algorithm can also be employed to identify current influence and current influencers (as defined by the firm).

7.2.2 Online review communities

The algorithms in this dissertation also provide benefits to SNSs that specifically have ORCs. These sites can employ the algorithms from this dissertation to locate the members that have the potential to write many reviews, connect with many members, and widely spread their messages, thereby incentivizing these members on that path. This is like Yelp's intent with their badges and elite status.

7.2.3 Yelp

While this study provides intelligence generalizable to ORCs, it also delivers valuable intelligence specific to Yelp and those businesses that look to exploit influence from the Yelp community.

Businesses must not depend on Yelp elite as their sole indication of influence. The results show that the Yelp elite designation is not a suitable predictor of whether a member will increase their network. We can posit that due to restrictions, specific to Yelp's process for designation, many deserving members may never be assigned the elite status. While these restrictive items may make a great Yelp member, for many businesses, this information is irrelevant to whether a person will visit and purchase from the establishment. Firms must identify which members are likely to increase their influence as well as those who are showing the tendency to increase their network, potentially allowing the company's messages to spread with the increase of the friend's network.

Also, the intelligence gathered from this study allows Yelp a way to identify members that are elite material without the use of a manual process. Bakhshi et al.'s (2015) study suggests that votes can tell us information about a member's involvement

and understanding of the community as well as how to write reviews that resonate with users. These items, understanding and involvement in the community, are relevant to who Yelp wants as their elite. Yelp desires active members that are writing quality reviews. Investigating what variable best predicts a CIV gives insight into who the members are that are either gaining an understanding and increasing their involvement or honing their understanding and involvement. The results from the study show that solely using a member's review count can predict for Yelp, members that may increase their votes. Yelp's use of the predictive power of the models from this study provides the company a way to identify which members to incentivize toward gaining the elite status or continuing to work to keep their status for the next cycle. By providing incentives along the way, Yelp is being proactive in its mission to bring people together with businesses. Furthermore, investigating which variable best predicts a CIF and what model to use to predict friends, gives insight into who the members are that are gaining friends and will potentially gain friends. Members that are gaining friends are members who are showing activity and have the potential to reach more people, which again, helps Yelp toward its goal of connecting people with businesses. The results from the study show that looking at a member's network review count activity in addition to their own review count activity can predict CIF. The intelligence gained from these results helps Yelp toward an automated process for identifying those members who should be designated elite and provide a way to proactively incentivize those members that have the most potential to increase their influence.

7.2.4 Seekers of influence

The models generated in this dissertation not only benefit firms and ORC SNSs, but also benefit those members of ORCs that seek to influence and or to become influencers. The models provide information that members can utilize to increase their chances of increasing friends and votes, and therefore, gaining and increasing influence within an ORC. Writing more reviews and increasing connections in one's network, aids in achieving the goal of being an influencer; thereby possibly being able to extract economic value from businesses.

7.3 Conclusion

The data generated from SNSs is rich and plentiful. The number of sites and users of these sites continues to increase. This continuous upsurge means that SNSs have much to offer researchers and practitioners. OSNs provide a plethora of data that is delivered in real-time and is of different varieties. This data has been termed big data. The analysis of big data has become an imperative in analyzing market opportunities. The intelligence derived from the OSNs within these sites provide insights that can assist decision-makers in making strategic decisions. This dissertation provides intelligence that can be used to identify members of OSNs who may likely influence others. By identifying these individuals, firms and SNSs can leverage these individuals in a way that will help the sites achieve their missions. Specifically, the results in this dissertation provide methods to identify members, within online review communities (a segment of OSNs), who can be targeted and motivated towards spreading word of mouth messages; thereby influencing others. Word of mouth, also known as reviews in an ORC, affects a firm's bottom line. Word of mouth also affects whether a SNS can be an effective ORC.

Twitter, a directed network, has been the environment of study in much of the present influence research; however, another type of network structure can and does exist. Depending on the SNS, the structure of the network can be undirected or directed. An undirected network is one in which there is a bilateral agreement to the connection and sharing of information flow. In a directed network, there is a clear distinction between an incoming connection and an outgoing connection and the route in which information flows. An undirected network's strategy is focused on maximizing information flow, whereas, a directed network's strategy is focused on maximizing useful information for each person.

In this dissertation, I utilize OSN data from an undirected network, Yelp. Yelp's primary focus is the connection of users to businesses through the public posting of reviews. Reviews, also known as word of mouth, are a leading source of influence. In order to connect users with businesses, Yelp allows all posts of a member to be viewed by other registered members of the community, as well as non-members. Members of ORCs can influence many others, within the community and outside of the community, with a few clicks of the keyboard. In a traditional OSN, which has a directed network structure, posts are restricted to only those registered to be a part of the community. In order to distinguish Yelp and its mission, focus, and structure from that of a traditional OSN, I have termed Yelp an online review community (ORC).

Through SNA, the relationships and exchanges of the members of Yelp are examined to uncover the patterns and circumstances of predicting influence in ORCs. This dissertation highlights the significant differences between OSNs and ORCs and approaches the examination of influence from the Yelp undirected network perspective.

The goal of the dissertation is to provide a utility to ORCs and firms as well as to fill a gap in academic research.

Researchers have framed the identification of influential members in an OSN as an influence maximization problem. The goal is to create or discover a method for which to identify the members of OSNs who should be targeted and motivated towards spreading a message so that the message spreads to as many people as possible. The influence maximization problem is slightly modified in this research; the approach in this dissertation is to identify members who will likely have influence instead of those who may currently have influence. A number of models have been introduced by researchers to solve the influence maximization problem: susceptible infective removed, threshold, cascade, and multi-level. While the models' approaches to solving the problem are different, they are all similar in their focus on identifying actual versus potential individuals that are influenced by messages. The approach used in this dissertation is the multi-level model. In this model, user-user and user-item relationships are examined to solve the problem. This approach enables one to examine the actions taken by members and users of ORCs and observe the effect on predicting influence. Several concepts and theories emphasize the importance of the individual and the network as well as actions or abilities that are important to influence: social influence, social influence and status characteristics theory, expectations theory, self-categorization theory, homophily, and exchange theory.

In approaching the influence maximization problem from a predictive perspective, two measures of prospective influence are examined, a measure of a member's change in friends (CIF) and a measure of a member's change in votes (CIV). CIF and CIV are

derived from subtracting data of interest at time point A (July 2014) from time point B (December 2015). The variables CIF and CIV are then coded as dichotomous variables. Members with a positive change are coded as a one and members with no change or a negative change are coded as a zero.

In line with the theories mentioned and using the multi-level model, the constructs, social structure, user engagement, overall review quality, and elite status are utilized to investigate and compare determinants of predicting CIF and CIV. Degree centrality (also referred to as friends count or TIS throughout the dissertation) is utilized as the measure of social structure, review count is utilized as the measure of user engagement, and votes count is utilized as the measure of overall review quality. The elite construct is a measure employed by Yelp. Yelp does not provide specifics on exactly what goes into the make-up of this construct; however, the construct is utilized as a measure of a member's importance. Thus, the elite status is examined as a predictor of future influence.

Two separate investigations are conducted to study the prediction of influence in ORCs. In the first study, friends count, review count, and votes count are examined at three different levels of the variables: the member, the member's network, and the member along with the member's network.

Degree centrality is used as a network measure of current influence and is examined as a predictor of future influence. Degree centrality is also used as the basis for creating modified network measures of influence using votes (VNS and VIS) and reviews (ENS and EIS). In the classic degree centrality calculation, each node is given equal weight (weight of one). Because the degree is counting the connections a member has, it

is a network measure. The modifications (ENS, EIS, VNS, and VIS) use the premise of looking at the direct network, but substitutes the count of friends, for a sum of a variable attributed to each friend. ENS and VNS are measures at the network level, while EIS and VIS are measures at the member along with the network level. The ENS measure sums the review count (number of reviews by a member) for each direct friend of a member. This becomes the member's influence score. The EIS measure sums the review count (number of reviews by a member) for each direct friend of a member and adds it to that member's review count. The VNS measure sums the votes received for all the reviews written by a member for each direct friend of a member. This becomes the member's influence score. VIS sums the votes received for all the reviews written by a member for each direct friend of a member and adds it that member's votes count.

After creating scores for all the measures, the data is then broken into four groups: members with a time frame of a year on Yelp, members with six months on Yelp, members with two months on Yelp and members with one month on Yelp, i.e., months since joining Yelp. Each member's friends count (or TIS), votes count, review count, elite status, ENS, VNS, EIS, VIS, CIF, and CIV are imported into SPSS. Using the ROC method in SPSS, an investigation is conducted into the determinants of predicting influence for each group. The predictors (friends count (or TIS), votes count, review count, elite status, ENS, EIS, VNS, and VIS) are plotted against the dichotomous variables CIF and CIV. This is done to obtain the area under the curve (AUC) statistic and the receiver operating characteristic (ROC) curve. The AUC value derived from the ROC curve is utilized to assess how well each predictor or each model is doing in distinguishing between the different dependent variable outcomes.

Results of the first study reveal that the elite measure employed by Yelp is not an appropriate measure to predict change in friends or change in votes for any of the four groups (one month, two months, six months, and one year as part of the Yelp community) examined in the study. The ROC curve and AUC values for each group for each dependent variable indicate that the elite measure has little to no predictive power. Yelp's manual method of assessing influence should be reassessed and businesses should not use the elite measure as an indicator of a member's future influence. Results indicate that the number of reviews written by members is the best way to predict influence.

The review count of a member along with the review counts of a member's friends (EIS) is the best individual predictor of change in friends for members who have been a part of the Yelp community for all groups in the study, one month to one year. The review count of a member is the best individual predictor of change in votes for members who have been a part of the Yelp community for two months or more. The review count of a member along with the review counts of a member's friends (EIS) is the best individual predictor of change in votes for members who have been a part of the Yelp community for one month.

Based on the findings from the first study, the network level is removed from the second study. The member, as well as the member along with the network, are the levels used in the second study. Due to severe nonlinearity, VIS and EIS are transformed into ranks for modeling purposes. Supervised machine learning methods (naive Bayes, NN, SVM) and logistic regression are used to generate models of prediction for the dependent variables, CIF and CIV. After gaining an understanding of the independent variables in the generated models from the logistic regression analysis, the variables are entered into

other supervised machine learning methods—NN using SPSS and naive Bayes and SVM using R. The AUC value for the model's probability in predicting the dependent variable is used to compare the predictive power of the models.

The friends count and rank of the review count of a member along with the review counts of a member's friends (REIS) are consistently generated predictors for change in friends using logistic regression. These predictors are found in the model for each group. For members who have been a part of the ORC for one month, REIS is the only predictor generated from the logistic regression for a change in votes. For members who have been a part of the Yelp community for two months or more, the number of reviews of a member is a consistently generated predictor for a change in votes using logistic regression.

The results of the logistic regression support the continued use of a member's social network in investigating influence in terms of friends, but also indicate other measures (such as the number of reviews) and levels of measures (such as the member along with his or her network) that need to be evaluated to get insights into prediction of influence. Results from both of the studies reveal to future researchers that the number of reviews written by members must be utilized to best predict influence. The research indicates that the number of reviews written is the key variable in predicting influence. The duration of time since joining an ORC has an impact on the prediction model for influence, but the number of reviews still plays a vital role.

Based on the findings from this dissertation, service firms and ORC SNSs have insights into the best predictor and best models for predicting influence. Seekers of influence also have a clear guide to establish influence in ORCs. The research in this

dissertation analyzes the effect that happens between actors in a specific type of OSN, an ORC. Examining the patterns of connections enables marketers to identify and target individuals who can be effective in spreading word of mouth messages. This dissertation provides a way to identify those to be targeted based on their likelihood to have an influence, in and outside of the network.

Whether influence is measured by its traditional means of social structure or by a more modern means of acquiring votes, this dissertation demonstrates that models of prediction are attainable. Whether a person is new in joining an ORC or a veteran, models of prediction are attainable. Results vary depending on the duration of time since joining an ORC and whether one measures influence by social structure or by votes; however, models for each group and each measure of influence are attainable.

Based on research and two studies with several investigations, this dissertation has created modified measures of influence, compared various measures of influence, highlighted the best individual predictors of influence, and generated models for predicting influence.

References

Alarcón-del-Amo, M. D. C., Lorenzo-Romero, C., & Gómez-Borja, M. Á. (2011). Classifying and profiling social networking site users: A latent segmentation approach. *Cyberpsychology, Behavior, and Social Networking*, *14*(9), 547–553.

AlFalahi, K., Atif, Y., & Abraham, A. (2014). Models of influence in online social networks. *International Journal of Intelligent Systems*, *29*(2), 161–183.

An introduction to Yelp metrics as of June 30, 2019. (2019, June 30). Retrieved from https://www.yelp.com/factsheet.

Anagnostopoulos, A., Kumar, R., & Mahdian, M. (2008, August). Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 7–15). ACM.

Aral, S. (2014). The problem with online ratings. *MIT Sloan Management Review*, *55*(2), 47.

Bakhshi, S., Kanuparthy, P., & Shamma, D. A. (2015, February). Understanding online reviews: funny, cool or useful?. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1270–1276). ACM.

Battiston, P., & Stanca, L. (2015). Boundedly rational opinion dynamics in social networks: Does indegree matter?. *Journal of Economic Behavior & Organization*, *119*, 400–421.

Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association*, *72*(357), 46-53.

Berger, J., Rosenholtz, S. J., & Zelditch, M. (1980). Status organizing processes. *Annual Review of Sociology*, 479–508.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. Hoboken, New Jersey: John Wiley & Sons.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly*, *4*(2011), 24–35.

Budalakoti, S., DeAngelis, D., & Barber, K. S. (2011). Unbiased trust estimation in content-oriented social networks. *Trust in Agent Societies (TRUST-2011)*, 13.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge University Press.

Comscore and The Kelsey Group. (2007, November 29). Online consumer-generated reviews have significant impact on offline purchase behavior. Retrieved from http://www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior.

Coulter, K. S., & Roggeveen, A. (2012). "Like it or not" Consumer responses to word-of-mouth communication in on-line social networks. *Management Research Review*, *35*(9), 878–899.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403.

Danaher, P. J., & Rust, R. T. (1996). Rejoinder-indirect financial benefits from service quality. *Quality Management Journal*, *3*(2).

Dens, N., De Pelsmacker, P., & Purnawirawan, N. (2015). “We (b) care” How review set balance moderates the appropriate response strategy to negative online reviews. *Journal of Service Management*, *26*(3), 486–515.

Dimensional Research. (2013, April). Customer service and business results: A survey of customer service from mid-size companies. Retrieved from https://d16cvnquvjw7pr.cloudfront.net/resources/whitepapers/Zendesk_WP_Customer_Service_and_Business_Results.pdf.

Dubois, E., & Gaffney, D. (2014). The multiple facets of influence identifying political influentials and opinion leaders on Twitter. *American Behavioral Scientist*, *58*(10), 1260–1277.

Economist Intelligence Unit. (2011). Big data: Harnessing a game changing asset. *The Economist,* 1–32.

Emerson, R. M. (1976). Social exchange theory. *Annual Review of Sociology*, *2*(1), 335–362.

Faletra, M., Palmer, N., & Marshall, J. S. (2014). Effectiveness of opinion influence approaches in highly clustered online social networks. *Advances in Complex Systems*, *17*(02), 1450008.

Fang, Q., Sang, J., Xu, C., & Rui, Y. (2014). Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. *Multimedia, IEEE Transactions on*, *16*(3), 796–812.

French, J. R., Raven, B., & Cartwright, D. (1959). The bases of social power. *Classics of Organization Theory*, *7*, 311-320.

Fruth, A., & Neacsu, M. (2014). Online consumer reviews as marketing instrument. *Knowledge Horizons. Economics*, *6*(3), 128.

Fu, F., Liu, L., & Wang, L. (2008). Empirical analysis of online social networks in the age of Web 2.0. *Physica A: Statistical Mechanics and Its Applications*, *387*(2), 675–684.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Godinho de Matos, M., Ferreira, P., & Krackhardt, D. (2014). Peer influence in the diffusion of the iPhone 3G over a large social network. *Management Information Systems Quarterly*, *38*(4), 1103–1133.

Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, *24*(1), 88–107.

GPS.gov. (2013). GPS Applications. Retrieved from http://www.gps.gov/applications/

Gross, R., & Acquisti, A. (2005, November). Information revelation and privacy in online social networks. *In Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society* (pp. 71–80). ACM.

Huang, B., Yu, G., & Karimi, H. R. (2014). The finding and dynamic detection of opinion leaders in social network. *Mathematical Problems in Engineering*, *2014*.

Jahoda, M. (1959). Conformity and independence. *Human Relations*, 12(2), 99–120.

Jindal, T. (2015). Finding local experts from Yelp dataset (Unpublished master thesis). Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/78499/JINDAL-THESIS-2015.pdf.

Kassarjian, H. H. (1971). Personality and consumer behavior: A review. *Journal of Marketing Research*, 409–418.

Kelman, H. C. (1961). Processes of opinion change. *Public Opinion Quarterly*, *25*(1), 57–78.

Kelman, H. C. (1974). Further thoughts on the processes of compliance, identification, and internalization. *Perspectives on Social Power*, 125–171.

Kempe, D., Kleinberg, J., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137–146). ACM.

Kim, E. S., & Han, S. S. (2009, July). An analytical way to find influencers on social networks and validate their effects in disseminating social games. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in* (pp. 41–46). IEEE.

Kiss, C., & Bichler, M. (2008). Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, *46*(1), 233–253.

Kuan, K. Y., Zhong, Y., & Chau, P. K. (2014). Informational and normative social influence in group-buying: Evidence from self-reported and EEG data. *Journal of Management Information Systems*, *30*(4), 151–178.

Kuechler, W. L. (2007). Business applications of unstructured text. *Communications of the ACM*, *50*(10), 86–93.

Kumar, R., Novak, J., & Tomkins, A. (2006, August). Structure and evolution of online social networks. In *Proceedings of the 12th ACM KDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Kuss, D. J., & Griffiths, M. D. (2017). Social networking sites and addiction: Ten lessons learned. *International Journal of Environmental Research and Public Health*, *14*(3), 311.

Lampe, C., & Johnston, E. (2005, November). Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 11–20). ACM.

Levi, A., & Mokryn, O. (2014, April). The social aspect of voting for useful reviews. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 293–300). Springer, Cham.

Li, F., & Du, T. C. (2011). Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. *Decision Support Systems*, *51*(1), 190–197.

Li, G., & Yang, X. (2014). Effects of social capital and community support on online community members' intention to create user-generated content. *Journal of Electronic Commerce Research*, *15*(3), 190.

Li, P., He, S., Wang, H., & Zhang, X. (2014). Competitive diffusion in online social networks with heterogeneous users. *International Journal of Modern Physics B*, *28*(22), 1450147.

Li, Y. M., Lai, C. Y., & Chen, C. W. (2011). Discovering influencers for marketing in the blogosphere. *Information Sciences*, *181*(23), 5143–5157.

Li, Y., Ma, S., Zhang, Y., & Huang, R. (2013). An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems*, *43*, 43–51.

Liu, R. R., & Zhang, W. (2010). Informational influence of online customer feedback: An empirical study. *Journal of Database Marketing & Customer Strategy Management*, *17*(2), 120–131.

Liu, S., Jiang, C., Lin, Z., Ding, Y., Duan, R., & Xu, Z. (2015). Identifying effective influencers based on trust for electronic word-of-mouth marketing: A domain-aware approach. *Information Sciences*, *306*, 34–52.

Livingstone, S. (2008). Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New Media & Society*, *10*(3), 393-411.

Lohr, S. (2012). The age of big data. *New York Times*, *11*(2012).

Lu, D., Li, Q., & Liao, S. S. (2012). A graph-based action network framework to identify prestigious members through member's prestige evolution. *Decision Support Systems*, *53*(1), 44–54.

Lu, X., Ba, S., Huang, L., & Feng, Y. (2013). Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Information Systems Research*, *24*(3), 596–612.

Luftman, J. N., Lewis, P. R., and Oldach, S. H. 1993. Transforming the enterprise: The alignment of business and information technology strategies. *IBM Systems Journal*, 32(1), 198–221.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute*, 1–143.

McAfee, A. and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60–66.

Mearian, L. (2014, March 14). Data storage – then and now. *Computerworld*. Retrieved from https://www.computerworld.com/article/2473980.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (pp. 29–42). ACM.

Momeni, N., & Rabbat, M. (2016). Qualities and inequalities in online social networks through the lens of the generalized friendship paradox. *PloS one*, *11*(2), e0143633.

Murugananthan, A., Gandhi, G. M. (2005). Ranking the influence users in a social networking site using an improved Topsis method. *Journal of Theoretical and Applied Information Technology*, *73*(1).

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, *61*, 47–58.

Nguyen, H., & Zheng, R. (2014, June). A data-driven study of influences in Twitter communities. In *Communications (ICC), 2014 IEEE International Conference on* (pp. 3938–3944). IEEE.

Oldmeadow, J. A., Platow, M. J., Foddy, M., & Anderson, D. (2003). Self-categorization, status, and social influence. *Social Psychology Quarterly*, *66*(2), 138–152.

Otterbacher, J. (2011). Being heard in review communities: Communication tactics and review prominence. *Journal of Computer-Mediated Communication*, *16*(3), 424–444.

Pal, S. K., Kundu, S., & Murthy, C. A. (2014). Centrality measures, upper bound, and influence maximization in large scale directed social networks. *Fundamenta Informaticae*, *130*(3), 317–342.

Patrizio, A. (2018, December 3). IDC: Expect 175 zettabytes of data worldwide by 2025. *NetworkWorld*. Retrieved from https://www.networkworld.com/article/3325397.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14.

Podium. (2017). Consumers get "buy" with a little help from their friends. Retrieved from http://learn.podium.com/rs/841-BRM-380/images/2017-SOOR-Infographic.jpg.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Rashotte, L. (2007). Social influence. *The Blackwell Encyclopedia of Social Psychology*, *9*, 562–563.

Rosenquist, J. N., Murabito, J., Fowler, J. H., & Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, *152*(7), 426.

Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109–127.

Stoppelman, J. (2013, January 17). Now on Yelp: Restaurant Inspection Scores [Web log post]. Retrieved from https://blog.yelp.com/2013/01/introducing-lives.

Subbian, K., & Melville, P. (2011, October). Supervised rank aggregation for predicting influencers in twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 661–665). IEEE.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009, June). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 807–816). ACM.

The most popular database for modern apps. (n.d.). Retrieved from https://www.mongodb.com.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Cambridge, MA: Basil Blackwell.

Vidmer, A., Medo, M., & Zhang, Y. C. (2015). Unbiased metrics of friends' influence in multi-level networks. *EPJ Data Science*, *4*(1), 1–13.

Walther, J. B., Van Der Heide, B., Kim, S. Y., Westerman, D., & Tong, S. T. (2008). The role of friends' appearance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep?. *Human Communication Research*, *34*(1), 28–49.

Wan, Y., & Nakayama, M. (2014). The reliability of online review helpfulness. *Journal of Electronic Commerce Research*, *15*(3), 179.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Review of Psychology*, *51*(1), 539–570.

Xu, W. W., Sang, Y., Blasiola, S., & Park, H. W. (2014). Predicting opinion leaders in twitter activism networks the case of the Wisconsin recall election. *American Behavioral Scientist*, 0002764214527091.

https://www.yelp.com/

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.701.4456&rep=rep1&type=pdf.

Zhang, Y., & Li, X. (2014). Relative superiority of key centrality measures for identifying influencers on social media. *International Journal of Intelligent Information Technologies (IJIIT)*, *10*(4), 1–23.

Zhang, Y., Li, X., & Wang, T. W. (2013). Identifying influencers in online social networks: The role of tie strength. *International Journal of Intelligent Information Technologies (IJIIT)*, *9*(1), 1–20.

Zhao, P., Li, Y., Xie, H., Wu, Z., Xu, Y., Ma, R. T., & Lui, J. (2016). Impact of online activities on influence maximization: A random walk approach. *arXiv preprint arXiv:1602.03966*.

Zhu, L., Yin, G., & He, W. (2014). Is this opinion leader's review useful? Peripheral cues for online review helpfulness. *Journal of Electronic Commerce Research*, *15*(4), 267.

Zhu, Y. X., Zhang, X. G., Sun, G. Q., Tang, M., Zhou, T., & Zhang, Z. K. (2014). Influence of reciprocal links in social networks. *PloS one*, *9*(7), e103007.

Appendices

Appendix A: Value Influence Factor (VIF)

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	ReviewCount	0.917	1.091
	FriendsCount	0.934	1.070
	Rank of VIS	0.326	3.070
	Rank of EIS	0.319	3.137

a. Dependent Variable: Total Votes

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	FriendsCount	0.791	1.264
	Rank of VIS	0.325	3.081
	Rank of EIS	0.326	3.071
	TotalVotes2014	0.809	1.237

a. Dependent Variable: Review Count

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	Rank of VIS	0.325	3.076
	Rank of EIS	0.32	3.126
	TotalVotes2014	0.837	1.195
	ReviewCount	0.804	1.244

a. Dependent Variable: FriendsCount

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	Rank of EIS	0.885	1.131
	TotalVotes2014	0.708	1.413
	ReviewCount	0.8	1.251
	FriendsCount	0.789	1.268

a. Dependent Variable: Rank VIS (RVIS)

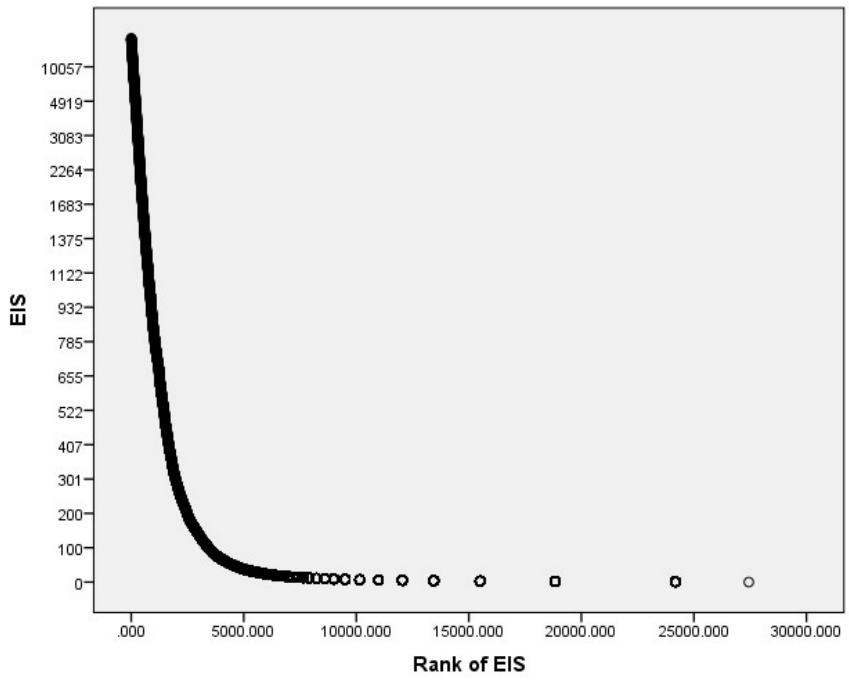
Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	TotalVotes2014	0.707	1.414
	ReviewCount	0.819	1.221
	FriendsCount	0.792	1.262
	Rank of VIS	0.903	1.107

a. Dependent Variable: Rank of EIS (REIS)

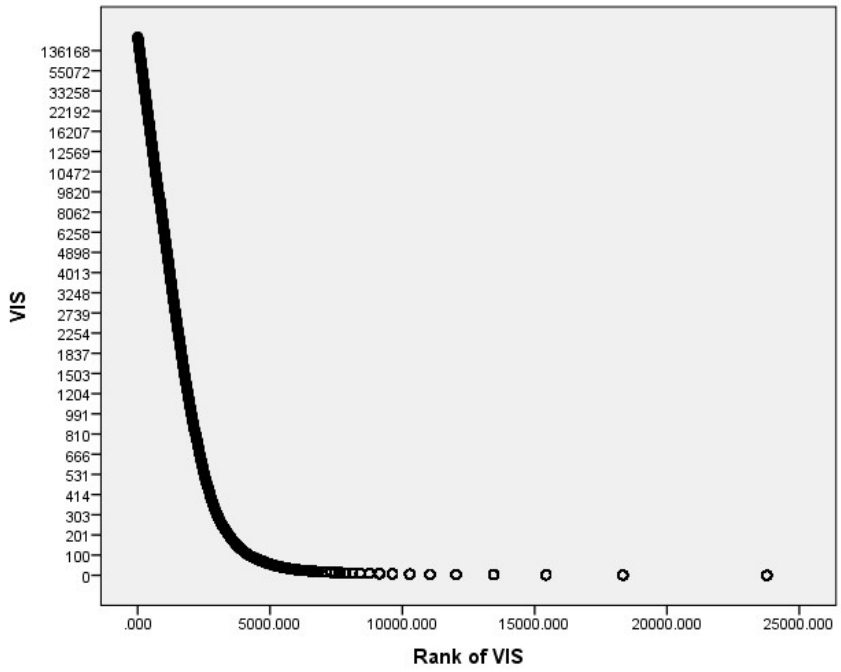
Appendix B: EIS and VIS tests for Linearity

A. EIS and VIS tests for linearity

EIS



VIS


VIS
136168
55072
33258
22192
16207
12569
10472
9820
8062
6258
4898
4013
3248
2739
2254
1837
1503
1204
991
810
666
531
414
303
201
100
0
.000
5000.000
10000.000
15000.000
20000.000
25000.000
Rank of VIS

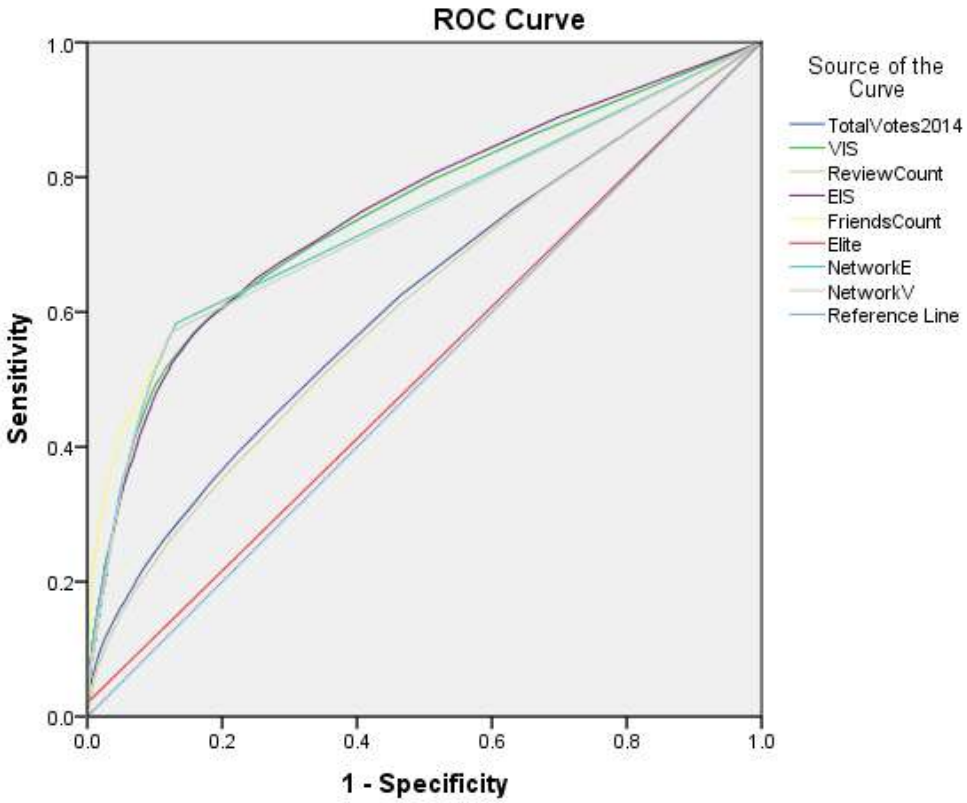
Appendix C: AUC ROC Data for Change in Friends

Group A: July 2013–July 2014 Change in Friends

Case Processing Summary

ChangeInFriends	Valid N (listwise)
Positive[a]	9115
Negative	18321

The positive actual state is 1.



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.616	.004	.000	.609	.623
VIS	.746	.003	.000	.740	.753
ReviewCount	.608	.004	.000	.601	.615
EIS	.751	.003	.000	.744	.757
FriendsCount	.743	.003	.000	.736	.750
Elite	.510	.004	.007	.503	.517
NetworkE	.736	.003	.000	.730	.743
NetworkV	.732	.003	.000	.725	.739

a. Under the nonparametric assumption

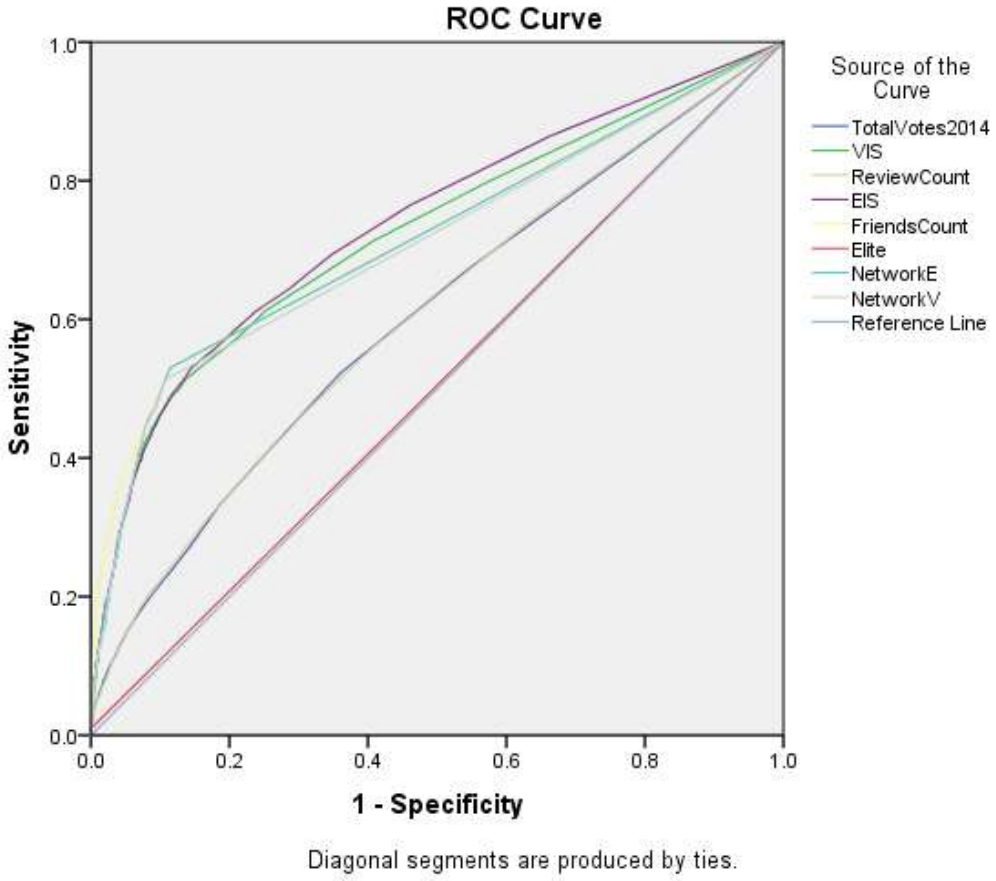
b. Null hypothesis: true area = 0.5

Group B: Feb 2014–July 2014 Change in Friends

Case Processing Summary

ChangeInFriends[a]	Valid N (listwise)
Positive[b]	3278
Negative	6280

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.603	.006	.000	.591	.615
VIS	.724	.006	.000	.713	.736
ReviewCount	.604	.006	.000	.591	.616
EIS	.738	.006	.000	.727	.749
FriendsCount	.721	.006	.000	.709	.732
Elite	.505	.006	.433	.493	.517
NetworkE	.716	.006	.000	.704	.727
NetworkV	.709	.006	.000	.698	.721

a. Under the nonparametric assumption

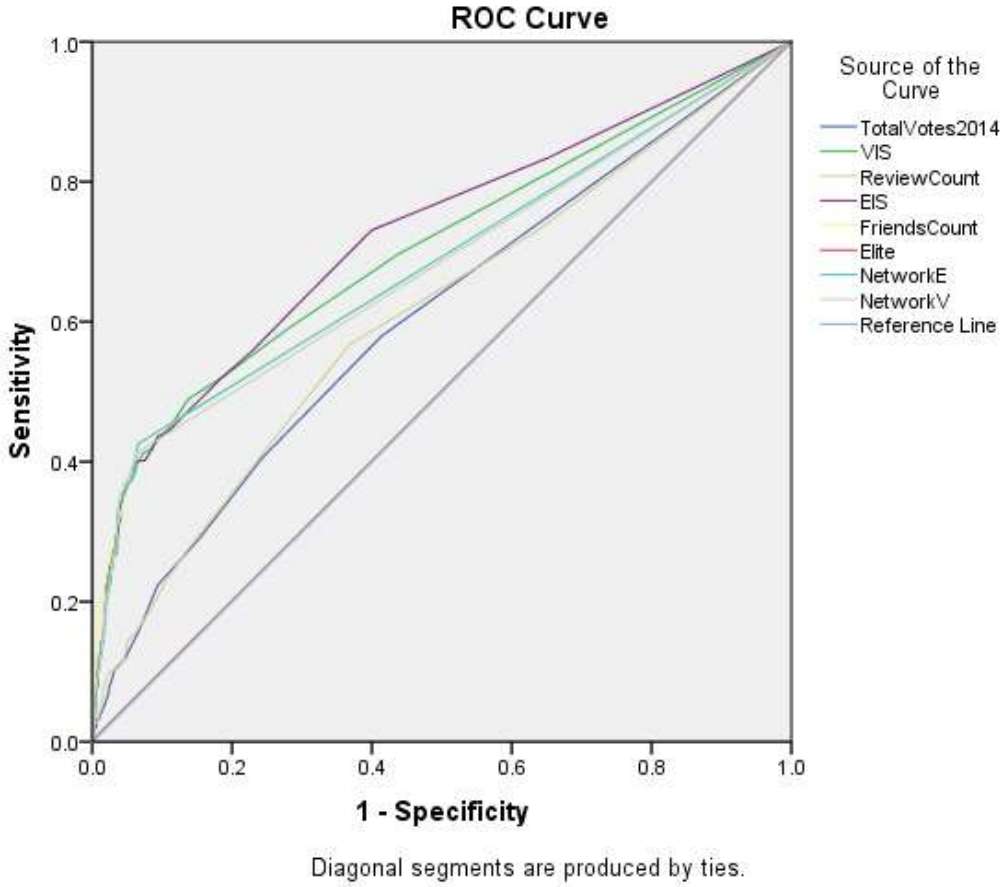
b. Null hypothesis: true area = 0.5

Group C: June 2014–July 2014 Change in Friends

Case Processing Summary

ChangeInFriends[a]	Valid N (listwise)
Positive[b]	587
Negative	1168

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.605	.015	.000	.576	.633
VIS	.703	.014	.000	.676	.731
ReviewCount	.608	.015	.000	.580	.637
EIS	.722	.014	.000	.696	.749
FriendsCount	.685	.015	.000	.656	.713
Elite	.501	.015	.954	.472	.530
NetworkE	.683	.015	.000	.655	.712
NetworkV	.678	.015	.000	.649	.706

a. Under the nonparametric assumption

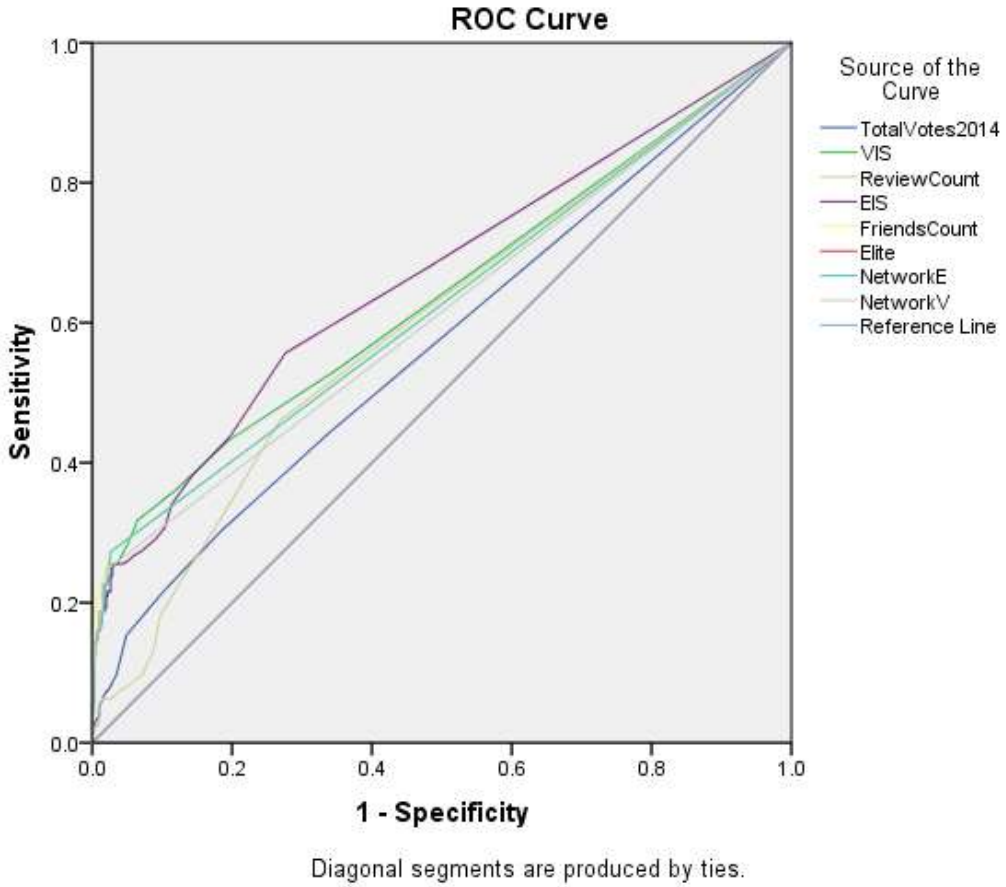
b. Null hypothesis: true area = 0.5

Group D: June 2014 Change in Friends

Case Processing Summary

ChangeInFriends[a]	Valid N (listwise)
Positive[b]	176
Negative	492

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.570	.026	.006	.518	.621
VIS	.637	.027	.000	.585	.689
ReviewCount	.599	.026	.000	.549	.649
EIS	.663	.026	.000	.613	.713
FriendsCount	.625	.027	.000	.573	.678
Elite	.500	.025	1.000	.450	.550
NetworkE	.624	.027	.000	.572	.677
NetworkV	.614	.027	.000	.562	.667

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

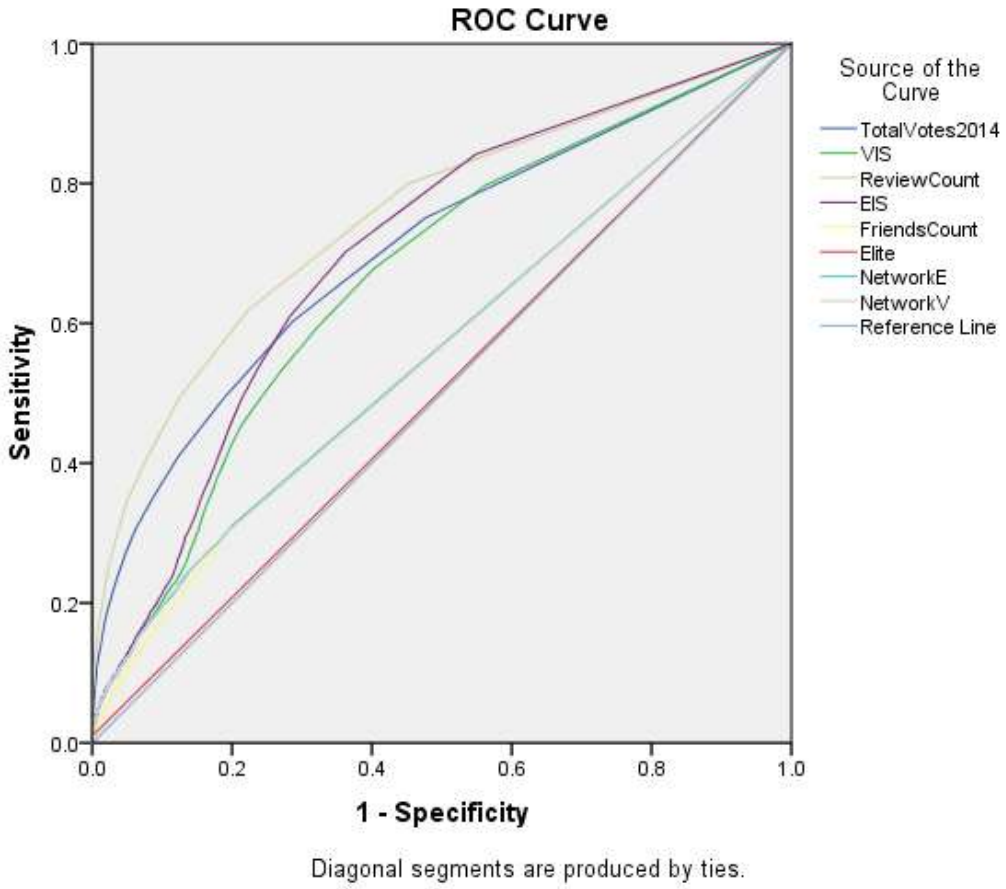
Appendix D: AUC ROC Data for Change in Votes

Group A: July 2013–July 2014 Change in Votes

Case Processing Summary

ChangeInVotes	Valid N (listwise)
Positive[a]	20024
Negative	7412

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.705	.003	.000	.698	.711
VIS	.668	.004	.000	.661	.676
ReviewCount	.754	.003	.000	.748	.760
EIS	.702	.004	.000	.695	.709
FriendsCount	.558	.004	.000	.551	.566
Elite	.505	.004	.210	.497	.513
NetworkE	.562	.004	.000	.554	.569
NetworkV	.561	.004	.000	.553	.568

a. Under the nonparametric assumption

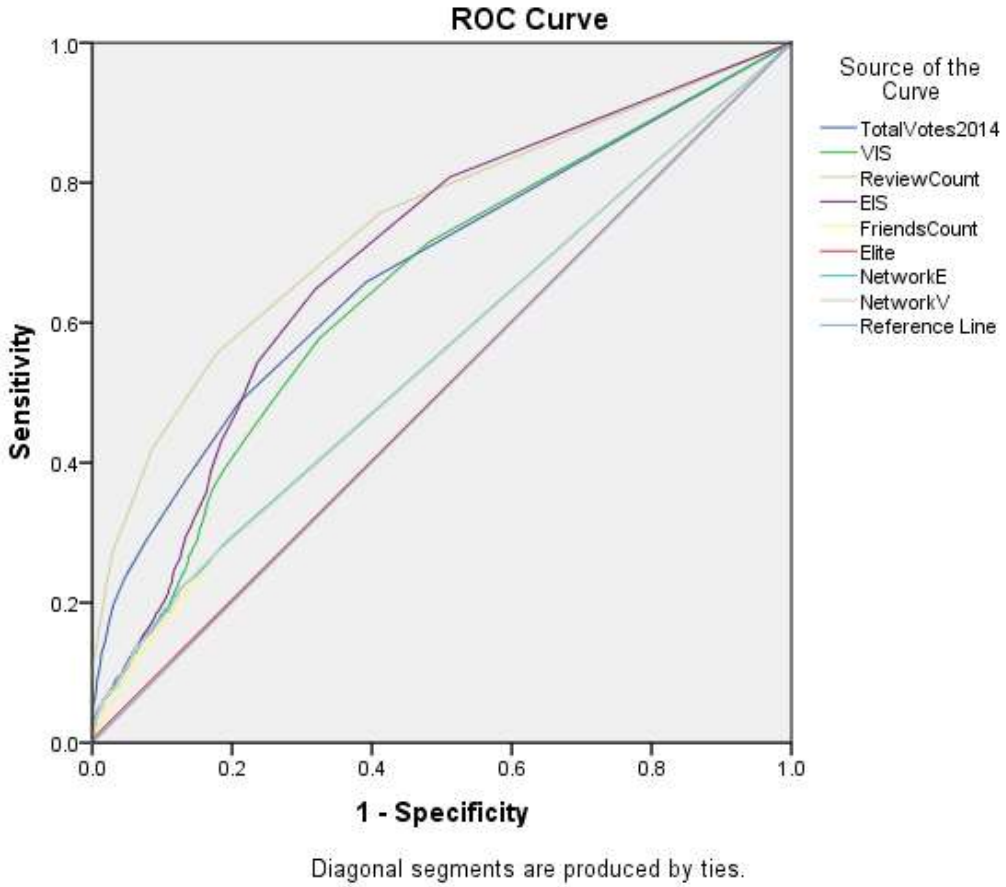
b. Null hypothesis: true area = 0.5

Group B: February 2014–July 2014 Change in Votes

Case Processing Summary

ChangeInVotes[a]	Valid N (listwise)
Positive[b]	7133
Negative	2425

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.675	.006	.000	.664	.687
VIS	.648	.006	.000	.635	.660
ReviewCount	.740	.005	.000	.730	.751
EIS	.694	.006	.000	.681	.706
FriendsCount	.550	.007	.000	.537	.563
Elite	.502	.007	.718	.489	.516
NetworkE	.553	.006	.000	.540	.565
NetworkV	.551	.007	.000	.538	.564

a. Under the nonparametric assumption

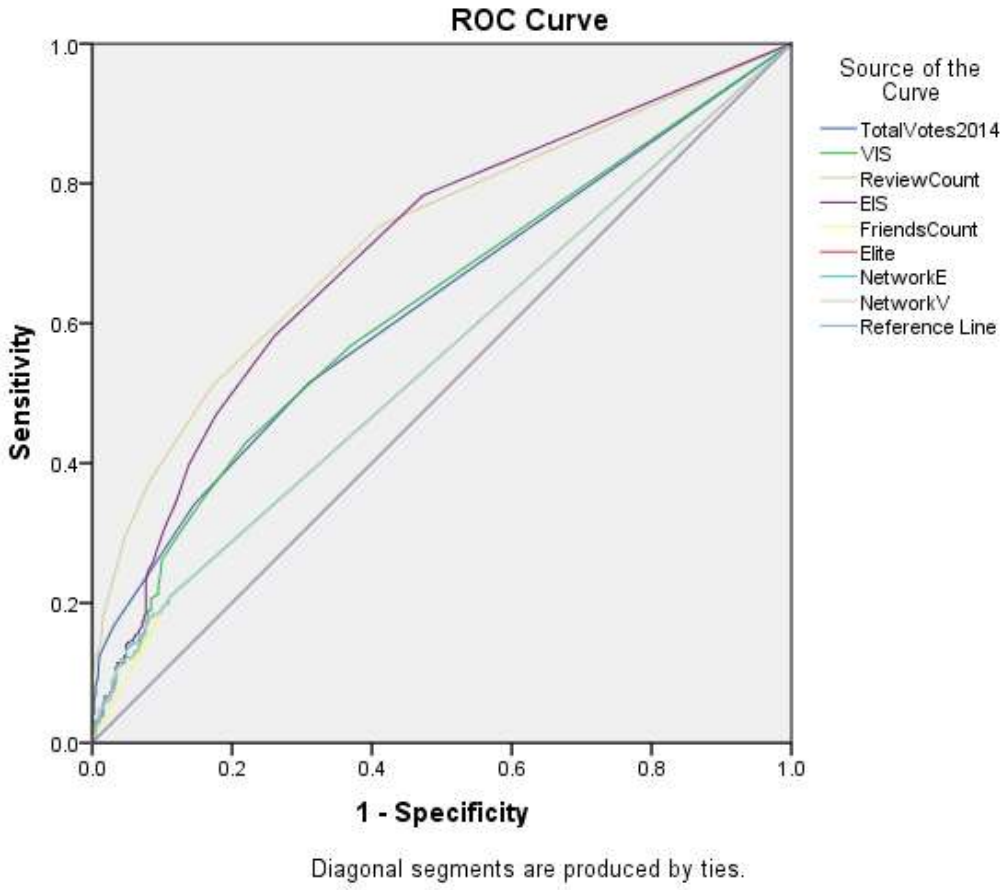
b. Null hypothesis: true area = 0.5

Group C: June 2014–July 2014 Change in Votes

Case Processing Summary

ChangeInVotes[a]	Valid N (listwise)
Positive[b]	1364
Negative	391

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.626	.015	.000	.597	.655
VIS	.623	.015	.000	.593	.652
ReviewCount	.723	.013	.000	.697	.749
EIS	.703	.015	.000	.674	.732
FriendsCount	.549	.016	.003	.518	.580
Elite	.500	.017	.982	.468	.533
NetworkE	.551	.016	.002	.520	.582
NetworkV	.550	.016	.003	.519	.581

a. Under the nonparametric assumption

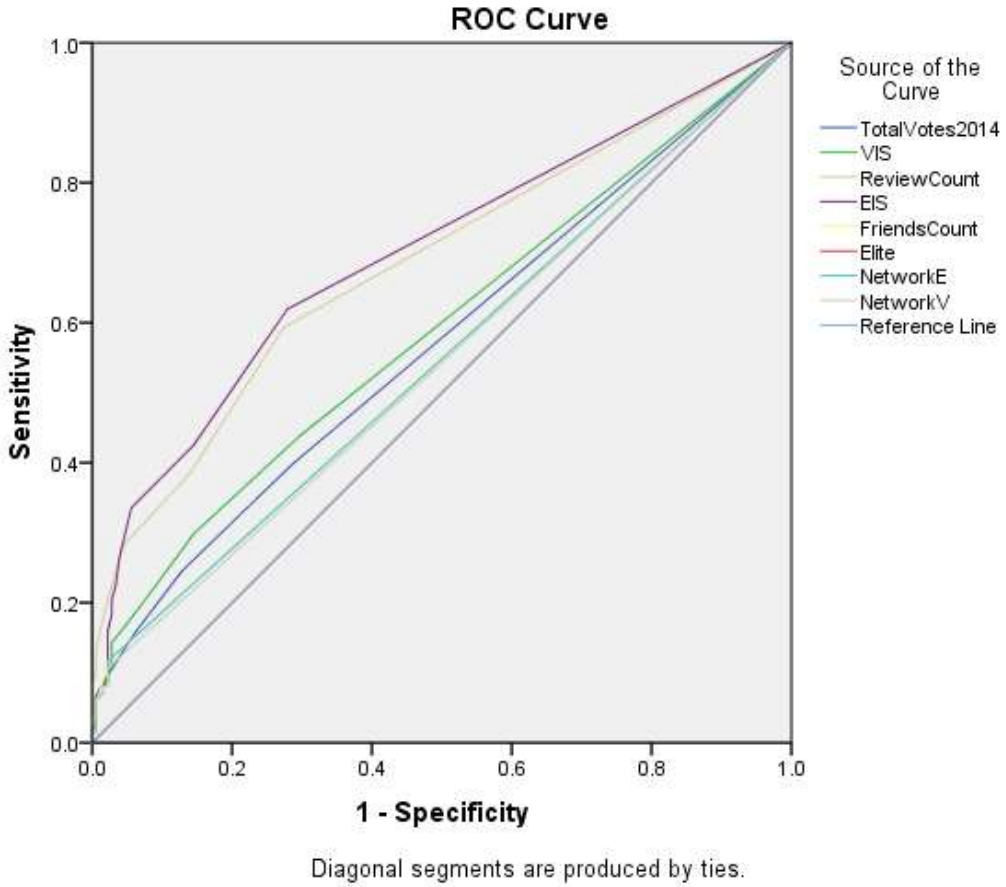
b. Null hypothesis: true area = 0.5

Group D: July 2014 Change in Votes

Case Processing Summary

ChangeInVotes[a]	Valid N (listwise)
Positive[b]	489
Negative	179

The positive actual state is 1.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TotalVotes2014	.570	.024	.006	.523	.616
VIS	.589	.023	.000	.543	.635
ReviewCount	.684	.021	.000	.643	.726
EIS	.698	.021	.000	.656	.739
FriendsCount	.547	.024	.060	.500	.595
Elite	.500	.025	1.000	.451	.549
NetworkE	.547	.024	.062	.500	.594
NetworkV	.542	.024	.095	.495	.590

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Appendix E: AUC Values of Elite

		Change in Friends	Change in Votes
Time Yelping	Group A: 1 Year	0.510	0.505
	Group B: 6 Months	0.505	0.502
	Group C: 2 Months	0.501	0.500
	Group D: 1 Month	0.500	0.500

Appendix F: Logit Model Data for Change in Friends

Group A: July 2013–July 2014 Change in Friends Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	20938.554[a]	.166	.231
2	20716.416[a]	.176	.244
3	19947.463[b]	.208	.289

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInFriends		Percentage	ChangeInFriends		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInFriends	0	11213	1599	87.5	4809	700	87.3
		1	3044	3346	52.4	1297	1428	52.4
	Overall Percentage				75.8			75.7
Step 2	ChangeInFriends	0	11220	1592	87.6	4806	703	87.2
		1	2970	3420	53.5	1289	1436	52.7
	Overall Percentage				76.2			75.8
Step 3	ChangeInFriends	0	11857	955	92.5	5095	414	92.5
		1	3431	2959	46.3	1467	1258	46.2
	Overall Percentage				77.2			77.2

a. The cut value is .500

b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

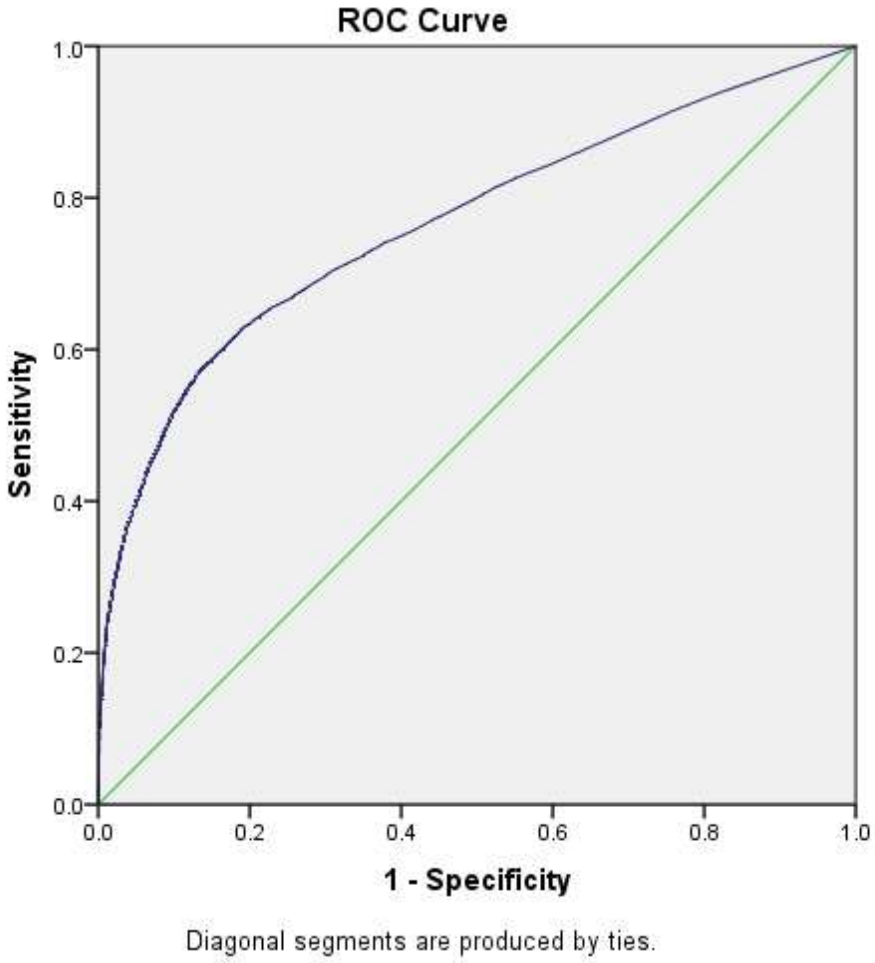
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0001271284	.000	2863.771	1	.000	1.000
	Constant	.8908114918	.032	783.501	1	.000	2.437
Step 2[b]	RVIS	-.0000582682	.000	217.598	1	.000	1.000
	REIS	-.0000776203	.000	376.143	1	.000	1.000
	Constant	1.0139425975	.033	934.411	1	.000	2.756
Step 3[c]	FriendsCount	.2603271016	.013	386.927	1	.000	1.297
	RVIS	-.0000338209	.000	69.301	1	.000	1.000
	REIS	-.0000531895	.000	168.249	1	.000	1.000
	Constant	.1525177309	.048	10.094	1	.001	1.165

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: RVIS.

c. Variable(s) entered on step 3: FriendsCount.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.766	.003	.000	.760	.772

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group B: February 2014–July 2014 Change in Friends Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7568.978[a]	.145	.201
2	7056.366[b]	.208	.288
3	7046.858[b]	.210	.289
4	7042.099[b]	.210	.290

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInFriends		Percentage	ChangeInFriends		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInFriends	0	3823	555	87.3	1674	228	88.0
		1	1153	1156	50.1	482	487	50.3
	Overall Percentage				74.5			75.3
Step 2	ChangeInFriends	0	4098	280	93.6	1789	113	94.1
		1	1375	934	40.5	565	404	41.7
	Overall Percentage				75.3			76.4
Step 3	ChangeInFriends	0	4064	314	92.8	1769	133	93.0
		1	1342	967	41.9	546	423	43.7
	Overall Percentage				75.2			76.3
Step 4	ChangeInFriends	0	4093	285	93.5	1781	121	93.6
		1	1361	948	41.1	557	412	42.5
	Overall Percentage				75.4			76.4

a. The cut value is .500

b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

Variables in the Equation

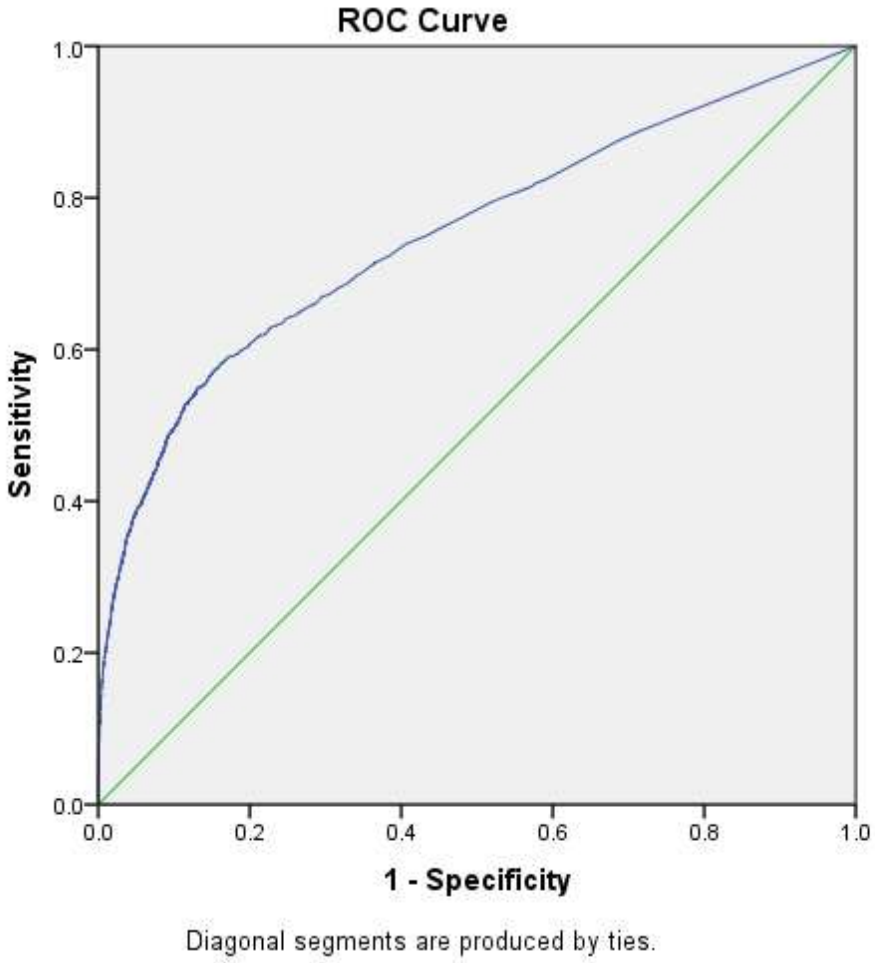
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0003309081	.000	893.550	1	.000	1.000
	Constant	.8192623055	.053	238.184	1	.000	2.269
Step 2[b]	FriendsCount	.5475344134	.035	243.104	1	.000	1.729
	REIS	-.0001456740	.000	110.370	1	.000	1.000
	Constant	-.3629231462	.080	20.823	1	.000	.696
Step 3[c]	FriendsCount	.5157880821	.036	206.845	1	.000	1.675
	REIS	-.0001116733	.000	40.056	1	.000	1.000
	RVIS	-.0000544710	.000	9.487	1	.002	1.000
	Constant	-.2483583221	.087	8.085	1	.004	.780
Step 4[d]	FriendsCount	.5334898008	.038	202.372	1	.000	1.705
	REIS	-.0000960457	.000	25.190	1	.000	1.000
	ReviewCount	.0097098204	.005	4.371	1	.037	1.010
	RVIS	-.0000529210	.000	8.915	1	.003	1.000
	Constant	-.3823552653	.108	12.436	1	.000	.682

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: FriendsCount.

c. Variable(s) entered on step 3: RVIS.

d. Variable(s) entered on step 4: ReviewCount.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.751	.006	.000	.740	.762

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group C: June 2014–July 2014 Change in Friends Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1328.990[a]	.160	.223
2	1269.962[b]	.200	.279

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted				
			Selected Cases[b]			Unselected Cases[c]	
			ChangeInFriends		Percentage	ChangeInFriends	Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInFriends	0	726	98	88.1	292	52	84.9
		1	208	190	47.7	107	82	43.4
	Overall Percentage				75.0			70.2
Step 2	ChangeInFriends	0	782	42	94.9	315	29	91.6
		1	241	157	39.4	118	71	37.6
	Overall Percentage				76.8			72.4

a. The cut value is .500

b. Selected cases InModel EQ 1

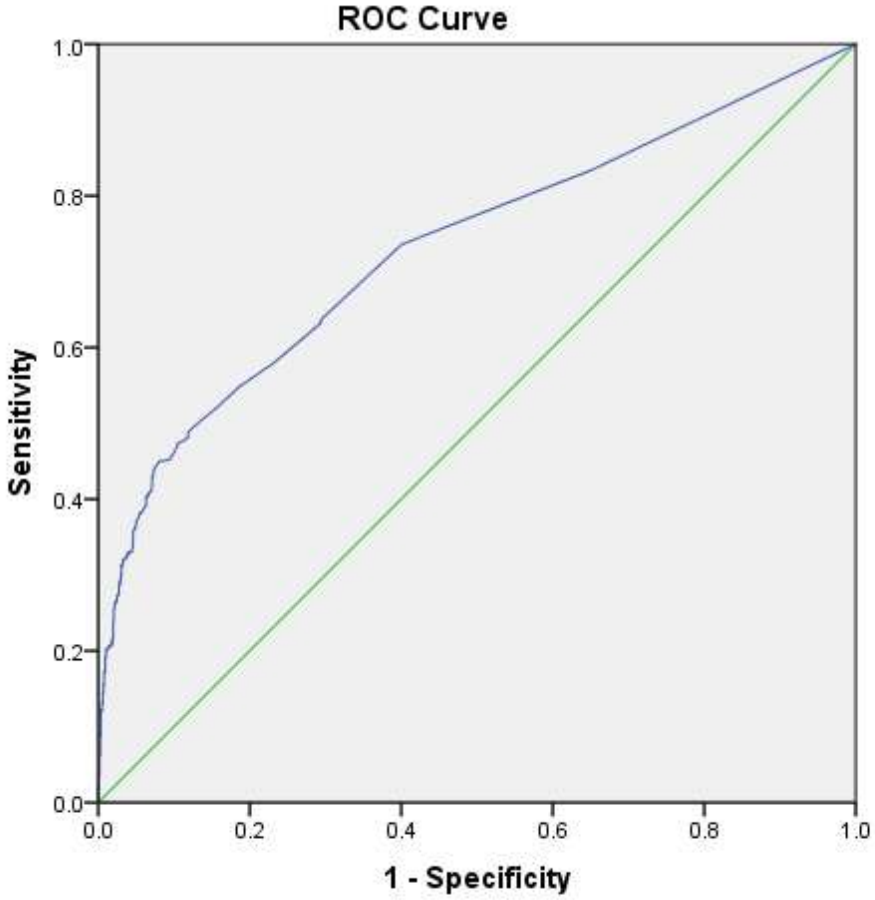
c. Unselected cases InModel NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0019462906	.000	177.568	1	.000	.998
	Constant	.8406042788	.126	44.251	1	.000	2.318
Step 2[b]	FriendsCount	.5009021471	.090	31.233	1	.000	1.650
	REIS	-.0012498318	.000	53.680	1	.000	.999
	Constant	.0615672203	.168	.134	1	.714	1.064

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: FriendsCount.



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.731	.014	.000	.704	.757

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group D: July 2014 Change in Friends Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	516.783[a]	.078	.114
2	508.655[a]	.093	.136
3	481.775[b]	.143	.208
4	482.455[b]	.141	.207

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted				
		Selected Cases[b]			Unselected Cases[c]		
		ChangeInFriends		Percentage	ChangeInFriends		Percentage
	Observed	0	1	Correct	0	1	Correct
Step 1	ChangeInFriends 0	354	1	99.7	137	0	100.0
	1	118	9	7.1	45	4	8.2
	Overall Percentage			75.3			75.8
Step 2	ChangeInFriends 0	338	17	95.2	136	1	99.3
	1	94	33	26.0	37	12	24.5
	Overall Percentage			77.0			79.6
Step 3	ChangeInFriends 0	345	10	97.2	136	1	99.3
	1	95	32	25.2	37	12	24.5
	Overall Percentage			78.2			79.6
Step 4	ChangeInFriends 0	344	11	96.9	136	1	99.3
	1	95	32	25.2	37	12	24.5
	Overall Percentage			78.0			79.6

a. The cut value is .500

b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

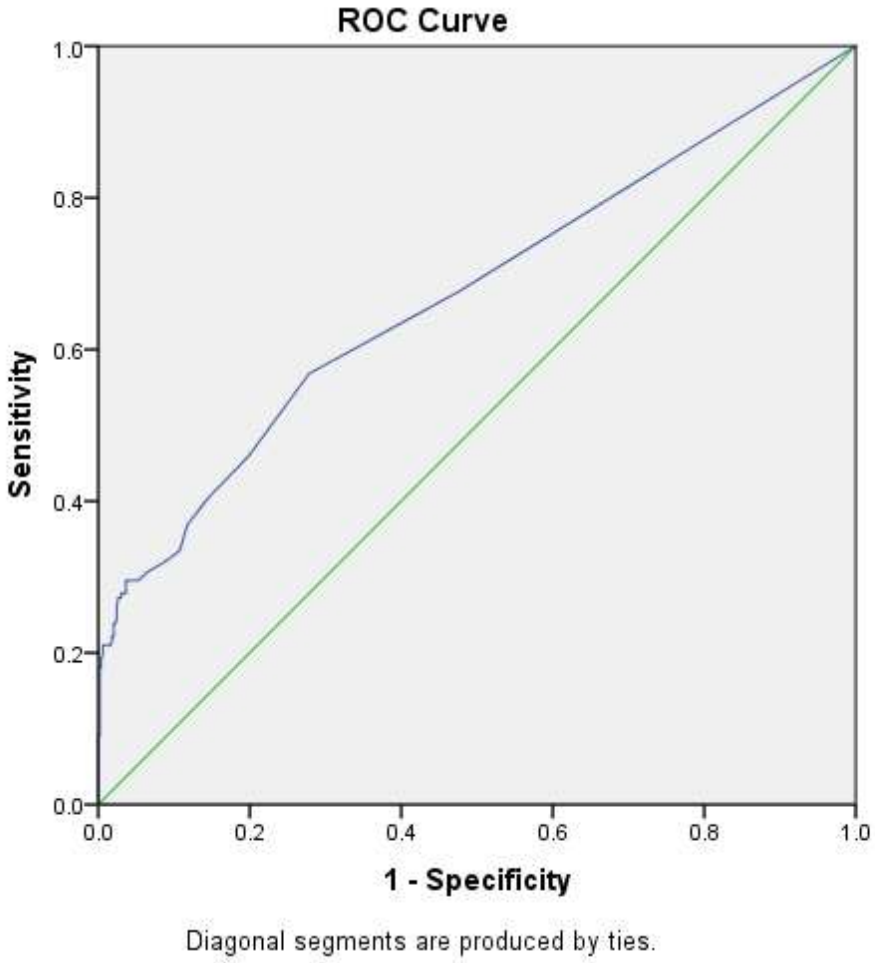
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0035727741	.001	36.514	1	.000	.996
	Constant	.0527238661	.195	.073	1	.787	1.054
Step 2[b]	RVIS	-.0019281100	.001	8.135	1	.004	.998
	REIS	-.0026986983	.001	16.763	1	.000	.997
	Constant	.3923646857	.230	2.906	1	.088	1.480
Step 3[c]	FriendsCount	.8685261301	.240	13.086	1	.000	2.383
	RVIS	-.0006236715	.001	.686	1	.408	.999
	REIS	-.0016442703	.001	5.400	1	.020	.998
	Constant	-.5345838382	.310	2.983	1	.084	.586
Step 4[c]	FriendsCount	.9251009198	.234	15.564	1	.000	2.522
	REIS	-.0018286586	.001	7.388	1	.007	.998
	Constant	-.6895643248	.249	7.700	1	.006	.502

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: RVIS.

c. Variable(s) entered on step 3: FriendsCount.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.671	.026	.000	.620	.721

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Appendix G: Logit Model Data for Change in Votes

Group A: July 2013–July 2014 Change in Votes Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	20476.907[a]	.097	.141
2	19167.385[b]	.156	.227
3	19105.521[b]	.159	.231
4	19105.538[b]	.159	.231
5	19086.676[c]	.160	.232
6	19082.186[b]	.160	.232

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.
b. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.
c. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInVotes		Percentage	ChangeInVotes		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2
Step 2	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2
Step 3	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2
Step 4	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2
Step 5	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2
Step 6	ChangeInVotes	0	0	5202	.0	0	2210	.0
		1	1	13999	100.0	0	6024	100.0
	Overall Percentage				72.9			73.2

a. The cut value is .500

b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0000984447	.000	1710.660	1	.000	1.000
	Constant	2.4729021670	.043	3385.594	1	.000	11.857
Step 2[b]	ReviewCount	.3074668040	.011	735.538	1	.000	1.360
	REIS	-.0000259711	.000	81.822	1	.000	1.000
	Constant	.4051023622	.068	35.336	1	.000	1.499
Step 3[c]	ReviewCount	.3282739135	.012	751.803	1	.000	1.389
	RVIS	-.0000303521	.000	61.968	1	.000	1.000
	REIS	.0000005844	.000	.017	1	.896	1.000
	Constant	.4012667856	.069	33.912	1	.000	1.494
Step 4[c]	ReviewCount	.3274639014	.010	1020.582	1	.000	1.387
	RVIS	-.0000299673	.000	143.260	1	.000	1.000
	Constant	.4066893162	.055	54.373	1	.000	1.502
Step 5[d]	TotalVotes2014	.0269514366	.007	16.201	1	.000	1.027
	ReviewCount	.3135801627	.011	855.406	1	.000	1.368
	RVIS	-.0000242109	.000	72.681	1	.000	1.000
	Constant	.2931586201	.061	22.894	1	.000	1.341
Step 6[e]	TotalVotes2014	.0348592645	.008	19.320	1	.000	1.035
	ReviewCount	.2948504568	.014	455.613	1	.000	1.343
	RVIS	-.0000154248	.000	9.295	1	.002	1.000
	REIS	-.0000108280	.000	4.467	1	.035	1.000
	Constant	.3602113295	.069	27.068	1	.000	1.434

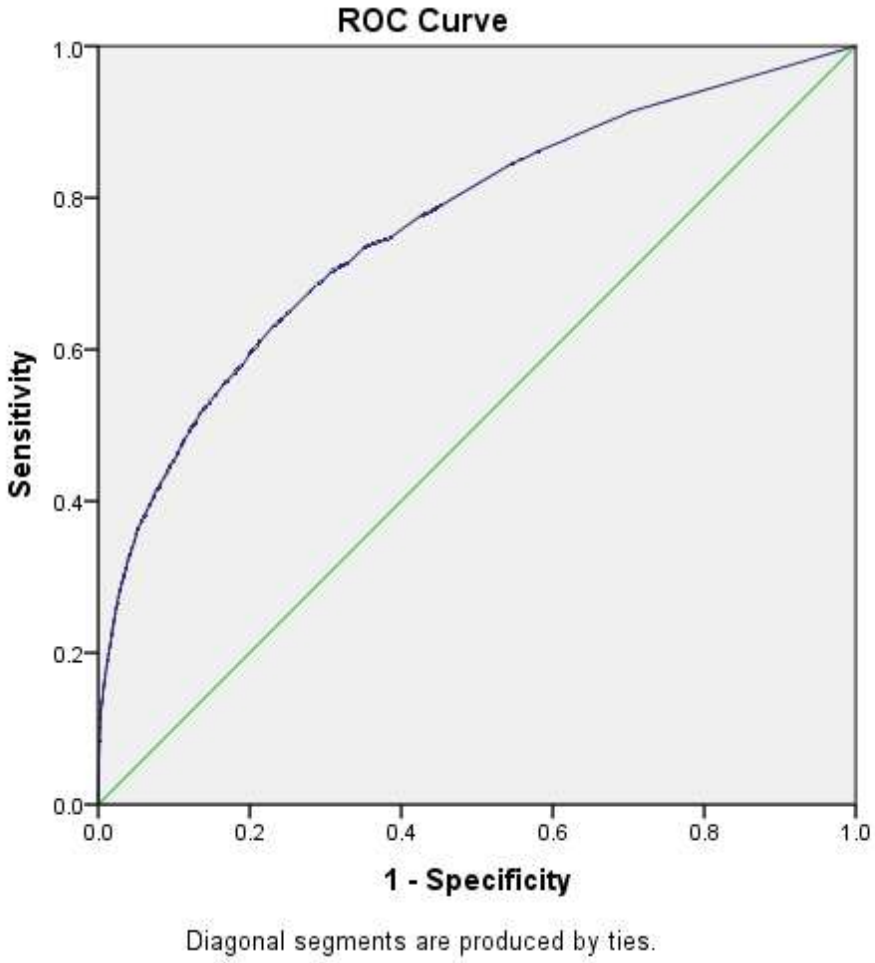
a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: ReviewCount.

c. Variable(s) entered on step 3: RVIS.

d. Variable(s) entered on step 5: TotalVotes2014.

e. Variable(s) entered on step 6: REIS.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.763	.003	.000	.757	.769

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group B: February 2014–July 2014 Change in Votes Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6926.211[a]	.087	.129
2	6517.514[b]	.141	.209
3	6480.456[c]	.146	.216

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

c. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInVotes		Percentage	ChangeInVotes		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInVotes	0	0	1679	.0	0	746	.0
		1	0	5008	100.0	0	2125	100.0
	Overall Percentage				74.9			74.0
Step 2	ChangeInVotes	0	0	1679	.0	0	746	.0
		1	0	5008	100.0	0	2125	100.0
	Overall Percentage				74.9			74.0
Step 3	ChangeInVotes	0	0	1679	.0	0	746	.0
		1	0	5008	100.0	0	2125	100.0
	Overall Percentage				74.9			74.0

a. The cut value is .500

b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

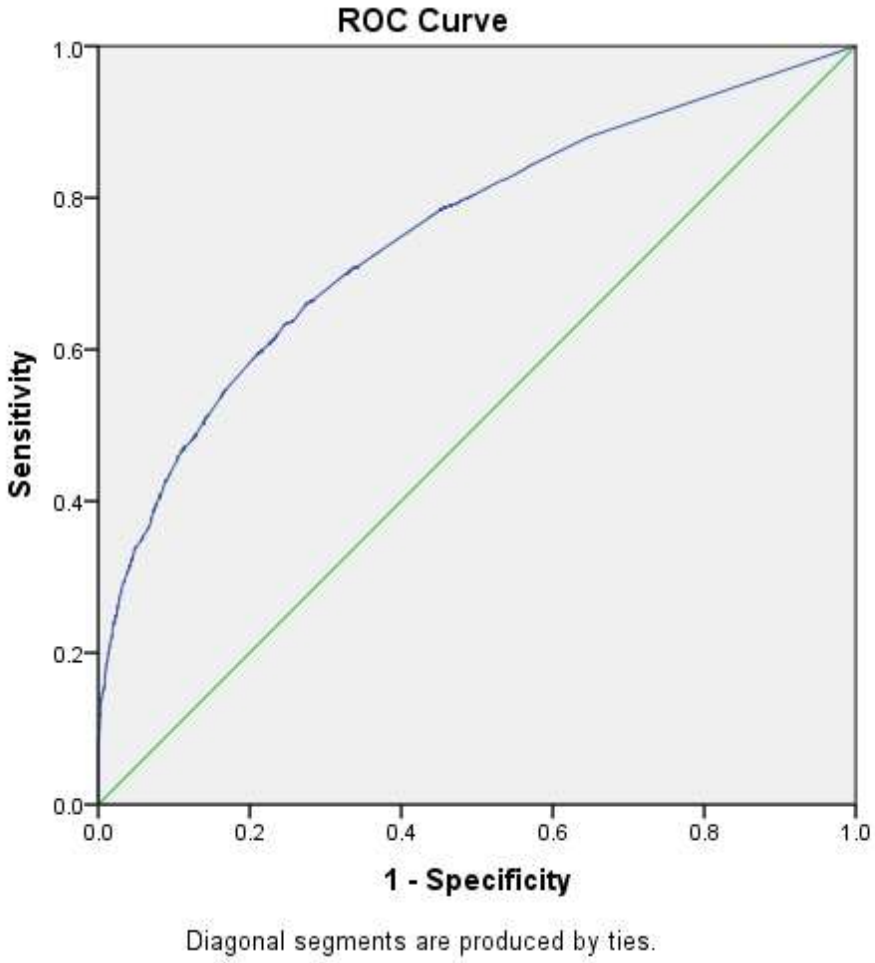
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0002747857	.000	535.204	1	.000	1.000
	Constant	2.5391527178	.074	1169.868	1	.000	12.669
Step 2[b]	REIS	-.0000697580	.000	24.014	1	.000	1.000
	ReviewCount	.3824356854	.025	240.588	1	.000	1.466
	Constant	.4346533886	.121	12.807	1	.000	1.544
Step 3[c]	REIS	-.0000619633	.000	18.637	1	.000	1.000
	TotalVotes2014	.0832226264	.016	27.979	1	.000	1.087
	ReviewCount	.3454507659	.025	185.556	1	.000	1.413
	Constant	.3567740047	.123	8.453	1	.004	1.429

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: ReviewCount.

c. Variable(s) entered on step 3: TotalVotes2014.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.752	.005	.000	.742	.762

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group C: June 2014–July 2014 Change in Votes Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1169.651[a]	.091	.139
2	1134.142[b]	.117	.179

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInVotes		Percentage	ChangeInVotes		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInVotes	0	0	268	.0	0	123	.0
		1	0	954	100.0	0	410	100.0
	Overall Percentage				78.1			76.9
Step 2	ChangeInVotes	0	0	268	.0	0	123	.0
		1	0	954	100.0	0	410	100.0
	Overall Percentage				78.1			76.9

a. The cut value is .500

b. Selected cases InModel EQ 1

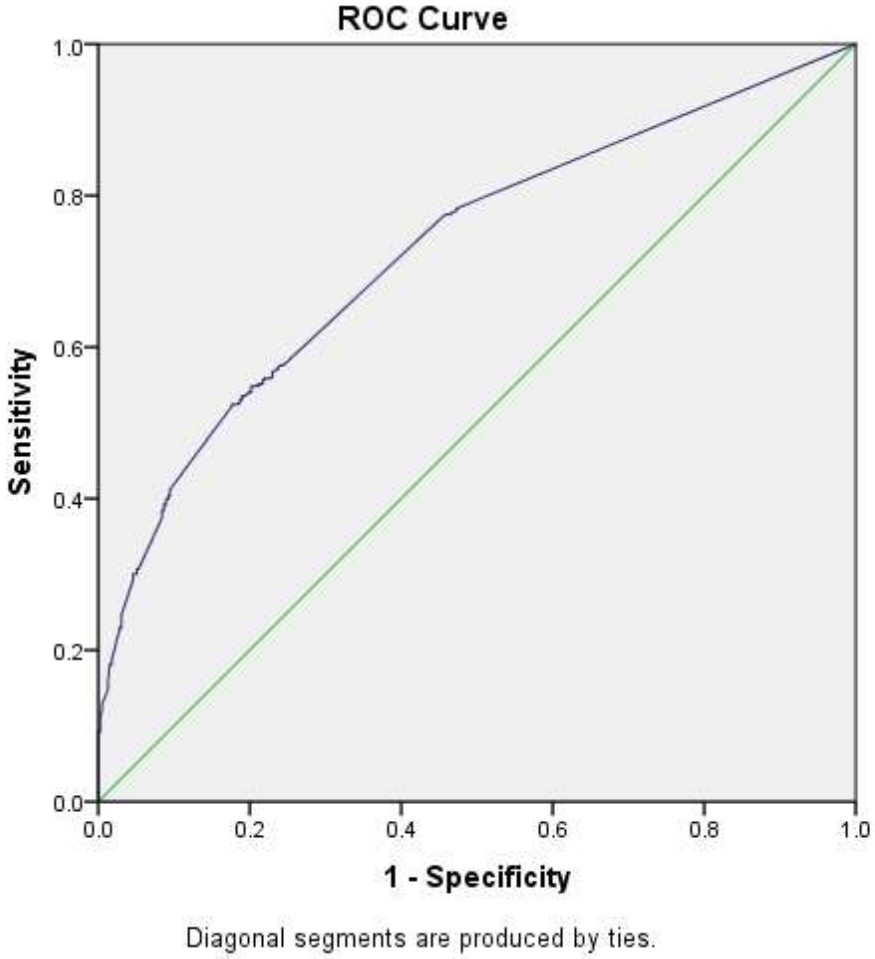
c. Unselected cases InModel NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0016339687	.000	98.383	1	.000	.998
	Constant	2.8963633852	.195	220.614	1	.000	18.108
Step 2[b]	ReviewCount	.3410057297	.070	23.913	1	.000	1.406
	REIS	-.0005574559	.000	6.114	1	.013	.999
	Constant	.9596715970	.363	6.980	1	.008	2.611

a. Variable(s) entered on step 1: REIS.

b. Variable(s) entered on step 2: ReviewCount.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.727	.013	.000	.702	.753

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Group D: July 2014 Change in Votes Logit Model (Forward Conditional)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	486.454[a]	.141	.206

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			ChangeInVotes		Percentage	ChangeInVotes		Percentage
	Observed		0	1	Correct	0	1	Correct
Step 1	ChangeInVotes	0	0	129	.0	0	50	.0
		1	0	353	100.0	0	136	100.0
	Overall Percentage				73.2			73.1

a. The cut value is .500

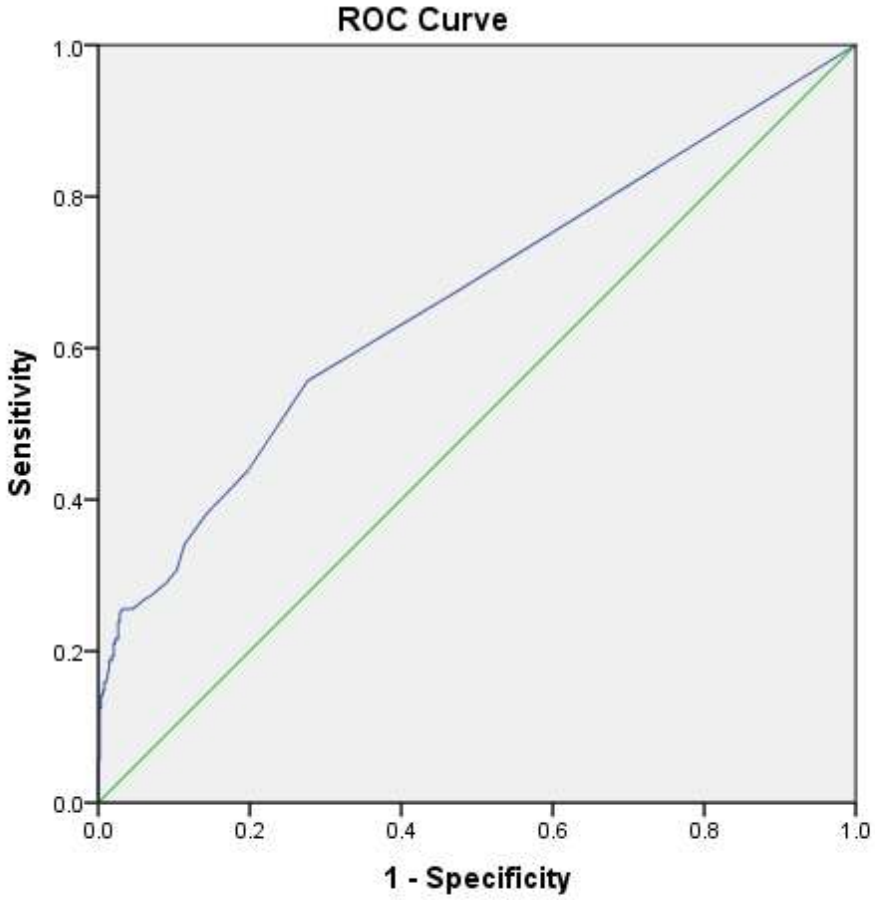
b. Selected cases InModel EQ 1

c. Unselected cases InModel NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REIS	-.0055873036	.001	55.907	1	.000	.994
	Constant	3.0990953153	.331	87.897	1	.000	22.178

a. Variable(s) entered on step 1: REIS.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.663	.026	.000	.613	.713

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5