ESSAYS ON OPERATIONAL PROBLEMS IN DIGITAL ECONOMY

A Dissertation

by

RAKESH REDDY MALLIPEDDI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Chelliah Sriskandajarah
Co-Chair of Committee,	Subodha Kumar
Committee Members,	Arvind Mahajan
	Bruce A. McCarl
	Rogelio Oliva
	Bala Shetty
Head of Department,	Richard D. Metters

May 2019

Major Subject: Business Administration

ABSTRACT

In this dissertation, I investigate operational issues in the context of online social networks and digital economy. The first essay analyzes the phenomenon of open technology in the context of resource allocation. In this study, based on evidence from prior literature and current business practices, I develop optimal control models to determine the optimal extent of technology openness and firm's effort levels for maintaining an existing technology and developing a newer version of the technology. I derive and discuss important insights and shed light on the business practices of major technology firms. In the second essay, I develop a data-driven prescriptive framework for conducting an influencer marketing campaign on online social networks. Influencer marketing involves hiring influencers to promote products on behalf of a firm. The effectiveness of an influencer marketing campaign depends on choosing the right set of influencers for seeding and scheduling of ads on social media platforms. I first develop an optimization model to select influencers and then propose a model to schedule their posts on social media. Next, I develop a polynomial time heuristic that provides a near-optimal solution for selecting influencers. Next, using actual data from a popular social network, I demonstrate the superior performance of our selection model against current industry practices. Finally, the third essay empirically analyzes the effects of social media content created by influencers on audience engagement. In particular, I focus on the tone of an influencer's content, an important emotional facet that plays a vital role in determining whether an audience engages with the content. Our results demonstrate that the tone of social media content affects the engagement levels of an influencer with the audience. In addition, the findings from this study establish the moderating role of an influencer's popularity and the tone of related brands on the relationship between an influencer's tone and engagement. The results of these essays provide prescriptive solutions that are easy to implement and several important managerial insights.

DEDICATION

To my Grandparents, Parents, Wife, and Daughter.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance of my advisors, dissertation committee, other faculty members in the department of information and operations management, and my family.

I would like to thank Prof. Subodha Kumar and Prof. Chelliah Sriskandarajah for their excellent supervision, guidance, encouragement, and patience throughout my doctoral education. They have given me all the necessary tools to have a successful career in academia. I feel honored and lucky to have them as my advisors and role models.

I wish to sincerely thank to Prof. Rogelio Oliva for his comments, thoughtful insights, and introducing me to behavioral science. I wish to thank Prof. Bala Shetty for constantly guiding me and motivating me throughout my time at Texas A&M University. I would like to thank Prof. Arvind Mahajan of the Department of Finance for his constant encouragement and insightful comments. I express my gratitude and thanks to Prof. Bruce McCarl of the Department of Agricultural Economics for stimulating my interest in optimization.

I also wish to thank Prof. Gregory Heim for his seminar on empirical research and constantly encouraging me and other PhD students. I would like to thank Prof. Michael Ketzenberg for his mentorship and his seminar class, which I thoroughly enjoyed. I would like to acknowledge Prof. Neil Geismar for his class and organizing the seminars, which helped me improve my research. I would like to express my sincere gratitude to Prof. Richard Metters, our department head, for his leadership and making the lives of PhD students enjoyable. Furthermore, I wish to thank all the faculty members of the INFO department for all the help that I received from them. Also, thanks to Serkan Akturk, Xinghzhi Jia, and Seokjun Youn for being such wonderful friends during my PhD. Finally, thanks to Ms. Donna Shumaker, Ms. Tammy Louther, and Ms. Veronica Stilley for all the wonderful things they do on a daily basis to make the lives of PhD students easier.

Lastly, I would like to thank my parents for their never ending support. Special thanks

to my wife, Sukrutha, and my daughter, Diya, who stood by me during all the tough times and provided all the support that I needed to finish this dissertation.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This dissertation work was primarily conducted by me with the support of my dissertation committee. I also acknowledge the inputs and suggestions of Prof. Emre Demirezen on Essay 1. Prof. Yunxia Zhu has provided valuable comments and suggestions while working on Essay 2. Finally, I sincerely thank Prof. Ramkumar Janakiraman for his valuable insights for Essay 3.

The data analyzed for Chapter 2 was provided by Professors Subodha Kumar and Chelliah Sriskandarajah. The data analyzed for Chapter 3 was provided by Professor Subodha Kumar.

All other work conducted for the dissertation was completed by the student under the supervision of the dissertation committee.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The rapid digitization of businesses presents several new challenges, yet, research in this literature stream is scarce. In this dissertation, I specifically focus on online social networking platforms such as Twitter and GitHub that have enabled firms to access massive social networks and directly interact with the users in these networks, thereby transforming how firms operate their businesses. My dissertation, through three essays, studies how firms can effectively utilize social networks to manage technology development, improve their service and marketing operations, and attract customers to their products and services.

In the first essay, we analytically examine the phenomenon of open innovation in the context of resource allocation between maintaining an existing technology and developing a newer version of the technology. Open innovation is an approach where firms open-up their technology to its customers, suppliers, and competitors and rely on them to spur innovation. We develop an optimal control model to determine the optimal extent of technology openness and effort levels. In our analysis, we consider various factors including: (i) the characteristics of the market (e.g., valuation of the technology, market competitiveness, network effects, and market expectations from the next version of the technology), (ii) the characteristics of the technology firm (e.g., costs for developing technology and maintaining existing technology), and (iii) the effects of making the technology fully or partially open source on the current and next versions of the technology. Based on the results, we derive and discuss several important insights and shed light into the business practices of different technology firms. Some of our key findings include: (i) the firm might keep its technology proprietary even in conditions that seem to favor an open source environment, and (ii) if the technology becomes more valuable, the technology firm might decide to decrease the level of effort for fixing defects but rely more on the open source technology community. Finally, we derive several managerially relevant insights and explain possible rationale on current business practices of major technology firms.

Explosive growth in the number of users on various social media platforms has transformed the way firms strategize their marketing activities. To take advantage of the vast size of social networks, firms have now turned their attention to influencer marketing. Influencer marketing strategy involves seeding (sending) ads through opinion leaders or influencers to promote a firm's products to an online audience. The effectiveness of an influencer marketing campaign depends on choosing the right set of influencers and scheduling of ads to be posted by influencers. While the topic of scheduling of ads in other mediums such as television, internet, and mobile has been extensively studied in the management sciences literature, the problem of scheduling of ads in social networks has not been rigorously studied in the literature. In this paper, we develop a data-driven framework to help a firm successfully conduct an influencer marketing campaign. We decompose the problem into two phases: (i) *selection of influencers*, and (ii) *scheduling of influencers'* ads. For the first phase, we develop an optimization model to select influencers for an ad campaign. We validate the assumptions that we make in the model for the first phase by conducting an empirical analysis using real data from Twitter. We then present structural properties for our model. Using these properties, we develop a polynomial time heuristic to provide near-optimal solutions to the problem of selecting influencers. For the second phase, we present a model to help firms with scheduling of ads to be posted by influencers (selected in the first phase) over a planning horizon in order to maximize the effectiveness of the campaign. Finally, using the results from extensive numerical experiments, we develop and present several managerially relevant insights to improve marketing operations of a firm.

In the third essay, I empirically analyze the operational effects of social media content created by influencers on audience engagement. Influencers or human brands in fields ranging from sports to art to politics use social media platforms to connect and engage with their audience. We analyze the effects of content generated by human brands (referred to as human generated content) in the popular social media platform, Twitter, on audience engagement. The first objective of this study is to examine the effect of tone (positive versus negative) of human generated content on the social media engagement measured by number of retweets. Next, we investigate the role of popularity of human brand on engagement. Furthermore, as human brands are often associated with a group, our third objective is to investigate the spillover effects of the content generated by human brands. In particular, we empirically examine if the content of brands associated with a focal human brand affects the engagement of the focal human brand. Set in the context of Indian general election 2014, our analysis based on a mixed effects model that accounts for user and group level heterogeneity reveals that the tone of social media content created by human brands significantly affects audience engagement. Specifically, we find that the main effect of negative toned content on social media engagement is significantly positive. In contrast, we find that the effect of positive toned content on engagement is greater only for popular candidates. Furthermore, our results also suggest that the popularity of human brand and the tone of related brands moderate the relationship between negative tone content and engagement. We supplement our core results with additional analyses that include alternative model specifications, estimation strategies, engagement metrics, and models that account for selection bias. Finally, we present implications of this study to practice as well as the literature on social media content.

In summary, my dissertation, which is motivated by current business practices, makes several theoretical and managerial contributions. I utilize various methodological approaches including optimal control theory, combinatorial optimization, and empirical modeling to generate actionable insights for firms and provide prescriptive solutions for problems that have emerged with the digitization of businesses.

The rest of this dissertation is organized as follows. The first essay of my dissertation is presented in Chapter 2 of the dissertation whereas essays two and three are presented in Chapters 3 and 4 respectively. Formal proofs and additional analyses are presented in the Appendix.

2. ESSAY 1: HOW MUCH TO OPEN, HOW FAST TO FIX AND DEVELOP? – IMPACTS OF TECHNOLOGY OPENNESS

2.1 Introduction

Firms are increasingly opening their proprietary technologies to accelerate innovation through both internal and external resources such as suppliers, customers, and competitors (Terwiesch and Xu 2008). The approach of opening up proprietary technologies to external resources and relying on them is often referred to as *open innovation*. Dominant technology firms have taken initiatives to make a portion of their technology open, and open innovation has become a key part of their strategies to provide more robust and secure products and services (e.g., Microsoft 2014; Apple 2015). Although open technology is widely popular in the software industry, firms from other industries such as automobile (e.g., Toyota and Tesla), aerospace (e.g., Airbus), and 3-D printing (e.g., MakerBot) have also embraced strategies that relate to open sourcing or open technology. In this study, we refer to technology as the design of a tangible (e.g., batteries and pharmaceutical drugs) or an intangible product (e.g., software and services).

Open technology has several advantages including (i) higher demand for the product because of network effects, (ii) faster improvement in quality, and (iii) lower costs related to maintenance of the current technology (i.e., fixing defects and perfecting the current technology) and development (i.e., upgrading the current technology or innovation of new technology) (Lerner and Tirole 2002; West 2003). On the other hand, opening the technology may increase market competition (the market competition refers to the ability of competitors in copying unique features of the technology) and the costs required to manage the collaboration between the firm and the external resources. Consequently, managing open innovation presents new operational challenges concerning managing a firm's internal resources (Lerner and Tirole 2002; West 2003). Against this background, in this essay (or hereafter also referred to as the paper), we seek to analyze the effects of technology openness on the optimal behavior of a technology firm (hereafter also referred to as simply the *firm*).

In today's markets, adding features to existing technologies and innovating new technologies to have reliable and up-to-date products are vital for firms to remain competitive by making their products or technologies more relevant to the ever-changing customer or business needs. An industry study by PricewaterhouseCoopers reports that major technology firms have increased their spending on research and development to innovate and develop new technologies or products (PWC 2017). However, several products are released into the market with defects (although not intentionally). In particular, defects are often identified by the external or internal resources after the product is rolled out into the market. For example, customers of Tesla cars often report problems that include broken door handles, or internal computer systems crashing (Matousek 2018). In software products alone, the detrimental effects of defects on the global economy is estimated to be more than $312 billion (Cambridge University 2013). Consequently, several technology firms, including Microsoft and Tesla, have announced intentions to improve the quality of their products continually even after the release by becoming more efficient and effective in fixing defects. Hence, it is necessary for the technology firms to "*continually*" fix defects in their technologies after they release their products (Arora et al. 2010) while developing new technologies. Therefore, firms need to optimally allocate their resources between *innovating or developing new technologies* and *maintaining existing technologies.*

Paulson et al. (2004) argue that open technology is more creative, better in quality, easier and cheaper to develop, and defects are identified and fixed faster. The following statements ratify this argument from the perspective of large technology firms:

"*The first generation hydrogen fuel cell vehicles, launched between 2015 and 2020, will be critical, requiring a concerted effort and unconventional collaboration between automakers, government regulators, academia and energy providers. By eliminating traditional corporate boundaries, we can speed the development of new technologies and move into the future of* *mobility more quickly, effectively and economically.*" – (Toyota 2015).

"Apple believes that using Open Source methodology makes Mac OS X a more robust, secure operating system, as its core components have been subjected to the crucible of peer review for decades. Any problems found with this software can be immediately identified and fixed by Apple and the Open Source community." – (Apple 2015).

While there is unanimity in the literature that making a technology open source reduces the time to find and fix defects (Arora et al. 2010) and accelerates innovation of newer technologies (Toyota 2015), its effect on the behavior of a technology firm (regarding the maintenance of current version and development of newer version) has not received enough attention (Schryen and Kadura 2009). Our study aims to fill this gap in the literature. In particular, we develop an optimal control model to determine: (i) the allocation of efforts between maintenance of current version and development of the next version of technology, and (ii) the extent of openness of its technology. In the next subsection, we discuss our key findings and contributions.

2.1.1 Contributions and Key Findings

Our study contributes to the open technology and resource allocation literature by examining the impact of open technology on the firm's maintenance and development efforts. With several technology firms opening up their technologies to external resources (whom we refer to as members of *open source community*, which may include suppliers, customers, and competitors), determining the optimal level of openness along with optimal effort levels is not only a relevant issue to literature that seeks to broaden the knowledge and theory of open innovation but also to industry. Our analysis provides critical insights to the firms by helping them decide the best strategy in determining the effort levels and the extent of technology openness. Furthermore, we use our analysis to explain real-world decisions of major technology firms such as Tesla, Microsoft, Apple, and IBM. Traditionally, technology openness has been examined in the literature as a binary variable (i.e., fully open source or proprietary), whereas we allow it to be a continuous variable between 0 and 1 (i.e., the firm may partially open the technology, which is more prevalent in the industry). In the remainder of this subsection, we highlight the key questions that are answered in this study.

To gain a deeper understanding of firms' open source strategy, the first key question that we analyze is: *When should the firm keep its technology proprietary, or make it partially or fully open?* More specifically, we examine how various key factors (i.e., the characteristics of the market, firm, and technology) affect the optimal choice of the technology firm regarding openness. We find that the firm might keep its technology proprietary even in conditions that seem to favor an open source environment. For example, as the network effect due to technology openness (more specifically, demand sensitivity of the next version to openness) increases, we should expect the firm to increase the extent of openness to drive more demand for the technology. However, our analysis reveals that this is not always the case.

Further, we ask the following question: *Should the firm increase its effort of fixing defects if the technology becomes more valuable?* We find that the answer to this question is not necessarily "yes." Under certain conditions, as the technology becomes more valuable, the firm should focus on increasing the extent of openness and rely on the open source community (i.e., external resources) in fixing defects. Therefore, firms should carefully manage their resources in response to changing market conditions.

Next, we analyze the following question: *should the firm increase the extent of openness if the effectiveness of the open source community in developing the next version increases?* The firm's decision in this scenario should depend on the trade-offs between the effectiveness and the collaboration cost. Intuitively, the firm should reduce the extent of openness as the collaboration cost between the firm and the open source community increases, in order to reduce additional costs. However, contrary to this intuition, our analysis reveals that the firm should sometimes increase the extent of openness despite higher collaboration costs. In fact, we find that the firm should sometimes switch to making the technology fully open source (from keeping it proprietary or partially open) as the collaboration cost increases.

In addition, we investigate other important business settings. For example, in some settings, the market might have a pre-determined quality requirement for the next version. Hence, we also examine such a setting in which the market has a minimum quality requirement for the next version of the technology. Although we mainly focus on an environment where the firm has sufficient resources required for the maintenance of the existing technology and the development of the next version, for completeness, we also analyze the scenario when the firm has a limited budget. In the next subsection, we briefly review the relevant literature, and highlight our contributions with respect to past studies.

2.1.2 Related Literature

Our work closely relates to the following two literature streams: (i) open innovation, and (ii) resource allocation. Here, we briefly discuss related studies in these streams and highlight our contributions with respect to those studies.

2.1.2.1 Open Innovation

The gradual increase in the number of firms making their technologies open source has raised challenging questions and a need for new models and theories to address these concerns (von Hippel and von Krogh 2003). Prior literature in this stream mainly focuses on the (i) motivations of users and contributors in participating in the development and maintenance of open source technology (e.g., Bonaccorsi and Rossi 2003), (ii) market competition between open source technology and proprietary technology (e.g., Casadesus-Masanell and Ghemawat 2006), or (iii) management of open source community (e.g., von Krogh and von Hippel 2006). In a more recent study, Hu et al. (2017) study the impact of competing firms' open technology strategy. They find that opening their technology can promote higher supplier investments. However, Casadesus-Masanell and Ghemawat (2006) call for research that explicitly examines the impact of technology openness on firm's effort levels. Our study responds to this call and contributes to this stream of literature by analytically analyzing the optimal extent of technology openness and its impact on firm's effort levels.

Open technology literature suggests that there is substantial tension between the two opposing strategies of keeping technology proprietary versus making it open. On the one hand, proponents argue that open technology will replace proprietary technology as the predominant mode of product development. For example, West (2003) argues that firms adopt open source methodology because of higher adoption rates and better network effects. Higher technology adoption rates, in turn, may attract more skilled resources and help the firm achieve higher levels of efficiency and effectiveness in development and maintenance (Lerner and Tirole 2002). On the other hand, opponents of the open source methodology agree with the following argument of Levy (2000): "*Sure, the code is available. But is anyone reading it?*" Researchers have analyzed this question by empirically examining whether the open source community is responsive in identifying and fixing defects, or whether defects are found and fixed faster in open source environments in software industry (e.g., see Paulson et al. 2004; Arora et al. 2010). Findings from these studies demonstrate that defects are identified and fixed faster in when the technology is open source compared to proprietary technology. To further validate some of the assumptions in our analytical model, we also supplement the past studies in this stream by empirically studying the impact of openness on the time it takes to fix defects.

Although it is clear that there are several advantages of making a technology open source, firms need to be aware of the costs associated with coordination between open source community and the firm's internal resources (West 2003), and the loss in competitive edge (Lerner and Tirole 2002). In order to balance this tension, firms have adopted a hybrid strategy in which they open a part of their proprietary technology to derive the benefits of open source technology, while minimizing the costs associated with it and maintaining their competitive edge (e.g., see Microsoft 2014). Besides, several past studies have recognized the need to examine hybrid technology development forms that combine the strengths of both open source and proprietary development (Mockus et al. 2000; Bonaccorsi and Rossi 2003). Our study builds on the findings of the open innovation literature and complements the literature by analytically modeling the impact of openness on the firm's operational strategies.

2.1.2.2 Resource Allocation

Our paper is also related to the literature on managing resources in maintaining existing technologies and developing new technologies. The operational issues related to technology development and maintenance has been studied extensively in both operations and information systems literature. The technology development literature examines the frequency and the timing of various activities involved in the "technology development stage." In the closest study to ours, Ji et al. (2005) formulate an optimal control model to study the problem of allocation of resources between new software construction and fixing defects before the software is released into the market. Their results suggest that integrating the decisions pertaining these activities leads to significant savings.

The technology maintenance literature mainly deals with finding optimal policies to minimize the overall cost of maintenance of existing technology (Kulkarni et al. 2009). This literature has its roots in the traditional machine maintenance literature. Furthermore, Lientz et al. (1978) argue that the maintenance of the software (i.e. fixing of software defects and bugs) along with development of new features take a high share of the total budget after the software is released. In the closest study to ours in this area, Kulkarni et al. (2009) examine the problem of optimal allocation of resources for technology (i.e., software) maintenance requests from the customer using variations of $M|G|1$ queues. In addition, Repenning (2001) examines the allocation of resources for fixing problems during the product development stage. Ji et al. (2011) complement the earlier studies in this stream by simultaneously examining the initial number of features in a technology and the enhancement effort after the technology is released. However, unlike our study, they do not consider the effects of making the technology fully or partially open.

In contrast to the above studies, to the best of our knowledge, our study is the first to find the optimal level of technology openness and investigate how it impacts the firm's behavior in allocating resources for maintenance (for the current version) and development of the next version of technology, an important managerial challenge in the presence of limited resources. In particular, we focus on the notion of technology openness that is not considered in Ji et al. (2011) together with fixing/maintenance and development efforts. In the next section, we present our analytical model.

2.2 Analytical Model

We consider a typical technology firm who releases a product (referred to as the "*current version*") and provides support for it (i.e., by fixing defects or maintaining the technology) while simultaneously working on developing a new version of this product (referred to as the "*next version*"). In the next subsection, we begin with discussing the business setting along with the objective function of the technology firm.

2.2.1 Objective Function

Even though several open source technology products (specifically in software industry) are offered for free to customers, technology firms employ different business models to generate revenue from them (e.g., see Lerner and Tirole 2002; Fitzgerald 2006). For example, Mozilla's Firefox is a free and open source software, and yet the revenues of Firefox were more than $300 million in 2013 (Mozilla 2013). Further, technology firms often generate value by providing complementary services or products to the open source systems (Lerner and Tirole 2002). In addition, there are products that are open source but not free (e.g., any software licensed under NASA Open Source Agreement) (Gnu.org 2012). We would like to emphasize that our focus in this paper is not on the revenue models, instead it is on the overall profit of the technology firm based upon the demand for the technology irrespective of being offered for free or not.

In the remainder of this subsection, we discuss the following components of the firm's objective function: (i) demand for current version of the technology, (ii) demand for the next version, (iii) costs associated with the effort (both maintenance and development), and (iv) costs related to collaboration with the open source community.

2.2.1.1 Demand for the Current Version of the Technology

Quality is a complex and multi-faceted concept that can be interpreted from different perspectives (Garvin 1984). Similarly, the definition of product quality is also ambiguous and there is no single metric to quantify the quality of product that is acceptable to everyone (Kitchenham and Pfleeger 1996). One realistic and universally accepted metric for measuring quality of a technology is the number of defects (Fenton and Neil 1999). Therefore, as the number of defects decreases, the quality of the technology improves. Fixing defects improves the quality of the current version (that we denote by $q(t)$), which in turn, increases the demand from the customers. Further, fixing defects also helps in improving customers' goodwill that translates into higher demand for the current version of the technology.

In addition, literature in open technology suggests that firms open their technology to get their products widely adopted (Lerner and Tirole 2002; West 2003). In software industry, open source technology also attracts skilled users such as engineers and developers as it provides them the opportunity to modify the code and work with other developers (Dahlander and Magnusson 2005). Besides, Hu et al. (2017) show that by making their technologies open, firms can induce higher supplier investments, which could in turn lead to wider adoption of the technology. Therefore, as the firm makes a higher portion of its technology open source, more users (i.e., customers) and suppliers become interested in the technology so that the demand for the current version increases. Hence, in line with the prior literature, we model the demand as a function of extent of openness "s." Here, s can also be considered the portion of the technology that is open. Unlike most of the past literature that considers the technology to be either fully open (i.e., $s = 1$) or fully closed or proprietary (i.e., $s = 0$), in line with current industry practices, we also acknowledge that the technology can be partially open (i.e., $0 < s < 1$).

To summarize, the demand for the current version of the technology (denoted by $D_q(t)$) primarily depends on: (i) the quality of the technology $q(t)$ (extent of this dependence is controlled with the sensitivity parameter, θ_1), and (ii) the extent of openness s (extent of this dependence is controlled with the sensitivity parameter, θ_2). In particular, we model the demand for the existing version of the technology at time t as follows: $D_q(t) = \theta_0 + \theta_1 q(t) + \theta_2 s$. If the customers are highly sensitive (resp., not highly sensitive) to the quality of the technology, then θ_1 is high (resp., low). On the other hand, as Lerner and Tirole (2002) argue, the rate of diffusion of an open source technology depends on the characteristics of the product (such as complexity of the technology) and the market (such as sophistication of its end users). Hence, we capture the extent of demand sensitivity to openness with parameter θ_2. Clearly, we can set θ_2 to 0 when there is no impact of openness on the demand of the existing technology.

One of the objectives of the firm is to increase the sales of the current version throughout the planning horizon that is expressed as $\int_0^T kD_q(t)dt$. We consider a continuous model, where T is the time between the release of the current version and the next version, and k is the value per unit of demand of the current version. Thus, the value of the current version to the firm is ongoing and dynamic, i.e., it is not realized only at the end of the project. By using such a model, we are able to analyze the dynamics of variables such as the effort levels and the quality levels of both versions of the technology throughout the planning horizon. Since the decision to release the technology is generally pre-determined based on external factors such as market characteristics, business plans of the competitors, and contractual obligations between the firm and the customer (Ji et al. 2005), we consider the release time of the next version (i.e., T) to be exogenous.

2.2.1.2 Demand for the Newer Version of the Technology

Older technologies need to be replaced with new versions beyond a point in order to provide the functions needed by the users (Aoyama 2002; Heales 2002). There are several reasons for switching to a next version of the technology. First, replacement might be necessary for economic reasons due to the high cost of maintaining the system since each maintenance effort deteriorates the structural integrity of the system (Bianchi et al. 2001). Secondly, replacement might be needed for technical reasons. For example, the development environment might become obsolete with no support from the firm, or business pressures or market needs might require new functions that are fundamentally different from the existing capabilities of the current version of the technology. Hence, we consider that, at the end of the useful life of the current version (i.e., T), the next version of the technology is released.

Similar to the demand for the existing version, the demand for the new version of technology depends on: (i) the quality of the technology (denoted by $r(t)$), and (ii) the extent of openness (s). First, the demand for the new version depends on the quality of the technology. In particular, higher quality at the time of release (i.e., T) would translate to higher demand. Next, as discussed earlier, prior literature suggests that technology openness increases the likelihood of attracting a wider audience, and hence the demand for the new version of the technology depends on the openness of the technology (West 2003). Hence, we model the demand for the next version of the technology as $D_r(t) = \rho_0 + \rho_1 r(t) + \rho_2 s$, where ρ_1 represents the demand sensitivity to quality of the technology, ρ_2 denotes the demand sensitivity to the extent of openness, and ρ_0 is the intercept. If the extent of openness of the existing version has no affect on the demand for the next version, we can set ρ_2 to 0.

While modeling the value from the next version, one needs to consider the negative effect of openness as well. In particular, as the large portions of the technology become open, the value of the next version decreases in the market. That is because, in an open source setting, the competitors have access to the technology, and they may utilize it for competitive advantage (Kumar et al. 2011). In addition, some users can alter the technology (if they are capable of doing so) to address their individualized needs. This modified version may be made available to other customers as well. Hence, some customers may not be willing to upgrade to the next version when it is released. Therefore, we consider that the extent of openness is directly proportional to the loss in competitiveness of the firm. More specifically, as the portion of technology that is made open source increases, the competition effect (from the rival firms and own customer) increases linearly (consideration of nonlinear structures does not change the findings qualitatively). Hence, we represent the value of the next version of the technology as $(h - es)D_r(T)$, where h is the value of the next version per unit level of quality, e represents the competitiveness of the market, and $D_r(T)$ is the demand for the next version of the technology when it is released (i.e., at time T). Note that the negative effect due to openness because of market competition is es, hence the degree of openness (i.e., s) affects the competition. In order to represent the settings with no increased competition effect due to openness, one can set the corresponding sensitivity parameter e to 0. In the following subsection, we now introduce the maintenance effort and development effort, and the corresponding costs.

2.2.1.3 Effort Levels

The firm strives for both maintaining the current version (i.e., by fixing defects) and developing the next version throughout the planning horizon. We define $u(t)$ as the effort exerted by the firm to maintain the current version by fixing defects and $v(t)$ as the effort allocated for developing the next version. In the literature, diseconomies of scale has been observed and verified during both the support and development phases of a technology product (Banker et al. 1994; Banker and Slaughter 1997). This phenomenon is due to the fact that adding more resources to the project increases the cost non-linearly due to increased complexity, training, and coordination issues. Therefore, analogous to the literature (e.g., Tsay and Agrawal 2000), we use a quadratic cost structure. In particular, we model the cost of fixing defects (in the current version) as $\int_0^T c_f u^2(t)dt$ and the cost of developing the next version as $\int_0^T c_d v^2(t)dt$. Hence, the rate of expenditure incurred by the firm (denoted by $\dot{x}(t)$) is as follows: $\dot{x}(t) = c_f u^2(t) + c_d v^2(t)$. This implies that the expenditure incurred by the firm at time T (i.e., throughout the project) is as follows: $x(T) = \int_0^T [c_f u^2(t) + c_d v^2(t)]dt$. Although our objective is to maximize the profits of the technology firm, in certain instances, the firm may be constrained by a limited budget (denoted by B). Hence, we consider that $x(T) \leq B$. The budget threshold is determined by the firm and hence is considered as an exogenous variable in our model.

2.2.1.4 Cost of Collaboration with Open Source Community

There are several sources of costs related to making the technology open source, such as the management of collaboration between external resources (i.e., open source community) and internal resources (i.e., in-house engineers). With users of the open source community scattered around the globe, the coordination between the internal team of the firm and the open source community is even more complicated and important from the firm's perspective (Koch 2009). In the context of software industry, the firm needs to examine every change in the source code (or every modified function) in order to make sure that the changes are meaningful and do not introduce new defects or for compatibility issues. As the extent of openness increases, more users adopt the technology and users have more access to the designs. This increases the number of interactions between the open source community and the firm. That is, if there is more contribution to fixing defects, then the extent of resources required for coordination also increases. Thus, management of collaboration becomes harder and the cost increases non-linearly due to increased complexity and coordination as the basis of interaction (i.e., extent of openness) increases as it is stated in the literature (Clemons and Row 1992). This is also in line with the findings presented in Scholtes et al. (2016) in which the authors empirically observe diseconomies of scale in the management of large-scale open source projects. That is, the cost of collaboration with the open source community is needed throughout the planning horizon and it increases with time. Hence, we model the direct cost of managing the collaboration in the quadratic form as as^2T. In this formulation, a is the unit collaboration cost of technology that is open.

Taking into account the value from the current and next versions of the technology, effort costs, and cost of the collaboration with the open source community (discussed above), we can write the objective function of the firm as:

$$\text{Max}_{u(t),v(t),s}\left[\int_0^T kD_q(t)dt - as^2T - x(T) + (h - es)\,D_r\,(T)\right].$$

In the next subsection, we discuss the state equations.

2.2.2 Quality Levels of the Current and the Next Version

The quality of the current version of the technology depends on: (i) the effort level of the firm, and (ii) the extent of openness. We first discuss the impact of technology openness on the evolution of the quality of the current version of technology over time t. Several studies have empirically shown that defects are fixed faster when firms employ open source methodology. Arora et al. (2010) use a proportional hazard model to investigate if technology openness impacts patch (i.e., fixes for software defects) release time. They find that patches for the defects in open source software are released faster compared to closed source software. Similarly, Paulson et al. (2004) find that the defects are found and fixed faster in open source projects.

As discussed in Section 2.2.1.1, as the number of defects decreases, the quality of technology improves. Therefore, fixing defects essentially improves the quality of technology. Hence, Arora et al. (2010) (and our empirical results) indicate that the quality of technology improves at a faster rate when firms a open portion of its design. We utilize this result in modeling the impact of technology openness on the quality of the current version of the technology and state that $\dot{q}(t)$ improves with bs, where $b \geq 0$ is the effectiveness of the open source community in fixing defects.

Since the quality of technology can also be improved by the maintenance effort of the firm, we can write the state equation for the quality of the current version of the technology as follows: $\dot{q}(t) = mu(t) + bs$. Here, $u(t)$ is the maintenance effort exerted by the technology firm at time t. Although the maintenance effort directly impacts the quality of the technology, it is important to note that the initial quality of the technology plays a vital role in how $q(t)$ evolves over time. In particular, when the initial number of defects is high, i.e., the initial quality of the technology is low, the firm needs to exert more effort than it would if the initial quality of the technology were better. That is, the impact of effort on the quality is proportional to the initial quality level of the technology. Hence, in the state equation, we employ m as the effectiveness of the firm's effort level. In particular, when the initial quality of the technology is low, the value of m is lower, and hence, the firm needs to exert more effort to improve the quality. Since m already captures the impact of the initial quality, for simplicity, we normalize $q(0)$ to zero.

The quality of the next version of the technology primarily increases with the development effort of the firm that is denoted by $v(t)$. Besides, since each version of the technology generally builds on the past releases (e.g., MacOS, Windows, and Matlab) by adding new features to the current version of the technology (Rahmandad 2005), the quality of the next version increases with the effort of fixing defects in the current version (i.e., $u(t)$) as well. The corresponding sensitivity term (i.e., w) can be set to zero if there is no such effect, i.e., the next version is built from scratch with no relationship to the current version. Similarly, the extent of openness may also help in improving the next version of the technology, because some members of open source community might initialize or undertake development activities. For example, while making its artificial intelligence engine TensorFlow open source, Google stated:

> *"What we are hoping is that the community adopts this as a good way of expressing machine learning algorithms of lots of different types, and also contributes to building and improving [TensorFlow] in lots of different and interesting ways." –* (Wired.com 2015).

Thus, Google expects that making TensorFlow open source will feed new projects (or new versions of its technology) as well (Wired.com 2015). In a similar vein, we model the rate of improvement in the quality of the next version of the technology as $\dot{r}(t) = wu(t) + nv(t) + ys$. Clearly, we can set $y = 0$ when there is no impact of openness on the quality of the next version. Besides, n is utilized in this formulation to represent the effectiveness of the firm's development effort in the quality of the next version. All parameters and variables are summarized in Table 2.1.

Table 2.1: List of Parameters and Variables

Symbol	**Definition**
Parameters	
a	Cost multiplier for managing the collaboration with the open source community
b	Sensitivity of the quality of the current version of the technology to openness (Effectiveness of openness on the rate of fixing defects)
B	Maximum budget available to the firm for maintenance and development effort (Overall budget/resource constraint)
c_f	Cost multiplier of the maintenance effort (Costliness of the maintenance effort)
c_d	Cost multiplier of the development effort (Costliness of the development effort)
e	Market competition parameter due to openness (Competitiveness of the market)
h	Value of the next version of the technology per unit level of quality (Valuation of the next version of the technology)
k	Value of the current version of technology per unit level of quality (Valuation of the current version of the technology)
m	Sensitivity of the quality of the current version to maintenance effort (Effectiveness of maintenance effort on the current version of the technology)
n	Sensitivity of the development of the next version to development effort (Effectiveness of development effort on the next version of the technology)
T	End of the planning horizon (Release time of the next version of the technology)
w	Sensitivity of the development of the next version to maintenance effort (Effectiveness of maintenance effort on the next version of the technology)
y	Sensitivity of the development of the next version to openness (Effectiveness of openness on the next version of the technology)
Z	Minimum quality level of the next version of the technology
θ_0	Intercept for the demand function of the current version
θ_1	Sensitivity of the demand of the current version to quality of the technology
θ_2	Sensitivity of the demand of the current version to openness
ρ_0	Intercept for the demand function of the next version
ρ_1	Sensitivity of the demand of the next version to quality of the technology
ρ_2	Sensitivity of the demand of the next version to openness
Variables	
$D_q(t)$	Demand for the current version of the technology at time t **(State Variable)**
$D_r(t)$	Demand for the next version of the technology at time t **(State Variable)**
$q(t)$	Quality of the current version of the technology at time t **(State Variable)**
$r(t)$	Quality of the next version of the technology at time t **(State Variable)**
s	Portion of the technology that is open (Extent of openness) **(Decision Variable)**
$u(t)$	Effort of fixing defects in the current version of the technology at time t (Maintenance effort) **(Control Variable)**
$v(t)$	Development effort of the next version of the technology at time t **(Control Variable)**
$x(t)$	Expenditure of firm on maintenance effort and the development effort at time t **(State Variable)**

2.2.3 The Optimal Control Model

Taking the objective function, state equations, and other constraints into consideration, the optimal control model for the firm can be written as below.

$$\Pi = \text{Max}_{u(t),v(t),s} \left[\int_0^T kD_q(t)dt - as^2T - x(T) + (h - es)\, D_r\,(T) \right] \tag{2.1}$$

subject to:

$$\begin{aligned}
\dot{q}(t) &= mu(t) + bs, && (2.2)\\
\dot{r}(t) &= wu(t) + nv(t) + ys, && (2.3)\\
D_q(t) &= \theta_0 + \theta_1 q(t) + \theta_2 s, && (2.4)\\
D_r(T) &= \rho_0 + \rho_1 r(T) + \rho_2 s, && (2.5)\\
\dot{x}(t) &= c_f u(t)^2 + c_d v(t)^2, && (2.6)\\
x(T) &\leq B, && (2.7)\\
0 &\leq s \leq 1, && (2.8)\\
q(0) &= 0,\ r(0) = 0,\ x(0) = 0,\ u(t) \geq 0,\ v(t) \geq 0. && (2.9)
\end{aligned}$$

In summary, Equation (2.1) states that the firm determines the optimal levels of (i) maintenance effort, (ii) development effort, and (iii) the extent of openness, with the objective of maximizing the profit from both versions of the technology. Next, Equations (2.2) and (2.3) depict the evolution of the quality levels of the current version and the next version, respectively. Equations (2.4) and (2.5) represent the demand for the current version and the next version of the technology respectively. Equations (2.6) and (2.7) track the expenditure incurred by the firm for the maintenance effort and the development effort and state that the total budget is capped at B. Equation (2.8) defines bounds on the extent of openness since it is defined as a portion. Finally, Equation (2.9) lists the technical constraints: the starting levels of the quality of current and next versions, initial budget use, and the non-negativity of the effort levels. When possible, we suppress the time indicator (t) from the variables for simplicity in notation. Before presenting the results, we briefly discuss the methodology used in our analytical model.

2.2.4 Methodology

The optimal control technique has been used in multiple disciplines including operations management to solve problems where the outcome depends not only on the current action, but also the on previous ones (Sethi and Thompson 2000). The optimized decision variables (here, the effort of fixing defects, the development effort, and the extent of openness) manage the evolution of the dynamic system through state variables (here, the quality levels of the current version and the next version of the technology, and the spending on maintenance and development) to achieve an optimal outcome (here, the profit of the firm) by the end of the planning horizon. For our dynamic problem, a traditional mathematical programming approach would use a multistage dynamic programming formulation and solve the problem numerically without providing analytical results and insights. Optimal control, however, decouples our dynamic problem over time into an infinite number of simpler static optimization problems for each time instance t, and enables us to derive managerial insights from the analytical results that we present in the following section.

2.3 Results and Managerial Insights

In this section, we first derive a closed-form solution to the problem presented in Equations 2.1-2.9. Next, we glean several results and managerial insights. As discussed in Section 2.2.1.3, we analyze the optimal effort levels and extent of openness for two different scenarios: (i) firm has sufficient budget (e.g., by borrowing from an external investor or a financial institution such as a bank) to cover the expenditure of the firm (i.e., $x(T) < B$), and (ii) firm has a limited budget and cannot borrow from elsewhere (i.e., $x(T) = B$). Since the primary objective of our study is to maximize the profits of the firm, our analysis mainly focuses on the first scenario, i.e., the firm has enough resources to maintain the current version and develop the next version. Nevertheless, we find that most of the insights remain the same for both scenarios. We present and discuss results that are different in these two scenarios. Below, we first present the solution of the optimal control model for an arbitrary level of openness. All the necessary proofs for lemmas and propositions that we present in the rest of this paper are provided in Appendix A.2

Lemma 1. *For an arbitrary level of openness (i.e., s), the optimal trajectories of the maintenance effort (i.e., $u(t)$) and the development effort (i.e., $v(t)$) are given by:*

$$u(t) = \frac{\rho_1 w(h-es) + \theta_1 km(T-t)}{2c_f}, \text{ and } v(t) = \frac{n\rho_1(h-es)}{2c_d}.$$

Lemma 1 implies that opening up a higher portion of the technology (i.e., increasing s) decreases both fixing and development efforts. This is due to the fact that opening up a portion of the technology has positive effects on the quality levels of both the current version (because there are more eyes on the defects, see Equation (2.2)), and the next version (because some members of open source community might initialize or undertake development activities, see Equation (2.3)). Furthermore, opening up the technology increases the demand for the existing and next versions (due to network effects, see Equations (2.4) and (2.5)). On the other hand, the objective function presented in Equation (2.1) implies that increasing the extent of openness is costly to the firm because of two main reasons: (i) complications in the management of collaboration between the parties (internal resources and the open source community), and (ii) the decrease in the competitiveness or value of the next version (since it becomes easier for: (a) the competitors to understand and copy features of the technology and gain competitive advantage, and (b) the competitors to tailor the technology for different needs and distribute the modified technology). By examining the interdependencies among the extent of openness and both types of effort levels, and the effects of openness on the current and next versions of the technology, the firm can determine the optimal level of openness. The results of this analysis, i.e., the quality of next version of the technology, the optimal level of openness, the corresponding optimal effort trajectories, and the profit of the firm are presented in Lemma 2.

Lemma 2. *The optimal levels of the openness (i.e., s^*), effort of fixing defects (i.e., $u^*(t)$), the development effort (i.e., $v^*(t)$), the quality of the next version of the technology (i.e., $r^*(T)$), and the profit of the firm (i.e., Π^*) are given by:*

$$
\begin{aligned}
s^* &= \frac{2b\theta_1 kT^2 c_d c_f - 4e\rho_0 c_d c_f - 2eh\rho_1^2 Tw^2 c_d - e\theta_1 km\rho_1 T^2 wc_d + 4h\rho_2 c_d c_f + 4h\rho_1 Tyc_d c_f + 4\theta_2 kTc_d c_f - 2ehn^2\rho_1^2 Tc_f}{2\left(4aTc_d c_f - e^2\rho_1^2 Tw^2 c_d + 4e\rho_2 c_d c_f + 4e\rho_1 Tyc_d c_f - e^2 n^2 \rho_1^2 Tc_f\right)}, \\
u^*(t) &= \frac{\begin{array}{c}[(c_d e^2\theta_1 km\rho_1^2 Tw^2 (2t-T) + 2c_f (e^2\theta_1 kmn^2\rho_1^2 T(t-T) + c_d (\theta_1 kT^2 (4am + e\rho_1(4my - bw)) + \\ T(2\rho_1 w(2ah + eh\rho_1 y - e\theta_2 k) - 4\theta_1 km(at - e\rho_2 + e\rho_1 ty)) + 2e(\rho_1 w(e\rho_0 + h\rho_2) - 2\theta_1 km\rho_2 t))))]\end{array}}{4c_f\left(c_f\left(4c_d(T(a + e\rho_1 y) + e\rho_2) - e^2 n^2\rho_1^2 T\right) - c_d e^2\rho_1^2 Tw^2\right)}, \\
v^*(t) &= \frac{n\rho_1\left(c_f(8ahT + 2e(-T(b\theta_1 kT - 2h\rho_1 y + 2\theta_2 k) + 2e\rho_0 + 2h\rho_2)) + e^2\theta_1 km\rho_1 T^2 w\right)}{4c_d\left(4c_f(aT + e\rho_2 + e\rho_1 Ty) - e^2\rho_1^2 Tw^2\right) - 4e^2 n^2\rho_1^2 Tc_f}, \\
r^*(T) &= \frac{\begin{array}{c}(T(c_f n^2\rho_1(T(4ah - 2e\theta_2 k) - be\theta_1 kT^2 + 2e(e\rho_0 + h\rho_2)) + \\ c_d(\theta_1 kT^2(2amw + 2bc_f y + e\rho_1 w(my - bw)) + 4c_f y(h\rho_2 - e\rho_0) + 2e\rho_1 w^2(e\rho_0 + h\rho_2) + \\ 2T(w(\rho_1 w(2ah - e\theta_2 k) + e\theta_1 km\rho_2) + 2c_f y(h\rho_1 y + \theta_2 k)))))\end{array}}{2c_f\left(4c_d(T(a + e\rho_1 y) + e\rho_2) - e^2 n^2\rho_1^2 T\right) - 2c_d e^2\rho_1^2 Tw^2}, \\
\Pi^* &= \frac{(Tc_d\left(3\theta_1 km\rho_1 Tw(h - es^*) + 3\rho_1^2 w^2 (h - es^*)^2 + \theta_1^2 k^2 m^2 T^2\right)}{12c_d c_f} \\
&\quad - \frac{6c_f c_d\left(2\left(as^{*2}T - (h - es^*)\left(\rho_0 + \rho_2 s^* + \rho_1 s^* Ty\right) - \theta_2 ks^* T\right) - b\theta_1 ks^* T^2 - 2\theta_0 kT\right)}{12c_d c_f} \\
&\quad + \frac{3n^2\rho_1^2 Tc_f (h - es^*)^2}{12c_d c_f}.
\end{aligned}
$$

Further examination of the optimal trajectories of the effort levels, i.e., $u(t)$ and $v(t)$ presented in Lemma 1 reveals that the firm needs to actively manage its resources. In particular, the firm should exert higher maintenance effort during the initial stages of the project and then gradually decrease its effort level. The intuition behind this result is that the firm gets a "cumulative" benefit from improving the quality of the current version because of higher demand for the technology in the early stages. Furthermore, because of higher maintenance effort in the early stages, more defects are fixed in the early stages, and hence the number of defects that need to be fixed in the later stages are reduced. Moreover, the benefit of fixing defects in the later stages is marginal as the current version is replaced by the next version. Thus, the firm can reduce the effort it exerts for maintaining the technology in the final stages of the project.

Lemma 1 also reveals that the trajectory of the development effort does not depend on time. This result is both interesting and important. The key takeaway for the managers is that the resources need to be actively managed for maintenance, but development of newer technology does not require micromanagement. Next, we present and analyze the conditions when the firm should keep its technology proprietary (i.e., $s = 0$) or make it fully open (i.e., $s = 1$) or make it partially open (i.e., $0 < s < 1$).

2.3.1 When to Keep the Technology *Proprietary* or *Fully* Open or *Partially* Open

In the following proposition, we begin with summarizing the conditions when the firm should choose extreme strategies (i.e., keep the technology proprietary or fully open). The expressions for thresholds in all the propositions are provided in Appendix A.1.

Proposition 1.

(a) The firm should keep its technology proprietary if the sensitivity of the demand of the next version to openness (i.e., sensitivity to network effect) is high (i.e., $\rho_2 > \zeta$) and the market is highly competitive (i.e., $e \geq \mathscr{E}$).

(b) The firm should make its technology fully open source if the sensitivity of the quality of the current version to openness is high (i.e., $b \geq \mathscr{B}$) and the collaboration cost is high (i.e., $a > \mathscr{A}$).

We first focus on part (a) of the proposition that analyzes the conditions when the firm should not rely on the open source community and keep the technology proprietary. Proposition 1(a) suggests that despite higher benefits from wider adoption (or network effects) due to openness, the firm should in fact not open any portion of the technology when the market is highly competitive (i.e., e is large). In this scenario, the benefit of increased demand due to openness (see Equation (2.5)) does not outweigh the loss in competitiveness. Therefore, it is optimal for the firm to keep its technology proprietary even when the sensitivity of the demand of the next version to openness (i.e., ρ_2) is high. The finding in Proposition 1(a) provides an important and interesting insight to the firms: Despite higher demand for the next version of the technology due to openness, it is, in fact, optimal for the firm to keep its technology proprietary when the competitiveness is sufficiently high. This result may provide a plausible explanation behind Microsoft's decision to keep the technology of some of their products proprietary. Hoffman (2015) argues that if Microsoft decides to make Windows open source "*competing companies could take Windows [Kernel] and use it to make a competing operating system.*" This suggests that the market is highly competitive (i.e., $e > \mathscr{E}$). Thus, in line with our finding, Microsoft's strategy is to keep its technology proprietary.

The finding in Proposition 1(a) is illustrated in Figure A.1. In Figure A.1, we consider a scenario where the characteristics of the firm, the technology, and the market are operationalized using the following parameter values: $k = \frac{1}{32}, h = 1, a = 1, T = 16, m = 2, b = 1, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2$. These parameter values are reasonable in the sense that they represent a typical technology firm. First, the sensitivities of the firm's effort (i.e., m and n) are higher than the sensitivities of the open source community (i.e., b and y). These values of the sensitivities are representative of real business settings because the firms are typically expected to be more productive and proactive compared to the open source community, and hence the effect of firm's effort on the quality of the technology is higher (i.e., the values of m and n are higher than b and y). Furthermore, industry reports suggest that the operational expenses in the technology industry are around 50% to 75% (e.g., Mergent Online 2017). Thus, we select the values of cost parameters (i.e., $a, c_f, c_d, e, h,$ and k) such that the expenses are in the range of 50% to 75% of the revenue generated the technology (however, we find that the results are consistent and do not change qualitatively for a wide range of values for all the parameters).

We now focus on the conditions when the firm makes its technology fully open source (see Proposition 1(b)). We utilize these key findings to explain the possible motivation behind the strategy of *IBM* to fully open some of their technology systems. This result reveals that the level of sensitivity of the quality of the current version on openness (i.e., b) has a very important role in the firm's decision to fully open the technology or not. In particular, it shows that the firm should fully open source its technology when b is above a critical value, despite high cost of collaboration (i.e., a). This result may not be intuitive since it is expected that higher cost of collaboration might discourage the firms to make their technology fully open.The intuition behind this result is that despite the negative effect of cost of collaboration on firm's profit, the firm derives greater benefits by the increased rate of improvement in the quality of the existing technology (see Equation (2.2)). This finding might provide a plausible explanation to IBM's decision to make the technology (i.e., software code) of *WebSphere* fully open. Despite, the high expected cost of collaboration (i.e., high a) (Wayner 2000), IBM recognized that large groups of open source community were very effective in fixing defects (i.e., high b) (West 2003). In line with our finding, IBM decided to make *WebSphere* fully open, IBM's first open source technology. We illustrate this finding in Figure A.2. The parameter values in this figure are the same as those in Figure A.1 except $e = 0.2$.

Before turning our attention to the case when the firm should keep the technology partially open, in the following proposition, we first present an integrated result regarding the decision of openness.

Proposition 2.

(a) In a highly competitive market (i.e., when $e \geq \mathscr{E}$), the firm should either fully open source its technology or keep it proprietary.

- *if the cost of collaboration is low (i.e., $a < \bar{\mathscr{A}}$), the firm should make its technology open source (i.e., $s = 1$),*
- *if the cost of collaboration is high (i.e., $a > \bar{\mathscr{A}}$), the firm should keep its technology proprietary (i.e., $s = 0$).*

(b) The firm should make its technology partially open source only if $e < \mathscr{E}$.

Proposition 2(a) suggests that when the market is very competitive (i.e., e is high), the firm should either keep its technology proprietary or make it fully open source depending on the level of the cost of collaboration (i.e., a). More specifically, our results reveal that if the market is very competitive, the firm should keep its technology proprietary when the cost of collaboration is high. On the other hand, in this case, if the cost of collaboration is relatively lower, the firm should fully open its technology without even considering to make it partially open. This result is somewhat counter-intuitive because if the market is very competitive (i.e., e is very high), one may expect that the firm should minimize its loss due to market competition by making its technology partially open under some circumstances. We now discuss the intuition behind this finding.

Note that higher values of parameter e suggest that the firm is likely to have several competitors that are keen to utilize the technology of the firm to develop their own technology. A highly competitive market might suggest that the customer demand for the technology is likely to be fragmented among several firms. In order to increase the demand, the firm may try to increase the quality of the technology and/or open its technology (to increase the demand for the technology via network effects, see Equations (2.4) and (2.5)). The new customer base will in turn increase the rate of improvement of the quality of both versions. Therefore, the loss due to e is compensated by higher demand and reduced effort levels. Further, if the firm only partially opens the technology in a highly competitive environment, the increase in demand because of openness is not very high compared to how much the competitors would gain from the released patents related to the technology (because of high e). Hence, the costs of opening the technology are not compensated by the benefits of openness. Thus, despite competitive market conditions, it is profitable for the firm to make its technology fully open source only if the cost of collaboration is low. Our result is consistent with the strategy of Sun Microsystems and its decision to open source its programming software *Java*. Martens (2006) summarizes the strategy of Sun Microsystems in the following statement: *"Sun is hoping [fully] open sourcing Java will help stop fragmentation in the* *market and instead drive convergence around Java."* Thus, an optimal strategy for the firm to operate in a highly competitive environment is to fully open its technology when the cost of collaboration is relatively low or keep the technology proprietary when the cost of collaboration is relatively high.

After discussing the conditions in which the firm chooses to either keep its technology proprietary (i.e., $s = 0$) or make it fully open (i.e., $s = 1$), we shift our focus to the condition when the firm should partially open the technology (i.e., $0 < s < 1$). The finding in Proposition 2(b) suggests that if the negative impact of competition is relatively low (i.e., $e < \mathscr{E}$), the firm may partially open up its technology. For example, in Figure A.1, as the value of e goes below the threshold $\mathscr{E} \approx 0.26$, the value of $s > 0$. The rationale behind this result is that, by opening up its technology, the firm's cost of collaboration is compensated by: (i) the open source community's effort in fixing defects and new technology development, and (ii) the additional demand for the technology due to network effects. Accordingly, Tesla's approach of opening up its battery technology seems very much in line with our results. In 2014, Tesla considered that the competition among electric car manufacturers was low and the technology would not be used by the big automotive players, and thus $e < \mathscr{E}$. Hence, loss to competition due to opening the technology behind batteries (by opening up its patents) was minimal and Tesla decided to open up part of its technology (Bessen 2014).

2.3.1.1 Discussion on Strategies of Technology Firms

In this subsection, we use our model based results to explain the open source strategies (i.e., to keep the technology of its products proprietary or make it fully open or make it partially open) adopted by four different technology firms, namely, Microsoft, Apple, IBM, and Sun Microsystems in the software industry.

We first discuss the rationale behind Microsoft's decision to keep its operating system (OS), Windows, proprietary. Given that Windows OS is a highly successful product in an extremely competitive marketplace, making Windows open source would not only allow the existing competitors to use the software code and create better competing software, but also allow newer competitors to enter the market place. Furthermore, Microsoft often uses Windows as a platform to promote several of its other services and products (e.g., Microsoft Edge). Moorhead (2017) argues that if Microsoft were to open source Windows, it would diminish the revenues from services and products that depend on Windows. Therefore, opening up its software has a two-fold effect on Microsoft: 1) lower revenue from the sales of Windows OS, and 2) reduced revenues of its other products and services that are based on Windows platform. In other words, the value of e (that captures the competitiveness of market) is extremely high for Windows. Proposition 1(a) suggests that if the market is highly competitive, despite the benefits of increased user based due to network effects, the firm should keep its technology proprietary, which explains Microsoft's Windows strategy to keep its OS code proprietary.

On the other hand, Wayner (2000) observes that large portions of Apple's *Mac OS X* software code came from the free versions of Berkeley Software Distribution (BSD) based operating systems. The project manager of Apple's first open source effort explains the motivation behind their strategy to partially open source *Mac OS X*:

"We realized that the pieces they're most interested in are the most commoditized. There wasn't really any proprietary technology added that we had to worry about them copying. There are people who know them better than we do like the BSD community. We started making the case, if we really want to partner with the universities we should just open the source code and release it as a complete BSD-style operating system" – (Wayner 2000).

This suggests that by opening up part of its technology pertaining to its software, the negative impact of the competition from Apple's perspective was relatively low, and moreover the benefit from the opening up an already commoditized part of the technology outweighed the negative impact. Hence, it was beneficial for Apple to partially open up its technology, in particularly, the code based on BSD OS. This reasoning is similar to our finding and rationale presented in Proposition 2(b), which suggests that the firm should partially open up its code when market competition is low. This provides one plausible rationale behind why Microsoft and Apple had two diverging strategies with respect to their operating systems.

Next, we discuss the motive behind the strategy of IBM to make *WebSphere* fully open source. West (2003) notes that the large open source community working with IBM to maintain *WebSphere* provided greater flexibility and faster implementation of bug fixes. This suggests that the value of b (that captures the sensitivity of the quality of the current version to openness) is high. However, in order to collaborate with the open source community, IBM had to hire permanent liaisons (West 2003), which suggests high value of a. Despite high cost of collaboration, Proposition 1(b) suggests that the firm should open source its technology when b is relatively high, a scenario that corresponds to *WebSphere*. Given that IBM's *WebSphere* open source strategy was a success, this industry evidence validates the robustness of our model based results.

Finally, we shift our focus to explaining the reasoning behind Sun Microsystem's strategy to make Java fully open source. As discussed earlier, Proposition 2(a) suggests that the firm should make its technology fully open source if the market is very competitive (i.e., e is high) and if the level of the cost of collaboration (i.e., a) is relatively low. In late 1990s and early 2000s, the marketplace for general purpose programming, which Java belongs to, was extremely competitive, i.e., e was high. Furthermore, Sun Microsystem's existing ties with the open source community kept the cost of collaboration considerably lower, i.e., low a. In such a scenario, our model based result can explain one plausible reason behind Sun Microsystem's decision to fully open source its software.

2.3.2 Should the Firm Focus on Maintenance or Development

Next, we present and analyze the conditions when the firm should focus more on the maintenance effort or the development effort. The results are summarized in the proposition below.

Proposition 3. *If $\frac{w}{c_f} > \frac{n}{c_d}$, the maintenance effort (i.e., u) is always greater than the development effort (i.e., v). However, if $\frac{w}{c_f} \leq \frac{n}{c_d}$, the maintenance effort is not always less than the development effort.*

We first define the productivity in developing the next version as the sensitivity to the effort levels on the development of the next version adjusted by the respective costs (i.e., w/c_f and n/c_d). The proposition reveals that if the productivity of the maintenance effort is higher than the productivity of development effort (i.e., $w/c_f > n/c_d$), then the firm should exert more effort on maintenance throughout the project. Hence, if the next version relies heavily on the current version (i.e., high w), the focus should be more on the maintenance effort. The reasoning behind this finding is as follows. Higher values of w implies that higher maintenance effort would lead to faster rate of development of next version. Hence, if the firm shifts its focus on fixing the defects in the existing version, it would eventually lead to increased quality of the next version as well. Thus, in this scenario, by focusing more on the maintenance effort, the firm is able to increase the inherent quality of both the current version and the next version. This finding is illustrated in Figure A.3(a). On the other hand, if $w/c_f < n/c_d$, interestingly, the firm's focus is not necessarily more on the development effort. Rather, we find that, in certain scenarios, the focus of the firm should be more on the maintenance effort in the earlier phases and the focus on the maintenance effort is reduced in the later stages of the project as illustrated in Figure A.3(b).

Therefore, the firm should carefully evaluate the characteristics of its technology (such as the cost of efforts and productivity levels) before allocating the resources to the maintenance effort and the development effort. In the following subsection, we now examine the behavior of the firm with respect to the changes in the valuation of the current version of the technology.

2.3.3 Impact of the Valuation of Current Version on Firm's Behavior

In this subsection, we discuss the firm's reaction to changes in the valuation of the current version of the technology (i.e., k). If the valuation increases, in order to benefit more from the higher valued technology, the firm has an incentive to focus on increasing the demand for current version. The demand for existing technology depends on the network effects from opening the technology and the quality of the current version of technology. Thus, to increase the quality of the current version of technology, the firm should increase: (i) the effort of fixing defects, and/or (ii) the portion of the technology that is open.

We find that if the valuation increases, the firm's effort levels and the extent of openness depend on the sensitivity of the quality of the current version to openness (i.e., b). In particular, we find that, if the current version is not much sensitive to openness (i.e., b is low), the firm reduces the openness but increases its maintenance and development efforts to boost its demand due to higher quality. However, for moderate levels of b, the firm decreases its development effort and shifts its focus towards improving the quality of the current version by increasing both openness and maintenance effort.

Furthermore, we find that if the current version is highly sensitive to openness (i.e., b is high), the firm relies more on the open source community by increasing the openness and decreasing both of its effort levels (maintenance and development). By doing so, the firm is able to save costs related to fixing and development. Note that to benefit from the increase in valuation, the firm does not decrease both the maintenance effort and the openness at the same time because it reduces the rate of improvement of the quality (i.e., $\dot{q}(t)$), which in turn affects the demand for the technology. We also find that whenever the firm increases its development effort, it also increases its maintenance effort. We outline these findings in the following proposition.

Proposition 4. *With an increase in the valuation of the current version (i.e., k)*

(a) If the sensitivity of the quality of the current version to openness (i.e., b) is low (i.e., $b < \beta_0$),

(i) both the development effort of the next version (i.e., v) and the maintenance effort (i.e., u) increase,

(ii) the portion of the technology that is open (i.e., s) decreases.

(b) If the sensitivity of the quality of the current version of the technology to openness is at a moderate level (i.e., $\beta_0 < b < \beta_1$),

(i) the development effort of the next version decreases but the maintenance effort increases,

(ii) the portion of the technology that is open increases.

(c) If the sensitivity of the quality of the current version to openness is high (i.e., $b > \beta_1$),

(i) both the development effort of the next version and the effort of fixing defects decrease,

(ii) the portion of the technology that is open increases.

The results of Proposition 4 indicate that the managers should be careful in planning their response to changing market conditions. According to parts (a) and (b), an increase in the valuation of the current version should be accompanied with an increase in the internal resources who are responsible for maintaining the technology if the quality of existing technology is not much sensitive to openness. On the other hand, when the quality of the current version is highly sensitive to openness, the firm should be prepared to increase the openness in order to benefit more from the open source community and reduce internal resources who are responsible for maintenance and development (see part (c)). As discussed earlier, the maintenance effort decreases with time, and therefore the firm may keep some part-time or contract workforce (for the maintenance project) to handle the varying resource requirement.

Furthermore, to ensure that the conditions in the above proposition are reasonable, we perform numerical experiments. One representative instance of our numerical study that illustrates the findings in the Proposition 4 is shown in Figure A.4 in Appendix A.3. Interestingly, the threshold β_1 in part (c) of Proposition 4 depends on time t (i.e., the stage of the project). In particular, we find that the value of β_1 is very high in the early stages of the project compared to the later stages of the project. In order to check the feasibility of the condition in part (c) of the proposition, we conduct a numerical study. Indeed, we find that the value of β_1 is extremely high relatively in the initial stages. Figure A.5 in Appendix A.3 illustrates the values of β_1 at various stages of project. As apparent in this figure, the values of b needs to be relatively high for the condition $b > \beta_1$ to be satisfied at the beginning of the project, a very unlikely case. In fact, this finding confirms the intuition that it is not reasonable to reduce the maintenance effort when the valuation of the technology increases. However, if the open source community is indeed very capable and effective (i.e., the open source community finds and fixes all the defects in the technology), then part (c) of the proposition applies. We also find that in the later stages of the project, it is indeed possible to reduce the maintenance effort despite a higher valued technology. The reason behind this finding is that the firm has a very little incentive to fix defects at the end of the project as the new version is about to be released.

As discussed earlier, in certain environments, the firm may have a limited budget and the overall spending on the workforce (or the effort levels) cannot exceed this budget (i.e., $X(T) \leq B$). We conduct an extensive numerical study to analyze how the firm reacts to increased valuation of the technology in such a setting. Interestingly, in contrast to the findings in part (a) of Proposition 4, we find that if the valuation of the current version of the technology increases, the extent of openness may increase despite a low level of sensitivity of the quality of the current version to openness. The intuition behind this result is that when the valuation increases, the firm has an incentive to increase quality, and consequently the demand for the technology. However, if the firm has a limited budget, the effort levels are already constrained. Hence, in this scenario, the firm's only leverage is openness. This is because openness might increase the demand (due to network effects, through parameters θ_2 and ρ_2) and quality (even though at a lower rate when b is low). The details of this numerical study are omitted for brevity, but we summarize this finding in the following observation.

Observation 1. *If the firm has a limited budget, as the valuation of the current version (i.e., k) increases, the firm might increase the extent of openness (i.e., s) even if the sensitivity of the quality of the current version to openness (i.e., b) is low.*

In the next subsection, we now analyze how the changes in openness sensitivity terms (i.e., b and y) affect the behavior of the firm.

2.3.4 Impact of Openness Sensitivity Terms on Firm's Behavior

We first examine the implications of a change in the sensitivity of the current version to openness (i.e., b) on firm's effort levels and the extent of openness. An increase in b implies that the open source community becomes more effective or helpful in fixing defects. In such a case, Equation (2.2) implies that the inherent quality of the current version increases at a faster rate with the extent of openness (i.e., s). Hence, we find that the firm increases the openness with b. On the other hand, the firm decreases both fixing and development efforts. Furthermore, we find that the profit of the firm increases with b. These results are formally presented in Proposition 5.

Proposition 5. *If the effectiveness of the open source community increases in fixing defects (i.e., b increases), the*

(a) portion of the software that is open source (i.e., s) increases,

(b) maintenance effort (i.e., u) decreases,

(c) development effort (i.e., v) decreases,

(d) profit of the firm increases.

In summary, it is advantageous for the firm to have a higher b since it is able to reduce its effort levels but increase its profit. Therefore, the managers should pro-actively seek ways to make the members of the open source community more effective and more interested in the technology. For example, the managers can consider providing the open source community with development tools, access to testing environments, educational materials, or any other resource that can increase their productivity and make them more engaged.

Next, we examine the change in the behavior of the firm when the sensitivity of the development rate of the next version to openness (i.e., y) increases. An increase in this sensitivity term implies that the members of the open source community become more effective or engaged in developing new features. Because of this higher level of engagement of the open source community, the firm can build the next version of the technology at a reduced level of development effort. In effect, the firm also reduces the maintenance effort. Hence, the firm benefits from the reduced costs of effort on both maintenance and development. Further, the marginal benefit of an increase in y is higher for the firm if the open source community can access and work on a larger portion of the technology. Hence, intuitively, the firm should increase the extent of openness as y increases. However, we find that this is not always the case. In particular, if the cost of collaboration (i.e., a) is low, the firm, in fact, decreases the extent of openness with y. This finding is summarized in Proposition 6(a) below and illustrated in Figure A.6(a) in Appendix A.3.

Proposition 6. *As the effectiveness of the open source community increases in developing the next version of the technology (i.e., y),*

(a) if the cost of collaboration with the open source community is low (i.e., $a < \alpha$),

(i) both the development effort of the next version (i.e., v) and the effort of fixing defects (i.e., u) increase,

(ii) the portion of the technology that is open (i.e., s) decreases;

(b) if the cost of collaboration with the open source community is high (i.e., $a > \alpha$),

(i) both the development effort of the next version and the effort of fixing defects decrease,

(ii) the portion of the technology that is open increases.

Clearly, the result presented in Proposition 6(a) is somewhat counter-intuitive, but can be explained as follows. It is apparent from Lemma 2 that $\frac{ds^*}{da} < 0$, hence the extent of openness actually decreases with the cost of collaboration. Therefore, when the cost of collaboration is low, the portion of the technology that is open is already at a high level. Hence, increasing the sensitivity of the next version on openness enables the firm to reduce the extent of openness and help it save both the direct cost (i.e., the collaboration cost between the firm and the open source community) and the indirect cost (due to competition). However, our numerical results suggest that the decrease in openness and increase in efforts is marginal (see Figure A.6(a)). On the other hand, if the cost of collaboration is high, the portion of the technology that is open is at a low level (see Figure A.6(b)). Hence, as the open source community becomes more effective in developing the next version, it is beneficial for the firm to increase the extent of openness despite the high collaboration cost (see Proposition 6(b)).

To summarize, the managers should be careful about the changes in the effectiveness of the open source community in developing the next version. This is because, although increasing the extent of openness seems to be a reasonable strategy (with an increase in the effectiveness of the open source community), the firm should actually decrease it if the collaboration cost is low. However, in this case, since the overall contribution of the open source community increases (regardless of the reduced level of openness), the firm is able to reduce costs by decreasing both fixing and development efforts. Furthermore, our numerical results (not presented for brevity) suggest that the above proposition holds true even when the firm has a limited budget. This indicates that the results presented in Proposition 6 are robust. In the next subsection, we now examine the behavior of the firm with respect to the changes in the valuation of the next version of the technology.

2.3.5 Impact of the Valuation of Next Version on Firm's Behavior

We now shift our focus to analyzing the firm's reaction to the changes in the valuation of the next version of the technology (i.e., h). When h increases, the firm has an incentive to increase the demand for the next version of the technology. In particular, the firm can increase the demand by: (i) increasing the quality of the next version, and/or (ii) increasing the extent of openness (see Equation (2.5)). To take advantage of the higher valuation, we find that it is optimal for the firm to increase its effort levels (both maintenance and development). However, the optimal extent of openness may increase or decrease. When the market competition is high, intuitively, the firm should decrease the openness. On the contrary, we find that the firm should increase the extent of openness when the market is highly competitive and the development time for the new technology (i.e., T) is relatively low. This result can be explained as follows. When the market competition is high, the extent of openness is low. Therefore, an increase in the valuation of the next version entices the firm to increase the extent of openness from its low baseline since the time to release the next version is also short. We outline these findings in the following proposition and illustrate them in Figure A.7 in Appendix A.3.

Proposition 7. *As the valuation of the next version (i.e., h) increases, the portion of the technology that is open (i.e., s) increases if the market competition is high (i.e., $e > \bar{\mathscr{E}}$) and the time to release is low (i.e., $T < \bar{\mathscr{T}}$).*

Proposition 7 indicates that the managers should be careful when the valuation of the next version increases. Under the conditions presented in the proposition, with an increase in the valuation of the next version, the firm should go "all-in" in the next version by increasing the openness as well as both maintenance and development efforts. On the other hand, when the firm has a limited budget, we numerically find that the openness may increase regardless of the values of e and T. This is because when the firm is constrained by a budget to achieve a higher quality level for the next version, the firm's only choice is to increase the openness. We summarize this finding in the following observation.

Observation 2. *When the budget is limited, as the valuation of the next version (i.e., h) increases, the firm may increase the portion of the technology that is open (i.e., s^*) regardless of the level of market competition (i.e., e) and the time to release the new version of the technology (i.e., T).*

We extend our discussion in the next section to the markets that have a pre-determined minimum quality level for the next version.

2.4 Minimum Quality Threshold for the Next Version

In this section, we analyze a scenario in which the firm has to meet a certain quality threshold before the next version of the technology is released. This quality threshold, Z, may be determined by customer expectations or market requirements, and hence considered as an exogenous variable in our model. In other words, the quality of the next version (that we denote by $r(t)$ at time T) needs to be at least Z at the time it is released (i.e., $r(T) \geq Z$). In this setting, the technology firm attempts to meet the required quality level for the next version by taking one (or more) of the following actions: (i) increase the development effort of the next version of the technology (direct effect), (ii) increase the effort of fixing defects in the current version of the technology (indirect effect), and (iii) increase the portion of the technology that is open source (indirect effect).

In the rest of this section, we consider the setting where the firm needs to meet the market expectations (i.e., $r(T) = Z$) (if the firm finds it profitable to exceed the quality threshold (i.e., $r(T)^* > Z$), then the analysis in Section 2.3 applies). Hence, in addition to the constraints presented in Equations (2.2) - (2.9), an additional constraint, $r(T) = Z$ needs to be considered in the model. Although most of the solutions and variables of interest are cumbersome to present, we glean several results and managerial insights. In the following lemma, we first outline the solution of the corresponding optimal control problem for an arbitrary level of openness.

Lemma 3. *If the market has a minimum quality requirement for the next version (i.e., $r(T) = Z$), the optimal trajectories of the maintenance effort (i.e., $u(t)$) and the development effort (i.e., $v(t)$) for an arbitrary level of openness (i.e., s) are given by:*

$$\begin{aligned} u(t) &= \frac{\theta_1 kmT\left(2c_f n^2\left(T-t\right)+c_d w^2\left(T-2t\right)\right)+4c_f c_d w\left(Z-sTy\right)}{4c_f T\left(c_f n^2+c_d w^2\right)}, \text{ and} \\ v(t) &= \frac{n\left(4c_f\left(Z-sTy\right)-\theta_1 kmT^2 w\right)}{4T\left(c_d w^2+c_f n^2\right)}. \end{aligned}$$

Lemma 3 reveals that opening up a higher portion of the technology (i.e., increasing s) decreases both fixing and development efforts. This result is similar to that in Lemma 1, and can be explained in a similar manner. By examining the interdependencies among the extent of openness and both types of efforts, and the effects of openness on the quality and demand of the current and next versions of the technology, the firm can determine the optimal level of openness. The results of this analysis, i.e., the optimal level of openness, corresponding optimal effort trajectories, and the profit of the firm are presented in Lemma 4.

Lemma 4. *The optimal level of openness (s^*), effort of fixing defects ($u^*(t)$), development effort ($v^*(t)$), and the profit of the firm (Π^*) are as follows if the market has a pre-determined quality requirement (Z):*

$$\begin{aligned}
s^* &= \frac{\begin{array}{c}[bc_f\theta_1kn^2T^2 + bc_d\theta_1kT^2w^2 - 2c_fen^2\rho_0 - 2c_fen^2\rho_1Z - 2c_de\rho_0w^2 - 2c_de\rho_1w^2Z \\ + 2c_fhn^2\rho_2 + 2c_dh\rho_2w^2 - c_d\theta_1kmT^2wy + 2c_f\theta_2kn^2T + 2c_d\theta_2kTw^2 + 4c_fc_dyZ]\end{array}}{4\left(ac_fn^2T + ac_fTw^2 + c_fen^2\rho_2 + c_de\rho_2w^2 + c_fc_dTy^2\right)}, \\
u^*(t) &= \frac{\begin{array}{c}\left[c_d\theta_1(-k)mTw^2\left(2t-T\right)\left(aT+e\rho_2\right) + c_f\left(-2\theta_1kmn^2T\left(t-T\right)\left(aT+e\rho_2\right) + \right.\right. \\ \left.\left. c_d\left(2Tw\left(2aZ + e\rho_0y + e\rho_1yZ - h\rho_2y\right) + \theta_1kT^3y(2my - bw) + 4e\rho_2wZ - 2kT^2y\left(\theta_1mty + \theta_2w\right)\right)\right)\right]\end{array}}{4c_fT\left(c_f\left(T\left(an^2 + c_dy^2\right) + en^2\rho_2\right) + c_dw^2\left(aT + e\rho_2\right)\right)}, \\
v^*(t) &= -\frac{n\left(c_f\left(T\left(y\left(kT\left(b\theta_1T + 2\theta_2\right) - 2e\left(\rho_0 + \rho_1Z\right)\right) - 4aZ\right) + 2\rho_2\left(hTy - 2eZ\right)\right) + \theta_1kmT^2w\left(aT + e\rho_2\right)\right)}{4T\left(c_f\left(T\left(an^2 + c_dy^2\right) + en^2\rho_2\right) + c_dw^2\left(aT + e\rho_2\right)\right)}, \\
\Pi_{r(T)=Z} &= k\left(\frac{\theta_1T\left(12c_f\left(wc_d(bs^*Tw + m(Z - s^*Ty)) + bn^2s^*Tc_f\right) + \theta_1km^2T^2\left(w^2c_d + 4n^2c_f\right)\right)}{24c_f\left(w^2c_d + n^2c_f\right)} + \theta_2s^*T + \theta_0T\right) \\
&\quad - \frac{\theta_1^2k^2m^2T^4\left(w^2c_d + 4n^2c_f\right) + 48c_dc_f^2(Z - s^*Ty)^2}{48Tc_f\left(w^2c_d + n^2c_f\right)} + \left(h - es^*\right)\left(\rho_0 + \rho_2s^* + \rho_1Z\right) - as^{*2}T.
\end{aligned}$$

The behavior of the variables presented in Lemma 4 with respect to changes in different parameters are mostly similar to the results presented in Section 2.3. Hence, for brevity, we present only a limited number of findings that are interesting and different from those in the earlier analyses.

2.4.1 Impact of the Valuation of the Current Version on Firm's Behavior

The valuation of the current version (i.e., k) might increase because of several reasons. For example, the technology might become more valuable to customers if business conditions or regulations change. In Proposition 4 (in Section 2.3), we find that the firm decreases its development effort (of the next version) with k only if the sensitivity of the quality of the current version to openness is above a certain threshold (i.e., $b > \beta_0$). However, interestingly, when the market requires a minimum quality threshold, we find that the firm always decreases its development effort with k and its focus tends to shift towards improving the quality of the current version. The reason why the development effort decreases lies in the fact that when the quality of the next version is pre-determined by the customer and is high, the firm already has a high level of development effort to meet the quality requirement. An increase in the valuation of the current version gives the firm an incentive to increase either the maintenance effort or openness in order to benefit more from the current version. Furthermore, since there are indirect benefits of the maintenance effort and openness on the development rate of next version, the firm can decrease the development effort and still meet the quality threshold Z. Also, similar to the results in Proposition 4, with an increase in the valuation of the current version, the firm finds it beneficial to increase either the maintenance effort or the portion of the technology that is open, and there is no case in which both decrease.

Since, with a pre-determined quality requirement, the impacts on the effort of fixing defects (i.e., u) and the portion of technology that is open (i.e., s) remain qualitatively similar to those in Proposition 4 (with different thresholds), we do not present them formally. Hence, in the following proposition, we present the impact of increasing k only on the the development effort of the next version (which is different from that in Proposition 4).

Proposition 8. *As the valuation of the current version (i.e., k) increases in a market that has a pre-determined quality requirement of Z, the development effort of the next version (i.e., v) decreases.*

When the firm has a limited budget, despite minimum quality requirement for the next version of the technology, we observe similar findings as in Observation 1.

2.4.2 Impact of the Valuation of the Next Version on Firm's Behavior

We now examine how the valuation of the next version of the technology (i.e., h) affects the extent of openness and the effort levels in the following proposition.

Proposition 9. *As the valuation of the next version (i.e., h) increases in a market that has a pre-determined quality requirement of Z, the portion of the technology that is open (i.e., s) increases but both the maintenance effort (i.e., u) and the development effort (i.e., v) decrease.*

When there is no minimum quality threshold, with an increase in the valuation of the next version (i.e., h), the firm actually increases both of the effort levels and sometimes decreases the extent of openness depending on the other factors (see Section 2.3.5). However, the above proposition shows that, with a pre-determined quality requirement, it is more profitable for the firm to decrease both maintenance and development effort levels and increase the extent of openness as the valuation of the next version increases. The reason for the difference in the firm's behavior is that, in the main model (i.e., Section 3), the quality of the next version is not fixed and the firm can decide to increase or decrease the quality level depending on the characteristics of the market such as the changes in the valuation of the technology. However, in the current model, to take advantage of an increase in the valuation of the next version, the firm would attempt to increase the demand. The demand for the next version depends on: (i) the quality of the next version, and (ii) the extent of openness. Since the goal of the firm is still meeting the minimum quality threshold, it is optimal to increase the demand by increasing openness, but not by exceeding the quality threshold. Therefore, the firm decreases both effort levels and saves cost while meeting only the quality threshold.

2.4.3 Impact of Changes in Market Requirements on Firm's Behavior

As discussed earlier, the technology systems need to be replaced with new versions beyond a point. One of the reasons for switching to a new version is to provide the functions needed or required by the market or existing customers. Hence, if such differences or requirements become more involved (i.e., Z increases), the firm would need to exert more effort in order to meet or exceed those expectations. In effect, in such a case, we find that not only the development effort of the next version increases, but also the maintenance effort (of the current version) increases (because of its indirect contribution to the quality of next version; see Equation (2.3)). In the next proposition, we summarize the impact of change in Z on the extent of openness and the effort levels.

Proposition 10. *As the market requirements (i.e., Z) increase,* *(a) the portion of the technology that is open (i.e., s) increases (resp., decreases) if the market competitiveness is low, i.e., $e < \eta$ (resp., $e > \eta$);*

(b) both the development effort of the next version (i.e., v) and the maintenance effort of the current version (i.e., u) increase.

If the market competitiveness is high (i.e., $e > \eta$), by reducing openness, the firm reduces the risk of its (i) rivals getting a competitive edge, and (ii) customers tailoring the technology to their needs and distributing it to other users. In addition, the firm is also able to reduce the costs of collaboration with the open source community. On the other hand, if the competitiveness is at a low level (i.e., $e < \eta$), the firm should increase the extent of openness in order to benefit more from the open source community. As discussed earlier, regardless of the competition level, the firm should increase the effort levels to meet the higher market requirement. These findings are illustrated in Figure A.8 in Appendix A.3.

On the other hand, when the firm has a limited budget, our numerical results suggest that if the market requirement increases, the firm may increase the extent of openness despite higher market competition. In this case, the effort levels may not increase because the firm already uses its available resources (as $x(T) = B$). However, to be able to meet the higher level of market requirement, the firm increases technology openness and expects the open source community to help them develop a higher quality product. We summarize this finding in the following observation.

Observation 3. *When the budget is limited, as the market requirements increase, the firm may increase the portion of the technology that is open (i.e., s^*) regardless of the level of market competition (i.e., e).*

2.5 Conclusions

This study is among the first of its kind to model the effects of making a portion of its technology open source on the quality and maintenance of the current version of the technology, and on the quality and development of the next version. In order to determine the best course of action in open source environments, the technology firms need to make several decisions. In particular, they need to decide on the level of maintenance effort in the current version and the level of development effort of the next version throughout the planning horizon. In addition, a firm should also decide whether to make the technology open source or keep it proprietary. If the managers decide to make the technology open source, they also need to determine the portion of the technology that is going to be open source. Our study aims to help managers in their decision making process and highlight the importance of various factors that affect the optimal course of action, such as the characteristics of the market (e.g., market competitiveness, market expectations from the next version of the technology, and the valuation of the technology), characteristics of the firm (e.g., the cost of fixing technology defects, the cost of development effort, and the available budget), and the effects of making the technology fully or partially open source on the quality and demand of current and next versions of the technology.

In this study, we derive and discuss several interesting and useful results and managerial insights. For example, we find that the firm might keep its technology proprietary even in conditions that seem to favor an open source environment. We also investigate the conditions when the technology should be made fully open source, and find that the firm might make its technology fully open source despite high collaboration costs. In addition, the sensitivities of the current and next versions of the technology on openness influence the firm in deciding whether to make the technology open or not. We further recommend managers to be careful in planning their response to changes in the cost of managing the collaboration with the open source community. This is because a drastic change in the behavior of the firm may be observed if this cost changes. In particular, we find that the firm might change its decision from keeping the technology proprietary to fully open (without any consideration of partial openness) when the cost of collaboration decreases.

Furthermore, we investigate how the valuation of the technology might affect firm's decisions, and find that the managers should be careful in planning their response to changing market conditions. An increase in the valuation of the current version of the technology might require the firm to reduce internal resources who are responsible for fixing defects and developing the next version of the technology. Instead, the managers should be conducive to increase the extent of openness in order to compensate for the reduction in fixing and development efforts.

We further examine other business settings. For example, in some environments, firms might be required to meet a specific quality threshold before the next version of the technology is released and the firm might have a limited budget. Hence, we also analyze such settings and provide useful managerial insights. For example, we find that if the firm has a limited budget, it might increase the extent of openness with increasing market requirements despite high levels of market competition. However, if the firm is not constrained by a budget and the market is highly competitive, the firm might decrease the extent of openness if the market requirements increase.

3. ESSAY 2: A FRAMEWORK FOR ANALYZING INFLUENCER MARKETING IN SOCIAL NETWORKS: SELECTION AND SCHEDULING OF INFLUENCERS

3.1 Introduction

The ubiquity of social media allows businesses to easily access massive online social networks and interact with users on these networks. The number of users on various social media platforms such as *Facebook, Instagram, LinkedIn, Twitter*, and *YouTube* has increased from 0.91 billion in 2010 to more than 2.46 billion in 2017 (eMarketer 2018). Consequently, the total spending on social media advertising is expected to increase by two-fold from $7.52 billion in 2014 to $15.36 billion in 2018 to leverage the potential of social media (Statista 2018b). However, Aral et al. (2013) note that several operational aspects of advertising through social networks are yet to receive a rigorous academic examination.

Although firms have a presence on social media, they often rely on opinion leaders or influencers (e.g., Cristiano Ronaldo, a soccer player), who share their experiences with products on social media, to market their products. Indeed firms have advertised their products through influencers for many decades. However, social media has enabled them to quantify the value that influencers bring and effectiveness of their influence on social networks. The focus of this study is on analyzing "*influencer marketing*" in social media for selection of influencers and scheduling the posts of influencers over a planning horizon. Influencer marketing involves hiring opinion leaders or influencers[1] (e.g., users can influence their followers) to seed (or post) ads on behalf of a firm on their social media platforms thereby disseminating the contents of the ad to their existing followers on the platforms.

Influencer marketing relies on two essential features of social media platforms. First, in addition to the massive size of social networks, social media enables influencers to directly communicate with the audience without an intermediary. Second, social media platforms facilitate message propagation, which plays a vital role in the success of influencer marketing.

[1] In the rest of this paper, we use the terms opinion leaders and influencers interchangeably.

In particular, when an influencer posts a message (or an ad) on his/her social media page, the message will directly reach all the followers of the influencer. Furthermore, the followers of influencer may choose to share the same message on their social media page. As a result, the message reaches the followers of followers who decide to share the message of an influencer. In other words, the ad posted by an influencer reaches the audience that includes direct followers of the influencer on social media and indirect followers of the influencer by virtue of message propagation.

Advertising through influencers on social media, commonly known as seeding, has become popular in the recent years. For example, Twitter (Karp 2016) reports that "*nearly 40% of Twitter users say they've made a purchase as a direct result of a Tweet from an influencer.*" In one study, it was found that 86% of the firms have utilized influencers to promote their products in 2017 and 92% of these firms found this strategy to be effective (Linquia.com 2017). Another finding from this study is that the spending on influencer marketing by firms (25% of firms on average spend $125,000-$250,000 per year) is set to increase by 39% in 2018. The main reason behind the rise in influencer marketing is primarily attributed to (i) the unique features of social media (as discussed earlier), and (ii) the findings of recent studies, which suggest consumers are more likely to trust ads from influencers compared to regular ads (Duran 2017). However, the literature on how to conduct an influencer marketing campaign on social media is scarce. In this study, we seek to bridge this gap by developing a data-driven optimization framework to conduct an influencer marketing campaign on social media successfully.

Prominent social media influencers are often celebrities including actors, writers, musicians, politicians, or athletes. Cristiano Ronaldo is one of the popular influencers with a ten-year $1 billion deal with Nike to promote their products on his social media. Hookit (2015) estimates that the tweets posted by Cristiano Ronaldo in 2016 to promote Nike brand generated an estimated media value of $499.6 million for the firm. In addition to celebrity influencers, there are a growing number of "*non-celebrity*" influencers on social media platforms. Micro influencers, as non-celebrity influencers are often referred to, are social bloggers with expertise in a particular product category with a considerable amount of followings on social media (e.g., 5,000 - 100,000 followers). Although micro influencers may not have large followings like celebrities (e.g., Cristiano Ronaldo, who has more than 70 million followers on Twitter), they are considered to (i) be more cost effective, (ii) generate higher social media engagement, and (iii) have knowledge in the product domain and hence are considered more trustworthy compared to celebrity influencers (Barker 2017). MarketingHub (2017) reports that most of the influencers (around 90%) that are hired by the firms (e.g., Ford, Dr.Pepper, and La Croix) belong to the category of micro-influencers. Furthermore, another study reports the average price per social media post by a micro-influencer is around $271 (Heald 2017). Moreover, this form of advertising is not limited to a few countries but is prevalent across the world. For example, a study estimates presence of more than 350,000 influencer in India (Das 2018).

3.1.1 Problem Statement

The primary setting of the problem that we study is as follows. A firm wants to launch an influencer marketing campaign on a social media platform, and the goal is to maximize the effectiveness of this ad campaign. Industry reports suggest that firms typically work with multiple influencers. For example, Linquia.com (2018) reports that on average 29% of firms typically work with 1-10 influencers per marketing campaign and 52% of the firms work with 10-25 influencers. Therefore, the success of an influencer marketing campaign depends on (i) selecting the right set of influencers, and (ii) proper scheduling of ads by multiple influencers during the planning horizon. The firms usually have a limited budget for social media marketing, and hence can only hire a certain number of influencers for the campaign and employ them throughout the planning horizon. The firm attains an expected weighted benefit when the followers of the influencers receive and engage with the ad. In this paper, we develop a comprehensive modeling framework to support decision making related to conducting an influencer marketing campaign on a social media platform (such as

Twitter). To do this, we break the problem into two phases.

In the first phase, we begin by developing an optimization model for selection of an optimal influencer set from a larger set of influencers (which we refer to as set W) who are appropriate for promoting the firm's products. This set W is obtained by marketers qualitatively based on various product attributes (York 2018). For example, if a firm is interested in promoting athletic shoes, it can identify a short list of influencers in the domains of fitness and sports. Marketers can use platforms such as www.buzzsumo.com or www.shoutcart.com to identify a list of influencers for set W. A simple web search revealed that there are approximately 150 micro-influencers in the domains of fitness and sports on one of the platforms. Marketers can use these 150 influencers as set W or the size of set W can be further reduced based on other attributes. Furthermore, according to Businesswire (2016), about 67% of marketers consider identifying the right set of influencers as one of their biggest operational challenges. Common industry practice is to hire popular influencers (i.e., influencers with large following). However, using real data collected from Twitter, we demonstrate that this strategy is sub-optimal (refer to Section 3.4). To model the problem of selection of an optimal influencer set from set W, we need to account for the following three unique features of online social networks.

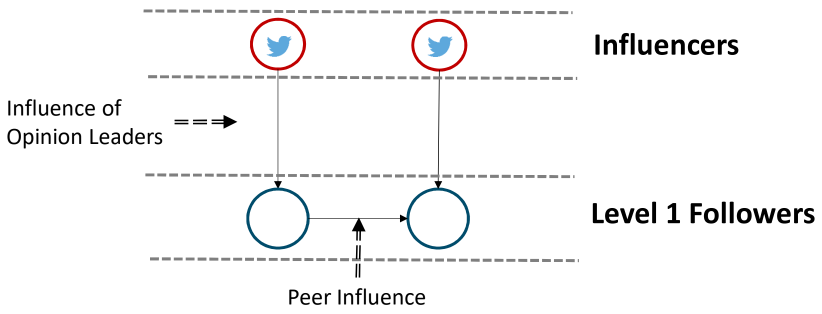
- First, different influencers have different levels of following (i.e., number of users following an influencer) and engagement (i.e., followers sharing the posts of influencers to their followers). For example, an influencer might have a large following but lower engagement levels, while another influencer might have a relatively lower following but higher engagement levels. As a result, selecting an influencer either by the number of followers or by engagement levels may not be optimal.

- Second, influencers might have a common subset of overlapping followers. Consequently, the effect of multiple exposures (i.e., a follower getting exposed to the same ad from multiple influencers) needs to be considered.

- Third, in the context of the current study, social media users may receive the same ad from multiple influencers (i.e., if a user follows multiple influencers) and also from followers of influencers. For example, let us assume that users Y and Z follow influencer X and user Z follows user Y. When influencer X tweets message, it reaches users Y and Z. Further, if user Y chooses to retweet, the message reaches user Z the second time. We define the influence of an influencer (i.e., X) on a follower as the *influence of an opinion leader* and the influence of user Y on user Z is referred to as the *peer influence* (See Figure 3.1).

Hence, the problem of selection of influencers who can seed a firm's ad such that the message reaches their followers and also influence their followers to propagate the message to other social network members is a classic combinatorial optimization problem.

Figure 3.1: Influence of an Opinion Leader and a Peer



Next, given that firms hire multiple influencers for their marketing campaign, it is also important to sequence the ads posted by influencers on social media over a planning horizon. For example, if a firm chooses to hire ten influencers for three weeks to promote a product, the following questions arise. Should the influencers randomly post the ad on their social media pages or should all the influencers post the ad simultaneously at the same time or same day or is there an optimal sequence in which the influencers should post the ad on social media? The problem of scheduling of ads to be posted by influencers is not trivial as one needs to take into account (i) the multiple exposure effect and (ii) the impact of time between each exposure. The multiple exposure effect is because of the following two reasons. First, an influencer is often hired to post the same ad several times, and hence the effect of a follower receiving the same ad multiple times from the influencer needs to be taken into account. Second, as influencers may have a common subset of followers, the firm needs to take into account ads reaching the audience from multiple influencers and multiple times during the planning horizon. Furthermore, findings from an empirical study that we conduct suggest that time between two exposures of the same ad also plays a vital role in determining whether a follower engages with the ad or not. Hence, the problem of ad scheduling by influencers is not trivial. Consequently, in the second phase, we address this issue by developing an optimization framework to optimally sequence the posting of ads by multiple influencers (who are selected in the first phase) over a planning horizon, a relevant and critical problem from firm's perspective.

3.1.2 Contributions

To the best of our knowledge, this study is the first to develop a comprehensive optimization framework to conduct an influencer marketing campaign in social media. As discussed earlier, our framework consists of two phases, namely, (i) selecting influencers or seeders in the context of influencer marketing on social media, and (ii) scheduling of ads by multiple influencers over a planning horizon.

With regard to the problem of selecting an influencer set, our contributions are as follows.

- We develop a mixed integer non-linear model (hereon, we refer to this model as the main model) based on observations and findings from an empirical study that we conduct using Twitter data. We show that the complexity of this model is NP-Hard. Hence, the need for fast and effective solutions to solve large-size problems is supported by the combinatorial explosion in the number of feasible influencer sets.

- To obtain a solution for the main model in realistic time, we develop an alternative model that provides a tight upper bound to the main model. Although the upper bound model is non-linear and intractable, we reformulate the model as a linear integer program using separable programming techniques. This enables us to obtain an optimal solution for the upper bound model for small and medium-size problems. Further, to deal with a large-size problem, we provide an LP-based largest fraction rounding heuristic.

- We find that the average gap is less than 6% between the upper bound model and the main model. Furthermore, we find that the solution provided by the upper bound model is an optimal solution for the main model in more than 94% of the instances.

- We develop a lower bound to the main model by ignoring peer influence. We find that ignoring peer influence could lead to solutions significantly worse than the optimal solutions. In particular, our numerical analysis reveals that the average gap between the lower bound and the optimal solution of the main model ranges from 5% to 81%. Hence, it is in the best interest of the firm to consider peer influence in modeling the influencer selection problem in promoting a product.

Next, we develop an optimization model to help firms with scheduling of ads by multiple influencers based on observations and findings from an empirical study. To summarize, our study provides a framework for managers to effectively conduct an influencer marketing campaign on social media. The model provides a guideline for the selection of influencers, the frequency of the ads to be posted by influencers, and the schedule for posting ads.

We also find that the objective function of the selection model is monotonic and there remains a threshold on the number of influencers that we can hire. This is because of diminishing marginal returns of hiring an additional influencer. We demonstrate how the selection model can be implemented via two case studies that we conduct using data from Twitter. We also validate the superiority of the selection model against current industry benchmarks via the case studies. The results of both case studies demonstrate that the model-based solutions have significant gains in the effectiveness over the current industry practice of selecting influencers. With regard to the scheduling model, we find that the cost of conducting an influencer marketing campaign increases non-linearly with minimum engagement requirements of the firm. Further, we provide insights on how frequently influencers are used throughout the planning horizon.

3.1.3 Related Literature

This research builds on the existing research in operations and computer science. There is a large body of related work on information diffusion in the computer science literature. The main focus of this literature is on the propagation of influence of active users (i.e., users who have received the information) on inactive users (i.e., users who have not received the information) (Domingos and Richardson 2001). In particular, two of the most extensively studied models are the linear threshold model (e.g. Kempe et al. 2003) and the independent cascade model (e.g. Nguyen and Zheng 2013). In the linear threshold model, an inactive user gets influenced based on the number of connections to the active nodes. On the other hand, in the independent cascade model, a user in the network is influenced with a certain probability by an active and connected node independently. That is, each node independently activates an inactive node (if they are connected) by a predetermined probability. In contrast, in our study, the active nodes have a cumulative influence effect on the inactive nodes, i.e., multiple active nodes activate the inactive nodes simultaneously (rather than independent influence). More importantly, the problem that we analyze in the current study is specialized in that it focuses on the specific problem of influencer marketing.

The problem of identification of network seeders shares some similarity with maximal covering location problem that has been extensively studied in the operations literature. In Table 3.1, we compare the models that we develop in our study to models that were analyzed in the prior literature. The main model of the current study (formulated as Model $M2L$) is rather unique to the context of influencer marketing on social media as we incorporate the effect of peer influence in the model. We show that firms can experience a significant profit loss when implementing a policy that is computed based on a model that ignores the peer effect. The profit loss is particularly significant when the online network has a high number of connections in the second level, and the influence of the peer effect is high. Moreover, the size of the problem that we study in this paper is quite large compared to the ones that were previously studied in the literature.

Table 3.1: Related Literature in Other Domains

- Set W denotes predetermined influencer set. Set V represents followers and followers of followers. t represents the number of influencers to be selected from set W.
- Model $M2L$ denotes problem of identification of influencers. Model $M1L$ denotes a relaxed version of Model $M2L$. More details regarding the models are provided in the sections that follow.

Our Models	**Current Study**	Chan et al. (2016)	Camm et al. (2002)
Main Model ($M2L$)	**Yes**	**No**	**No**
Upper Bound Model ($M2L^{UB}$)	**Yes**	**No**	**No**
Lower Bound Model ($M1L$)	Yes	Yes	Yes
Structural Properties	Yes	Yes	No
Problem size	$\lvert W\rvert = 25$ - 100; $\lvert V\rvert \geq 25,000$; $t = 5$ to 50	$\lvert W\rvert = 10000$; $\lvert V\rvert = 5000$; $t = 5$ to 200	$\lvert W\rvert = 50$ - 400; $\lvert V\rvert = 300$ to 450; $t = 1$ to 5

The other literature stream that our paper contributes to is ad scheduling. Prior literature has analyzed how ads are scheduled in different contexts such as television, websites, and mobiles has been extensively studied in the operations and information systems literature (e.g., Bollapragada et al. 2004; Kumar et al. 2006; Sun et al. 2017). However, our work is among the first, to the best of our knowledge, to address the problem of scheduling of ads by multiple influencers over a planning horizon.

The rest of the paper is organized as follows. In Section 3.2, we formally present the framework. In Section 3.3, we present the main model and its relaxations. In Section 3.4, we present the results of our case studies. We present the scheduling framework for influencer marketing in Section 3.5. Finally, we conclude with future research directions in Section 3.6.

3.2 Problem Background

We start by focusing on developing a model to determine an optimal influencer set in the context of influencer marketing on social media. To model this problem, we first understand the underlying structure of online social networks. While there are several online social networks such as Facebook, Instagram, and LinkedIn, we develop our framework in the context of Twitter, a popular social media platform.[2] Over the last decade, Twitter emerged to be an effective social broadcasting tool with more than 335 million active users (Statista 2018a). Twitter users can post short messages (also referred to as *tweets*) with the length of a message not exceeding 280 characters.

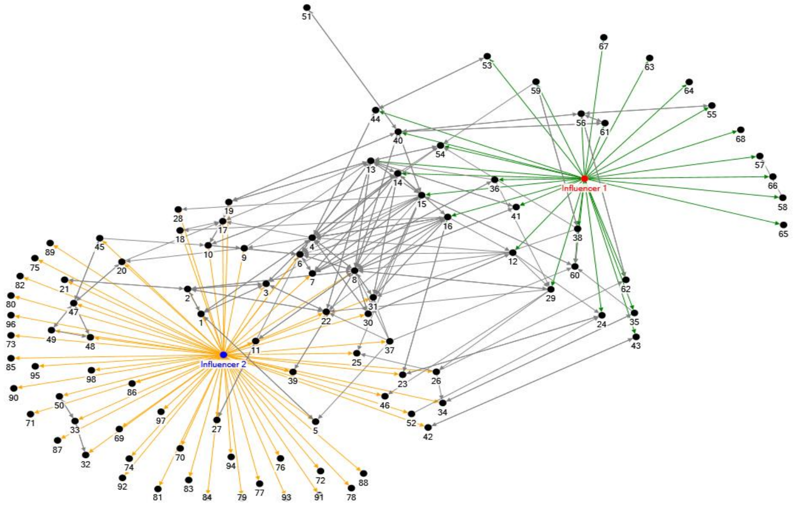
Twitter users can follow other users on the network (unless the users are private). Further, if user X follows user Y, then X immediately receives all the tweets from user Y. Another unique feature of Twitter is the ability to share or "retweet" the tweets of other users. Retweeting is a simple function of Twitter and users only need to click a single button to retweet someone's message. By retweeting user Y's tweet, user X is sharing the original message (of user Y) to his/her followers. Specifically, retweeting can be seen as a form of message propagation.

To learn about the underlying structure of the online social networks, we collect data on influencers and their followers from Twitter. Using this data, we construct a network graph to understand the relationship among Twitter users. Figure 3.2 illustrates one such network graph of two influencers (denoted by Influencer 1 and Influencer 2) and their followers who are connected by a directional edge (influencer → follower). From this graph, we can observe that these two influencers have several common followers (e.g., nodes 23 and 30) and the followers of these influencers follow each other as well (e.g., nodes 13 and 19). Thus, we need to account for the following crucial characteristics of online social networks. First, influencers

[2] Although the context of this study is Twitter, our framework can be applied to other popular social media platforms including Facebook, Instagram, and YouTube.

may have a common subset of followers. Hence, when two influencers tweet the same ad, the common subset of followers would receive the ad multiple times. As a result, we need to account for the multiple exposure effect of an ad. Second, directional edges exist not only between an influential user and others but also between the followers of the influential users. In particular, a follower might get an ad from the influencer that s/he follows, in addition to another follower who retweets the ad of an influencer. From these observations, we proceed to develop the model for determining an optimal influencer set. We now formally state the problem of determining an optimal influencer set along with the parameters and variables used. We refer to this problem as the main problem.

Figure 3.2: A Example of Online Social Network



Created with NodeXL Basic (http://nodexl.codeplex.com) from the Social Media Research Foundation (http://www.smrfoundation.org)

3.2.1 Problem Definition

Technical Definition:

<u>Instance:</u> A social network associated with users of a domain of interest is represented by a directed bipartite graph $G(N, E)$, where N is the set of vertices representing the users in the domain of interest, $N = W \cup V$. Set W consists of potential network seeders or influential

users whereas set V contains the direct followers of influential users in W and the followers of direct followers of influencers in set W. Set E denotes the set of directed arcs within V, and W to V that represent the follower relationship. A user $j \in V$ retweets a message of user $i \in W$ with a probability of $p_{ij} \geq 0$. In addition, a user $j \in V$ retweets a message of user $k \in V$ with a probability of $p_{kj} \geq 0$. The retweeted message reaches the followers of followers who retweeted. The number of followers of followers who retweet is denoted by F_j. We are interested in determining a set of t influential users from set W such that the firm maximizes the expected benefit from the marketing campaign. The variables and parameters used in this model are summarized in Table 3.2.

SOLUTION: Find t influencers from set W. The value of t is given.

OBJECTIVE: Maximize the expected benefit to the firm for placing ads on social media via influencers.

3.2.2 Model Parameters

Expected Benefit to Firm: Prior literature in marketing suggests that advertising provides visibility to the firm, and thus positively affects the sales of a firm. However, our conversations with marketers along with recent literature indicate that in addition to the reach of the ad (i.e., number of followers who receive the message), follower's engagement with the message is often used to measure the effectiveness of a social media ad campaign (Lambrecht et al. 2018). Consistent with recent literature, we operationalize engagement via number of retweets (e.g., Long 2015a; Lambrecht et al. 2018). By retweeting the message of influencers, the user is spreading or transmitting the message to their followers and thus increasing the reach of the message beyond the followers of the influencers (Kwak et al. 2010) and validating the content of the tweet (Suh et al. 2010). Consequently, in line with industry practices and prior academic literature, we assume that a firm attains benefit (denoted by b_j) only when user $j \in V$ retweets a message from an influencer. Although in the current study, we operationalize engagement via number of retweets, other engagement metrics including number of comments, number of likes, and number of clicks on the links in the tweet can be

Table 3.2: List of Parameters and Variables

Symbol		Description
Parameters:		
a_{ij}	$=$	$\begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } i \in W, \\ 0, & \text{otherwise.} \end{cases}$
a_{kj}	$=$	$\begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } k \in V, \\ 0, & \text{otherwise.} \end{cases}$
p_{ij}	$=$	probability of user $j \in V$ retweeting the message of influencer $i \in W$; $p_{ij} = 0$ if $a_{ij} = 0$.
p_{kj}	$=$	probability of user $j \in V$ retweeting the message of user $k \in V$; $p_{ij} = 0$ if $a_{jk} = 0$.
b_j	$=$	benefit to the firm when the tweet reaches user $j \in V$ through $i \in W$.
d_j	$=$	benefit to the firm when the tweet reaches one follower of user $j \in V$.
F_j	$=$	number of followers of user $j \in V$.
Decision Variables:		
x_i	$=$	$\begin{cases} 1, & \text{if } i \in W \text{ is chosen as a seeder,} \\ 0, & \text{otherwise.} \end{cases}$
g_j	$=$	the cumulative probability of user $j \in V$ retweeting a message.
y_j	$=$	the cumulative probability of user $j \in V$ retweeting a message from users in W.
δ_j	$=$	the cumulative probability of user $j \in V$ not retweeting a message from users in W.
α_j	$=$	the cumulative probability of user $j \in V$ not retweeting a message from users in V.

used to operationalize engagement.

Furthermore, when followers in set V decide to share the message with their followers, there is an additional benefit to the firm as the ad reaches the second level followers. We assume that a firm attains a benefit of d_j per follower of follower j. Based on the data that is available to firms from Twitter, we cannot identify the number of times a follower of follower retweets. Hence, we do not consider engagement beyond the second level. To summarize, the total expected benefit that a firm gets is the sum of benefits derived when a user gets the message from multiple users (i.e., from users in set W and set V).

Network Connections and Followers: The parameters a_{ij} and a_{kj} denote the links between an influencer and a follower or between two followers. This data can be obtained from the social networking data. If $\sum_j a_{ij}$ is high for influencer i, it denotes that the influencer has large number of followers. Likewise, If $\sum_j a_{kj}$ is high for a follower k, it implies that user k has high number of followers who also follow other influencers. Furthermore, F_j denotes the total number of followers of followers, which can be obtained from the profile of followers on social media.

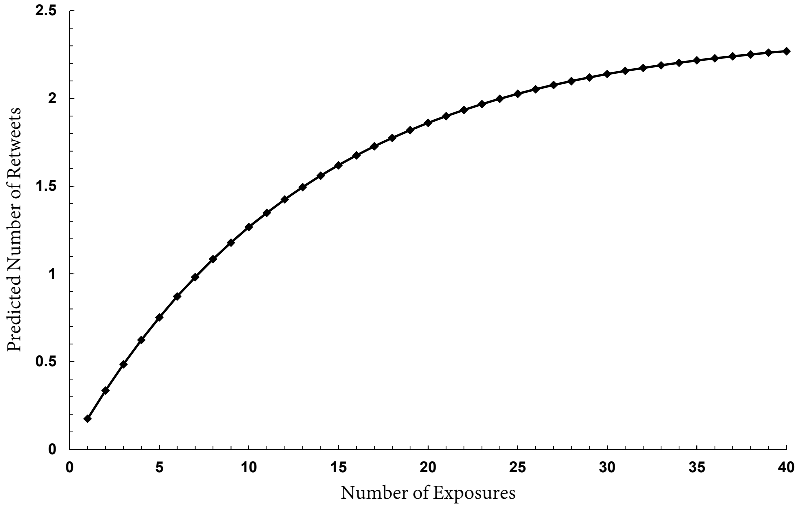
Probability of Retweet: When user $j \in V$ receives a tweet from influencer $i \in W$, user j may decide to retweet the tweet or not retweet. In other words, there is a probability (which we denote by p_{ij}) with which user j retweets a tweet of influencer i. This probability is similar to *adoption probability* in the diffusion literature and *click probability* in the context of web and mobile ads. Prediction of these probabilities has received some attention in the recent literature. For example, Mookerjee et al. (2016) develop a Logit model using web user's characteristics to predict the probability of a user clicking on an ad. Fang et al. (2013) develop a Bayesian learning procedure to predict adoption probability of a follower while Goyal et al. (2010) predict the probability of influence in social networks using existing network data. Therefore, in this paper, we consider that the firms can calculate p_{ij} and it is known to the firm.

Multiple Exposure Effect: The effect of multiple exposure on the total probability of retweet can be modeled as follows. Let p_1 be the probability that follower j retweets a message from influencer X_1 and p_2 be the probability that follower j retweets a message from influencer X_2. The probability that follower j retweets the message of *either* influencer X_1 *or* influencer X_2 is $p_1 + p_2 - p_1 p_2 = [1 - (1 - p_1)(1 - p_2)]$. More generally, the effect of multiple exposure on follower j from multiple influencers in set W is modeled as $[1 - \prod\limits_{i \in W} (1 - p_{ij})]$.

Our method of modeling total probability to incorporate the effect of multiple exposures implies that a follower is likely to retweet the ad only once and the marginal probability of a user retweeting (probability per exposure) an influencer's message decreases with the number of exposures. This way of modeling effect of multiple exposures is consistent with existing literature on diffusion and mobile advertisement (e.g., Kumar et al. 2007; Goyal et al. 2010; Sun et al. 2017). Furthermore, marketing literature has long studied the impact of ad repetition on consumer attitudes. In particular, literature in this stream argues that the probability of consumer interacting with an ad depends on number of times that the user has been exposed to the same ad (i.e., number of times that the user has displayed on his/her social media page) and the marginal benefit of an ad is highest at the first exposure and decreases with additional exposures.

To further validate this assumption, we conduct an empirical analysis using data collected from Twitter. To do this, we collect data related to 18,571 tweets posted by 37 influencers. The details of this empirical study are provided in Appendix B.3. The main finding of this empirical study is that the total number of retweets of an ad increases with the number of exposures. Figure 3.3 depicts the empirically predicted relationship between total number of retweets and number of exposures. As can be seen in the figure, the first exposure is predicted to get the maximum number of retweets and the number of retweets tend to decrease with additional number of tweets.

Figure 3.3: Empirically Predicted Relationship Between Total Number of Retweets and Exposure



3.3 Selection of Influencers: The Main Model

In this section, we first present the problem of selecting an optimal influencer set, and then the relaxations of this problem. Hereon, we refer to this problem as the main problem or Problem $M2L$ as it considers two-level influence. The mathematical formulation for Model $M2L$ is as follows.

Mathematical Formulation - Model $M2L$:

$$\text{Max} \;\; \Psi_2 = \sum_{j \in V} b_j y_j + \sum_{j \in V} d_j g_j F_j \tag{3.1}$$

subject to:

$$\sum_i x_i \leq t, \tag{3.2}$$

$$\delta_j = \prod_{i \in W} (1 - p_{ij} x_i), \quad \forall j \in V \tag{3.3}$$

$$y_j = 1 - \delta_j, \qquad \forall j \in V \tag{3.4}$$

$$\alpha_j = \prod_{k \in V} (1 - p_{kj} y_k), \quad \forall j \in V \tag{3.5}$$

$$g_j = 1 - \alpha_j \delta_j, \qquad \forall j \in V \tag{3.6}$$

$$x_i \in \{0, 1\} \qquad \forall i \in W. \tag{3.7}$$

In the above formulation, the objective function, Ψ_2 (i.e., Equation 3.1), maximizes the total benefit to the firm when tweets reach the users on Twitter. Constraint (3.2) ensures that at most t number of influencers are selected. Constraint (3.3) denotes the expected probability of user $j \in V$ not retweeting message from influencer $i \in W$, while Constraint (3.4) represents the probability that user j retweets the message from influencer $i \in W$. Constraint (3.5) represents the expected probability that follower $j \in V$ does not retweet the message from $k \in V$. Constraint (3.6) is the expected probability of user $j \in V$ retweeting a message.

Finally, Constraint (3.7) imposes the binary constraint on the decision variable x_i. Note that Constraints (3.3), (3.5), and (3.6) make the problem non-linear and difficult to solve. In Theorem 1, we formally prove the complexity of the problem. Theorem 1 asserts that the problem to identify an optimal influencer set is strongly NP-hard.

Theorem 1. *The decision problem corresponding to the multiplicative two-level influence problem (i.e., model $M2L$) is strongly NP-complete.*

The theorem is formally proved in Appendix B.1.1. For our reduction, we choose 3-Satisfiability (3SAT) problem, which is a well-known strongly NP-complete problem (Garey and Johnson 1979).

The first proposition deals with a scenario where an optimal solution for Model $M2L$ can be obtained in polynomial time. In particular, when the social network is disjointed, i.e., a user $j \in V_i$ follows only one influencer i in set W or other followers of influencer i, we find that the solution can be found by sorting the influencers based on the value of $\Lambda_i = \sum_{j \in V_i} b_j p_{ij} + \sum_{j \in V_i} d_j F_j [1-(1-p_{ij}) \prod_{k \in V_i} (1-p_{kj} p_{ik})]$, V_i represents the set of followers of user i.

Proposition 11. *If the social network is disjointed (i.e., any user $j \in V$ follows only one influencer i in Set W and some followers of influencer i), then there exists an optimal solution to Problem $M2L$ that includes the first t users in Set W after sorting them in the descending order of Λ_i, in which $\Lambda_i = \sum_{j \in V_i} b_j p_{ij} + \sum_{j \in V_i} d_j F_j [1-(1-p_{ij}) \prod_{k \in V_i} (1-p_{kj} p_{ik})]$, V_i represents the set of followers of user i.*

We present the proof in Section B.1.2 of the Appendix.

Proposition 11 has an important practical relevance. It provides a condition when the solution to Model $M2L$ is analytically tractable. Nevertheless, when the social network is not disjointed, our plan to solve Model $M2L$ is as follows. We first reformulate this model as Model $M2L^{UB}$. Next, we provide a linear approximation to Model $M2L^{UB}$, which is a mixed integer non-linear program. By converting the non-linear integer program to a linear integer program, we now are able to solve Model $M2L^{UB}$ for small and medium-size problems. Furthermore for large-size problems, we provide a rounding heuristic to obtain solution for Model $M2L^{UB}$. Finally, we show that Model $M2L^{UB}$ provides a tight upper bound to Model $M2L$, hence can provide us with near-optimal solutions for the main problem. We present Model $M2L^{UB}$ in the next subsection.

3.3.1 $M2L^{UB}$ - Relaxation of Model $M2L$

As discussed earlier, the decision problem corresponding to the multiplicative two-level influence problem (i.e., model $M2L$) is strongly NP-complete. Also, constraints (3.3) and (3.5) make the problem non-linear (and difficult to solve). We start by defining a new model $M2L^{UB}$, by replacing constraint (3.5) in Model $M2L$ with $\bar{\alpha}_j = \prod_{k\in V}(1-p_{kj})^{y_k}$. The mathematical formulation of Model $M2L^{UB}$ is presented below:

Mathematical Formulation - Model $M2L^{UB}$:

$$\text{Max } \ \Psi_2^{UB} = \sum_{j\in V} b_j y_j + \sum_{j\in V} d_j \bar{g}_j F_j \tag{3.8}$$

subject to:

Constraints (3.2) - (3.4)

$$\bar{\alpha}_j = \prod_{k\in V}(1-p_{kj})^{y_k}, \quad \forall j \in V \tag{3.9}$$

$$\bar{g}_j = 1 - \bar{\alpha}_j \delta_j, \qquad \forall j \in V \tag{3.10}$$

$$x_i \in \{0,1\} \qquad \forall i \in W. \tag{3.11}$$

Replacing constraint (3.5) with (3.9) will allow us to linearly approximate the non-linear integer program, which we discuss shortly (in subsection (3.3.1.1)). Before doing so, we first characterize the relationship between Models $M2L$ and $M2L^{UB}$ in Theorem 2.

Theorem 2. *The objective value of Model $M2L^{UB}$ (i.e., Ψ_2^{UB}) is an upper bound to the objective value of Model $M2L$ (i.e., Ψ_2).*

We present the proof in Section B.1.3 of the Appendix. For this proof, we rely on Taylor's expansion of the following two logarithmic functions: (i) $\ln(1-py)$, and (ii) $ln(1-p) \times y$.

Theorem 2 implies that an optimal solution to Model $M2L^{UB}$ (i.e., Ψ_2^{UB}) provides an upper bound to Ψ_2. Next, in the proposition below, we show that the solution for Model $M2L^{UB}$ is an optimal solution for the main model in certain instances.

Proposition 12. *The objective value of Model $M2L^{UB}$ (i.e., Ψ_2^{UB}) tends to the objective value of Model $M2L$ (i.e., Ψ_2) when the peer influence (i.e., p_{kj}, probability of user $j \in V$ retweeting the message of user $k \in V$) tends to zero.*

This proposition is proved using the binomial expansion of $(1-x)^n$. We formally prove this proposition in Section B.1.4 in the Appendix.

Proposition 12 implies that if the peer influence or the probability of users in set V retweeting the messages from other users in set V tends to zero (i.e., $p_{kj} \to 0$), $\Psi_2^{UB} \to \Psi_2$. Alternatively, when the number of edges between followers in set V tends to zero (i.e., followers do not follow each other), by solving Model $M2L^{UB}$, we can obtain an optimal solution for Model $M2L$. In Theorem 3, we formally prove the complexity of the problem.

Theorem 3. *The decision problem corresponding to the relaxation of the main problem (i.e., Model $M2L^{UB}$) is strongly NP-complete.*

The theorem is formally proved in Appendix B.1.1. For our reduction, we choose 3-Satisfiability (3SAT) problem, which is a well-known strongly NP-complete problem (Garey and Johnson 1979).

3.3.1.1 Linear Approximation of Model $M2L^{UB}$

Clearly, we see that constraints (3.3), (3.9), and (3.10) make Model $M2L^{UB}$ a non-linear mixed integer program. We reformulate the problem to transform the model to be separable as proposed by Camm et al. (2002). Let $W_j^{<1} = \{i | 0 < p_{ij} < 1\}$ and $W_j^{=1} = \{i | p_{ij} = 1\}$. We can now replace Equations (3.3) and (3.4) with the following new constraints:

$$h_j = \prod_{i \in W_j^{<1}} (1 - p_{ij})^{x_i}, \quad \forall j \in V \tag{3.12}$$

$$y_j \leq 1 - h_j + \sum_{i \in W_j^{=1}} x_i, \quad \forall j \in V \tag{3.13}$$

Proposition 13. *Following Camm et al. (2002), the mathematical formulation for model $M2L^{UB}$ is equivalent to Equation (3.8) subject to constraints (3.2), (3.12) to (3.13), and (3.9) to (3.11).*

Further, since h_j and $\bar{\alpha}_j$ are strictly positive, we can now take log of each side of the non-linear constraints. This results in the following optimization model.

$$\text{Max} \quad \Psi_2^{UB} = \sum_j b_j y_j + \sum_j d_j \bar{g}_j F_j \tag{3.14}$$

subject to:

$$\sum_i x_i \leq t, \tag{3.15}$$

$$\ln(h_j) = \sum_{i \in W_j^{<1}} \ln(1 - p_{ij}) x_i, \quad \forall j \in V \tag{3.16}$$

$$y_j \leq 1 - h_j + \sum_{i \in W_j^{=1}} x_i, \quad \forall j \in V \tag{3.17}$$

$$\ln(\bar{\alpha}_j) = \sum_{k \in V} \ln(1 - p_{kj}) y_k, \quad \forall j \in V \tag{3.18}$$

$$\bar{g}_j = 1 - \alpha_j \delta_j, \quad \forall j \in V \tag{3.19}$$

$$0 \leq \bar{g}_j \leq 1, \ \forall j \in V; \quad h_j \geq 0, \ \forall j \in V; \quad x_i \in \{0, 1\}, \ \forall i \in W. \tag{3.20}$$

We now present a method to approximate the non-linear term $\ln(h_j)$ by taking a convex combination of points, λ, on the curve on similar lines as Bradley et al. (1977) and Camm et al. (2002). Let R_j be the set of break points used to approximate the interval from 0 to

1. Let B_{jt} be the t^{th} breakpoint and λ_{jt} the weighting of the t^{th} breakpoint. We then make the below mentioned linear approximation substitutions:

$$\begin{aligned}
h_j &= \sum_{t\in R_j} B_{jt}\lambda_{jt}, && (3.21)\\
\ln(h_j) &= \sum_{t\in R_j} \ln(B_{jt})\lambda_{jt}, && (3.22)\\
\sum_{t\in R_j} \lambda_{jt} &= 1, && (3.23)\\
\bar{\alpha}_j &= \sum_{t\in R_j} \beta_{jt}\gamma_{jt}, && (3.24)\\
\ln(\bar{\alpha}_j) &= \sum_{t\in R_j} \ln(\beta_{jt})\gamma_{jt}, && (3.25)\\
\sum_{t\in R_j} \gamma_{jt} &= 1. && (3.26)
\end{aligned}$$

Next, let $\rho = \bar{\alpha}_j\delta_j$. Taking log on both sides, we get $\ln(\rho) = \ln(\bar{\alpha}_j) + \ln(\delta_j)$. we can now make the following linear approximations of $\ln(\rho)$.

$$\begin{aligned}
\rho_j &= \sum_{t\in R_j} \upsilon_{jt}\omega_{jt}, && (3.27)\\
\ln(\rho_j) &= \sum_{t\in R_j} \ln(\upsilon_{jt})\omega_{jt}, && (3.28)\\
\sum_{t\in R_j} \omega_{jt} &= 1. && (3.29)
\end{aligned}$$

Substituting the above values of h_j, $\ln(h_j)$, $\bar{\alpha}_j$, $\ln(\bar{\alpha}_j)$, ρ_j, and $\ln(\rho_j)$ into Equations 3.16 to 3.19, we get the linearized approximation of problem $M2L^{UB}$ that is presented in model $M2L^{UB} - AP$.

Mathematical Formulation - $M2L^{UB} - AP$:

$$\text{Max:}\quad \Psi_2^{AP} = \sum_j b_j y_j + \sum_j d_j \bar{g}_j F_j \qquad (3.30)$$

subject to:

$$\sum_i x_i \leq t, \tag{3.31}$$

$$\sum_{t\in R_j} \ln(B_{jt})\lambda_{jt} = \sum_{i\in W_j^{<1}} \ln(1-p_{ij})x_i, \qquad \forall j \in V \tag{3.32}$$

$$y_j \leq 1 - \sum_{t\in R_j} B_{jt}\lambda_{jt} + \sum_{i\in W_j^{=1}} x_i, \qquad \forall j \in V \tag{3.33}$$

$$\sum_{t\in R_j} \lambda_{jt} = 1, \qquad \forall j \in V \tag{3.34}$$

$$\sum_{t\in R_j} \ln(\beta_{jt})\gamma_{jt} = \sum_{k\in V} \ln(1-p_{kj})y_k \qquad \forall j \in V \tag{3.35}$$

$$\sum_{t\in R_j} \ln(\upsilon_{jt})\omega_{jt} = \sum_{t\in R_j} B_{jt}\lambda_{jt} + \sum_{t\in R_j} \beta_{jt}\gamma_{jt} \qquad \forall j \in V \tag{3.36}$$

$$\bar{g}_j \leq 1 - \sum_{t\in R_j} \upsilon_{jt}\omega_{jt} + \sum_{i\in W_j^{=1}} x_i, \qquad \forall j \in V \tag{3.37}$$

$$\sum_{t\in R_j} \gamma_{jt} = 1, \qquad \forall j \in V \tag{3.38}$$

$$\sum_{t\in R_j} \omega_{jt} = 1, \qquad \forall j \in V \tag{3.39}$$

$$0 \leq y_j \leq 1,\ \forall j \in V;\quad 0 \leq \bar{g}_j \leq 1,\ \forall j \in V;\quad \lambda_{jt} \geq 0,\ \forall j \in V\ \ \forall t \in R_j \tag{3.40}$$

$$x_i \in \{0,1\}, \qquad \forall i \in W. \tag{3.41}$$

We note that as the number of breakpoints (t) increases, the accuracy of the approximation increases. Clearly, the number of breakpoints can also vary for different approximations (i.e., Constraints (3.33), (3.36), and (3.37)). Although we provide a formulation to approximate Model $M2L^{UB}$, the problem is still a mixed integer program. While the current computational capacities allow us to solve small and medium-size problems directly using the available solvers such as *CPLEX* and *Gurobi*, they fail to provide an optimal solution for large size problems. We now present an iterative heuristic based on the fractional rounding principle to solve large-size problems of Model $M2L^{UB} - AP$.

3.3.1.2 Iterative Rounding Heuristic

In this heuristic, we first relax the integer constraints for variables x_i in Model $M2L^{UB} - AP$. Relaxing the integer constraints transforms the model to a linear program, thus allowing us to solve the problem in polynomial time. However, this could give us fraction solutions for all the decision variables. We round-off the largest fraction of x_i, obtained from the first run of linear programming model, to one. We repeat this procedure until t number of x_i's are equal to one. This algorithm is formally presented below.

Algorithm 1 Iterative Rounding Heuristic:

```
$\hat{W} \leftarrow W$, $U \leftarrow \emptyset$.
while $|U| < t$ do
    Solve Model $M2L^{UB} - AP$ as a linear program and obtain solution $X = \{x_i | i \in W\}$.
    Denote $\hat{i} = \arg\max_{i \in \hat{W}} \{x_i\}$, then update $U \leftarrow U \cup \{\hat{i}\}$, $\hat{W} \leftarrow \hat{W} \backslash \{\hat{i}\}$.
    Add an additional constraint $x_{\hat{i}} = 1$ to Model $M2L^{UB} - AP$.
end while
```

To demonstrate the improvement of solving time of the approximate model, we conduct a numerical study. In this study, we run ten test instances to compare the CPU run time to solve Model $M2L$ using a non-linear solver (i.e., BONMIN) and the Model $M2L^{UB} - AP$ using a linear solver (i.e., CPLEX). The results are presented in Table 3.3. As expected, we can see the huge difference in run times between these two models.

3.3.2 Performance of Upper Bound Model

In this section, we assess the quality of the upper bound that we develop in the previous subsection (i.e., Model $M2L^{UB}$) with respect to the optimal solution for the main problem (i.e., Model $M2L$). To do this, we perform an extensive numerical analysis. As the measure of quality of the upper bound, we use the percentage optimality gap, which is calculated

Table 3.3: Computational Evaluation of Upper Bound Model ($M2L^{UB}$)

Problem Size			Run Time in CPU secs	
$\lvert W\rvert$	$\lvert V\rvert$	t	$M2L$	$M2L^{UB}-AP$
25	250	10	15.64	0.92
25	500	10	321.56	6.83
25	750	10	483.58	29.92
25	1000	10	1169.22	91.84
25	1250	10	2583.66	8.91
50	250	20	1.05	1.05
50	500	20	211.34	36.19
50	750	20	745.81	2.50
50	1000	20	1782.54	5.08
50	1250	20	3225.08	9.28

by % *Optimality Gap* $= \frac{\Psi_2^{UB}-\Psi_2}{\Psi_2}$, where Ψ_2^{UB} is the objective function value of the upper bound problem and Ψ_2 is the objective function value of the main problem. A commercial non-linear solver was used to solve both the models.[3] Basic Open-source Nonlinear Mixed Integer programming (BONMIN) is an open-source solver for general mixed integer non-linear programming problems. BONMIN solver is run on Intel Core i7-7700 CPU with 3.60 GHz (2 cores) with 64GB ram.

To capture different real-life instances, we generate an experimental test bed by varying the key parameters of the model that could potentially affect the quality of solutions. The test bed is generated as follows. The number of influencers is 25 (i.e., $|W| = 25$) and Size of set V is 250 (this represents a relatively small problem size. This assumption is required to ensure that the non-linear solver is able to provide us with an *optimal* solution for Model $M2L$). Benefit, b_j is drawn from continuous uniform distribution, $Uniform(0, 10)$ and d_j is drawn from a continuous uniform distribution, $Uniform(0.1, 1)$. The number of followers of followers in set V, F_j is drawn from continuous uniform distribution, $Uniform(100, 1000)$.

[3] We note that the solution provided by current non-linear solvers may not be globally optimal. However, since the same solver is being used for solving both the problems the results pertaining to % optimality gap are appropriate.

The first level influence was set to $p_{ij} \in$ {Low, High}.[4] The second level influence was set to $p_{kj} \in$ {Low, High}. Next, the number of edges between influencers in set W and followers in set V was set to, $a_{ij} \in$ {Low, Medium, High}. Finally, the number of edges between the followers of influencers is set to $a_{kj} \in$ {Low, Medium, High}. To summarize, we mainly vary the following parameters in each problem instance: a_{ij}, a_{kj}, p_{ij}, and p_{kj}. Thus, we generated $3 \times 2 \times 2 \times 3 = 36$ combinations. For each of the 36 instances, we conducted runs for cardinality $t \in \{1, 2, ..10\}$. Further, to estimate the average performance of solutions provided by the models, for each of the 360 instances, we simulate the model for 30 new instances by randomly drawing the values for the following parameters: b_j, d_j, and F_j. Thus combining all parameter settings, we generated a set of 10,800 instances to ensure that our test bed covers different scenarios.

The average performance of Model $M2L^{UB}$ against $M2L$ in terms of % *Optimality Gap* for each of the 36 core instances are summarized in Table B.8 (in the online appendix). The average % *Optimality Gap* presented in the table represents the mean of 300 instances and the maximum % *Optimality Gap* denotes the maximum gap over 300 instances. As can be seen from the table, the gap is less than 5% (both average and maximum) across all the instances. Figure 3.4 illustrates the performance of the upper bound model against the main model over different values of a_{kj} and p_{kj}. The gap between Model $M2L$ and Model $M2L^{UB}$ decreases with the magnitude of second level influence and the number of edges within set V. This leads us to our first managerial insight.

Insight 1. *In a social network with relatively low second level influence (i.e., p_{kj}) and low number of edges between the followers of the influencers (i.e., a_{kj}), the benefit estimated by Model $M2L^{UB}$, i.e., Ψ_2^{UB} is approximately equal to the benefit estimated by Model $M2L$, i.e., Ψ_2.*

The intuition behind Insight 1 is as follows. The main difference between Models $M2L$ and $M2L^{UB}$ is the constraint which calculates the gain in total probability as a result of

[4] The distributions of Low and High are presented in Appendix B.4.

second level influence. Hence, when the second level influence is low, the solutions of these two models should be close to each other. This observation is consistent with the finding in Proposition 1.

Figure 3.4: % Optimality Gap – Models $M2L$ vs $M2L^{UB}$

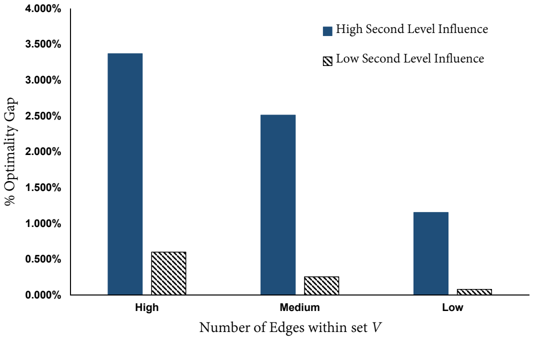
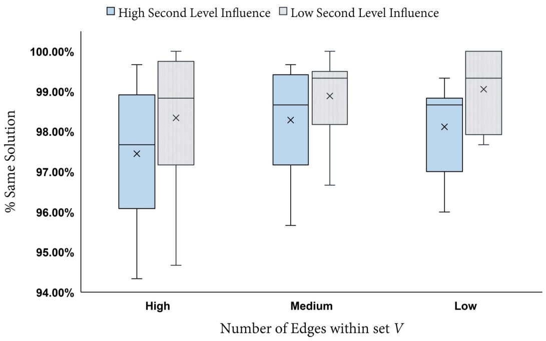


Figure 3.5: % Same Solution – Models $M2L$ vs $M2L^{UB}$



Next, the last column in Table B.8 denotes the % of times that the solution set obtained from the upper bound model (i.e., $M2L^{UB}$) was the optimal solution set for the main model (i.e., $M2L$) over 300 simulations. As can be seen, the solution set (i.e., optimal influencer set) obtained from the upper bound problem is an optimal solution for the main model in excess of 94% of instances in our test bed. Figure 3.5 summarizes the average performance, in terms of an optimal solution set, of Model $M2L^{UB}$ against Model $M2L$, over different values of a_{kj} and p_{kj}. The interquartile box denotes 25%, 50%, and 75% percentiles, the whiskers indicate the minimum and maximum % same solution over 10,800 instances. The mean of means is represented by the cross mark in the figure. From the figure, we can infer that with lower values of a_{kj} and p_{kj}, the optimal solution of Model $M2L^{UB}$ is an optimal solution of Model $M2L$ in increasing number of instances.

To summarize, the above computational results suggest that average gap between the objective values of Model $M2L^{UB}$ and Model $M2L$ is less than 4.52%. Furthermore, the solutions for Models $M2L^{UB}$ and $M2L$ is same for over 94.33% of the instances. Given that our test bed covers a wide variety of scenarios, the results suggest that Model $M2L^{UB}$ provides a tight upper bound to Model $M2L$ and near-optimal solution for the main model.

3.3.3 Single Level Influence Model

In this subsection, we present the second relaxation of the main model (i.e., Model $M2L$) by considering only single level influence. In particular, we ignore the influence that users in set V can exert on other users in set V, we now assume that $p_{kj} = 0$ (that is there is no peer effect). In such a scenario, $y_j = g_j$. We prove that this relaxed model provides the lower bound for the main problem. The main purpose of this model is to demonstrate the drawbacks of ignoring the second level influence (i.e., peer influence). We proceed by presenting the formulation for this relaxed problem referred to as Model $M1L$.

Mathematical Formulation - M1L:

$$\text{Max} \quad \Psi_1 = \sum_j b_j g_j + \sum_j d_j g_j F_j \tag{3.42}$$

subject to:

$$\begin{aligned} \sum_i x_i &\leq t, && (3.43) \\ g_j &= 1 - \prod_{i \in W} (1 - p_{ij} x_i), \quad \forall j \in V && (3.44) \\ x_i &\in \{0, 1\}. && (3.45) \end{aligned}$$

The following theorem discusses the computational complexity of Model $M1L$.

Theorem 4. *The decision problem corresponding to the above problem (Model $M1L$) is strongly NP-complete.*

The reduction is from the NP-Complete 3-Satisfiability problem (Garey and Johnson 1979). See Section B.1.1 in the Appendix for the formal proof.

Lemma 5. *The solution of Model $M1L$ is a lower bound to the solution for Model $M2L$.*

The proof for this theorem is presented in Appendix B.1.5. Lemma 5 formally characterizes the relationship between Model $M2L$ and Model $M1L$. Furthermore, note that Model $M1L$ is a non-linear mixed integer program. However, this problem can be approximated linearly through separable programming method that we used to linearize Model $M2L^{UB}$. The detailed description of the approximation is provided in Section B.2 of the Appendix. The ordering of objective functions Models $M2L, M2L^{UB}$, and $M1L$ is presented in Lemma 6.

Lemma 6. *The ordering of objective functions of Models $M2L, M2L^{UB}$, and $M1L$ is as follows: $\Psi_1 \leq \Psi_2 \leq \Psi_2^{UB}$.*

3.3.4 Comparing Main Model and Lower Bound Model

We conduct numerical analysis to observe the impact of ignoring the second level influencer or the peer effect (i.e., $p_{kj} = 0$). We utilize the same experimental test bed that as in Section 3.3.2.

Table B.9 (in the appendix) summarizes the performance of Model $M1L$ against Model $M2L$ on this test bed. This table is illustrated in Figure 3.6. As can be seen in the figure, the optimality gap is quite high in most of the instances. Furthermore, we observe that the gap between the lower bound model and the main model is relatively high (i.e., greater than 20%) when the probability of retweet in the second level (i.e., p_{kj}) is relatively high.

Insight 2. *Ignoring peer influence in the main model to select influencers could lead to significant loss in objective value.*

Furthermore, the last column of Table B.9 represents the percentage of instances where the solution set (i.e., optimal influencer set) obtained from the lower bound model is same as the solution set of the main problem. This result is summarized in Figure 3.7. As can be seen, the solution set (i.e., optimal influencer set) obtained from the lower bound problem is not an optimal solution for the main model in the majority of the instances. These results

Figure 3.6: % Optimality Gap – Models $M2L$ vs $M1L$

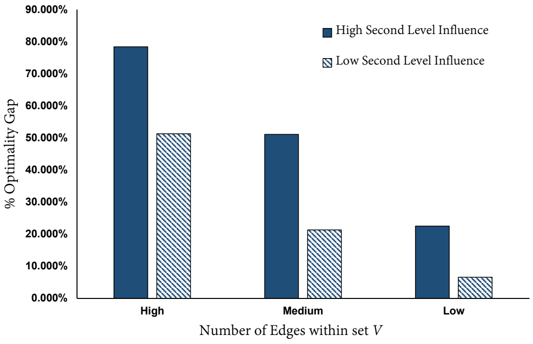
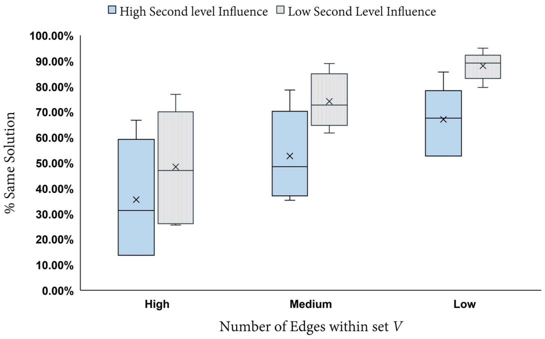


Figure 3.7: % Same Solution – Models $M2L$ vs $M1L$



indicate that ignoring the second level influence (which most of the prior literature does) could lead to a sub-optimal solution. This result is summarized in the next insight.

Insight 3. *High second-level influence as a result of high values of p_{kj} and a_{kj} increases the optimality gap between Models $M2L$ and $M1L$. In addition, the likelihood of Model $M1L$ providing an optimal solution to Model $M2L$ decreases as p_{kj} and a_{kj} increase.*

Insight 3 concludes that ignoring second level influence could lead to a higher likelihood of not obtaining an optimal solution to the main problem of selection of influencers. Further, this likelihood will increase with higher values of p_{kj} and a_{kj}. As can be seen in the box and whisker plot (in Figure 3.7), the performance of Model $M1L$ in terms of providing an optimal solution to the main problem reduces for high a_{kj} and p_{kj}. This insight is important not only to the firms using influencer marketing but also the social media platforms. In order to improve the effectiveness of influencer marketing campaigns, social media platforms can share necessary data with the firms so that the firms identify the right set of influencers for their ad campaign.

3.4 Case Study - Influencers on Twitter

This section describes the two case studies that we conduct to illustrate how our optimization framework for selection of influencers can be applied to a real-world application and to quantify the value of model-based solutions against the current industry benchmarks of selecting influencers. For the purpose of this case study, we obtain data from Twitter. For the first case study, we randomly identify 18 influencers who actively promote products on Twitter. The Federal Trade Commission act requires adding a hashtag $'\#ad'$ in the tweet that is specifically designed for advertisement or for endorsing a product. We search for users who tweet $\#ad$ to identify the influencers who actively tweet advertisements from their Twitter account. Due to data restrictions from Twitter, we restrict our focus on influencers with less than 25,000 followers. The data for this case was collected over a period of 10 months.

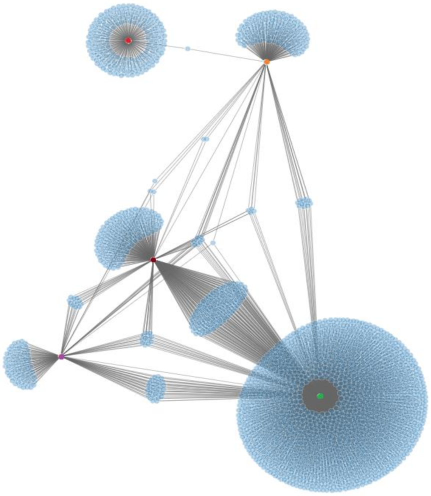
Our data consists of: (i) tweets posted by these 18 influencers, (ii) how many times the tweet was retweeted, and (iii) by which follower was it retweeted. First, we estimate the probability of retweet of a follower, i.e. p_{ij}. Although there are several ways to estimate the probability of a user retweeting the message of an influencer, we estimate the probability of retweet using historical data of an influencer on Twitter. In particular, we estimate the probability of a follower retweeting the message of an influencer using the following formula: $\frac{\text{Total number of times a follower retweeted an influencer's tweet}}{\text{Total number of tweets from an influencer}}$. From our dataset we obtain information on the number of followers, i.e., F_j, of each follower $j \in V$. For the purpose of numerical experiments, we assign the value of $b_j = 10\ \forall j \in V$ and $d_j = .001\ \forall j \in V$.[5] Furthermore, due to current data limitations from Twitter, we are unable to estimate p_{kj} (i.e., peer influence) for this case study. However, provided data availability, this can be easily estimated. The estimation procedure is as follows. The first step is to get the follower list of all retweeters. Next, we identify if the followers of retweeters are also following an influencer. If so, we keep such followers on the list and drop the other followers of retweeters. Once we have the network graph of followers of followers who also follow an influencer, we can estimate the probability of retweet using historical data. Hence to show the performance of our model for a wide variety of instances, we run the experiment with different values of second level influence. In particular, for high second level influence scenario, we have high values of a_{kj},

[5] The results remain the same qualitatively with different values of b_j and d_j. In practice, the marketing firm should be able to quantify the value of each retweet.

i.e., connection between two peers (or edges between followers). Likewise, for low second level influence scenario, we have low values of a_{kj}. The descriptive statistics of data collected from Twitter for Case Study-I are provided in Table 3.4.[6] A representative network graph of 5 influencers from this dataset is presented in Figure 3.8. For Case Study-II, we follow the similar procedure, but with a new set of influencers (i.e., set W). In particular, for Case Study-II, we identify 37 new influencers. The data for the second case study was collected for 5 months. The descriptive statistics for this influencer set are provided in Table 3.5.

Figure 3.8: Case Study Problem-I: Network Graph.

- In the network graph below, influencers are represented by dark colored dots.



3.4.1 Optimization Results

Given the problem size of both the case studies, it is computationally not feasible to obtain an optimal solution for Model $M2L$ using any of the existing non-linear solvers. In the earlier sections, we demonstrated the accuracy of the upper bound model in estimating a near-optimal solution for the main problem (i.e., Model $M2L$). Therefore, for this case study, we first obtain the solution for the upper bound model (i.e, an influencer set for

[6] The actual influencer names are anonymized.

Table 3.4: Case Study Problem - I: Descriptive Statistics

Influencer	No. of	No. of	% of Followers	Probability of Retweet (p_{ij})			Followers of users who retweeted (F_j)		
Identification	retweeters	Followers	who retweeted	Mean	Min	Max	Mean	Min	Max
INFLUENCER A	354	5676	6.24%	0.053%	0.017%	0.886%	7060.22	5	190595
INFLUENCER B	152	6621	2.30%	1.276%	0.089%	26.810%	6671.92	16	76495
INFLUENCER C	216	10444	2.07%	0.111%	0.066%	1.586%	6163.75	10	274763
INFLUENCER D	180	8359	2.15%	0.060%	0.043%	0.346%	6158.71	23	577415
INFLUENCER E	2028	15253	13.30%	0.515%	0.038%	14.869%	9178.47	0	523939
INFLUENCER F	4	737	0.54%	1.370%	1.370%	1.370%	29151.50	276	84816
INFLUENCER G	1	65	1.54%	7.692%	7.692%	7.692%	36.00	36	36
INFLUENCER H	582	7967	7.31%	0.150%	0.028%	2.344%	8993.99	0	208535
INFLUENCER I	81	9637	0.84%	0.287%	0.215%	1.720%	10496.46	1	484365
INFLUENCER J	1	25	4.00%	8.333%	8.333%	8.333%	876.00	876	876
INFLUENCER K	5	19	26.32%	0.595%	0.595%	0.595%	2552.00	97	7375
INFLUENCER L	2416	15551	15.54%	0.407%	0.028%	16.312%	9062.02	0	1149153
INFLUENCER M	789	9446	8.35%	0.291%	0.037%	3.862%	10704.77	4	155137
INFLUENCER N	175	9962	1.76%	0.559%	0.201%	6.036%	19636.61	45	152606
INFLUENCER O	46	2483	1.85%	1.236%	0.625%	8.750%	3785.65	7	94171
INFLUENCER P	273	10591	2.58%	0.714%	0.040%	18.508%	15439.40	2	228928
INFLUENCER Q	63	12917	0.49%	0.126%	0.074%	0.442%	8568.89	0	163404
INFLUENCER R	1802	10770	16.73%	0.632%	0.046%	30.159%	7892.67	0	158161

Table 3.5: Case Study Problem - II: Descriptive Statistics

Influencer	No. of	No. of	% of Followers	Probability of Retweet (p_{ij})			Followers of users who retweeted (F_j)		
Identification	retweeters	Followers	who retweeted	Mean	Min	Max	Mean	Min	Max
INFLUENCER 1	13	14168	0.09%	0.980%	0.800%	1.600%	25739.90	0	274910
INFLUENCER 2	74	14174	0.52%	1.173%	0.870%	20.870%	24263.94	0	274910
INFLUENCER 3	40	6051	0.66%	0.570%	0.275%	6.061%	21661.97	0	167424
INFLUENCER 4	9	7876	0.11%	0.086%	0.064%	0.254%	18478.68	1704	112451
INFLUENCER 5	115	20583	0.56%	2.047%	1.538%	52.308%	23728.92	0	229605
INFLUENCER 6	51	14997	0.34%	4.131%	0.787%	37.795%	20230.39	0	229605
INFLUENCER 7	46	9715	0.47%	4.431%	3.333%	10.000%	25483.31	0	229605
INFLUENCER 8	66	15853	0.42%	15.688%	9.091%	54.545%	23949.99	0	167424
INFLUENCER 9	16	23205	0.07%	3.142%	2.500%	30.000%	18323.86	0	108549
INFLUENCER 10	63	7496	0.84%	3.954%	2.778%	16.667%	23193.78	0	229605
INFLUENCER 11	19	15506	0.12%	2.956%	2.703%	5.405%	21766.29	0	229605
INFLUENCER 12	17	11794	0.14%	1.541%	0.389%	11.284%	25191.47	0	274910
INFLUENCER 13	3	8481	0.04%	3.125%	3.125%	3.125%	21776.19	0	165649
INFLUENCER 14	16	20101	0.08%	0.282%	0.188%	0.942%	25679.69	136	229605
INFLUENCER 15	14	19945	0.07%	4.575%	2.941%	14.706%	26928.93	1039	144313
INFLUENCER 16	9	7941	0.11%	6.190%	1.124%	52.809%	22154.54	305	127778
INFLUENCER 17	58	20222	0.29%	1.829%	1.471%	7.353%	25817.63	0	229605
INFLUENCER 18	209	24195	0.86%	2.796%	0.474%	54.028%	18983.69	0	144313
INFLUENCER 19	95	19145	0.50%	2.788%	1.961%	45.098%	21952.27	0	229605
INFLUENCER 20	19	22130	0.09%	5.000%	5.000%	5.000%	18465.59	0	229605
INFLUENCER 21	16	7507	0.21%	3.947%	0.377%	7.170%	17231.02	233	167424
INFLUENCER 22	75	6680	1.12%	4.247%	0.565%	20.904%	11201.24	19	123670
INFLUENCER 23	5	18711	0.03%	3.171%	2.439%	7.317%	20297.06	0	93202
INFLUENCER 24	135	11079	1.22%	1.929%	1.087%	9.783%	18081.28	0	167424
INFLUENCER 25	7	17686	0.04%	0.046%	0.027%	0.269%	35791.65	150	167424
INFLUENCER 26	17	15530	0.11%	0.933%	0.408%	4.490%	28179.06	211	274910
INFLUENCER 27	45	11952	0.38%	1.159%	0.549%	5.495%	20678.04	0	123670
INFLUENCER 28	56	9883	0.57%	5.119%	1.190%	27.381%	23049.24	0	229605
INFLUENCER 29	23	15519	0.15%	4.348%	4.348%	4.348%	24805.46	0	229605
INFLUENCER 30	26	16038	0.16%	2.018%	0.115%	51.491%	34919.26	18	165649
INFLUENCER 31	11	7854	0.14%	0.945%	0.212%	7.203%	17931.20	0	167424
INFLUENCER 32	28	21437	0.13%	0.750%	0.114%	21.396%	20454.51	0	167424
INFLUENCER 33	31	24446	0.13%	10.000%	10.000%	10.000%	24378.27	0	274910
INFLUENCER 34	55	9731	0.57%	1.763%	0.334%	42.475%	16129.28	0	167424
INFLUENCER 35	80	20926	0.38%	0.229%	0.123%	3.567%	23585.28	0	274910
INFLUENCER 36	6	10414	0.06%	0.049%	0.027%	0.382%	27987.17	257	127778
INFLUENCER 37	83	17002	0.49%	0.627%	0.274%	3.836%	23560.73	0	229605

Model $M2L^{UB}$) and use this solution set to compute the objective value of Model $M2L$. CPLEX solver was used for computing the solutions for the mixed-integer optimization program. The results of Case Study-I and II are presented in Tables B.10 and B.11 (in Appendix B.6) respectively. The tables report (i) the objective values for high and low a_{kj}, (ii) the benefit that a firm gets from an influencer (see column - Benefit per Influencer), and (iii) % improvement in objective value for an additional influencer (i.e., marginal benefit of an additional influencer). Each row in the table represents the solution corresponding to problem instance t, i.e., number of influencers to be selected. For example, in Case Study-I, $t = 1$ denotes that out of 18 influencers the firm wants to hire only one influencer.

Figures 3.9 and 3.10 illustrate the value of each influencer with increasing t for Case study-I and II respectively. Figures 3.11 and 3.12 report the % marginal benefit of an influencer for Case study-I and II respectively. The downward trend in all these figures suggests that as the firms choose to hire multiple influencers, the benefit per influencer decreases. In addition, the benefit per influencer flattens out faster when the second level influence is lower. Furthermore, we report the total benefit of hiring influencers for Case Study-I and II in Figures 3.13 and 3.14 respectively. These figures consistently indicate that the benefit from influencers does not increase linearly with number of influencers. This leads us to our next insight.

Insight 4.: *The objective function of the main problem to select an influencer set is monotonic but at a decreasing rate with the number of selected influencers.*

The main takeaway from the above insight is that the firms should carefully decide on how many influencers to hire. In particular, hiring too many influencers might lead to higher overall costs than the benefit. This can be seen in Figures 3.11, 3.12, 3.13, and 3.14, wherein the curves flatten out with high values of t. Whereas, hiring too few influencers leads to decreased effectiveness of the ad campaign. Using Figures 3.9 and 3.10, the firm can perform a cost benefit analysis and decide to hire the right number of influencers. For example, industry reports suggest that the average cost of hiring micro-influencers is around \$271

Figure 3.9: Case Study Problem-I: Value per Influencer

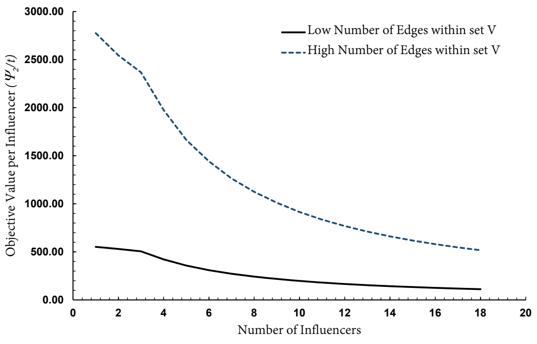


Figure 3.10: Case Study Problem-II: Value per Influencer

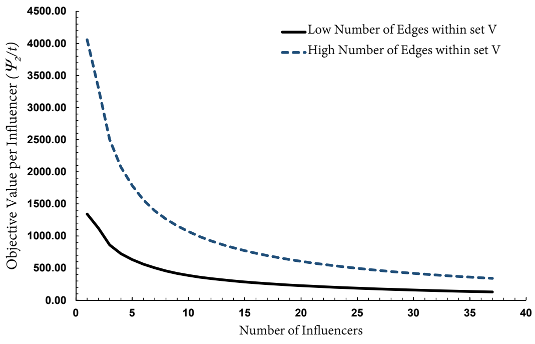


Figure 3.11: Case Study Problem-I: % Increase in Benefit of an Influencer

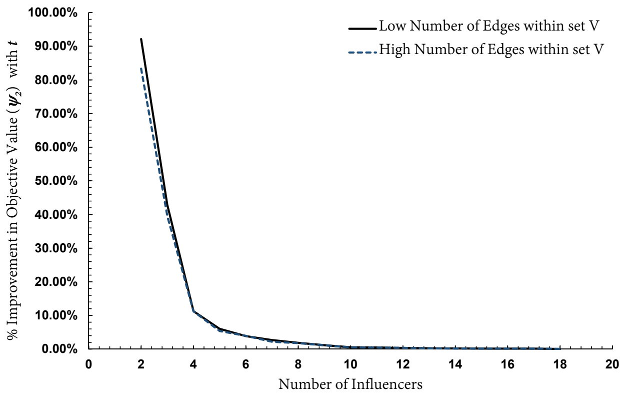
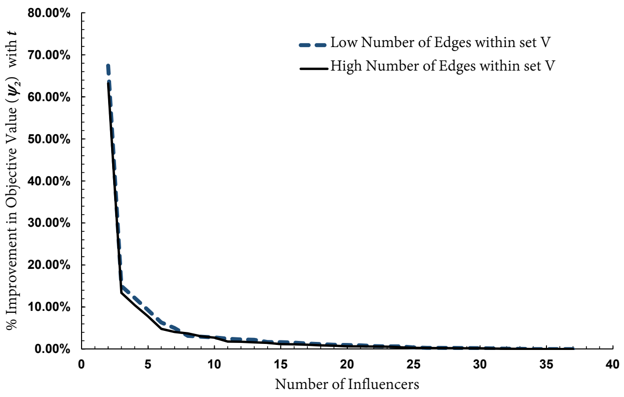


Figure 3.12: Case Study Problem-II: % Increase in Benefit of an Influencer



(Heald 2017). In the context of Case Study-II, if the cost of hiring an influencer is around \$271, the firm is better off with hiring only 7 influencers as the benefit per influencer goes below \$271 with 7 influencers (in the case of low second level influence).

3.4.2 Comparison of Proposed Solutions to Current Industry Practices

Next, we assess the performance of our solution framework against current industry practices. For this, we first reached out to several firms that facilitate influencer marketing.

Figure 3.13: Case Study Problem-I: Marginal Benefit of an Influencer

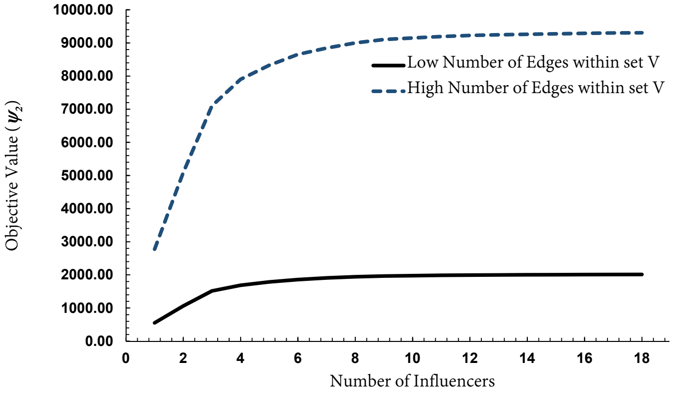
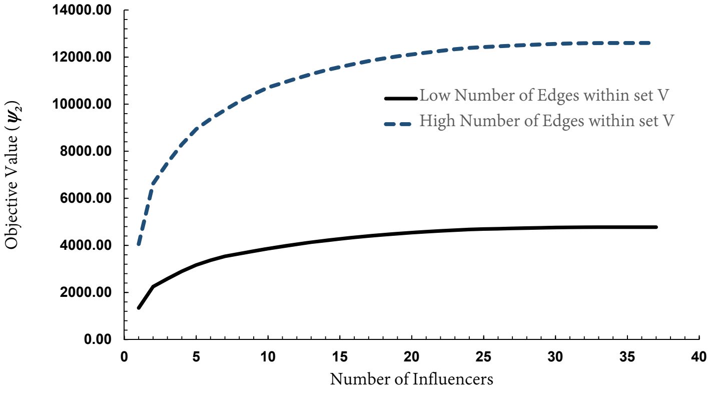


Figure 3.14: Case Study Problem-II: Marginal Benefit of an Influencer

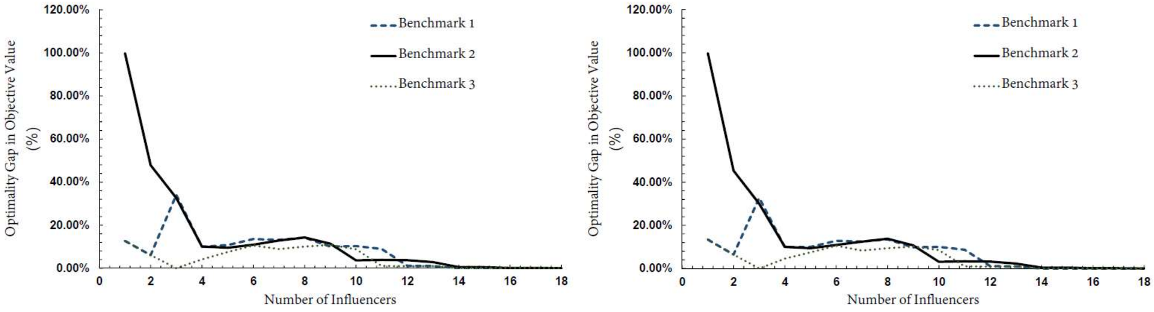


Our discussions with these firms indicate that the selection of influencers is done mostly in an ad-hoc manner. They communicated that the decisions are made based on either (i) number of followers of an influencer (we refer to this procedure as Benchmark 1), or (ii) % active followers of an influencer (we refer to this procedure as Benchmark 2), or (iii) number of active followers of an influencer (we refer to this procedure as Benchmark 3). To check the performance of our framework against the current practices of a marketing firm, we compare the objective values obtained from our model against the objective values computed by Benchmark models 1, 2, and 3. We report the results of Case Study-I and Case Study-II in Tables B.12 and B.13 (in Appendix B.6) respectively. The tables report (i) the objective values computed by our model and Benchmark models 1, 2, and 3, and (ii) the performance metric quantified by % optimality gap between our model and the Benchmark models, where $\%Optimality\ Gap = \frac{\Psi_2^o - \Psi_2^B}{\Psi_2^o}$, where Ψ_2^0 denotes model-based solution and Ψ_2^B is the solution obtained by the Benchmark models. Each row in the table represents the solution corresponding to problem instance t, i.e., number of influencers to be selected. For example, in Case Study-I, $t = 1$ denotes that out of 18 influencers the firm wants to hire only one influencer. Further, we assess the performance for both high and low values of second level influence.

In Figures 3.15 and 3.16, we report the % gaps between our model-based solution and the benchmark models for case studies I and II respectively. In all the scenarios, our model-based solution outperforms the benchmark models. In particular, in Case Study-II, our model-based solution outperforms the benchmark models in the range of 38%-80% when $t \leq 7$. With \$271 being the average cost of hiring an influencer, we had found that firms benefit is maximized at $t = 7$. Hence, an increase in benefit by $> 38\%$ is significantly high.

Insight 5. *Our optimization based framework outperforms benchmark methods consistently in the plausible range of number of influencers (i.e., t). As expected, the gap between them decreases as t increases.*

Figure 3.15: Case Study Problem-I: Sensitivity of Optimality Gaps w.r.t. Number of Influencers



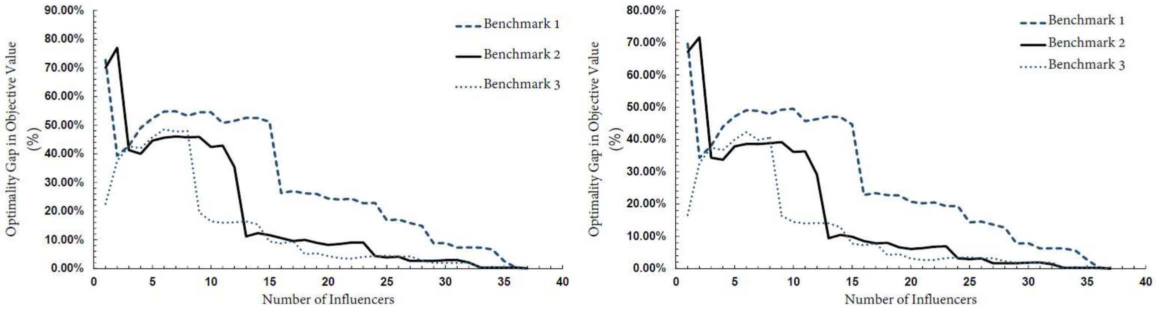
(a) Low a_{kj}

(b) High a_{kj}

3.5 Scheduling Influencers' Ads

We now shift our focus to the second phase, i.e., scheduling of ads to be posted by influencers that are selected in the first phase of our framework. More specifically, we now present a framework to help firms with scheduling of ads to be posted by influencers on their social media over a planning horizon. Although scheduling of ads in different contexts such

Figure 3.16: Case Study Problem-II: Sensitivity of Optimality Gaps w.r.t. Number of Influencers



(a) Low a_{kj}

(b) High a_{kj}

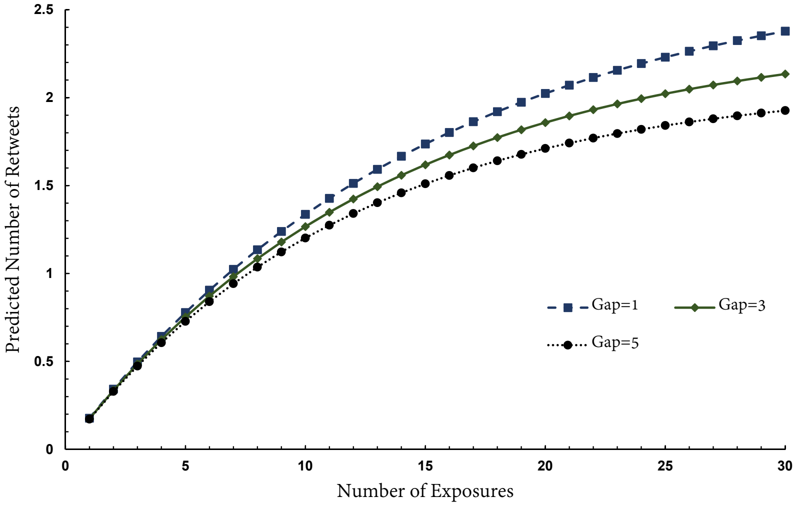
as television, websites, and mobiles has been studied extensively in the operations literature (e.g., Bollapragada et al. 2004; Kumar et al. 2006; Sun et al. 2017), our study is the first, to the best of our knowledge, to explore scheduling of ads for an influencer marketing campaign.

In the previous sections, we developed a modeling framework to help firms select a set of influencers for their ad campaign. The modeling assumption in Model $M2L$ may suggest that it may be optimal if all the influencers (that are selected) post at the same time to take advantage of the multiple exposure effect. However, scheduling in such a way is not feasible due to several reasons. First, firms, in general, plan to ensure that the ads consistently reach the audience throughout the planning horizon. This assumption is consistent with the earlier literature in scheduling of ads in different mediums such as television. For example, Bollapragada et al. (2004) suggest that advertisers typically want ads to be evenly placed over the planning horizon. Second, following Seshadri et al. (2015), who suggest "the contracts between the advertisers and the [television] network require that the network delivers a target viewership." In the context of this study, a typical contract between the firm and influencer would require that an influencer generates a certain level of *engagement* (measured by the number of retweets which can be predetermined by the firms). Further, the firms may want to consistently generate a certain level of engagement throughout the planning horizon (e.g., if the planning horizon is one month, the firm may want to set weekly engagement targets rather than a single monthly engagement target level). Finally, firms have a limited budget, and hence they may not have enough budget to let all influencers post the ads throughout the planning horizon. Consequently, it is necessary for firms to sequence the ads by influencers in such a way that influencers generate engagement consistently throughout the planning horizon while ensuring that they do not exceed their budget levels.

Prior literature on advertising has examined the effects of message spacing, i.e., should the ads be evenly spaced out (e.g., one ad every two days for two weeks) or should the ads be pulsed (e.g., 20 ads in 2 days), however, there are two schools of thoughts. On the one hand, Schmidt and Eisend (2015) demonstrate that message spacing has a positive effect on the relationship between exposures and attitude towards brand. This is because evenly spacing ads (instead of publishing all the ads at once) will prevent the audience develop boredom about the ad. For example, Heflin and Haygood (1985) demonstrate that massed repetition of ads led to a negative response from the customers. On the other hand, Schmidt and Eisend (2015) also find that the evenly spaced messages have a negative effect on the relationship between exposures and recall of a brand. To clear this ambiguity on the impact of message spacing on exposure effect, we empirically examine how message spacing (i.e., days between each ad with similar content) affects the relationship between the number of exposures of an ad (i.e., number of times the ad with similar content was tweeted) and the total number of retweets.

To empirically examine the effect of message spacing, we collect data from Twitter related to 18,571 tweets posted by 37 influencers over a period of 10 months. We present the details of our empirical study in Section B.3 of the Appendix. We briefly summarize the main finding of this empirical study. Our results indicate that message spacing, which we operationalize as the number of days between each identical ad (denoted by *Gap* in the empirical model) moderates the relationship between exposure (denoted by *Exposure*) and number of retweets (denoted by *Retweet Count*). We plot the relationship between *Exposure* and *Retweet Count* at different values of *Gap* in Figure 3.17. This graph accounts for all the effects including number of exposures, message spacing, influencer individual heterogeneity, and shows the total effect of exposure. As seen in the figure, the non-linear effect of exposures on engagement varies with message spacing. More specifically, with higher gap between each tweet, the effect of exposure lessens. This leads us to the following insight.

Figure 3.17: Empirically Predicted Relationship Between Total Number of Retweets, Days Between, and Exposure



Insight 6. *Message spacing moderates the non-linear relationship between number of exposures and number of retweets.*

This insight leads us to an important trade-off that needs to be taken into account when developing the scheduling framework. Although the empirical finding suggests that firms should reduce the number of days between each message (e.g., post the ad every day), doing so will, however, increase the costs to the firm as they need to pay influencers for posting the ad everyday. Hence, this trade-off needs to be accounted in our model. We utilize this empirical finding to develop the scheduling framework for ads to be posted by influencers. Before presenting our model, we present the assumptions that we make for our model.

ASSUMPTION 1: We assume that the diminishing effect of the number of exposures on the total number of retweets depends on the gap between two messages of similar content. In particular, we assume that the diminishing effect is greater when the gap between two messages is high.

ASSUMPTION 2: We assume that the influencers can be hired for posting ads as per the requirement of a firm. That is, the message frequency and quantity can be controlled by the firm.

ASSUMPTION 3: We assume that the firm would like to meet a certain engagement level, $\mathscr{E}$, every n time periods throughout the planning horizon.

ASSUMPTION 4: For analytical tractability, we assume that the probability that a follower retweets is same for all the influencers. In particular, we assume that $p_{ij} = p_j$. Further, p_j can be estimated by taking the mean of p_{ij}, $\forall i \in W$.

ASSUMPTION 5: To maintain a manageable computational complexity, we ignore peer influence (i.e., we set $p_{kj} = 0$). This assumption is reasonable because in the first phase (i.e., selection model), we already considered the peer influence, and thus the influencers who got selected in this phase might have high peer effect. Therefore, we believe that ignoring the peer effect in the second phase will not affect the solution (i.e., our model would provide us with a reasonably close solution).

We now formally state the problem of scheduling of ads to be posted by influencers. We refer to this problem as the scheduling problem.

PROBLEM INSTANCE: A firm would like to schedule the ads to be posted by multiple influencers over T time periods (e.g., days). For illustration purpose, in our problem instance, we assume $T = 12$ days. However, T can be larger or smaller than 12 days. The firm wants to categorize the influencers into 5 groups. The influencers in the first group post the same ad every day for 12 consecutive days. The influencers in the second group (denoted by modes

2 and 3) post the same ad once every two days. The influencers in the third group (denoted by modes 4, 5, and 6) post the same ad once every three days. Next, influencers in the fourth group (denoted by modes 7, 8, 9, 10, 11, and 12) post once in six days. Finally, influencer in the fifth group (denoted by modes 13-24) post at most one time during the planning horizon. Figure 3.18 illustrates the problem instance, where **X** denotes the day on which the influencer will post an ad. Further, in each of these groups, there could be several modes. For example, in group four, there are six different modes depending on what day the first ad is posted on Twitter. Note that number of modes can be varied based on the firm's requirement. Our empirical findings have indicated that the effect of the ad is highest when the users get an ad every day. In other words, when a user does not get an ad on one of the days, the probability of retweet decreases. We incorporate this loss in the probability of retweet by introducing a discount factor.

SOLUTION: Decision to assign influencers to a particular mode and group.

OBJECTIVE: Maximize the expected benefit to the firm for placing ads on social media via influencers.

3.5.1 Scheduling Model

The mathematical formulation is formally presented in Appendix B.5. The model is a mixed integer program. We briefly describe our model.

Objective Function: The objective of the firm in this model is to maximize the benefit it attains. As in Model $M2L$, the benefit, $B_j = b_j + d_j F_j$. However, there are costs involved in the hiring of influencers. Since we assume different modes, the costs of hiring influencers depend on how many ads that he/she is scheduled to post. In summary, the objective function for this model is to maximize the difference between the benefit and the costs.

Constraints: The first set of constraints determine whether an influencer is hired or not and if hired, which mode and group he/she is hired into. The second set of constraints estimate the expected marginal benefit in terms of number retweets. The third set of constraints keep track of how many consecutive times a follower, $j \in V$ receives a message. If the follower does

Figure 3.18: Scheduling Problem Framework

	Day	1	2	3	4	5	6	7	8	9	10	11	12
Group 1	Mode 1	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖
Group 2	Mode 2	✖		✖		✖		✖		✖		✖	
	Mode 3		✖		✖		✖		✖		✖		✖
Group 3	Mode 4	✖			✖			✖			✖		
	Mode 5		✖			✖			✖			✖	
	Mode 6			✖			✖			✖			✖
Group 4	Mode 7	✖						✖					
	Mode 8		✖						✖				
	Mode 9			✖						✖			
	Mode 10				✖						✖		
	Mode 11					✖						✖	
	Mode 12						✖						✖
Group 5	Mode 13	✖											
	Mode 14		✖										
	Mode 15			✖									
	Mode 16				✖								
	Mode 17					✖							
	Mode 18						✖						
	Mode 19							✖					
	Mode 20								✖				
	Mode 21									✖			
	Mode 22										✖		
	Mode 23											✖	
	Mode 24												✖

✖ - denotes that an ad is scheduled on this day

not receive a message on consecutive days, there is a penalty based on our empirical findings. This penalty is referred to as a discount factor. Finally, a set of constraints ensure that the posts by influencer generate a certain number of retweets. Since, we are maximizing the profit (i.e., benefit to the firm – cost to the firm), we initially ignore the budget constraint. However, if the cost of running the campaign is likely to be more than the available budget, we can easily add a budget constraint. Further, we perform sensitivity analysis with respect to budget to understand the effect of other parameters on budget. We note that several of these constraints turn out to be non-linear in nature. However, we present alternative formulations to linearize these constraints. Refer to Section B.5.1 in the Appendix. As a result, the optimization model for scheduling of ads is a mixed integer program.

3.5.2 Insights

To generate recommendations related to the scheduling problem, we perform extensive numerical analysis. The model was solved using a commercial linear solver, CPLEX, which is run on Intel Core i7-7700 CPU running at 3.60 GHz (2 cores) with 64GB ram. we present two sets of insights. First, we present insights in a scenario where there is no requirement for minimum engagement levels or the levels are extremely small, i.e., $\mathscr{E}$ is ignored. Next, we present insights when there is a minimum requirement for engagement levels. The details of the experiments are presented in Sections B.5.2 and B.5.2.1 of the Appendix.

3.5.2.1 Low Requirements of Engagement Levels

First, in a scenario where an influencer has less number of followers and the average probability the followers retweeting is low, we observe that it is not beneficial for the firm to hire him/her at a high price. However, if the retweet probability of the followers is high, we observe that it is beneficial for the firm to hire that influencer even when the cost of hiring him/her is high and the influencer has less number of followers. The intuition behind this result is that the expected benefit overcomes the high cost of hiring that the firm attains when the followers retweet. The results pertaining to this observation are presented in Table B.14 (in the appendix). The observation is summarized in the insight below.

Insight 7. *For an influencer with relatively less number of followers and who is relatively expensive to hire, we observe the following:*

- *If the probability of retweet of followers is relatively low, do not hire the influencer.*
- *If the probability of retweet of followers is relatively high, hire the influencer but use him sparingly.*

Next, we observe that an influencer with a large number of followers should be hired for posting an ad if he/she is less costly to hire despite low probability that his/her followers engage with the message. The rationale is that the although the probability of retweet is low since the message is reaching a larger audience the expected benefit is supposed to compensate for the low cost of influencer. The results pertaining to this insight are presented in Table B.14. The insight is formally presented below.

Insight 8. *If the relative cost of hiring an influencer with high number of followers is low, we observe that the firm should hire the influencer and assign him/her to tweet the ad everyday even if the average probability of followers retweeting the influencer's ad (i.e., p_j) is relatively low.*

Next, when an influencer with high following is costly to hire, it is beneficial to hire him but how to use him/her depends on characteristics of other influencers. Likewise, when an unpopular influencer is less costly, the firm can hire him/her but assign him/her to a mode based on characteristics of other influencers. The results related to these observations are presented in Table B.15 and summarized below.

Insight 9.

- *If an influencer with comparatively high following is costly to hire, we observe that it is cost effective to hire the influencer, however assignment to a certain mode depends on the characteristics of other influencers.*
- *If an influencer with comparatively low following is less costly to hire, we observe that it is cost effective to hire the influencer, however assignment to a certain mode depends on the characteristics of other influencers.*

Insight 10. *When the overlap of followers between influencers is relatively high, we observe the following:*

- *If the average retweet probability of followers is comparatively high, then the firm should hire all the influencers when the influencers have relatively high following on social media, even though they are costly to hire.*

- *If the average retweet probability of followers is comparatively low,*
    - *the firm should not hire any influencer if they have relatively low following on social media, even though they are less costly to hire.*
    - *the firm should hire all influencers if they have relatively high following.*

The results pertaining to the above insight are reported in Table B.16. The above insight highlights the importance of taking into account the overlap of followers between influencers. In particular, when the influencers have large followings and have high overlap, the above observation suggests that the firm should hire most of the influencers for posting ads.

3.5.2.2 High Requirements of Engagement Levels

The results pertaining to the numerical analysis used in this section are presented in Table B.17. We observe that when the firms need to generate high engagement levels by the end of the planning horizon, they need to assign influencers to post more number of ads (see Figure 3.19, Type denotes different network types). This observation is presented in the insight below.

Insight 11. *In a scenario where the firm needs to achieve its target engagement level ($\mathscr{E}$) by the end of the planning horizon, we observe that as the value of $\mathscr{E}$ increases the total number of ads posted by all the influencers increases.*

Next, we observe that the budget required by the firm increases non-linearly with the engagement levels (See Figure 3.20). This observation makes an interesting connection between budget and engagement levels, where the budget increases at an increasing rate.

Insight 12. *In a scenario where the firm needs to achieve its target engagement level ($\mathscr{E}$) by the end of the planning horizon, we observe that as the value of $\mathscr{E}$ increases the total budget required to run the campaign increases non-linearly at an increasing rate.*

The above insight indicates that a firm needs to have realistic expectations with regard to their engagement levels. If the firm has relatively high engagement levels, the cost of

Figure 3.19: Sensitivity of Engagement with respect to Number of Ads

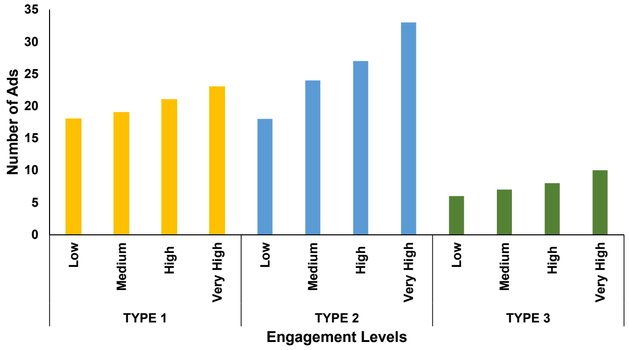
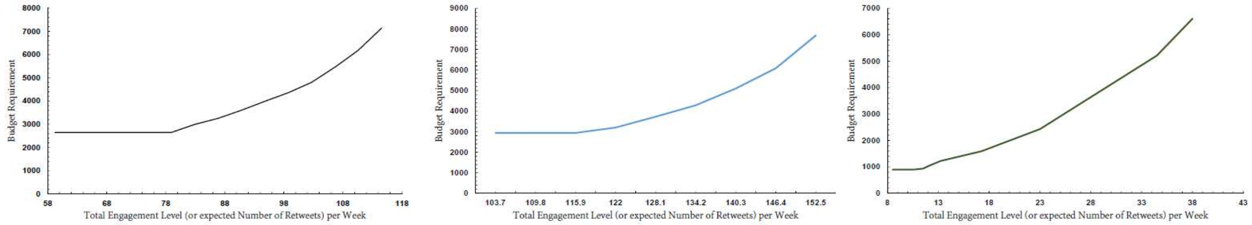


Figure 3.20: Sensitivity of Engagement with respect to Budget



(a) Type 1 Network

(b) Type 2 Network

(c) Type 3 Network

generating that engagement through influencers increases rapidly. In the next subsection, we briefly discuss how our framework can be utilized by firms interested in influencer marketing.

3.5.3 Dynamic Implementation of Scheduling Model

In phase one of our framework, we choose an optimal influencer set. The solution that we obtained in the first phase is based on the assumption that the retweet probability can be estimated and likely to remain the same throughout the ad campaign. However, when the planning horizon is long (e.g., one year), the retweet probability may increase or decrease. Therefore, the retweet probability needs to be updated dynamically. In particular, we need to continually learn about the parameters and update them to changing needs and re-estimate the influencer set. Hence, there is a need for updating the different sequence of ads frequently. The firm may start of with a yearly schedule by assuming the retweet probability based on historical averages, whereas for the monthly and weekly plan the probabilities can be updated dynamically based on the performance of the influencers and a new weekly or monthly scheduled can be generated. Since the data related to dynamic learning can be obtained only from practice, we leave this for future research in this area.

3.6 Conclusion and Future Research Directions

Although firms have relied on social networks to advertise their products, the ubiquity of social media has enabled easier access to large social networks. This study is among the first of its kind to model the problem of conducting an influencer marketing campaign. We provide a solution framework for analyzing the problem of selecting and scheduling influencers during the planning horizon for marketing campaigns. Our framework is based on interactions with marketers, exploratory analysis, and empirical analysis performed on data from Twitter. We decompose the problem into two phases: (i) selection of influencers, (ii) scheduling the placements of ads by influencers. For the first phase, we present the mathematical formulation to the problem (i.e., the main model) of choosing influencers. We show that firms can experience a significant profit loss when implementing a policy that is computed based on a model (namely single level model) that ignores the peer effect. The profit loss is particularly significant when the online network has a high number of connections among the peer influencers, and the influence of the peer effect is high. However, finding an optimal influencer set in main model with the presence of high peer effect is challenging. We propose an alternative formulation that provides a near-optimal solution for the main model. This alternative formulation is also non-linear, but we provide an approximation procedure to transform the problem into a mixed integer linear program. Furthermore, through computational analysis, we find that the gap between the upper bound model and the main model slightly increases with peer effect. Our analysis highlights the importance of peer effect on a firm's benefit. More specifically, we find that the gap between the single level model and the main model can be as high as 85% in some instances.

In the second phase of our framework, we propose an optimization model for scheduling the ads to be posted by the influencers (selected in the first phase) during the planning horizon of a marketing campaign. We provide several managerially relevant insights based on extensive numerical experiments. For example, our analysis suggests that an influencer with relatively large number of followers should be hired for posting an ad if he/she is relatively less costly to hire despite low probability of his/her followers engaging with the message (in terms of number of retweets). Furthermore, we find that the cost of influencer marketing campaign increases non-linearly with engagement. By laying out a framework, the current work provides a foundation for future research in influencer marketing campaigns. Future research can look at analyzing the impact of the interaction between budget and cost of influencers on optimal influencer set, i.e., introduce budget constraint in the main model. Another possible extension to the problem of the selection influencer set is to take into account the spatial considerations. Although we do not tackle these two extensions in the present study, we present the models in the appendix that incorporate budget and spatial requirements and leave further analysis for future research.

4. ESSAY 3: THE EFFECTS OF SOCIAL MEDIA CONTENT ON ENGAGEMENT

4.1 Introduction

Social media in the recent times has become one of the primary mediums for individual celebrities or human brands, including politicians, CEOs, top management executives, celebrities, musicians, artists, and athletes, to directly connect, communicate, and engage with their online fans and followers. Human brands, also referred to as personal brands, are individuals who desire to market and promote themselves to their desired audience. A recent article in Wall Street Journal suggests that human brands are the most influential users on social media and firms are increasingly trying to adopt human brands' strategies to build their brands on social media (Seetharaman 2015). Indeed, human brands rather than traditional brands (i.e., firms and organizations) have larger social media following, in terms of number of followers, compared to traditional "non-human" brands. For example, while a popular brand Coca-Cola has around 3.42 million followers on Twitter, Katy Perry has more than 109.5 million followers on Twitter and Cristiano Ronaldo's page has more than 120 million followers on Facebook.[1]

The recent $1 billion Nike's deal with Cristiano Ronaldo is an appropriate example for the economic value of human brand on social media. Badenhausen (2017) argues that "Cristiano is one of the top influencers on the planet who has effectively leveraged his social following and engagement into a media powerhouse to drive tremendous value for his sponsor," which suggests that the high brand valuation of Ronaldo is mainly attributed to his ability to generate engagement with his followers and being visible on social media. Despite the vast popularity of human brands, research on the social media strategy of human brands, in particular the effects of social media content created by human brands is scarce. Our study seeks to fill this gap.

[1] https://twittercounter.com/pages/100; https://twitter.com/CocaCola; https://twitter.com/katyperry; and https://www.facebook.com/Cristiano/ (last accessed: June 19, 2018).

We define the human brand-generated content (HGC) as messages, blogs, comments, Facebook posts, or tweets created and posted by celebrities (or human brands) on social media designed to interact, connect, communicate, and more importantly engage with their fans, followers, and online community in general. Although the recent studies in the area of social media analyze the effects of user-generated content (UGC) and firm-generated content (FGC) on social media (Goh et al. 2013; Lee et al. 2018), no study to our knowledge has systematically examined the effect of social media content created by human brands. In contrast to the relationship of social media users with firms, they build emotional connect with human brands because of virtual face-to-face relationship. Thus, users are likely to have different engagement patterns with content created by human brands compared to FGC. Indeed, research on social media activities of human brands has received attention in recent time. For example, Petrova et al. (2017) find that social media adoption led to an increase of 2-3% in contributions received by politicians (i.e., human brands). In a closely related study, Saboo et al. (2016) study the impact of social media activities of human brands (in the context of music industry) on product sales. Our study contributes to the literature by analyzing the effects of HGC, i.e., social media content generated by human brands. A unique characteristic of social media platforms is that it facilitates direct communication and more importantly engagement between human brands and online audience or customers. Recent business reports have suggested that engagement is one of the key metric to quantify the performance of social media strategies (Stelzner 2016). Recent studies suggest that higher customer social media engagement plays an important role in increasing sales (Kumar et al. 2016) and brand's profitability (Rishika et al. 2013). While anecdotal evidence suggests that human brands have been successful in creating and sustaining social media engagement, little is known if (and how) the content generated by human brands (i.e., HGC) affects engagement with the online audience.

There are several reasons why people engage with the content generated by human brands. One reason could be that the content has some handy information, and hence the audience would like to share the content of the human brand (Berger and Milkman 2012). While there are several reasons why audience may choose to engage with the content, Heath et al. (2001) argue that the emotional facet of the content plays a vital role in determining whether audience engages with the content. Yet, there is no consensus on the effects of the tone and no study to our knowledge has examined the effects of tone of HGC. Against the above background, we examine the effect of HGC on social media engagement with a focus on the differential effects of positive versus negative toned HGC.

4.1.1 Research Questions

More formally, the first research question that we examine is: (1) What is the impact of the tone of HGC on social media audience engagement? First, we analyze the impact of positive tone of HGC on social media engagement. On the one hand, one could argue that the audience is more likely to engage with positive toned HGC on social media. The rationale behind this argument is that the audience is more likely to interact with a positive toned content as it reflects the positive attitude of the person engaging with the article. On the other hand, based on the work of Fiske (1980), one can make a competing argument that positive content, especially in social media, is uninteresting and does not entice the curiosity of audience. Thus, audience may likely choose not to engage with positive toned HGC. With respect to the negative tone HGC, literature in psychology suggests that negative toned messages are likely to be perceived as engaging and may evoke more interest (Fiske 1980). Thus, we argue that the social media audience is more likely to interact or engage with candidate's negative toned tweets. However, negative tone of the content could have detrimental effects, as the audience may not want to be associated with negative toned content due to the fear of unintended backlash. Given this lack of empirical clarity on the relationship between the tone of content (both positive and negative) and engagement, we seek to answer the first research question, i.e., are the audience likely to engage more with positive toned content or with negative toned content? Our study contributes to this growing literature on the tone of content by providing an empirically driven answer to the first research question.

Human brands with large followers will likely have higher engagement levels and thus the effect of the tones of HGC by audience may vary depending on the popularity of the candidate. More specifically, will audience engage differently with content created by popular human brands in contrast to less popular human brands? Given that audience have different expectations from a popular human brand compared to a less popular human brand, understanding the interplay between popularity and the HGC tones becomes important as human brands pursue strategies to increase their engagement online. Although prior literature has examined the impact of negative publicity on a brand's popularity, what is not clear from the existing literature is the impact of social media popularity on the relationship between HGC tone and engagement. Therefore, our second research question is: (2) Does social media popularity moderate the relationship between the content tone (positive and negative) and engagement? Analyzing this question will provide important insights on whether human brands need to use different strategies, with respect to creating content with a certain tone, depending on how popular they are among the audience. In addition, our findings will contribute to the leadership and brand popularity literature.

A unique aspect of human brands is that they often belong to a group or a community. For instance, athletes, politicians are associated with teams and political parties respectively. We define related brands as individual brands (e.g., Corn Flakes and Froot Loops) belonging to the same corporate or umbrella brand (e.g., Kellogg's). Literature in psychology suggests that when forming an opinion about a person or a focal brand, the audience may look for traits that are common between the focal brand and other related or similar brands (Bazerman and Moore 2008). In a similar vein, if audience were to perceive some human brands as related to each other, the tone of HGC of other brands that are related to a focal brand can moderate the relationship between the HGC and engagement of a focal human brand. We refer to this phenomenon to as perception spillover effects between brands. We capture this by the moderating effect of related brands' HGC tone on the association between a focal human brand's HGC tone and engagement. Although spillover effects have been documented in the context of traditional brands, no study, to the best of our knowledge, has examined the role of perception spillover effects in the context of HGC and social media platforms. Our findings on spillover effects provide important insights to human brands as to how related brands' content can affect their social media strategy to increase engagement. To summarize, the third key research question that we analyze is: (3) What are the spillover effects (if any) of related human brands HGC tone on focal human brand's engagement levels in social media platforms? In particular, does the tone of related brands' HGC tone moderate the relationship between focal human brand's HGC and its engagement levels on social media platforms? By providing an empirically driven answer to the third research question, our study adds to the growing literature on perception spillovers.

4.1.2 Research Context

To empirically answer our research questions, we identify four critical requirements of the research setting that serves as the basis for our study. First, we require a context where human brands actively use social media platforms to communicate their ideas and messages to the target audience with an aim to increase their engagement levels. Second, we need a context where human brands are likely to use both positive and negative toned content on social media with an aim to engage with their audience. Third, we require a context where human brands belong to a group and the tone of HGC by a focal human brand is likely to affect the engagement of related human brands (i.e., human brands belonging to the same group). Finally, since we are investigating the effect of popularity, we need human brands with varying levels of popularity among the social media audience.

To accomplish our objectives, we leverage a novel dataset in the context of Indian General Elections 2014. The context of elections in which political candidates are human brands and political parties as the umbrella brands satisfy the above critical requirements and aid us in examining our research questions. Although brands were mainly associated with products and organizations, the notion of branding is popular in political science (Smith and French

2009). In recent years, alike corporate brands, both the political parties and the politicians have increasingly embraced social media platforms as a means to promote their brands and to engage with the target audience or their constituents. Several popular news outlets and industry studies have also emphasized the role of social media in election campaigns. For instance, a recent study reports that 66% of the online users have used social media to engage in political activities (Duggan and Smith 2016). Bode and Dalrymple (2015) report that all the major candidates during the 2010 U.S. elections have extensively used social media to engage with their constituents and audience in general. Furthermore, recent anecdotal evidence and evidence from prior literature suggest the pivotal role played by social media in the elections worldwide. For example, the widespread use of social media during the 2008 U.S. Presidential election campaign has been cited as one of the key reasons behind success of a presidential candidate (Carr 2008). The trend of widespread use of social media during the elections is prevalent across the world. For instance, the 2014 Indian general election was famously labeled as “India’s first social media election” (Khullar and Haridasani 2012) because of extensive use of social media by politicians and political parties.

Political candidates often post positive content to promote their manifesto, agenda, and ideas on various topics to their constituents and negative toned content to attack the reputation of their opponents. Furthermore, political candidates are not only concerned about their own victory, but also need to make sure that they do not hurt audience engagement with other members of their political party (or related candidates). This is because political candidates have an incentive to ensure that the related candidates are also elected. For example, in the case of a parliamentary system that is followed in India, where the elected members determine the Prime Minister, it is likely that social media content of candidates who belong to the same political party as that of a focal candidate affects how audience reacts to the social media content posted by the focal candidate. Further, in the context of politics and politicians, we have human brands with varying levels of popularity on social media, which allows us to analyze the impact of popularity on engagement. The above-discussed characteristics of political candidates and elections in general provides us with an ideal empirical setting to examine our research questions.

Set in the context of Indian general election held in 2014, we assemble a novel dataset using data collected from a popular social media platform (Twitter.com) and the Election Commission of India (ECI). Our data consists of detailed information on each political candidate's tweets, number of times each tweet was retweeted (i.e., shared), and the popularity of each candidate on Twitter. We combine this information with other offline characteristics of political candidates and their respective political parties obtained from ECI. We employ the Naive Bayes sentiment classification algorithm (Antweiler and Frank 2004) to categorize the tone of candidates' tweets into positive sentiment and negative sentiment. Several studies have identified multiple challenges associated with examining social network effects from observational data. In particular, we face the problem of differentiating the effect of HGC on engagement from different social interaction effects, i.e., engagement is consequence of users interacting with each other on social media rather than the actual HGC. To account for the social interaction effects in social media, we follow the prescriptions from recent literature on social networks (Ghose and Han 2011; Park et al. 2018) and propose a multi-level mixed effects model to answer the three primary research questions. In addition, we perform several robustness checks to rule out alternative explanations of our results. Specifically, we supplement our main analysis with different model specifications, estimation strategies, metrics for measuring engagement levels, and accounting for potential selection bias. In the next subsection, we briefly discuss our key results and contributions.

4.1.3 Key Findings and Contributions

Towards the first research question, we investigate the impact of positive toned HGC and negative toned HGC on social media engagement. We find that negative toned tweets by human brands/candidates are associated with higher levels of engagement as measured by number of retweets. Next, our results indicate that there is no empirical evidence for presence of relationship between positive toned tweets and audience engagement levels. Whereas, prior literature has analyzed the impact of tone of content (e.g. Berger and Milkman 2012), our study is the first to examine and document the relationship between the tone of social media content of human brands and audience engagement on social media. By demonstrating that the engagement levels are higher when the tone of HGC is negative, our study provides important insights on drivers of engagement to human brands, an important measure of success.

Towards the second research question, i.e., the moderating role of popularity of focal human brand on the relationship between human brand's HGC and engagement level, our results indicate that the human brand's popularity negatively moderates the positive relationship between negative tone and engagement. We also find that the marginal effect of positive toned HGC is higher for popular human brands compared to less popular human brands. In other words, our findings reveal that negative toned HGC is beneficial for less popular human brands while positive toned HGC is more favorable for human brands that are more popular.

Finally, with respect to spillover effects, we find evidence for asymmetric spillover effects of content created by brands related to the focal brand. In particular, we find that the positive effect of negative toned HGC is greater (resp., lower) when the tone of related candidates is negative (resp., positive). In other words, the increased audience engagement from negative toned HGC reduces when the tone of related brands is positive and increases when the tone of the related brands is negative. This study, to the best of our knowledge, is the first to analyze and document the moderating role of related brands' HGC. Our results demonstrate that the human brand's engagement levels depend not only on his/her tone but also on the tone of other related human brands. Finally, to demonstrate the effect of social media engagement, we extend our analysis to provide evidence on the how engagement increases the buzz and visibility of the candidate on social media.

Our work makes several important contributions to the academic literature and practice. First, our study complements the growing literature on social media content (e.g. Lee et al.

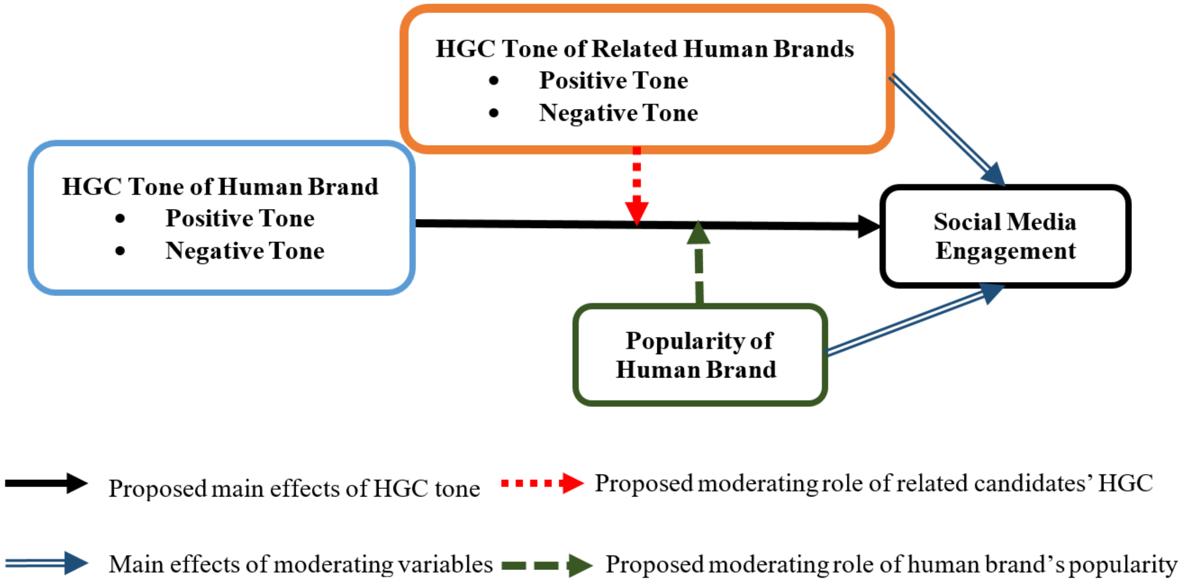
2018). Although the focus of the literature has been on the content generated by users and firms, we contribute to the information systems (IS) and marketing literature on social media engagement and social media content by studying the impact of tone of content generated by human brands, who are the most widely followed on social media, on engagement. Second, we identify and demonstrate the role of human brand's popularity as a moderator. Ignoring this moderating role may have an undesirable effect on the engagement levels of popular human brands. Third, we establish the presence of spillover effect, i.e., the human brand's engagement is affected by the content of the related brands. These findings provide new insights, thus help advance of current knowledge in research on social media content. From the practice perspective, we demonstrate the need for a social media strategy that takes into account the popularity of the human brand, the spillover effects of related brands' content, and the moderating role of related brands' content. Although the current study is in the context of human brands from the field of politics, our findings should extend to human brands in the corporate world. Top management executives of businesses are also under constant pressure to engage with the social media audience in an effort to influence the perception of their businesses (Saboo et al. 2016). Industry studies suggest that top management executives who engage on social media are found to be better equipped than their peers who are not active on social media in mitigating risks, enhancing the credibility of the business, and handling crisis (BrandFrog.com 2016). Our findings provide actionable insights to the executives on strategies to improve their social media engagement.

The remainder of the paper is organized as follows. In Section 4.2, we present the conceptual model to explain the rationale behind the effects of HGC tone on social media engagement. In Section 4.3, we present the data and model development. We discuss the results of our model in Section 4.4, followed by the discussion on how engagement affects social media visibility Section 4.5. In Section 4.6, we provide discussion on supplementary analysis that we conduct to ensure the robustness of our results. In Section 4.7, we discuss the contributions and implications of our study to theory and practice. Finally, we conclude by presenting the limitations of our study and future research opportunities in Section 4.8.

4.2 Conceptual Framework

In this section, we develop and present a conceptual framework (see Figure 4.1) that explains the rationale behind the effects of tone of HGC on social media engagement levels. We build upon existing literature from the areas of IS, psychology and marketing to explain and theorize the possible directions of the relationship between the tone of the human brand and the social media engagement.

Figure 4.1: Conceptual Framework


HGC Tone of Related Human Brands
Positive Tone
Negative Tone
HGC Tone of Human Brand
Positive Tone
Negative Tone
Social Media Engagement
Popularity of Human Brand
Proposed main effects of HGC tone
Proposed moderating role of related candidates' HGC
Main effects of moderating variables
Proposed moderating role of human brand's popularity

4.2.1 Social Media Content, Tone, and Engagement

Van Doorn et al. (2010, p. 254) define engagement as "customer's behavioral manifestation towards a brand" and argue, "engagement is behavioral construct that goes beyond purchase behavior alone." In the context of social media, retweeting or sharing allows a user to share social media content and helps the content reach beyond the followers and fans of original source thereby amplifying the impact of the original content (Kumar et al. 2010;

Kwak et al. 2010). Brand marketers acknowledge that the number of retweets is an appropriate measure of audience engagement on Twitter (Long 2015b). Consequently, consistent with more recent studies (e.g. Lambrecht et al. 2018; Lee et al. 2018), we operationalize engagement by the number of retweets.

Researchers in psychology and marketing have long been interested in understanding why audience engage or share information (e.g., news article or advertisement content) with others (e.g. Berger and Milkman 2012). Among all the factors that can persuade a user to engage with the content, emotional aspect of the content has often been cited as one of the critical driving forces (Heath et al. 2001). In the context of our study, i.e., social media content created by political candidates, emotions can be expressed by varying the tones of message. For instance, disappointment or anger against political opposition or about a policy can be expressed in the form of negative toned message, while the positive tone of content reveals politician's enthusiasm and his/her strengths compared to the opponents. Although the findings in the literature consistently demonstrate that users are more likely to engage with highly emotional content (Rime et al. 1991), scholars have called for more rigorous research in understanding the differential effects of positive vs negative tone on a user's decision to share (Godes et al. 2005). Our study responds to this call by analyzing the effects of positive and negative toned social media content created by human brands on engagement.

Positive toned social media content, in the context of elections, serves to communicate the policies and plans of the political candidate and his/her political party to the online audience in addition to emphasizing the strengths of the candidate. According to Cialdini (1984), people who express positive emotion are perceived more positively, and hence considered more likable than people who express negative mood. Further, Berger and Milkman (2012) argue that most users want to be identified as someone who endorse optimistic feel-good messages or messages from likable people rather than someone who share negative information, which embodies pessimism. The aforementioned reasons suggest that the online users may be more driven to share positive content rather than negative content. Based on the assumption that the online audience want to be perceived positively among their peers, we expect that the audience is likely to engage more with positive toned content of political candidates by sharing the content with their friends on social media.

Negative toned content in politics is used to attack an opponent by highlighting his/her weaknesses. Although negative campaigning is often associated with politics, it is quite common to see negative or comparative advertising in brand/product promotion. The impression formation literature in psychology, which studies how consumers process and integrate information, explains one of the rationales behind the effect of negative advertising. Within the impression formation literature, the novelty theory proposed by Fiske (1980) argues that negative information is novel and highly revealing as it distinguishes itself from other more common information and thus requires greater attention. As a result, individuals pay more attention to negative information and value it more when compared to positive information (Hamilton and Zanna 1972; Baumeister et al. 2001). This effect of negative information has been documented in both product evaluation and person perception (Wright 1974; Herr et al. 1991). Furthermore, negative information is deemed more diagnostic when compared to positive information, and hence given more weight by the audience (Maheswaran and Meyers-Levy 1990; Herr et al. 1991). Based on the aforementioned research, we argue that a human brand's negative toned HGC will evoke the interest of the online community, and therefore the online users are more likely to engage with the human brands by retweeting more number of negative toned HGC. In summary, our study builds upon the prior literature that focuses on the effects of the tone of content on the perception of audience; we extend the literature and understand the effects of negative and positive toned HGC on engagement with the audience on social media platforms.

4.2.2 Moderating Role of Popularity

Our context of elections brings up the role of leadership or popularity of political candidates. Literature on leadership suggests that followers view leaders as their role models and have certain expectations in terms of moral behavior of leaders (Johnson 2009). Aronson (2001) suggests that it is essential for leaders to set a "moral example" to their followers by emphasizing on values, beliefs, and honorable behavior, which in turn results in positive follower enthusiasm. Indeed, prior literature on leadership has demonstrated that followers expect charismatic leaders to propagate positive emotion (e.g. Lewis 2000). In the context of elections, audience anticipate charismatic leaders such as presidential candidates or other popular political candidates to behave in a manner befitting their stature. For example, the audience may not want charismatic leaders to indulge in mudslinging; rather expect them to create positive toned HGC consistent with his/her stature. On the one hand, the audience will have comparatively lower ethical expectations standards from not so popular candidates (i.e., regular political party members). Therefore, we postulate that the online audience will evaluate the tone of content based on the popularity and stature of the leader. As a result, we expect that the online audience will engage more when popular human brands create positive toned HGC compared to negative HGC.

On the other hand, recent literature suggests that negative publicity is good for firms who are not well known, while it negatively affects the firms that are familiar to the customers (Berger et al. 2010). Grounded on this reasoning, we propose that the negative tone of content will help human brands who are not very popular compared to popular human brands. An alternative explanation behind our postulation that negative tone may have a larger impact on engagement for less popular brands compared to brands that are more popular is as follows. Human brands that are more popular may already have high engagement levels given their fame among the audience, and hence we expect lower marginal effect of the tone of content on engagement for popular human brands. In summary, our study seeks to identify and document how the effect of HGC tone on engagement differs for different human brands depending on the popularity of the human brands.

4.2.3 Spillover Focal Brands and Related Brands

Spillover occurs when existing information and perception about a brand influences the beliefs and opinions of another brand (Janakiraman et al. 2009). Prior studies in the area of marketing have found evidence for presence of spillover, i.e., the perception of brands related to a focal brand influencing the perception of the focal brand (Lei et al. 2008). For example, in the context of marketing, information or perception about Corn Flakes can affect consumers' perception of Froot Loops, as they belong to the same corporate or umbrella brand (i.e., Kellogg's). Similar to the spillover of perception between brands related to an umbrella brand, we argue that spillover of perception can happen between political candidates related to the same political party. Analogous to corporate brands, political parties represent an umbrella brand with several individual brands, i.e., individual political members of a political party (Singer 2002). Hence, we argue that the perception of a focal political candidate can affect how audience evaluate the other candidates related to the focal candidates. In this study, we seek to examine whether (and how) perception spills over in the context of social media content created by human brands.

Drawing on the accessibilitydiagnosticity theory, we explain the underlying rationale behind the spillover of perceptions between human brands belonging to the same political party. The central argument of the theory is that if an individual considers that brand X is informative (diagnostic) about brand Y, s/he will use the information and perceptions of brand X to form an opinion about brand Y (Janakiraman et al. 2009). However, this is effective only if both the brands are accessible (or retrieved) in the memory of the individual at the same time (Roehm and Tybout 2006). In other words, perception spills over between human brands only when the association between the brands is strong and thereby they are associated in the individual's memory. In our context, the political members are often grouped together based on their party lines, and hence we argue that the candidates belonging to the same party are accessed in the memory of audience at the same time. Based on the arguments of accessibility-diagnosticity theory, it is likely that audience view the content created by the candidates related to the focal candidate as diagnostic of the focal political candidate.

Further, the spreading activation framework suggests that information about brands and their attributes reside in customers' knowledge network as network nodes and network linkages between such nodes promote accessibility (Anderson 1983; Janakiraman et al. 2009). Therefore, we argue that when a follower sees a tweet on Twitter from a focal candidate, the node of the focal candidate is activated; in addition, all the associated nodes (i.e., related candidates) are activated as well causing retrieval of memory of the user. Therefore, a user utilizes all the information that is available in his/her memory (i.e., retrieves and evaluates the HGC of focal and related candidates) when making a choice to engage with the social media content of a focal human brand.

Whereas the spreading activation framework proposes that the HGC tone of related brands will influence the decision of a user to engage with HGC of focal human brand, no study to our knowledge has analyzed how the effect of HGC tone of focal brand on engagement changes based on related brands' HGC. In particular, we propose that the related brands tone will moderate the relationship between the focal brand's tone and audience engagement, which we refer to as perception spillover effects within human brands. As we are interested in the positive and negative tone of content created by both the focal human brand and brands related to the focal brand, we have four possible moderating effects, namely (i) the positive tone of related brands moderating the positive tone of focal brand, (ii) the negative tone of related brands moderating the negative tone of focal brand, (iii) the positive tone of related brands moderating the negative tone of focal brand, and (iv) the negative tone of related brands moderating the positive tone of focal brand.

The multiple source effect proposed by Harkins and Petty (1981) suggests that when the audience is exposed to similar kind of information from multiple sources, they are more likely to process and trust that information. In the context of elections, when both the political candidate and the other members his/her political party use a similar toned content (e.g., all the members of a political party use similar negative toned content to attack the policies of opposing political party), it reinforces the credibility of the information in the content. Consequently, audience perceive the information to be trustworthy, which results in higher likelihood of users engaging with the content of a focal brand. Thus, we expect greater engagement levels when the tones of the focal brand and the related brands are similar. Specifically, we argue that the engagement levels will increase when the tones of both the focal and the related brands are positive (i.e., the positive tone of related brands is expected to positively moderate the effect of focal human brand's positive tone on engagement). Likewise, we expect increased engagement levels when the tones of both the focal and the related brands are negative.

In contrast, when the tones of focal human brands and the related brands are different (i.e., the tone of one is positive while the tone of other is negative), we argue that the opposite toned messages may work against each other. In other words, the opposite tones of focal and related brands can lead to lack of credibility of their content. Furthermore, as the tones of both the focal and related brands are different, we argue that the messages are significantly different from each other and hence lacks for validation of information from multiple sources. As a result, we expect the engagement levels to decrease when the tones of focal brand and related brands contrast with each other. Specifically, we posit that engagement levels will lessen when (i) the tone of related brands is positive and the tone of focal brand is negative or (ii) the tone of related brands is negative and the tone of focal brand is positive.

4.3 Data and Model

In this section, we first summarize our data. Next, we describe the variables of interest. Finally, we present our proposed empirical model to examine the effects of HGC on audience engagement.

4.3.1 Data Description

As discussed earlier, the context of our study is the Indian general elections held in 2014. India is the largest democracy in the world. The elections in India are widely followed around the world as it has a widespread impact on global economy (Agrawal 2017). India follows a parliamentary form of government where the parliament is the supreme legislative body comprising two houses: Rajya Sabha (or the Upper House) and Lok Sabha (or the Lower House). Constituents or people from a pre-defined geographical region (or a constituency) elect a representative (called as Member of Parliament or MP in short) to represent them in the Lok Sabha. The members of the Lok Sabha in turn elect the Prime Minister (PM) of India. The PM is generally a member of the majority political party in the Lok Sabha. The MPs not only elect the PM of India, but also act a representative of the people of the constituency. Lok Sabha consists of 543 elected representatives of people and the elections are held every five years. The general elections in India were last held in 2014 from April 7, 2014 to May 12, 2014. Several political parties and politicians have taken advantage of the enormous size of network on various social media platforms to reach out to their target constituents. Given the parliamentary structure of Indian elections, candidates belonging to the same political party are typically viewed as similar to each other and this makes the issue of perception spillover across politicians belonging to the same political party important.

Given this unique structure of the Indian election, we believe that the context is well suited to study our research questions for the following reasons. First, evidences from popular news media suggest that the use of social media by the political candidates in the Indian general election 2014 was high (Khullar and Haridasani 2012). Second, the primary objective of the candidates is to engage with the target audience and to propagate their messages. Third, given the parliamentary nature of elections, several candidates contest the elections, thus, this context provides us with the right setting to understand the spillover effects between the members of a political party. Finally, the period surrounding the Indian general election provides us with the right context to examine our research questions, as the candidates and the political parties are typically more active on social media trying to leverage the huge network of online users and hence the impact of social media tone on engagement can be examined.

Our main data come from a popular micro-blogging website, Twitter. Our data consist of individual level tweets related to the 2014 Indian general election. We collected tweets (content on Twitter) using various keywords, such as the names of political parties, politicians, and other terms related to the Indian elections from December 1, 2013 to May 31, 2014 (i.e., before, during, and after elections). Our data consists of two main components: (i) individual tweets posted by political candidates (i.e., human brands), prominent political party members, and the official Twitter accounts of political parties (i.e., related human brands); and (ii) tweets by users of social media mentioning the human brands and the related brands. Regarding the first component, we collect the content of tweets generated by 297 Twitter accounts belonging to the candidates (i.e., human brands) and important political party members (who are not contesting the elections) and political parties (i.e., related human brands). Since the focus of our study is to analyze the relationship between the tone of content and engagement, it is important to focus on candidates who are actively using social media platform to engage with the online audience. In other words, only when the candidates are active on social media, can the online audience engage with the candidate. As a result, we drop candidates with fewer than twenty-six tweets (i.e., at least one tweet per week).[2] Therefore, our main analysis focuses on the content generated by 297 Twitter accounts belonging to the candidates, important political party members (who are not contesting the elections), and political parties.

To summarize, our primary data consists of 63 candidates contesting the elections and 234 Twitter accounts belonging to the political parties and their members (i.e., related candidates). In order to confirm that these accounts were indeed official (i.e., the content is tweeted by the human brands), we manually verified the Twitter accounts of the Indian

[2] Not all candidates start tweeting in the first week of our dataset.

politicians. In our dataset, we found that a few candidates had more than one official Twitter account, in such an instance, we combined the data from all accounts. The above criteria leave us with 16,471 distinct tweets by 63 candidates and 98,237 tweets by 234 related candidates. Our second data source is the website of Election Commission of India.[3] The data from this source comprises of the personal attributes of a candidate including age, education, gender, result of the election, political party, total net assets, and their respective regions. This data helps us to control for candidate specific heterogeneity and fixed effects. In the following subsections, we discuss the operationalization of the variables we use.

4.3.2 Dependent Variable: Number of Retweets

The unit of our analysis is at the candidate-weekly level. The dependent variable in our study is the number of retweets (denoted by RT_{ipt}) received in week t by candidate i belonging to the political party p. When a human brand posts a message on Twitter, the message appears on the timeline of the followers. When the followers login to their Twitter account, they see the recent tweets from the human brands they follow. However, if users are not very active on Twitter (for example, they open the Twitter once a month), they will generally see only the most recent tweets. Thus, we assume that a user has access to only recent tweets and retweets tweets that are generated within the same week. We identify the number of retweets for each of the 16,471 distinct tweets posted by 63 candidates contesting the general elections.

4.3.3 Main Independent Variables

Our main explanatory variables deal with the sentiment of tweets by the individual candidates and their respective political party. Following recent studies, we employ a commonly used sentiment classification algorithm - Naive Bayes algorithm (Antweiler and Frank 2004) to classify tweets into three categories: positive sentiment, negative sentiment, and neutral/mixed sentiment. This technique is proven to be highly reliable and has been widely

[3] See http://eci.nic.in.

used in prior studies in management science (e.g. Das and Chen 2007; Gu et al. 2014). As the name suggests, this algorithm is based on the Bayes theorem. The Naive Bayes model is based upon the following principles: words in a tweet are independent to each other, each word has a pre-determined tone or sentiment, and the overall sentiment or tone of a sentence (or a tweet) is the collective sentiment of all the words in the tweet. For more details, please refer to (Antweiler and Frank 2004).

In addition to classifying the tweets into positive tone and negative tone, the sentiment classification algorithm categorizes tweets into mixed or neutral tone. We define mixed tone tweets as those that contain either both positive and negative tone or neither of these tones. Among the 16,471 tweets by candidates contesting the elections, our classification algorithm classified 3,965 tweets as positive toned, 1,710 tweets as negative toned, and the remaining tweets as mixed toned. The number of positive toned tweets in week t posted by candidate i belonging to political party p is denoted by $PosT_{ipt}$. $NegT_{ipt}$ represents the number of negative toned tweets by candidate i belonging to political party p during week t. Finally, $MixT_{ipt}$ represents the number of mixed toned tweets by candidate i belonging to political party p during week t.

Next, we capture the number of tweets of different tones of all the related candidates. The related candidates of the human brand i of the political party p consist of: (i) all the candidates of the political party p, (ii) the prominent members of the political party p who are not participating in the elections, and (iii) the official Twitter accounts of the political party p at the state level and the national level. The number of positive toned tweets by the related candidates (denoted by $PartyPosT_{ipt}$) is the sum of all the positive toned tweets by the candidates and prominent members belonging to party p not including the positive toned tweets of the focal candidate during week t. Likewise, $PartyNegT_{ipt}$ and $PartyMixT_{ipt}$ represent the number of negative and mixed/neutral toned tweets by related candidates. Among the 98,237 tweets by related candidates, the Naive Bayes algorithm classified 19,933 tweets as positive toned, 9,729 tweets as negative toned, and the rest as mixed toned.

Our final important explanatory variable is candidate's popularity on social media, which we measure by number of followers (denoted by Fol_{ipt}). We note that the number of followers is dynamic (i.e., it could increase or decrease over a period). Therefore, we operationalize this variable by calculating the number of followers on a weekly basis.

4.3.4 Control Variables

We include various candidate-specific variables in our model to improve the precision of our estimates and to control for heterogeneity at the candidate level. Age_i is the actual age of the candidate in years. The age of the candidate is a broad indicator for the experience of the candidate. $Male_i$ is a dummy variable set equal to one if the candidate is male. In addition, we control for the candidate's spending power measured by candidate's net worth in logarithmic scale ($assets_i$). We also control the level of education through three dummy variables (i.e., high school or below, college degree, and higher degree). Further, we control for the candidate's geographical area through the dummy variables $Region_i$ (i.e., North India, East India, West India, and South India). We obtain the data on candidate-specific time-invariant control variables from the website of the election commission of India. We also control for the expenditure by the candidate's party during the election ($partyExp_i$). To control for temporal shocks that represent major developments, such as announcements of election dates, and phases of elections, we include time fixed effects in our model.

Berger and Schwartz (2011) suggest that the visibility of the candidate on social media platform stimulates interest in the candidate, and it is positively correlated with engagement. Therefore, we control for the visibility of the candidate through the variable $mentions_{ipt}$, which represents log-transformed number of mentions of a candidate in tweets by all the users on Twitter at time t. Our data consists of entire set of Twitter activity on the topic of elections. Hence, mention of a candidate by their Twitter username is a good proxy for his/her social media visibility.[4] In other words, this variable captures how many times a candidate i was tagged by all the Twitter users in their tweets. Tagging a candidate on Twitter implies that the audience on social media is interested in discussing about the candidate. Thus, we expect that higher chatter about a candidate in week $t-1$ will motivate users to follow the HGC (or the tweets) of the candidate. That is, the higher chatter in week $t-1$ may affect the engagement in week t. Next, we control for the retweets and mentions received by the related brands ($PartyRT_{ipt}$ and $PartyMC_{ipt}$). Finally, we include the tone of opposition candidates ($OppPosT_{ipt}$, $OppNegT_{ipt}$, and $OppMixT_{ipt}$) to remove any unobserved confounding factors. We summarize the operationalization of all the variables and their descriptive statistics in Table 4.1.

4.3.5 Model Development

In the rest of this subsection, we discuss model development to answer our research questions: (i) the effect of human brand's HGC tone on social media engagement levels; (ii) the moderating role of popularity on the relationship between HGC tone and social media engagement; and (iii) the moderating role of related human brand's HGC tone on the relationship between HGC tone and social media engagement. Before presenting our model, we first present the identification challenges associated with using observable dataset and then provide discussion on how we overcome these challenges.

4.3.5.1 Identification Challenges and Strategies

We recognize that we work with observational data to establish the effect of HGC tone on social media engagement and that there are some potential challenges that need to be overcome. In particular, since we examine how audience engages with the content generated by human brands, we face the challenge of differentiating the audience engagement effect from the social interaction effect. Social interaction effect exists when a user directly influences

[4] Users may use the first or last name of the candidate in their tweet rather than using their Twitter username. However, we do not consider tweets that do not use the official Twitter username in a tweet. One reason is because several Indian names are quite common and hence it would not be possible to differentiate the candidates from non-candidates.

Table 4.1: Variable Operationalization and Summary Statistics

Variable	Variable Operationalization	Mean	SD
Dependent Variable			
RT_{ipt}	Total number of retweets of HGC of candidate i of party p during week t	380.46	1975.51
Independent Variables			
$PosT_{ipt}$	Total number of positive toned tweets by candidate i belonging to party p during week t	2.88	6.22
$NegT_{ipt}$	Total number of negative toned tweets by candidate i belonging to party p during week t	1.24	4.19
$MixT_{ipt}$	Total number of mixed or neutral toned tweets by candidate i belonging to party p during week t	7.85	20.66
$PartyPosT_{ipt}$	Total number of positive toned tweets by related brands of candidate i of party p during week t	331.1	212.83
$PartyNegT_{ipt}$	Total number of negative toned tweets by related brands of candidate i of party p during week t	127.02	96.77
$PartyMixT_{ipt}$	Total number of mixed/neutral toned tweets by related brands of candidate i of party p during week t	1122.38	893.7
$\log(Fol_{ipt})$	Number of followers of candidate i on Twitter during week t (on log scale)	9.29	2.73
Control Variables			
$\log(Mentions_{ipt})$	Number of mentions of candidate i by on Twitter by all users during week t (on log scale)	3.51	2.51
$\log(Assets_i)$	Net worth of candidate i (on log scale)	18.07	2.04
Age_i	Age in years of candidate i	50.9	10.59
$Male_i$	Gender of candidate i; Female=0, Male=1		
$Region_i$	Geographical region to which candidate i belongs (i.e., North, East, West, and South India)		
$Education_i$	Education level of candidate i (i.e., High School, Graduate, and Post Graduate)		
$\log(PartyExp_i)$	Expenditure of candidate i's Party (on log scale)	20.17	2.62
$\log(PartyRT_{ipt})$	Number of retweets received by related brands of candidate i of party p during week t (on log scale)	8.85	2.73
$\log(PartyMC_{ipt})$	Number of mentions of related brands of candidate i of party p during week t (on log scale)	9.19	2.79
$OppPosT_{ipt}$	Total number of positive toned tweets by related brands of candidate i of party p during week t	906.88	268.68
$OppNegT_{ipt}$	Total number of negative toned tweets by related brands of candidate i of party p during week t	414.49	216.5
$OppMixT_{ipt}$	Total number of mixed/neutral toned tweets by related brands of candidate i of party p during week t	3275.77	1836.1

the choices/outcomes of other users in a social network (Hartmann et al. 2008). In our context, a follower's decision to engage with the content created by a human brand could depend on the actions of other followers of the focal human brand or the focal political party. Prior literature has identified two primary confounding factors that could hinder uncovering causal effects accurately in social networks (i) correlated unobservables and (ii) reflection problem (Manski 1993; Hartmann et al. 2008). Before presenting our model and estimation methods, we first discuss how we overcome these identification challenges.

The first challenge associated with proper identification of HGC effects is the issue of correlated unobservables. This problem arises when the users' behavior is influenced by unobserved characteristics that may be correlated (Durlauf and Young 2004). In the context of our study, major political events (such as large political meetings, release of manifesto, and major news related to elections), could influence the decision of the audience to engage with the human brand. If ignored, this increase in engagement levels may be mistaken for the actual effect of HGC tone on engagement. To mitigate this effect, we include time-period fixed effects to account for any unobserved exogenous time specific shocks (and to capture important events) such as announcement of election notification, dates, and enforcement of election poll conduct. In line with existing literature, we also account for location fixed effects to account for any region specific effects (through region specific dummy variables) that allow us to control for time invariant spatially correlated unobservables.

The second problem that often arises when dealing with observational data is the issue of simultaneity or what is referred to as *"reflection problem"* in the social networks literature (Manski 1993). This problem arises when individuals belonging to the same group can influence each other's behavior simultaneously. For example, in our context, number of party retweets, and party level mentions could at the same time influence the number of retweets of the focal candidates. In other words, the audience behavior is not fixed rather it varies depending on the behavior of other group members simultaneously. To alleviate this concern, we use lagged values for number of mentions, party mentions, party retweets, and opposition

HGC tone. This method of using lagged values is often employed in the social networks literature to overcome the concern of simultaneity (Hartmann et al. 2008; Park et al. 2018). Hence, in our empirical model, we use lagged values for these variable, i.e., $mentions_{ipt-1}$, $partyMC_{ipt-1}$, $partyRT_{ipt-1}$, $OppPosT_{ipt-1}$, $OppNegT_{ipt-1}$, and $OppMixT_{ipt-1}$.

In the next subsection, we provide detailed discussion on our econometric model and the identification strategy to mitigate the problems discussed above.

4.3.5.2 Candidate-Level Model

To answer our first research question, we develop a candidate-level model where we examine the effect of individual human brand's HGC tone (i.e., $PosT_{ipt}$, $NegT_{ipt}$, and $MixT_{ipt}$) on audience engagement measured via number of retweets (i.e., RT_{ipt}). In line with existing literature (He et al. 2017), we log-transform the dependent variable (i.e., number of retweets, RT_{ipt}) to control for skewness and high dispersion of the variable. To summarize, the dependent variable in our model is the logged number of retweets received by the human brand i belonging to party p in week t (denoted as $\log(RT_{ipt})$). In addition to the effects of HGC tone, we also examine the popularity of human brand (i.e., Fol_{ipt}) on his/her social media engagement. We log transform this variable to control for skewness and high dispersion of the variable.[5] Controlling for the candidate-level heterogeneity through control variables (that we discuss in 3.4), the candidate-level model allows us to examine the relationship between tone and engagement. The candidate-level model is as follows.

$$\log(RT_{ipt}) = \beta_{0pt} + \beta_{1pt}PosT_{ipt} + \beta_{2pt}NegT_{ipt} + \beta_{3pt}MixT_{ipt} + \beta_{4pt}fol_{ipt} + \eta_{ipt} \tag{4.1}$$

where the variable $PosT_{ipt}$ (resp., $NegT_{ipt}$) denotes the number of positive (resp., negative) toned tweets of candidate i belonging to political party p during the time-period t. Likewise, $MixT_{ipt}$ represents the number of mixed or neutral toned tweets and fol_{ipt} denotes

[5] For brevity, from here on, we denote all log-transformed variables in lower case.

the logged number of followers of candidate i belonging to party p during the time-period t.

Estimating the candidate-level model presented in Equation 4.1 using the pooled ordinary least squares (OLS) method will lead to biased and inefficient estimation as candidates and their related candidates belonging to the same party tend to be more similar to each other. Because of this, the errors (η_{ipt}) may no longer be independent (Raudenbush and Bryk 2002). In other words, if the party level unobserved heterogeneity is unaccounted for, the effect of HGC tone that we estimate is likely to have an upward bias. To alleviate this concern and to separate party specific heterogeneity from the random error term, we decompose η_{ipt} into two components. The first component of η_{ipt} consists of random error at the party level. The component is common to all the candidates who belong to the same party (this random error is identical for all the candidates belonging to the same political party). The second term denotes random errors across all candidates and we assume that these errors are independent to each other. Following the literature (e.g., Lee et al. 2015), in order to efficiently estimate the model taking into account random errors at two-levels (i.e., party-level and candidate-level), we employ random coefficients model. This estimation procedure, also referred to as mixed effects estimator, is widely used in the IS and marketing literature when the data is at candidate-level and candidates belong to different groups (e.g., Mithas et al. 2006; Boh et al. 2007). In this estimation procedure, the party-level random errors are captured by introducing party specific random coefficients of main independent variables. Specifically, we allow the coefficients of interest β_{1pt}, β_{2pt}, β_{3pt}, and β_{4pt} (i.e., the slopes of $PosT_{ipt}$, $NegT_{ipt}$, $MixT_{ipt}$, and fol_{ipt}) to account for party specific characteristics. In addition, we also allow the intercept (β_{0pt}) to vary across different political parties. We now present the equations for β_{0pt}, β_{1pt}, β_{2pt}, β_{3pt}, and β_{4pt} as a "party-level" model.

4.3.5.3 Party-Level Model

In the party-level model, the estimates of β_{1pt} and β_{2pt} are modeled as a function of candidate's respective political party (i.e., cumulative effect of all human brands belonging to the same political party). In other words, the party specific coefficients allow us to better control for unobserved party level heterogeneity. The coefficients β_{1pt} and β_{2pt} are modeled as follows:

$$\beta_{1pt} = \gamma_{10} + \gamma_{11} PartyPosT_{ipt} + \gamma_{12} PartyNegT_{ipt} + u_{1pt} \tag{4.2}$$

$$\beta_{2pt} = \gamma_{20} + \gamma_{21} PartyPosT_{ipt} + \gamma_{22} PartyNegT_{ipt} + u_{2pt} \tag{4.3}$$

where $PartyPosT_{ipt}$ ($PartyNegT_{ipt}$) denotes the number of positive (negative) toned tweets of related brands of the focal candidate. We take into account the unobserved heterogeneity of political parties (which are treated as random) by including u_{1pt} and u_{2pt} in each of the coefficients as follows. The random error terms u_{1pt} and u_{2pt} are the same for all the candidates belonging to the same political party and unobserved by us. By allowing the coefficients to vary across different parties, we allow for the impact of key independent variables in Equation 1 to interact with the party-specific HGC behavior as presented in Equations 4.2 and 4.3. In addition, we allow the unobserved heterogeneity across political parties with a random coefficient on the intercept (i.e., β_{0pt}), $MixT_{ipt}$ (i.e., β_{3pt}), and fol_{ipt} (i.e., β_{4pt}) denoted as follows.

$$\beta_{0pt} = \gamma_{00} + \gamma_{01} PartyPosT_{ipt} + \gamma_{02} PartyNegT_{ipt} + u_{0pt} \tag{4.4}$$

$$\beta_{3pt} = \lambda_3 + \nu_{3pt} \tag{4.5}$$

$$\beta_{4pt} = \lambda_4 + \nu_{4pt} \tag{4.6}$$

where u_{0pt}, ν_{3pt}, and ν_{4pt} are random errors that denote unobserved heterogeneity at party level, and λ_3 and λ_4 denote the coefficients of Mix_{ipt} and fol_{ipt}, respectively, after accounting for random errors at the party level.

4.3.5.4 Full Model

Substituting the coefficients β_{1pt}, β_{2pt}, β_{0pt}, β_{3pt}, and β_{4pt} (presented in Equations 4.2, 4.3, 4.4, 4.5, and 4.6 respectively) in Equation 4.1, we obtain:

$$\begin{aligned}
\log(RT_{ipt}) &= \gamma_{00} + \gamma_{01} PartyPosT_{ipt} + \gamma_{02} PartyNegT_{ipt} \\
&+ \gamma_{10} PosT_{ipt} + \gamma_{11} PartyPosT_{ipt} PosT_{ipt} + \gamma_{12} PartyNegT_{ipt} PosT_{ipt} \\
&+ \gamma_{20} NegT_{ipt} + \gamma_{21} PartyPosT_{ipt} NegT_{ipt} + \gamma_{22} PartyNegT_{ipt} NegT_{ipt} \\
&+ \lambda_3 MixT_{ipt} + \lambda_4 fol_{ipt} + u_{0pt} + u_{1pt} PosT_{ipt} + u_{2pt} NegT_{ipt} \\
&+ \nu_{3pt} MixT_{ipt} + \nu_{4pt} fol_{ipt} + \eta_{ipt} \qquad (4.7)
\end{aligned}$$

The model in Equation 4.7 consists of main effects (i.e., $PosT_{ipt}$, $NegT_{ipt}$, and $MixT_{ipt}$) and the interaction of focal candidate's HGC tone and related candidates' HGC tone (i.e., $PosT_{ipt} \times PartyPosT_{ipt}$, $PosT_{ipt} \times PartyNegT_{ipt}$, $NegT_{ipt} \times PartyPosT_{ipt}$, and $NegT_{ipt} \times PartyNegT_{ipt}$). The interactions between the HGC tones of the focal brand and the related brands capture the spillover effects of related brands HGC. In particular, having an interaction term between $PartyPosT_{ipt}$ and $PosT_{ipt}$ captures the spillover effect of positive toned HGC of related brands when the tone of the focal brand is positive. The interaction between $PartyPosT_{ipt}$ and $NegT_{ipt}$ captures the spillover effect of positive toned HGC of related brands when the tone of the focal brand is negative. Similarly, the term $PartyNegT_{ipt} \times PosT_{ipt}$ ($PartyNegT_{ipt} \times NegT_{ipt}$) captures the spillover effect of negative toned HGC of related brands when the tone of focal brand is positive (negative). We thus have four interaction terms to measure the spillover effect as follows: (i) $PartyPosT_{ipt} \times PosT_{ipt}$, (ii) $PartyNegT_{ipt} \times NegT_{ipt}$, (iii) $PartyPosT_{ipt} \times NegT_{ipt}$, and (iv) $PartyNegT_{ipt} \times PosT_{ipt}$. The signs of the coefficients of these interaction terms will provide us with the direction of the spillover effects. For example, if the sign associated with the interaction term $PartyPosT_{ipt} \times PosT_{ipt}$ is positive, there is a positive spillover effect, i.e., positive tone of related brands positively moderates the relationship between the number of positive toned content of focal human brand and engagement.

$$
\begin{aligned}
\log(RT_{ipt}) &= \gamma_{00} + \gamma_{01} PartyPosT_{ipt} + \gamma_{02} PartyNegT_{ipt} \\
&+ \gamma_{10} PosT_{ipt} + \gamma_{11} PartyPosT_{ipt} PosT_{ipt} + \gamma_{12} PartyNegT_{ipt} PosT_{ipt} \\
&+ \gamma_{20} NegT_{ipt} + \gamma_{21} PartyPosT_{ipt} NegT_{ipt} + \gamma_{22} PartyNegT_{ipt} NegT_{ipt} \\
&+ \lambda_3 MixT_{ipt} + \lambda_4 fol_{ipt} + Z\mathscr{C}_{ipt} \\
&+ \lambda_5 fol_{ipt} PosT_{ipt} + \lambda_6 fol_{ipt} NegT_{ipt} + \lambda_7 fol_{ipt} MixT_{ipt} \\
&+ u_{0pt} + u_{1pt} PosT_{ipt} + u_{2pt} NegT_{ipt} + \nu_{3pt} MixT_{ipt} + \nu_{4pt} fol_{ipt} + \eta_{ipt} \qquad (4.8)
\end{aligned}
$$

We include all the control variables discussed in Section 4.3.4 to account for any observed candidate heterogeneity. $\mathscr{C}_{ipt}$ is a vector that consists of all the control variables to account for candidate-level heterogeneity. The list of control variables used in this model is presented in Table 4.2. The parameter estimates of all control variables are denoted by the vector Z. The impact of negative toned HGC on engagement is captured through the coefficient γ_{20} and the effect of positive tone is captured by γ_{10}. Next, the moderation effects of popularity are captured by λ_5, λ_6, and λ_7. The spillover effects between the focal human brand's HGC tone and related brands' HGC tone are captured by the interaction coefficients γ_{11}, γ_{12}, γ_{21}, and γ_{22}.

To summarize, the candidate-level variables are influenced by the party to which they belong. Specifically, the candidate-level data is in the first level nested within the political parties in level two. This structure of data is typically referred to as uniquely nested data structure (Raudenbush and Bryk 2002). However, this nested data structure of our data violates one of the key assumptions of OLS estimation, that is, independence of random errors (Kreft and De Leeuw 1998). To overcome this, following the previous literature, we employ the random coefficients model. This estimation procedure allows us to model the variation in slopes and intercepts across different political parties through level two (party level) variables.

We acknowledge that a controlled or natural experiment is the most appropriate way to establish a causal effect of tones of HGC on audience engagement. However, such an approach may not feasible in our context. We address the challenges (i.e., correlated unobservables and simultaneity) associated with establishing a causal effect in social network settings using observable data by relying on the solutions established in the social networks and panel data econometrics literature. Following the arguments presented in Ghose and Han (2011), we interpret the set of estimates of our model as an upper bound on the causal effect of HGC tone on social media engagement. In the next section, we discuss the results of our proposed model.

4.4 Results

We present the parameter estimates of the coefficients of our proposed model in Table 4.2. Before estimating the results of our proposed model, we estimate a series of alternative null models. First, we estimate a model of twitter engagement as a function of the intercept and all the control variables discussed in Section 4.3.4. Model 2 adds the proposed independent variables to Model 1. Model 3 builds on Model 2 by adding the proposed interaction terms. The proposed model (presented in Equation 4.8) includes the main effects (i.e., candidate's HGC tone and related candidate's HGC tone), interaction terms, and all the control variables. We estimate the proposed model (denoted by Model 3 in Table 4.2) as a random coefficients model using the full maximum likelihood via the EM algorithm (Dempster et al. 1977). The AIC, BIC and log likelihood of the models are presented in Table 4.3. We find that the proposed model has the best fit. All the models account for controls to address the identification challenges discussed in Section 4.3.5.1, time-period dummies, and party-level and individual-level random errors.

Furthermore, to address the concerns about failure to meet standard regression assumptions (e.g., i.i.d. errors), we account for heteroskedastic random errors by using the White's sandwich covariance matrix, which is robust to heteroscedasticity (Wooldridge 2010). We also cluster at political party level to account for within-cluster (i.e., candidates belonging to the same political party) correlation of errors. This clustering of random errors allows the errors to be correlated across all observations of all candidates belonging to the same political party. This includes observations belonging to the same candidate (Cameron and Miller 2015), thus further improving the efficiency of the estimator. We report the robust-clustered standard errors in column Model 4 of Table 4.2.[6] The results for Models 1-4 are presented in Table 4.2. As Model 4 has the best fit and accounts for heteroscedasticity clustered at party level, we focus our discussion on the parameter estimates obtained from this model (i.e., our proposed model).

4.4.1 Effect of Tone of HGC on Engagement

The estimates of the proposed model indicate that the negative toned HGC (i.e., Neg_T) has a positive and significant effect on the number of retweets ($\gamma_{20} = 0.425$ with $p-value < 0.01$). The coefficient associated with positive toned HGC (i.e., the conditional effect of positive HGC) is found to be positive but statistically insignificant.[7] However, we find that the net marginal effect of positive tone (i.e., $\gamma_{10} + \gamma_{11} + \gamma_{12} + \lambda_5$) is significantly positive at mean values of all other variables in the proposed model. The conditional marginal estimates of number of positive tone HGC and their standard errors for different levels of popularity and all other variables at mean are provided in Table C.1. These results indicate that the audience engagement with human brand tends to increase with negative toned HGC, while it is likely to increase with positive toned HGC only for popular human brands. We also find that the coefficient associated with mixed toned HGC ($\lambda_3 = 0.128$ with $p - value < 0.01$) is significantly positive.

[6] For the remainder of the analyses in the paper, we employ robust-clustered standard errors when estimating the parameters.

[7] As we have interaction terms, the interpretation of the estimate of $PosT_{ipt}$ is conditional on all interaction terms with $PosT_{ipt}$ equal to zero. In this case, when fol_{ipt}, $PartyPosT_{ipt}$, and $PartyNegT_{ipt}$ are zero, estimate of $PosT_{ipt}$ is 0.0359 and statistically insignificant.

Table 4.2: Parameter Estimates: Effect of HGC on Social Media Engagement

		Model 1	Model 2	Model 3	Model 4
	DV	***ln(RT)***	***ln(RT)***	***ln(RT)***	***ln(RT)***
		(SE)	(SE)	(SE)	(Robust clustered SE)
Main Effects of HGC	*PosT*		0.0612***	0.0359	0.0359
			(0.0136)	(0.069)	(0.0551)
	NegT		0.0668	0.425***	0.425***
			(0.0572)	(0.102)	(0.159)
	MixT		0.0409***	0.128***	0.128***
			(0.0128)	(0.0186)	(0.044)
	fol		0.177***	0.275***	0.275***
			(0.0375)	(0.0339)	(0.0246)
	PartyPosT		1.56E-03***	1.94E-03***	1.94E-03***
			(0.541E-03)	(0.533E-03)	(0.303E-03)
	PartyNegT		-0.267E-03	-1.79E-03*	-1.79E-03*
			(1.07E-03)	(1.08E-03)	(1.05E-03)
	PartyMixT		-1.36E-05	-5.62E-05	-5.62E-05
			(1.19E-04)	(0.111E-03)	(9.95E-05)
Interaction Effects	*fol* x *PosT*			0.901E-03	0.901E-03
				(5.1E-03)	(3.8E-03)
	fol x *NegT*			-0.0304***	-0.0304***
				(7.33E-03)	(0.0115)
	fol x *MixT*			-7.38E-03***	-7.38E-03**
				(1.3E-03)	(3.25E-03)
	PosT x *PartyPosT*			1.44E-05	1.44E-05
				(7.49E-05)	(4.02E-05)
	PosT x *PartyNegT*			0.99E-05	0.99E-05
				(0.152 E-03)	(0.115E-03)
	NegT x *PartyPosT*			-0.278E-03**	-0.000278**
				(0.13E-03)	(0.136E-03)
	NegT x *PartyNegT*			0.524E-03*	0.000524*
				(0.295E-03)	(0.306E-03)
Controls	*mentions(lag)*	0.688***	0.420***	0.371***	0.371***
		(0.0206)	(0.0326)	(0.0315)	(0.0691)
	assets	-0.107***	-0.0362	-0.0418*	-0.0418
		(0.0238)	(0.0239)	(0.0227)	(0.0915)
	Age	-0.0101*	-1.25E-03	5.02E-03	5.02E-03
		(5.35E-03)	(5.28E-03)	(4.99E-03)	(6.1E-03)
	Male	0.273**	0.313**	0.344***	0.344***
		(0.136)	(0.134)	(0.129)	(0.112)
	partyMC(lag)	-0.21	0.0321	-0.0602	-0.0602
		(0.131)	(0.126)	(0.119)	(0.0486)
	partyRT(lag)	0.382***	0.052	0.129	0.129***
		(0.138)	(0.131)	(0.125)	(0.0283)
	partyExp	-0.100***	-0.213***	-0.196***	-0.196***
		(0.0246)	(0.0548)	(0.0417)	(0.0537)
	Intercept	3.834***	3.449***	2.081**	2.081
		(0.742)	(1.155)	(0.947)	(2.272)
	N	1313	1313	1313	1313

Region/Education /Time/Opposition HGC Controls were included in all the models. The parameter estimates for these controls are not reported for brevity. Standard errors in parentheses for Models 1-3. Robust SE clustered at party level are in parentheses for Model 4. The coefficients of models 2, 3, and 4 are estimated as mixed effects model using full maximum likelihood model via EM algorithm.

* p<0.10, ** p<0.05,*** p<0.01

Table 4.3: Model Fit

Model	Description	Log-Likelihood	AIC	BIC
Model 1	Social media engagement as a function of control variables only	-2515.928	5071.856	5175.458
Model 2	Variables in Model 1 + Main Effects	-2397.009	4858.018	5023.78
Model 3	Variables in Model 2 + Moderating effects of popularity + Spillover effects	-2332.617	4731.234	4902.177
Model 4	Variables in Model 3 (estimated with robust cluster standard errors)	-2332.617	4731.234	4902.177

4.4.2 Moderating Effect of Popularity on the Impact of HGC Tone on Engagement

Turning our attention to the moderating role of popularity, we find that the coefficient of interaction between Neg_T and fol is significantly negative ($\lambda_6 = -0.0304$ with $p - value < 0.01$). To illustrate the moderating effect of popularity on how negative toned HGC affects engagement, we conduct a sensitivity analysis. In this analysis, we plot the predicted engagement levels corresponding to the number of negative toned HGC at different levels of popularity (i.e., number of followers). The results are shown in Figure 4.2. This figure highlights a higher slope for the less popular human brands (as compared to more popular human brands), which suggests that the (positive) effect of negative HGC on engagement diminishes as the popularity of human brand increases.

Although we find that the coefficient of the interaction between the number of positive toned HGC (i.e., $PosT$) and number of followers is insignificant, we perform additional analyses to demonstrate how the marginal effect of $PosT$ varies at different levels of popularity. Figure 4.3 illustrates the conditional marginal effect of positive tone posts on engagement for different levels of popularity at 90% significance level. From this figure, we can infer that the marginal effect of positive tone is statistically positive when $fol8$ at 90% significance level. This implies that the effect of positive tone HGC is significantly greater for popular human brands as opposed to less popular human brand for whom the effect is not signif-

Figure 4.2: Moderating Effect of Popularity (All other variables are at mean)

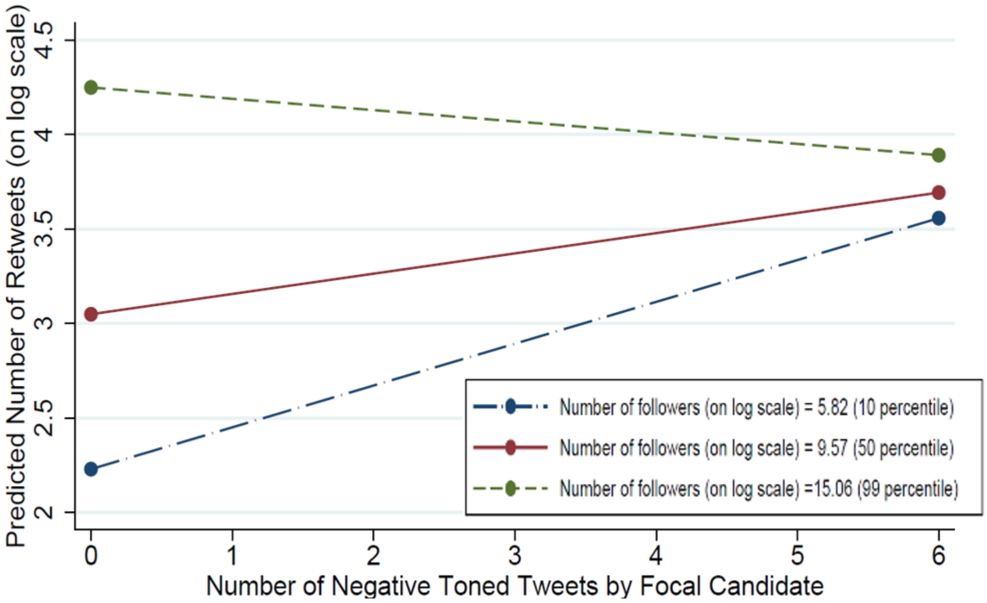
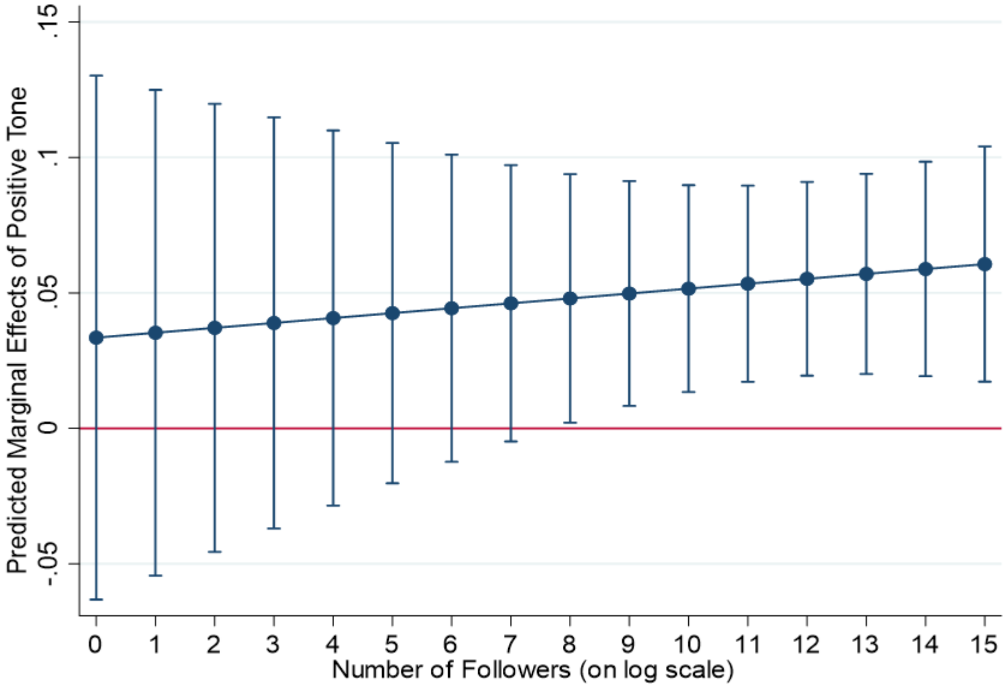


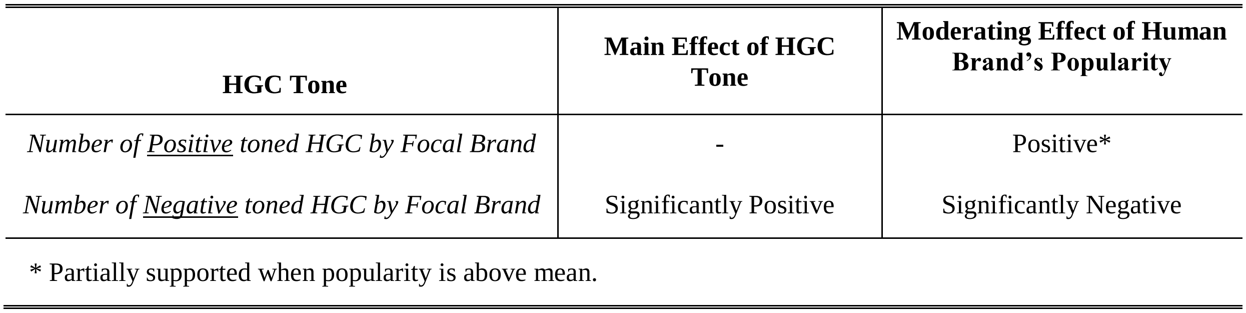
Figure 4.3: Conditional Marginal Effects of Positive Tone on Engagement at Different Levels of Popularity (All other variables are at mean)



icant.[8] Next, the coefficient of interaction between the number of mixed toned HGC and number of followers is significantly negative.

Finally, our results indicate that the conditional effect of popularity (i.e., fol) on audience engagement is significantly positive ($\lambda_4 = 0.275$ with $p - value < 0.01$). To summarize the moderating effect of human brand's popularity, our findings reveal that whereas human brand's popularity (measured via number of followers) lessens the positive impact of negative and mixed toned tweets on engagement, the popularity of a human brand strengthens the positive impact of positive toned tweets on engagement. These findings are summarized in Table 4.4. In the next subsection, we discuss the results related to the spillover effects.

Table 4.4: Summary of Results



HGC Tone	**Main Effect of HGC Tone**	**Moderating Effect of Human Brand's Popularity**
Number of Positive toned HGC by Focal Brand	-	Positive*
Number of Negative toned HGC by Focal Brand	Significantly Positive	Significantly Negative

* Partially supported when popularity is above mean.

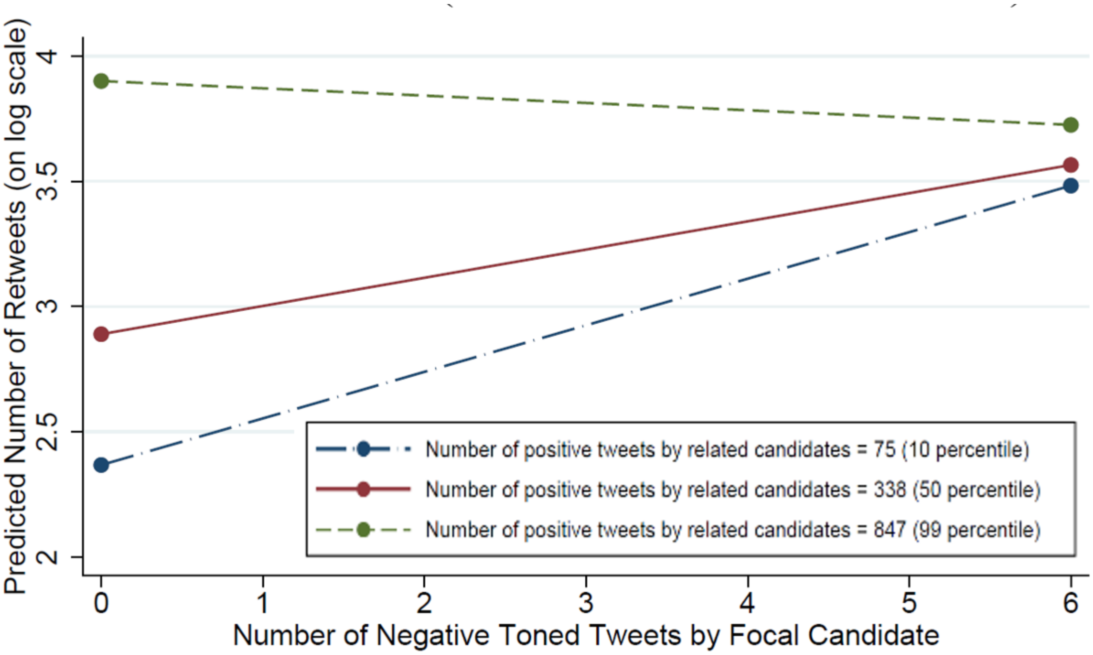
4.4.3 Spillover Effects of Related Brands' HGC Tone

With regards to the spillover effects, the estimates of our model suggest that the interaction between $NegT$ and $PartyPosT$ is negative and significant. That is, the effect of candidate's negative toned HGC on audience engagement decreases with the higher number of positive toned HGC from the related candidates. Next, we find that the interaction between $NegT$ and $PartyNegT$ is positive and significant. This result suggests that the effect of candidate's negative toned HGC on audience engagement is greater when the tone of related candidates' HGC is negative. Figure 4.4 illustrates the interaction effect of $NegT$

[8] We acknowledge that these results are not to be interpreted as a full-blown counterfactual analysis, rather, they are simulations conducted for illustration purposes to understand the dynamics of Pos_T and Neg_T on audience engagement.

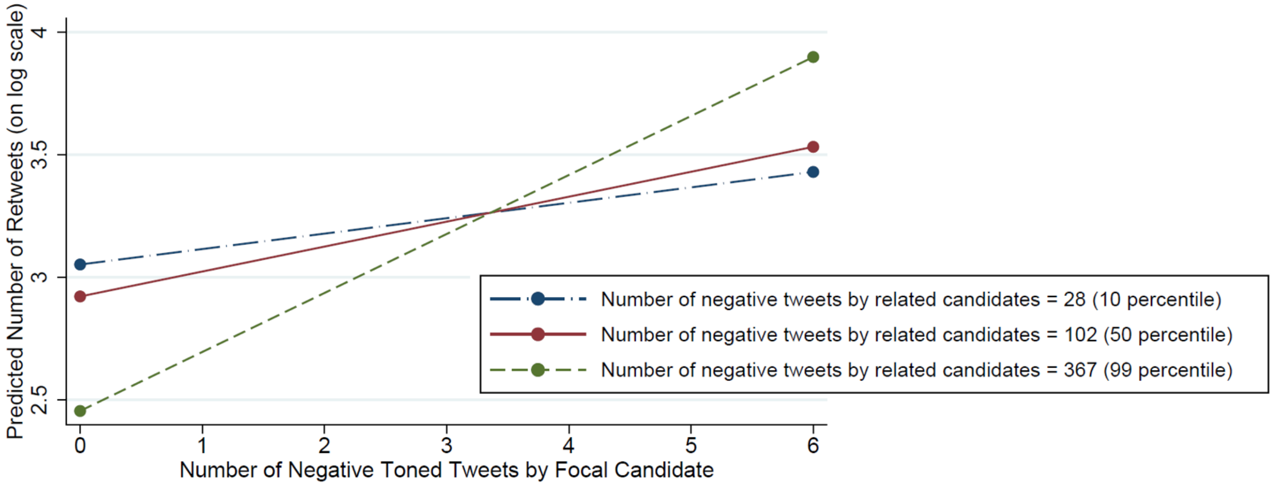
and $PartyPosT$, which hints at the decreasing effect of negative HGC on engagement with increased number of positive HGC by related brands (slopes are higher for lower number of $PartyPosT$). Figure 4.5 illustrates the interaction effect of $NegT$ and $PartyNegT$, which depicts that the percent change in engagement corresponding to higher number of negative HGC by related brands is higher than that corresponding to lower number of negative HGC by related brands (i.e., the slopes are higher for higher number of $PartyNeg$).

Figure 4.4: Moderating Effect of Positive Toned Tweets by Related Candidates (All other variables are at mean)



We find that the relationship between the focal candidate's positive toned HGC and the related candidate's positive (and negative) toned HGC is insignificant. However, the marginal analysis that we conduct demonstrates that the effect of $PosT$ varies at different levels of the tone of related brands (i.e., $PartyPosT$ and $PartyNegT$). Figure 4.6 illustrates the marginal effect of $PosT$ at different values of $PartyPosT$. We find that the effect of $PosT$ becomes positive and significant when $PartyPosT$ is around 250 (a little below mean). Figure 4.7 shows that the marginal effect of $PosT$ is positively significant when $PartyNegT$

Figure 4.5: Moderating Effect of Negative Toned Tweets by Related Candidates (All other variables are at mean)



is less than or equal to about 175 (which refers to around 69 percentile). The findings related to spillover effects are summarized in Table 4.5. Finally, the coefficients of the main effects of related brands' tone on focal brand's engagement. Our findings suggest that the coefficient of number of positive toned HGC by related brands (i.e., $PartyPosT$) is positive and significant, while the coefficient of number of negative toned HGC by related brands (i.e., $PartyNegT$) is negative and significant. For brevity, we do not discuss the results related to the control variables.

Table 4.5: Summary of Spillover Effects

	Spillover Effects	
	Number of Positive Toned HGC by Related Brands	*Number of Negative Toned HGC by Related Brands*
Number of Positive toned HGC by Focal Brand	Positive Spillover Effect[a]	Positive Spillover Effect[b]
Number of Negative toned HGC by Focal Brand	Significantly Negative Spillover Effect	Significantly Positive Spillover Effect

[a] Partially supported when *PartyPosT* is high (i.e., *PartyPosT* is above mean)
[b] Partially supported when *PartyNegT* is low (i.e., *PartyNegT* is below mean)

Figure 4.6: Conditional Marginal Effects of Positive Tone on Engagement at Different Levels of Number of Positive Toned Tweets of Related Candidates

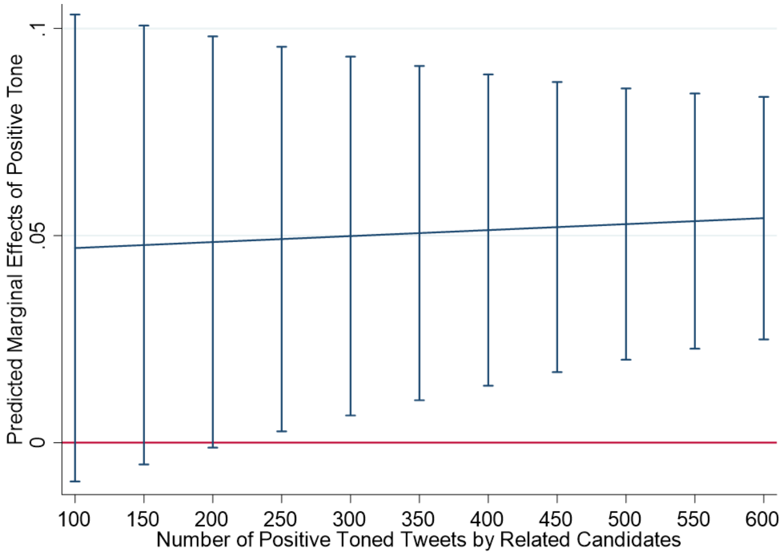
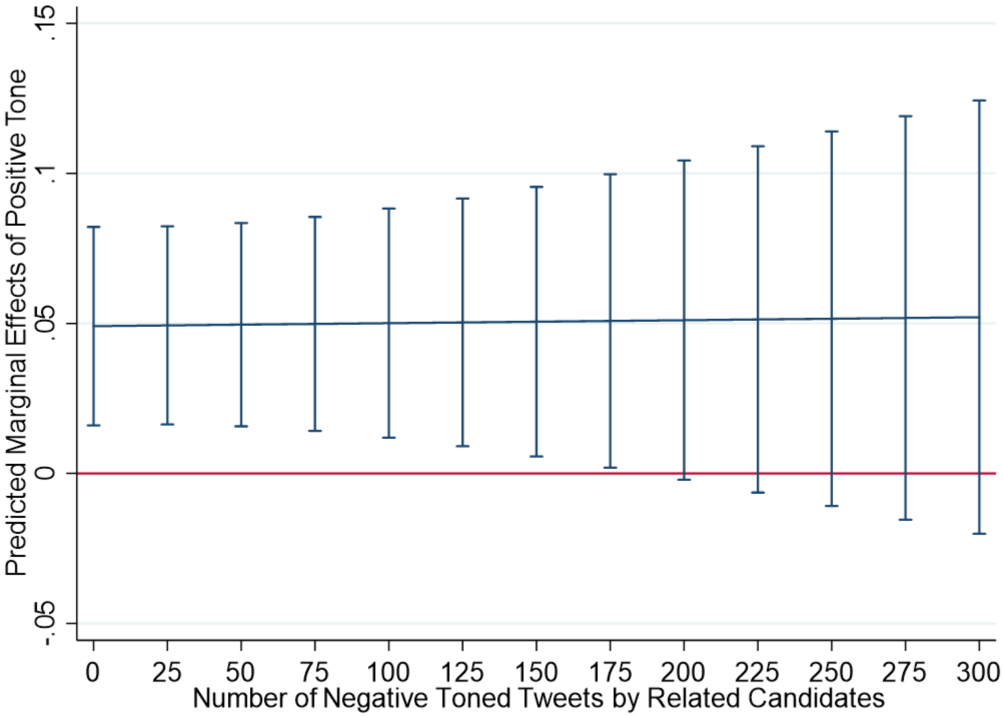


Figure 4.7: Conditional Marginal Effects of Positive Tone on Engagement at Different Levels of Number of Negative Toned Tweets of Related Candidates



4.5 Impact of Engagement on Social Media Visibility

Whereas earlier studies have demonstrated that higher social media engagement would lead to higher sales, in the context of engagement of human brands, quantifying the impact of engagement is challenging. This is because the value generated by audience engagement of human brands is not often observable. We had relied on the number of retweets as a measure of engagement. One would expect that higher engagement levels would lead to more discourse on the candidate's views. In this section, we assess the impact of HGC on "buzz" about the candidates. This buzz on social media platforms implies higher visibility of the candidate, which is one of the primary aims of human brand's social media strategy. In the context of Twitter, we can measure visibility of the candidate based on how many times that candidate was mentioned on social media platforms (Cha et al. 2010). Cha et al. (2010) argue that the number of mentions denotes response of audience to the candidate's HGC.

On Twitter, number of mentions represents the number of times the candidate was tagged or cited by other users in their tweets. In particular, it represents the interest that Twitter users have on the candidate. From our data, we count the number of times the candidate was tagged by other users. Using this information, we empirically test the argument that the higher engagement is associated with greater amount of chatter about the candidate by the audience on social media. To do this, we run the following model using fixed effects estimation strategy to control for individual level heterogeneity.

$$\begin{aligned}\log(Mentions_{it}) &= \beta_0 + \beta_1 \log(RT_{it}) + \beta_2 Tweets_{it} + \beta_3 fc_{it} + \beta_4 PartyTweets_{it} \\ &+ \beta_5 \log(PartyMC_{it-1}) + \beta_6 \log(PartyRT_{it-1}) + u_i + v_t + \eta_{it} \quad (4.9)\end{aligned}$$

In the above model, the dependent variable is the buzz around the focal candidate i. The main variable of interest is number of retweets (denoted by RT). Further, we control for candidate i's total number of tweets (i.e., $Tweets_{it} = PosT_{it} + NegT_{it} + MixT_{it}$), popularity of the candidate (fc_{it}), total number of tweets by related candidates (i.e., $PartyTweets_{it}$), number of mentions of related candidates in the previous period (i.e., $PartyMC_{it-1}$), and the engagement levels of related brands in the previous period (i.e., $PartyRT_{it-1}$). Table 4.6 reports the estimation results. We find that the coefficient of number of retweets is positive and significant. This validates our argument that higher engagement is indeed positively associated with chatter about the candidate on social media, which ratifies the importance of social media engagement.

Table 4.6: Parameter Estimates: Model of Social Media Visibility

DV	Model 7 ***log(Mentions$_{ipt}$)*** (Robust SE)
log*(RT)*	0.505***
	(0.0283)
Total Number of Tweets	0.00338
	(0.00224)
fol	0.233***
	(0.0563)
Party Tweets	0.000133*
	(0.0000601)
partyMentions(lag)	0.0339
	(0.0779)
partyRetweets(lag)	-0.0346
	(0.0805)
Intercept	-0.472
	(0.564)
Time Controls	[Yes]
N	1313
R-squared	0.8347

Robust standard errors clustered at party level are in parentheses.
The coefficients are estimated using fixed effects model.
* p<0.10, ** p<0.05,*** p<0.01

4.6 Supplementary Analysis

Although the estimation strategy that we discussed in Section 4.3.5 addresses the potential bias from unobserved confounding factors and simultaneity, we conduct a series of robustness tests using alternative estimation methods, specifications, and samples to rule out alternate explanations. We find that our key results are substantively robust to these alternative specifications. In the following subsections, we provide discussion on the supplementary analysis used to further validate our empirical results.

4.6.1 Alternative Model Formulation: Conditional Fixed Effects Negative Binomial Model

Although count data models (e.g., negative binomial models) may be considered more suitable for our data, He et al. (2017) argue that linear models are more robust to distribution misspecification. Hence, in our main analysis, we use a log-transformed linear model. Nevertheless, we supplement our results from the linear model with count data models to ensure consistency of results. Most commonly used count data models in the literature are Poisson model and negative binomial model. Literature suggests that negative binomial models compared to Poisson model provide un-biased estimates for over dispersed data, thus we use this model (He et al. 2017). Further to control any unobserved confounding factors due to time-invariant candidate level heterogeneity, we use the conditional fixed effects negative binomial (FE-NB) model for panel data as suggested by Hausman et al. (1984). Results of FE-NB model are presented in column Model R1 in Table C.2 (in Appendix C). These results suggest that our findings from the main model are robust to different estimation strategies, providing further empirical support to our results.

4.6.2 Robustness Check: Fixed and Random Effects

Next, we compare the estimates of our proposed estimation procedure for our model in Equation 4.8 to the conventional estimation procedures (fixed effects and random effects). As discussed earlier, the past literature suggests that the mixed effects model provides efficient estimates. Although it may be argued that fixed effects models can control for all unobserved candidate level heterogeneity, the fixed effects models are often considered inefficient because of the issue of over controls. Our model in Equation 4.8 employs various controls to account for any candidate level heterogeneity while accounting for party level heterogeneity, and hence we believe that our main estimated strategy, i.e., mixed effects model, provides the most efficient standard errors. However, we provide estimates and their standard errors from fixed effects model. If the estimates of the conventional fixed effects model are found to be consistent (even with not so efficient estimators), it will suggest that our qualitative results are robust to inefficient estimator. Further, we also compare the estimates from random effects model wherein we do not account for party level randomness. We present the results of both the conventional models (i.e., random and fixed effects) in Table C.2 in Appendix C (columns Model R2 and Model R3). The parameter estimates and their standard errors suggest that all the results are consistent.

4.6.3 Robustness with Alternate Sample

Next, we re-estimate our full model with an alternative sample. As the focus of our study is to understand the drivers of engagement based on human brand's social media activity, we had previously employed twenty-six tweets per candidate (i.e., on average at least one tweet per week) as a cut-off for our main model. In order to examine the consistency of our results, we re-estimated our model by including candidates with lower levels of activity on Twitter. Specifically, we lowered the cut-off for number of tweets to six tweets for the six-month period instead of twenty-six tweets (i.e., on average at least one tweet per month). To demonstrate the robustness of our results, we re-estimated all the models discussed in the previous sections. In particular, we estimated our models using four different estimation strategies - the mixed effects model, FE-NB model, random effects model, and fixed effects model to confirm robustness. The results, shown in Table C.3 (in Appendix C), are consistent with the key findings regardless of different levels of activity on social media platforms.

4.6.4 Accounting for Self-Selection Bias

Next, it is likely that human brands who are more likely to receive engagement from the online audience are expected to tweet. To test the impact of the selection of human brands in our sample, we estimated the model in Equation 4.8 using Heckman two-step selection model. In the first step, we develop a mixed effects Probit model (to account for party level heterogeneity) that models human brand's decision to either tweet or not tweet in a particular week.[9] The Heckman (1977) selection correction term is obtained from the estimates from the first step. In the second step, the selection term is then included in the main model presented in Equation 4.8. The results of the second step are presented in Table C.4 (in Appendix C). The results suggest that the findings are consistent with the results of the main model. This further validates our empirical results. Furthermore, these results indicate that the selection term is significant.

4.6.5 Alternative Measure of Engagement

In the next section, in line with the literature, we argue (and provide empirical evidence) that the number of mentions is in fact a measure for visibility and number of retweets may affect the visibility of the candidate on Twitter. However, one could argue that the number of mentions in addition to number of retweets may be seen as a combined measure for engagement. To alleviate this concern, we re-run our main model (as proposed in Equation 4.8) using an alternative construct for engagement. In particular, the new dependent variable is the sum number of retweets and number of mentions for each candidate i in week t. The results are presented in Table C.5 (in Appendix C). The results are qualitatively the same compared to our proposed measure for engagement, i.e., number of retweets.

4.7 Discussion

With human brands favoring social media over other traditional media modes to communicate and engage with audience, it is critical to analyze how the content generated by human brands and related brands affects audience engagement. Although recent literature has analyzed the relationship between the content generated by the users/firms and brand engagement on social media, there is a lack of research in IS that explicitly and systematically studies how human brands, who are widely followed on social media, generate engagement on social media. Prior studies have argued that the interaction between audience and human brands is of fandom and worship, which is different from how audience associate with corpo-

[9] Unlike the previous cases, we do not drop inactive candidates as we account for selection bias.

rate brands (Saboo et al. 2016). Hence, we argue that the effect of HGC is not the same as the effect of FGC. Our research provides insights into the impact of various tones of content generated by human brands and their related brands on engagement. Using a unique dataset obtained from Twitter, we examine the effects of the tone of human brand's social media content on audience engagement levels measured through the number of retweets. Moreover, as industry experts vouch for firms to emulate human brands' social media strategy, our results could provide insights not only to human brands but also to non-human brands (i.e., firms) on how to design the social media content to generate audience engagement. In the rest of this section, we first summarize our key findings and the intuition behind these findings, and then present theoretical contributions of our study. Finally, we discuss the insights that human brands, with respect to their social media strategy, can glean from our study.

4.7.1 Summary of Findings

Our empirical findings indicate that the negative tone of social media content generated by human brands positively impacts engagement. We find that the popularity negatively moderates the effect of negative tone on engagement, i.e., the impact of negative tone reduces with the popularity of human brands. In contrast, the positive tone of social media content has a positive effect on engagement only in case of highly popular human brands. These findings suggest that audience do have different expectations in terms of ethical behavior of their leaders. The audience expect highly popular leaders (e.g., Elon Musk, co-founder and CEO of Tesla) to create more positive toned content rather than negative toned content.

With respect to the spillover effects, we find that the HGC tones of related human brands affect how the content generated by human brands impacts social media engagement. In particular, we find that the engagement levels are higher when the tones of the focal brand and the related brands are negative. This indicates that frequent exposure to negative toned HGC increases the likelihood of audience engaging with the audience. One plausible explanation for this result is that when the tone of both the human brand and the related brands is negative, the audience may find the content of all the brands belonging to the same group intriguing and therefore engage with all of them at the same time. The principle of social proof could provide an alternative reasoning behind high engagement when both the focal and related brands' HGC tones are negative. Cialdini (2009) argues that audience tends to determine if an information is correct by examining what the other audience perceives to be correct. In particular, audience looks for consistency of the information and as the audience see more news about a failed policy (i.e., negative toned HGC) they may choose to engage with the content. For example, it is likely that all the candidates from the same political party (i.e., focal candidate and all the related candidates) post content about a similar topic in that particular week. For instance, they could be talking about a failed policy of an opposition candidate or political party, and hence the tone of both the focal brand and the related brands is negative. In this scenario, the audience perceives the negative toned HGC of related brands as an endorsement to the negative toned HGC of focal human brand and thus retweets the post. Further, our findings indicate partial support for positive spillover effect of positive toned HGC of related brands when the tone of human brand is also positive.

In contrary, our findings suggest that the effect of negative tone of focal candidate is lower when the tone of related candidate is positive. In effect, the positive impact of number of negative tone HGC on audience engagement is weakened with higher number of positive tone related candidates' HGC. Finally, we extend our analysis to empirically analyze the significance of social media engagement from a human brand's perspective. We find that the engagement is positively associated with visibility on social media platforms. This result indicates that higher engagement is indeed an important metric for human brands to evaluate their social media strategy. To summarize the findings of our work (i) the tone of HGC affects engagement; (ii) the effect of HGC on engagement is different for more popular human brands as compared to less popular human brands; (iii) tone of related brands' HGC affects social media engagement; and (iv) higher engagement is associated to increased visibility on social media. Overall, these findings provide interesting and important takeaways in addition to making a strong case for human brands (or their managers) to have a robust social media strategy.

4.7.2 Implications

Our findings have several important implications for the academic literature and practice. Our work presents, to the best of our knowledge, a first attempt to examine the engagement of human brands, who are the most widely followed on social media, while the majority of prior literature has mainly focused on the content of firms or non-celebrity users (e.g., Kumar et al. 2016). The content of human brands is usually more polarizing in terms of their tone. Hence, the effects of human brand's HGC would vary from the effects of firm's social media content. Besides, even if the results from the prior work on user generated content and firm generated were applicable to our context, there is no evidence for clear consensus on how the tone of content affects the engagement on social media. Thus, from theoretical perspective, our study takes the lead in understanding the consequences of the content created by human brands.

Our study is one of the first attempts to systematically document the moderating role of social media popularity on the effect of HGC, whereas Berger et al. (2010) have documented the impact of brand popularity on the relationship between negative publicity and product sales. The findings of our study indicate that the popularity of the human brand plays a vital role in moderating the effect of HGC tone on engagement. This implies that failing to account for popularity will likely result in overestimation of the effects of tone on engagement. These novel findings on how popularity affects social media engagement contribute to the growing literature on social media.

The results from the prior work demonstrate the presence of the spillover effects of content generated by other firms on engagement. Borah and Tellis (2016), for example, demonstrate the presence of the spillover of perception based on the content generated by online audience. However, this result does not answer our question, that is, perception spillover of related brands' HGC tone on the audience engagement levels of a focal human brand. We establish the presence of spillover and the direction of spillover of HGC on social media engagement. If this moderating role is ignored, the related brands' HGC could have an adverse effect on engagement. Our work contributes to the limited yet growing literature examining the spillover effects of related brands on focal brand.

Unlike most of the earlier studies that have operationalized engagement in terms of sales or other similar constructs, we focus a behavioral construct of engagement. Specifically, we operationalize engagement via number of retweets. Further, we also empirically demonstrate the significance of social media engagement. Our findings indicate that the engagement on social media is associated with increased online visibility of the content creator. Our findings demonstrate that the engagement can be seen as an important element for human brands who are interested in increasing the visibility of their brand. This unique perspective will enrich the stream of research examining the effects of the tone of social content.

Our work also has several significant managerial implications for the social media managers of human brands who seek to increase their engagement with the audience. Although we have shown the impact of HGC content on engagement in the context of politics and Twitter, the implications of this study are potentially relevant to other contexts (e.g., top management executives of corporations) and other platforms (such as Facebook and LinkedIn). Given that the human brands are increasingly using social media to engage with the online community (e.g., CEOs), the question of how to generate social media engagement is of importance. Furthermore, in certain industries, social media engagement of human brands can have direct economic implications. For example, Saboo et al. (2016) demonstrate that the social media activities of musicians have a positive impact on product sales. We discuss the managerial implications of our study below.

Established brands should actively use positive tone, while less popular brands should employ negative tone. An important insight that we glean from our study is that less popular human brands should focus on creating negative toned content to generate higher engagement and greater visibility on social media platforms. We believe that our findings can be extended to the context of traditional brands (or firms). In particular, our findings suggest that less popular firms (such as small businesses and startups) could increase their audience engagement (and thus visibility) by means of more number of negative toned content. It is often argued that smaller brands need to create their social media content so that they can stand out from the crowd. Our results indicate that one way of standing out is by creating negative toned content, which could give the visibility needed for a smaller size firm. On the other hand, in case of bigger firms (or more popular firms), offensive tone on social media could backfire and will lower the audience engagement. Hence, our prescription for larger firms is to create more number of positive toned content.

Have a comprehensive social media strategy. Our findings on presence of spillover effects help marketing managers of a human brand understand the impact of tone of both the human brands and its related brands on engagement, and thereby providing novel insights for better decision making of managing the content of human brands. Given that we find evidence for asymmetric spillover effects for different tones of HGC, the marketing managers of human brands need to coordinate their social media strategy with each other so that the tone of one human brand does not hurt the engagement of other human brands. Further, given the similarities of human brands and traditional brands, we speculate our findings may be applied to traditional brands as well. For example, the negative advertising strategy of one nameplate (e.g., Toyota Corolla) might increase its engagement levels. However, this could have negative impact on its related brands (e.g., Toyota Camry, Rav4, and Highlander). Thus, as a first step, brands should have a comprehensive social media strategy across different nameplates.

In summary, our research highlights the importance of social media tone, related brands' social media tone, and popularity on online engagement and visibility. Unlike traditional marketing strategies, social media allows human brands to implement its marketing plan instantaneously. Hence, an effective social media strategy from the perspective of a focal brand entails integrating the social media behavior of the related brands and its popularity.

By integrating these external aspects into one's strategy, he/she can adjust his/her strategy to retain or increase audience engagement and visibility on social media platforms.

4.8 Conclusions and Future Research Directions

In this paper, we develop a multi-level hierarchical model to examine the impacts of tone of a celebrity's social media post on the resulting engagement with the audience. In order to achieve our goals, we assemble a novel data comprising of candidate-level Twitter data collected during the Indian elections 2014. Our empirical findings provide guidance to human brands on an all-inclusive social media strategy while contributing to the literature. Although the findings of our study provide several interesting and important insights on the impact of HGC, we acknowledge the fact that our work has a few limitations. First, we recognize the fact that our study is limited to examining the impact on only one particular social media platform. Future research can explore and re-validate our findings on other platforms such as Facebook, Instagram, and Snapchat. Second, we restricted our focus to political candidates. Future studies can investigate the generalizability of our results with human brands from different context. Although there are several similarities among political candidates across the world, it would be interesting to check if our insights hold for candidates in other countries as well to account for difference in cultures. Whereas our research focuses only on one aspect of HGC content, i.e., tone, we hope that future research can build upon our study to examine the impact of other aspects (such as including emotion and humor) of HGC content. Finally, we use number of retweets as measure of engagement. It would be worthwhile to employ other measures of engagement and revisit this issue. Future research can build upon our study to explore other issues related to the content generated by human brands in social media.

5. SUMMARY AND CONCLUSIONS

The ubiquity of digital social networking platforms has enabled organizations to access massive social networks and directly interact with the users in these networks, thus transforming how organizations operate their businesses. For example, social networking platforms such as GitHub facilitate firms to collaborate with external parties including users of the products, competitors, and partners to build new products, test existing products, and manage open source projects. Likewise, platforms such as Twitter, LinkedIn, and Facebook have revolutionized the way firms market their products and interact with the customers. Not surprisingly, social media platforms have become an integral part of everyday business operations.

My research investigates the impact of online social networks on firms from an operational perspective. In particular, my dissertation research demonstrates how firms can effectively utilize online social networks to manage technology (Chapter 2), optimize their marketing operations (Chapter 3), and attract customers to their products and services (Chapter 4). Our findings suggest that the decision to either engage or not with the social networks is not trivial. For example, we demonstrate that seeding through the most popular influencer, which might seem to be an obvious strategy for firms, is not always an optimal strategy. Likewise, our results suggest that firms might choose not to engage with the open source community, in certain scenarios, even though it might seem to be a good strategy. Hence, firms need to carefully evaluate the necessary trade-offs associated with the social networks for improved decision-making. To accomplish my research goals, I take a multi-methodological approach, which includes using optimal control theory, combinatorial optimization, and empirical models. This multi-methodological approach allows me to conduct research and provide solutions at both strategic and operational levels.

While my current research sheds light on the impact of social networks on technology management and marketing operations, I foresee several new research directions studying the impact of social networks on various other aspects of a business such as product recalls, reverse logistics, product returns, demand forecasting, and policy implementation. Furthermore, whereas the literature is quite rich in the domain of advertising on internet and mobile platforms, research on optimizing social media advertising is limited. In this regard, I plan to build on my current work in the domain of social media marketing. For example, the problem of ad scheduling across multiple platforms is an apparent extension for the scheduling model that I developed in Chapter 3.

My dissertation makes several important contributions to both theory and practice. From theoretical perspective, the study presented in the first essay is among the first to analyze the effects of open source community on managing technology, and more importantly determine the optimal extent of openness, i.e., the extent of collaboration between the firm and open source community. Furthermore, the theoretical contributions pertaining to the second and third essays lie in providing a data-driven framework for conducting an influencer marketing campaign on online social networks and understanding what drives audience to engage with influencers. In addition, the results of the three essays presented in this dissertation provide several actionable insights for firms contemplating on using social networks. Below, I conclude by briefly summarizing the three essays of my dissertation.

In the first essay, we model the impacts of making a portion of a firm's technology open source on the quality and maintenance effort of the current version of the technology, and on the quality and development effort of the next version. The first study aims to help managers evaluate the impact of various characteristics of the market, firm, and technology on decisions related to optimal extent of openness and firm's effort levels. This study complements the existing operations literature that has studied the issue of resource allocation (i.e., effort levels of maintaining existing technologies and developing new technologies) and innovation separately, and has ignored the impact of technology openness on the behavior of the firm. More importantly, ours is the first study, to the best of our knowledge, to model partial openness - an increasingly popular business strategy.

The results of the first study provide several interesting and important managerial insights. For example, we find that the firm might open its technology even in conditions that do not seem to favor an open source environment. In particular, when the loss to market competition (a negative consequence of open technology) is high, our results suggest that the firm should in fact fully open its technology, which is counter-intuitive. We also determine the conditions when the technology should be kept proprietary, and find that the firm should keep its technology fully closed despite high demand sensitivity of openness. In addition, we find that the demand and quality sensitivities of the current and next versions of the technology on openness influence the firm in deciding whether to make the technology open or not.

In the second essay, we model the problem of conducting an influencer marketing campaign. In particular, we develop a solution framework for the problem of selecting and scheduling influencers during the planning horizon for marketing campaigns. This framework is based on the interactions with industry, and empirical analysis performed on data from Twitter. We first solve the problem of selection of influencers and then extend our analysis to scheduling the placements of ads by influencers. With regard to the selection model, we present the mathematical formulation to the problem (i.e., the main model) of choosing influencers. We show that the problem of selection of influencers is NP-complete. Thus to solve the problem in realistic time, we propose an alternative formulation that provides a near-optimal solution for the main model. Besides, we demonstrate that firms can experience a significant profit loss when implementing a policy that is computed based on a model (namely single level model) that ignores the peer effect. Results from two case studies that we conduct show that our model based solution consistently outperforms current industry practices. Next, we propose an optimization model for scheduling the ads to be posted by the influencers (selected in the first phase) during the planning horizon of a marketing campaign. We provide several managerially relevant insights based on extensive numerical experiments. For example, we find that the cost of influencer marketing campaign increases non-linearly with engagement. Our study is among the first to propose a scheduling model in the context of influencer marketing and to consider peer effect in the selection model.

In the third essay paper, we develop a multi-level hierarchical model to empirically investigate the effects of influencer's social media tone on the audience engagement. To do this, we assemble a novel data from Twitter collected during the Indian elections 2014. Our empirical findings provides several important insights to human brands while contributing to the growing literature on social media. From theoretical perspective, this study systematically examines the moderating role of popularity and the tone of related brands on the relationship between social media tone of an influencer and engagement. Our empirical findings suggest that, to increase audience engagement, influencers should vary the tone of their social media content based on their popularity levels. Finally, our findings on spillover effects demonstrate the importance of a comprehensive social media strategy across brands.

REFERENCES

Agrawal, R. (2017). *Why these Indian state elections matter to the whole world.* Available at http://www.cnn.com/2017/03/11/asia/uttar-pradesh-elections/index.html (last accessed: May 06, 2017).

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Aoyama, M. (2002). Metrics and analysis of software architecture evolution with discontinuity. In *Proceedings of the International Workshop on Principles of Software Evolution, Orlando*, pages 103–107.

Apple (2015). *Open at the source.* Available at http://www.apple.com/opensource/ (last accessed: September 24, 2018).

Aral, S., Muchnik, L., and Sundararajan, A. (2013). Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science*, 1(02):125–153.

Aronson, E. (2001). Integrating leadership styles and ethical perspectives. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*, 18(4):244–256.

Arora, A., Krishnan, R., Telang, R., and Yang, Y. (2010). An empirical analysis of software vendors' patch release behavior: Impact of vulnerability disclosure. *Information Systems Research*, 21(1):115–132.

Badenhausen, J. (2017). *Cristiano Ronaldo Generated $500 Million in Value for Nike in 2016.* Available at https://www.forbes.com/sites/kurtbadenhausen/2017/02/16/cristiano-ronaldo-generated-500-million-in-value-for-nike-in-2016/#13305c3ac3e9 (last accessed: January 31, 2018).

Banker, R. D., Chang, H., and Kemerer, C. F. (1994). Evidence on economies of scale in software development. *Information and Software Technology*, 36(5):275–282.

Banker, R. D. and Slaughter, S. A. (1997). A field study of scale economies in software maintenance. *Management Science*, 43(12):1709–1725.

Barker, S. (2017). *How Much Should You Pay Social Media Influencers?* Available at https://shanebarker.com/blog/influencer-marketing-celebrity-endorsements/ (last accessed: September 12, 2018).

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

Bazerman, M. H. and Moore, D. A. (2008). Judgment in managerial decision making.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Berger, J. and Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *Journal of Marketing Research*, 48(5):869–880.

Berger, J., Sorensen, A. T., and Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, 29(5):815–827.

Bessen, J. (2014). *History backs up Teslas patent sharing.* Available at https://hbr.org/2014/06/history-backs-up-teslas-patent-sharing (last accessed: September 24, 2018).

Bianchi, A., Caivano, D., Lanubile, F., and Visaggio, G. (2001). Evaluating software degradation through entropy. In *Proceedings of Seventh International Software Metrics Symposium, London*, pages 210–219.

Bode, L. and Dalrymple, K. E. (2015). Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing*, pages 1–22.

Boh, W. F., Slaughter, S. A., and Espinosa, J. A. (2007). Learning from experience in software development: A multilevel analysis. *Management Science*, 53(8):1315–1331.

Bollapragada, S., Bussieck, M. R., and Mallik, S. (2004). Scheduling commercial videotapes in broadcast television. *Operations Research*, 52(5):679–689.

Bonaccorsi, A. and Rossi, C. (2003). Why open source software can succeed. *Research Policy*, 32(7):1243–1258.

Borah, A. and Tellis, G. J. (2016). Halo (spillover) effects in social media: do product recalls of one brand hurt or help rival brands? *Journal of Marketing Research*, 53(2):143–160.

Bradley, S., Hax, A., and Magnanti, T. (1977). Applied mathematical programming.

BrandFrog.com (2016). *CEOs, social media, and brand reputation*. Available at http://brandfog.com/BRANDfog2016CEOSocialMediaSurvey.pdf (last accessed: May 06, 2018).

Businesswire (2016). *The seperation of influence: A view of influence from influencers and influence marketers*. Available at https://mms.businesswire.com/media/20160726005961/en/536548/5/AltimeterInfographicFinal.jpg?download=1 (last accessed: September 12, 2018).

Cambridge University (2013). *Cambridge University study states software bugs cost economy* $312 *billion a year*. Available at http://www.prweb.com/releases/2013/1/prweb10298185.htm (last accessed: September 24, 2018).

Cameron, A. C. and Miller, D. L. (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

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

Camm, J. D., Norman, S. K., Polasky, S., and Solow, A. R. (2002). Nature reserve site selection to maximize expected species covered. *Operations Research*, 50(6):946–955.

Carr, D. (2008). How Obama tapped into social networks' power. *New York Times*, 9.

Casadesus-Masanell, R. and Ghemawat, P. (2006). Dynamic mixed duopoly: A model motivated by Linux vs. Windows. *Management Science*, 52(7):1072–1084.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30.

Chan, T. C., Demirtas, D., and Kwon, R. H. (2016). Optimizing the deployment of public access defibrillators. *Management Science*, 62(12):3617–3635.

Cialdini, R. B. (1984). *Influence: How and why people agree to things*. Quill New York.

Cialdini, R. B. (2009). *Influence: Science and practice*, volume 4. Pearson Education Boston, MA.

Clemons, E. K. and Row, M. C. (1992). Information technology and industrial cooperation: The changing economics of coordination and ownership. *Journal of Management Information Systems*, 9(2):9–28.

Dahlander, L. and Magnusson, M. G. (2005). Relationships between open source software companies and communities: Observations from nordic firms. *Research Policy*, 34(4):481–493.

Das, M. (2018). *From Dombivli to Zambia, desi social media influencers make likes count*. Available at https://timesofindia.indiatimes.com/india/from-dombivli-to-zambia-desi-social-media-influencers-make-likes-count/articleshow/64434668.cms (last accessed: September 12, 2018).

Das, S. R. and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66. ACM.

Duggan, M. and Smith, A. (2016). *The political environment on social media*. Available at http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/ (last accessed: January 06, 2017).

Duran, H. B. (2017). *10 Essential stats for influencer marketing in 2017*. Available at https://www.ion.co/essential-stats-for-influencer-marketing-in-2017 (last accessed: September 12, 2018).

Durlauf, S. N. and Young, H. P. (2004). *Social Dynamics*. MIT Press.

eMarketer (2018). *eMarketer updates worldwide social network user figures*. Available at https://www.emarketer.com/Article/eMarketer-Updates-Worldwide-Social-Network-User-Figures/1016178 (last accessed: September 12, 2018).

Fang, X., Hu, P, J., Li, Z., and Tsai, W. (2013). Predicting adoption probabilities in social networks. *Information Systems Research*, 24(1):128–145.

Fenton, N. E. and Neil, M. (1999). A critique of software defect prediction models. *IEEE Transactions on Software Engineering*, 25(5):675–689.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, 38(6):889.

Fitzgerald, B. (2006). The transformation of open source software. *MIS Quarterly*, 30(3):587–598.

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability: A guide to NP-completeness*. WH Freeman New York.

Garvin, D. A. (1984). What does product quality really mean? *Sloan Management Review*, 26(1).

Ghose, A. and Han, S. P. (2011). An empirical analysis of user content generation and usage behavior on the mobile internet. *Management Science*, 57(9):1671–1691.

Gnu.org (2012). *NASA open source agreement*. Available at http://www.gnu.org/licenses/license-list.html#NASA (last accessed: September 24, 2018).

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M., and Verlegh, P. (2005). The firm's management of social interactions. *Marketing letters*, 16(3-4):415–428.

Goh, K., Heng, C., and 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.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. S. (2010). Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 241–250. ACM.

Gu, B., Konana, P., Raghunathan, R., and Chen, H. M. (2014). The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, 25(3):604–617.

Hamilton, D. L. and Zanna, M. P. (1972). Differential weighting of favorable and unfavorable attributes in impressions of personality. *Journal of Experimental Research in Personality*, 6(2–3):204–212.

Harkins, S. G. and Petty, R. E. (1981). The multiple source effect in persuasion: The effects of distraction. *Personality and Social Psychology Bulletin*, 7(4):627–635.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., and Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. *Marketing letters*, 19(3-4):287–304.

Hausman, J. A., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship.

He, S., Rui, H., and Whinston, A. B. (2017). Social media strategies in product-harm crises. *Information Systems Research*.

Heald, E. (2017). *How much should you pay social media influencers?* Available at https://sproutsocial.com/insights/paying-social-media-influencers (last accessed: September 12, 2018).

Heales, J. (2002). A model of factors affecting an information system's change in state. *Journal of Software Maintenance and Evolution: Research and Practice*, 14(6):409–427.

Heath, C., Bell, C., and Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology*, 81(6):1028.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Heflin, D. T. and Haygood, R. C. (1985). Effects of scheduling on retention of advertising messages. *Journal of Advertising*, 14(2):41–64.

Herr, P. M., Kardes, F. R., and Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of Consumer Research*, pages 454–462.

Hoffman, C. (2015). *Don't hold your breath: Why Windows won't be open-source any time soon*. Available at http://www.pcworld.com/article/2907278/dont-hold-your-breath-why-windows-wont-be-open-source-any-time-soon.html (last accessed: September 24, 2018).

Hu, B., Hu, M., and Yang, Y. (2017). Open or closed? technology sharing, supplier investment, and competition. *Manufacturing & Service Operations Management*, 19(1):132–149.

Janakiraman, R., Sismeiro, C., and Dutta, S. (2009). Perception spillovers across competing brands: A disaggregate model of how and when. *Journal of Marketing Research*, 46(4):467–481.

Ji, Y., Kumar, S., Mookerjee, V. S., Sethi, S. P., and Yeh, D. (2011). Optimal enhancement and lifetime of software systems: A control theoretic analysis. *Production and Operations Management*, 20(6):889–904.

Ji, Y., Mookerjee, V. S., and Sethi, S. P. (2005). Optimal software development: A control theoretic approach. *Information Systems Research*, 16(3):292–306.

Johnson, S. K. (2009). Do you feel what i feel? Mood contagion and leadership outcomes. *The Leadership Quarterly*, 20(5):814–827.

Karp, K. (2016). *New research: The value of influencers on Twitter*. Available at https://blog.twitter.com/marketing/en_us/a/2016/new-research-the-value-of-influencers-on-twitter.html (last accessed: September 12, 2018).

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

Khullar, A. and Haridasani, A. (2012). *Politicians slug it out in India's first social media election.* Available at http://www.cnn.com/2014/04/09/world/asia/indias-first-social-media-election/ (last accessed: June 25, 2016).

Kitchenham, B. and Pfleeger, S. L. (1996). Software quality: The elusive target. *IEEE Software*, (1):12–21.

Koch, S. (2009). Exploring the effects of sourceforge .Net coordination and communication tools on the efficiency of open source projects using data envelopment analysis. *Empirical Software Engineering*, 14(4):397.

Kreft, I. G. and De Leeuw, J. (1998). *Introducing multilevel modeling.* Sage.

Kulkarni, V. G., Kumar, S., Mookerjee, V. S., and Sethi, S. P. (2009). Optimal allocation of effort to software maintenance: A queuing theory approach. *Production and Operations Management*, 18(5):506–515.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., and Kannan, P. K. (2016). From social to sale: The effects of firm-generated content in social media on customer behavior. *Journal of Marketing*, 80(1):7–25.

Kumar, S., Dawande, M., and Mookerjee, V. (2007). Optimal scheduling and placement of internet banner advertisements. *IEEE Transactions on Knowledge and Data Engineering*, 19(11).

Kumar, S., Jacob, V. S., and Sriskandarajah, C. (2006). Scheduling advertisements on a web page to maximize revenue. *European Journal of Operational Research*, 173(3):1067–1089.

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., and Tillmanns, S. (2010). Undervalued or overvalued customers: Capturing total customer engagement value. *Journal of Service Research*, 13(3):297–310.

Kumar, V., Gordon, B. R., and Srinivasan, K. (2011). Competitive strategy for open source software. *Marketing Science*, 30(6):1066–1078.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM.

Lambrecht, A., Tucker, C., and Wiertz, C. (2018). Advertising to early trend propagators: Evidence from Twitter. *Marketing Science*, 37(2):177–199.

Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising content and consumer engagement on social media: evidence from facebook. *Management Science*.

Lei, J., Dawar, N., and Lemmink, J. (2008). Negative spillover in brand portfolios: Exploring the antecedents of asymmetric effects. *Journal of Marketing*, 72(3):111–123.

Lerner, J. and Tirole, J. (2002). Some simple economics of open source. *The Journal of Industrial Economics*, 50(2):197–234.

Levy, E. (2000). *Wide open source.* Available at http://www.securityfocus.com/news/19 (last accessed: September 24, 2018).

Lewis, K. M. (2000). When leaders display emotion: How followers respond to negative emotional expression of male and female leaders. *Journal of Organizational Behavior*, 21(2):221–234.

Lientz, B. P., Swanson, E. B., and Tompkins, G. E. (1978). Characteristics of application software maintenance. *Communications of the ACM*, 21(6):466–471.

Linquia.com (2017). *New Linqia Survey Uncovers Key Influencer Marketing Trends in 2018.* Available at http://www.linqia.com/about-linqia/newsroom/press-releases/linqia-influencer-marketing-trends-2018/ (last accessed: September 12, 2018).

Linquia.com (2018). *How much should you pay social media influencers?* Available at http://www.linqia.com/wp-content/uploads/2017/12/Linqia-The-State-of-Influencer-Marketing-2018.pdf (last accessed: September 12, 2018).

Long, J. (2015a). *Increase retweets and improve engagement on Twitter with these 12 tips.* Available at https://www.entrepreneur.com/article/242123 (last accessed: September 12, 2018).

Long, J. (2015b). *Increase retweets and improve engagement on Twitter with these 12 tips.*

Available at https://www.entrepreneur.com/article/242123 (last accessed: June 02, 2018).

Maheswaran, D. and Meyers-Levy, J. (1990). The influence of message framing and issue involvement. *Journal of Marketing research*, pages 361–367.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

MarketingHub (2017). *50 Influencer Marketing Statistics, Quotes and Facts.* Available at https://influencermarketinghub.com/influencer-marketing-statistics-quotes-facts/ (last accessed: September 12, 2018).

Martens, C. (2006). *It's official: Sun open sources Java.* Available at https://www.javaworld.com/article/2077658/core-java/it-s-official--sun-open-sources-java.html (last accessed: September 24, 2018).

Matousek, M. (2018). *Some new Tesla cars are being delivered with flaws, and owners say getting them fixed is a painful process.* Available at https://www.businessinsider.com/some-tesla-customers-experience-service-problems-2018-8 (last accessed: September 24, 2018).

Mergent Online (2017). *Microsoft Corporation: Ratios.* Available at http://www.mergentonline.com/companyfinancials.php?pagetype=ratios&compnumber=46247 (last accessed: September 24, 2018).

Microsoft (2014). *Microsoft takes .NET open source and cross-platform, adds new development capabilities with Visual Studio 2015, .NET 2015 and Visual Studio Online.* Available at https://news.microsoft.com/2014/11/12/microsoft-takes-net-open-source-and-cross-platform-adds-new-development-capabilities-with-visual-studio-2015-net-2015-and-visual-studio-online/ (last accessed: September 24, 2018).

Mithas, S., Ramasubbu, N., Krishnan, M. S., and Fornell, C. (2006). Designing web sites for customer loyalty across business domains: A multilevel analysis. *Journal of Management Information Systems*, 23(3):97–127.

Mockus, A., Fielding, R., and Herbsleb, J. (2000). A case study of open source software development: The Apache server. In *Proceedings of the 22nd International Conference on Software Engineering, Limerick, Ireland*, pages 263–272.

Mookerjee, R., Kumar, S., and Mookerjee, V. (2016). Optimizing performance-based internet advertisement campaigns. *Operations Research*, 65(1):38–54.

Moorhead, P. (2017). *Microsoft goes after Google with new education products and services.* Available at https://www.forbes.com/sites/patrickmoorhead/2017/05/03/microsoft-goes-after-google-with-new-education-products-and-services/#3f00c1ef267f (last accessed: September 24, 2018).

Mozilla (2013). *Mozilla audited financials.* Available at https://static.mozilla.com/moco/en-US/pdf/Mozilla_Audited_Financials_2013.pdf (last accessed: September 24, 2018).

Nguyen, H. and Zheng, R. (2013). On budgeted influence maximization in social networks. *IEEE Journal on Selected Areas in Communications*, 31(6):1084–1094.

Park, E., Rishika, R., Janakiraman, R., Houston, M. B., and Yoo, B. (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing*, 82(1):93–114.

Paulson, J. W., Succi, G., and Eberlein, A. (2004). An empirical study of open-source and closed-source software products. *IEEE Transactions on Software Engineering*, 30(4):246–256.

Petrova, M., Sen, A., and Yildirim, P. (2017). Social media and political donations: New technology and incumbency advantage in the united states. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2908221 (last accessed: January 01, 2019).

PWC (2017). *Corporate R &D spending hits record highs for the top 1000, despite concerns of economic protectionism.* Available at https://press.pwc.com/News-releases/corporate-r-d-spending-hits-record-highs-for-the-top-1000–despite-concerns-of-economic-protectionis/s/f9b38af9-7235-4360-a13f-048daa517a64 (last accessed: September 24, 2018).

Raffo, J. (2017). Matchit: Stata module to match two datasets based on similar text patterns.

Rahmandad, H. (2005). Dynamics of platform-based product development. In *Proceedings of the 23rd International Conference of the System Dynamics Society, Boston.*

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Repenning, N. P. (2001). Understanding fire fighting in new product development. *Journal of Product Innovation Management*, 18(5):285–300.

Rime, B., Mesquita, B., Boca, S., and Philippot, P. (1991). Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition & Emotion*, 5(5-6):435–465.

Rishika, R., Kumar, A., Janakiraman, R., and Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: An empirical investigation. *Information Systems Research*, 24(1):108–127.

Roehm, M. L. and Tybout, A. M. (2006). When will a brand scandal spill over, and how should competitors respond? *Journal of Marketing Research*, 43(3):366–373.

Saboo, A. R., Kumar, V., and Ramani, G. (2016). Evaluating the impact of social media activities on human brand sales. *International Journal of Research in Marketing*, 33(3):524–541.

Saydam, C. and McKnew, M. (1985). A separable programming approach to expected coverage: An application to ambulance location. *Decision Sciences*, 16(4):381–398.

Schmidt, S. and Eisend, M. (2015). Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, 44(4):415–428.

Scholtes, I., Mavrodiev, P., and Schweitzer, F. (2016). From Aristotle to Ringelmann: A large-scale analysis of team productivity and coordination in open source software projects. *Empirical Software Engineering*, 21(2):642–683.

Schryen, G. and Kadura, R. (2009). Open source vs. closed source software: Towards measuring security. In *Proceedings of the 2009 ACM Symposium on Applied Computing, Hawaii*, pages 2016–2023.

Seetharaman, D. (2015). *What Celebrities Can Teach Companies*

About Social Media. Available at https://www.wsj.com/articles/what-celebrities-can-teach-companies-about-social-media-1444788220 (last accessed: January 31, 2018).

Seshadri, S., Subramanian, S., and Souyris, S. (2015). Scheduling spots on television. *Working Paper.*

Sethi, S. P. and Thompson, G. L. (2000). *Optimal Control Theory: Applications to Management Science and Economics. Springer.*

Singer, C. (2002). Political branding in both political campaiging and consumer marketing, branding is more art than science. *Brandweek - New York*, 43(34):19–19.

Smith, G. and French, A. (2009). The political brand: A consumer perspective. *Marketing Theory*, 9(2):209–226.

Statista (2018a). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions).* Available at https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (last accessed: September 12, 2018).

Statista (2018b). *Social media marketing spending in the United States from 2014 to 2019.* Available at https://www.statista.com/statistics/276890/social-media-marketing-expenditure-in-the-united-states/ (last accessed: September 12, 2018).

Stelzner, M. A. (2016). *2016 Social media marketing industry report.* Available at http://www.socialmediaexaminer.com/wp-content/uploads/2016/05/SocialMediaMarketingIndustryReport2016.pdf (last accessed: May 06, 2017).

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE.

Sun, Z., Dawande, M., Janakiraman, G., and Mookerjee, V. (2017). Not just a fad: Optimal sequencing in mobile in-app advertising. *Information Systems Research*, 28(3):511–528.

Terwiesch, C. and Xu, Y. (2008). Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9):1529–1543.

Toyota (2015). *Toyota opens the door and invites the industry to the hydrogen future.* Available at https://corporatenews.pressroom.toyota.com/releases/toyota+fuel+cell+patents+ces+2015.htm (last accessed: September 24, 2018).

Tsay, A. A. and Agrawal, N. (2000). Channel dynamics under price and service competition. *Manufacturing & Service Operations Management*, 2(4):372–391.

von Hippel, E. and von Krogh, G. (2003). Open source software and the private-collective innovation model: Issues for organization science. *Organization Science*, 14(2):209–223.

von Krogh, G. and von Hippel, E. (2006). The promise of research on open source software. *Management science*, 52(7):975–983.

Wayner, P. (2000). *Free for all: How Linux and the free software movement undercut the high-tech titans.* Harper Business New York.

West, J. (2003). How open is open enough?: Melding proprietary and open source platform strategies. *Research Policy*, 32(7):1259–1285.

Wired.com (2015). *Google just open sourced Tensorflow, its artificial intelligence engine.* Available at http://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/ (last accessed: September 24, 2018).

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

Wright, P. (1974). The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology*, 59(5):555.

York, A. (2018). *How to identify social media influencers & collaborate on campaigns.* Available at https://sproutsocial.com/insights/social-media-influencers/ (last accessed: September 12, 2018).

APPENDIX A

APPENDIX FOR CHAPTER 2: HOW MUCH TO OPEN, HOW FAST TO FIX AND DEVELOP? – IMPACTS OF TECHNOLOGY OPENNESS

A.1 Expressions for Thresholds

$$\begin{aligned}
\mathscr{E} &\equiv \frac{2c_dc_f\left(T\left(b\theta_1kT+2h\rho_1y+2\theta_2k\right)+2h\rho_2\right)}{c_d\left(4\rho_0c_f+\rho_1Tw\left(2h\rho_1w+\theta_1kmT\right)\right)+2hn^2\rho_1^2Tc_f}\\
\zeta &\equiv \frac{1}{4}T\left(-\frac{4a}{e}+e\rho_1^2\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-4\rho_1y\right)\\
\mathscr{B} &\equiv \frac{2c_f\big(2c_d\left(T\left(2a+\rho_1y(2e-h)-\theta_2k\right)+\rho_2(2e-h)+e\rho_0\right)+en^2\rho_1^2T(h-e)\big)+c_de\rho_1Tw\left(2\rho_1w(h-e)+\theta_1kmT\right)}{2c_fc_d\theta_1kT^2}\\
\mathscr{A} &\equiv \frac{1}{4}e\left(e\rho_1^2\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-\frac{4\rho_2}{T}-4\rho_1y\right)\\
\bar{\mathscr{A}} &\equiv \frac{c_d\big(2c_f\big(b\theta_1kT^2-2(e-h)\left(\rho_2+\rho_1Ty\right)-2e\rho_0+2\theta_2kT\big)+e\rho_1Tw\left(\rho_1w(e-2h)-\theta_1kmT\right)\big)+en^2\rho_1^2Tc_f(e-2h)}{4Tc_dc_f}\\
\beta_0 &\equiv \frac{em\rho_1w}{2c_f}-\frac{2\theta_2}{\theta_1T}\\
\beta_1 &\equiv \frac{c_d\big(4c_f\left(2\theta_1m(t-T)\left(aT+e\rho_2+e\rho_1Ty\right)+e\theta_2\rho_1Tw\right)+e^2\theta_1m\rho_1^2Tw^2(T-2t)\big)+2e^2\theta_1mn^2\rho_1^2Tc_f(T-t)}{2e\theta_1\rho_1T^2wc_dc_f}\\
\alpha &\equiv \frac{e\big(c_d\left(e\rho_1Tw\left(h\rho_1w+\theta_1kmT\right)-2c_f\left(kT\left(b\theta_1T+2\theta_2\right)-2e\rho_0\right)\right)+ehn^2\rho_1^2Tc_f\big)}{4hTc_dc_f}\\
\bar{\mathscr{E}} &\equiv \frac{2c_fc_dy}{\rho_1\left(c_fn^2+c_dw^2\right)}\\
\bar{\mathscr{T}} &\equiv \frac{2c_fc_d\rho_2}{c_fen^2\rho_1^2+c_de\rho_1^2w^2-2c_fc_d\rho_1y}\\
\eta &\equiv \frac{2c_fc_dy}{c_fn^2\rho_1+c_d\rho_1w^2}
\end{aligned}$$

A.2 Proofs and Additional Propositions

Proofs of Lemmas 1 and 3

The Hamiltonian (Sethi and Thompson 2000) of the optimal control problem is given by:

$$\begin{aligned}
H(u,v,s,\lambda_1,\lambda_2,\lambda_3,q,r,x,t) \;=\;& \lambda_1(t)(bs+mu(t))-\lambda_3(t)\left(c_dv(t)^2+c_fu(t)^2\right)+k\left(\theta_0+\theta_1q(t)+\theta_2s\right)\\
&+\lambda_2(t)(nv(t)+sy+wu(t))
\end{aligned} \tag{A.1}$$

Control variables $u(t)$ and $v(t)$ presented in the lemma are derived from solving the following set of equations:

$$\begin{aligned}
\frac{d\lambda_1}{dt} &= -\frac{\partial H}{\partial q} = -k\theta_1, && \text{(A.2)}\\
\lambda_1(T) &= k\theta_1(T-t), && \text{(A.3)}\\
\frac{d\lambda_2}{dt} &= -\frac{\partial H}{\partial r} = 0, && \text{(A.4)}\\
\lambda_2(T) &= \rho_1(h-es), && \text{(A.5)}\\
\lambda_2(T) &= \beta, && \text{(A.6)}\\
\frac{d\lambda_3}{dt} &= -\frac{\partial H}{\partial x} = 0, && \text{(A.7)}\\
\lambda_3(T) &= 1, && \text{(A.8)}\\
\lambda_3(T) &= \gamma, && \text{(A.9)}\\
\frac{\partial H(t)}{\partial u(t)} &= -2c_f u(t)\lambda_3(t) + m\lambda_1(t) + w\lambda_2(t) = 0, && \text{(A.10)}\\
\frac{\partial H(t)}{\partial v(t)} &= n\lambda_2(t) - 2c_d\lambda_3(t)v(t) = 0, && \text{(A.11)}\\
\frac{\partial^2 H(t)}{\partial u(t)^2} &= -2c_f\lambda_3(t), && \text{(A.12)}\\
\frac{\partial^2 H(t)}{\partial v(t)^2} &= -2c_d\lambda_3(t). && \text{(A.13)}
\end{aligned}$$

Note that, for Lemma 1, Equation (A.5) is active, whereas in when customer has specific quality requirements (Lemma 3), i.e., when $r(T) = Z$, Equation (A.6) is active. Next, when the firm has sufficient budget to cover the expenditure for the effort levels, Equation (A.8) is active else Equation (A.9) is active. The presented results are derived by solving the optimal control problem with the corresponding active constraints.

In scenario 1, i.e., when there is no pre-set quality requirement and there is sufficient budget, substituting Equations (A.3), (A.5), and (A.8) into Equation (A.10) and solving for $u(t)$, we get: $u(t) = \frac{\rho_1 w(h-es)+\theta_1 km(T-t)}{2c_f}$. Similarly, substituting Equations (A.5) and (A.8) into Equation (A.11) and solving for $v(t)$, we obtain: $v(t) = \frac{n\rho_1(h-es)}{2c_d}$. Hence the results in Lemma 1.

In a similar way, we derive the effort levels for scenario 2, where customer has pre-set quality requirements, we know that $r(T) = Z$. Substituting Equations (A.3), (A.6), and (A.8) into Equation (A.10) and solving for $u(t)$, we get: $u(t) = \frac{\theta_1 km(T-t)+\beta w}{2c_f}$. Similarly, substituting Equations (A.6) and (A.9) into Equation (A.11) and solving for $v(t)$, we obtain: $v(t) = \frac{\beta n}{2c_d}$.

Thus, by substituting the above derived effort levels ($u(t)$ and $v(t)$) into Equation (2.3) and solving for β, we get: $\beta = -\frac{c_d\left(4c_f(sTy-Z)+\theta_1 kmT^2 w\right)}{2T\left(w^2c_d+n^2c_f\right)}$. Finally, replacing β in $u(t)$ and $v(t)$ with $-\frac{c_d\left(4c_f(sTy-Z)+\theta_1 kmT^2 w\right)}{2T\left(w^2c_d+n^2c_f\right)}$, we obtain: $u(t) = \frac{\theta_1 kmT\left(w^2c_d(T-2t)+2n^2c_f(T-t)\right)+4wc_dc_f(Z-sTy)}{4Tc_f\left(w^2c_d+n^2c_f\right)}$, and $v(t) = \frac{n\left(4c_f(Z-sTy)-\theta_1 kmT^2 w\right)}{4T\left(w^2c_d+n^2c_f\right)}$. Hence, the results in Lemma 3. In a similar way, we can derive the effort levels for scenarios with low budget levels. ■

Proofs of Lemmas 2 and 4

By substituting the efforts levels derived in Lemmas 1 and 3, the profit of the firm can be written as a function of s in both scenario 1 (no pre-set quality requirement) and scenario 2 (i.e., $r(T) = Z$). These are, respectively, given by:

$$\begin{aligned}\Pi \quad &= \frac{c_d\left(T\left(3\theta_1 km\rho_1 Tw(h-es)+3\rho_1^2w^2(h-es)^2+\theta_1^2k^2m^2T^2\right)-6c_f\left(2\left(as^2T-(h-es)\left(\rho_0+\rho_2 s+\rho_1 sTy\right)-\theta_2 ksT\right)-b\theta_1 ksT^2-2\theta_0 kT\right)\right)}{12c_dc_f}\\ &+\frac{3n^2\rho_1^2Tc_f(h-es)^2}{12c_dc_f}, \end{aligned} \tag{A.14}$$

$$\begin{aligned}\Pi_{r(T)=Z} \quad &= k\left(\frac{\theta_1 T\left(12c_f\left(wc_d(bsTw+m(Z-sTy))+bn^2sTc_f\right)+\theta_1 km^2T^2\left(w^2c_d+4n^2c_f\right)\right)}{24c_f\left(w^2c_d+n^2c_f\right)}+\theta_2 sT+\theta_0 T\right)\\ &-\frac{\theta_1^2k^2m^2T^4\left(w^2c_d+4n^2c_f\right)+48c_dc_f^2(Z-sTy)^2}{48Tc_f\left(w^2c_d+n^2c_f\right)}+(h-es)\left(\rho_0+\rho_2 s+\rho_1 Z\right)-as^2T. \end{aligned} \tag{A.15}$$

Hence, by taking the derivative of these functions with respect to s, we find that the optimal portion of the technology that is open source in both the scenarios are given by:

$$s^* = \frac{2b\theta_1 kT^2c_dc_f-4e\rho_0c_dc_f-2eh\rho_1^2Tw^2c_d-e\theta_1 km\rho_1T^2wc_d+4h\rho_2c_dc_f+4h\rho_1Tyc_dc_f+4\theta_2 kTc_dc_f-2ehn^2\rho_1^2Tc_f}{2\left(4aTc_dc_f-e^2\rho_1^2Tw^2c_d+4e\rho_2c_dc_f+4e\rho_1Tyc_dc_f-e^2n^2\rho_1^2Tc_f\right)},$$

$$s^*_{r(T)=Z} = \frac{c_d\left(w\left(\theta_1 kT^2(bw-my)+2w(-e(\rho_0+\rho_1 Z)+h\rho_2+\theta_2 kT)\right)+4yZc_f\right)+n^2c_f(kT(b\theta_1 T+2\theta_2)-2e(\rho_0+\rho_1 Z)+2h\rho_2)}{4c_d\left(w^2(aT+e\rho_2)+Ty^2c_f\right)+4n^2c_f(aT+e\rho_2)}.$$

After substituting these optimal values of s^* into the effort levels presented in Lemmas 1 and 3 and the objective function of the firm, we derive the reported effort levels in Lemmas 2 and 4. Similarly, the optimal solutions for Π^* in Lemma 2 and Lemma 4 can be obtained by substituting the respective values of optimal s in to Equations (A.14) and (A.15). However, for brevity, we do not present the expressions of Π^* in Lemma 2 and Lemma 4.

Note that in scenario 1, for the second-order condition to be satisfied, we need:

$$\frac{1}{2}\left(e\left(\rho_1 T\left(e\rho_1\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-4y\right)-4\rho_2\right)-4aT\right)<0.$$

Further, since $c_f > 0$ and $c_d > 0$, we can restate this condition as

$$c_d\left(e^2\rho_1^2Tw^2-4c_f\left(aT+e\rho_2+e\rho_1Ty\right)\right)+e^2n^2\rho_1^2Tc_f<0.$$

We use this result later in deriving Propositions 1-7. Next, in scenario 2, the second-order condition in the derivation of s^* is always satisfied. ■

Proof of Proposition 1

From Lemma 2, in order to have the second order condition satisfied, we need

$$\frac{1}{2}\left(e\left(\rho_1 T\left(e\rho_1\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-4y\right)-4\rho_2\right)-4aT\right)<0.$$

When this condition is satisfied, the objective function is strictly concave with respect to s. Therefore, when $s^* \leq 0$ and the second order condition holds, the objective function is decreasing in the feasible range and objective function is maximum at $s = 0$. By analyzing the conditions for $s^* \leq 0$ and satisfying the second order condition (and also ensuring that the feasibility conditions for e are satisfied), we find that the firm should keep its technology proprietary when $e \geq \frac{2c_dc_f(T(b\theta_1kT+2h\rho_1y+2\theta_2k)+2h\rho_2)}{c_d\left(4\rho_0c_f+\rho_1Tw(2h\rho_1w+\theta_1kmT)\right)+2hn^2\rho_1^2Tc_f} \equiv \mathscr{E}$, and $\rho_2 > \frac{1}{4}T\left(-\frac{4a}{e}+e\rho_1^2\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-4\rho_1y\right) \equiv \zeta$.

When the above conditions are satisfied, the optimal value of the extent of openness (i.e., s^*) is equal to zero. Hence part (a) of the proposition. Likewise, if the solution suggests that $s^* \geq 1$, and the second order condition is satisfied, it implies that the objective function value is maximum at $s = 1$. Therefore, by analyzing the conditions for the inequality $s^* \geq 1$, we find that the firm chooses to make its technology fully open source under the following conditions: $b \geq \frac{2c_f\left(2c_d(T(2a+\rho_1 y(2e-h)-\theta_2 k)+\rho_2(2e-h)+e\rho_0)+en^2\rho_1^2T(h-e)\right)+c_d e\rho_1 Tw(2\rho_1 w(h-e)+\theta_1 kmT)}{2c_f c_d\theta_1 kT^2} \equiv \mathscr{B}$ and $a > \frac{1}{4}e\left(e\rho_1^2\left(\frac{n^2}{c_d}+\frac{w^2}{c_f}\right)-\frac{4\rho_2}{T}-4\rho_1 y\right) \equiv \mathscr{A}$. Hence part (b) of the proposition. ■

Proof of Proposition 2

Using the results from Lemma 2, in order to have the second-order condition satisfied we need to have the following condition satisfied: $c_d\left(e^2\rho_1^2Tw^2-4c_f\left(aT+e\rho_2+e\rho_1Ty\right)\right)+e^2n^2\rho_1^2Tc_f<0$. By simple algebra, we obtain the following condition when the above mentioned second-order condition is not satisfied: $a<\mathscr{A}$.

Hence, when the second-order condition is not satisfied (so that the objective function is convex), we should have boundary conditions: either $s=0$ or $s=1$. Given that the second-order condition is not satisfied, we first solve the optimization problem in Equation (2.1) at $s=0$. We get $\Pi_{s=0}=\frac{3h^2\rho_1^2T\left(w^2c_d+n^2c_f\right)+c_d\left(12c_f(h\rho_0+\theta_0 kT)+\theta_1^2k^2m^2T^3\right)+3h\theta_1 km\rho_1T^2wc_d}{12c_dc_f}$. Next, we solve the optimization problem in Equation (2.1) at $s=1$. The expression for $\Pi_{s=1}$ is omitted for brevity. By comparing the profits at $s=0$ and $s=1$, we obtain:

$$(\Pi_{s=1})-(\Pi_{s=0})=\frac{c_d\left(c_f\left(2b\theta_1kT^2-4\left(aT+(e-h)\left(\rho_2+\rho_1Ty\right)+e\rho_0-\theta_2kT\right)\right)+e\rho_1Tw\left(\rho_1w(e-2h)-\theta_1kmT\right)\right)+en^2\rho_1^2Tc_f(e-2h)}{4c_dc_f}.$$

This expression is greater than zero only when:

$$a<\frac{c_d\left(2c_f\left(b\theta_1kT^2-2(e-h)\left(\rho_2+\rho_1Ty\right)-2e\rho_0+2\theta_2kT\right)+e\rho_1Tw\left(\rho_1w(e-2h)-\theta_1kmT\right)\right)+en^2\rho_1^2Tc_f(e-2h)}{4Tc_dc_f}\equiv\bar{\mathscr{A}}.$$

Therefore, when the second-order condition is not satisfied, i.e., $a<\mathscr{A}$ and $a<\bar{\mathscr{A}}$, we have a boundary solution, $s=1$. Further analysis of these two thresholds (i.e., $\mathscr{A}$ and $\bar{\mathscr{A}}$) suggests that $\bar{\mathscr{A}}<\mathscr{A}$ if and only if $e\geq\mathscr{E}$. Therefore, when $e\geq\mathscr{E}$ and $a<\bar{\mathscr{A}}$, we have $s=1$. Next, when $\bar{\mathscr{A}}<a<\mathscr{A}$ and $e\geq\mathscr{E}$, $s=0$. Furthermore, as we did in the proof of Proposition 1, by analyzing the conditions for $s^*\leq 0$ and second order condition is satisfied (and also ensuring that the feasibility conditions for e are satisfied), we find that the firm should keep its technology proprietary when $e\geq\mathscr{E}$, and $a>\mathscr{A}$. Hence, part (a) of the proposition.

From part (a) of Proposition 1, we know that $s \leq 0$ when $e \geq \mathscr{E}$. Therefore, when $e < \mathscr{E}$ (and the second order condition is satisfied), then $s > 0$, i.e., partially open. Hence, part (b) of the proposition. ■

Proof of Proposition 3

Analyzing the optimal trajectories of effort levels presented in Lemma 2, we have $u(t)^*$ strictly decreasing in t. Hence, we have $u(0) > u(T)$. Therefore, when $u(T) > v(t)$, the maintenance effort is always greater than the development effort throughout the planning horizon. We find that $u(T) > v(t)$ holds true if and only if $w > \frac{nc_f}{c_d}$. Hence the proposition.

Proof of Proposition 4

Using the results from Lemma 2 and taking the derivative of optimal level of openness and effort levels with respect to k, we get:

$$\begin{aligned}
\frac{dv(t)^*}{dk} &= \frac{en\rho_1 T\left(e\theta_1 m\rho_1 Tw - 2c_f\left(b\theta_1 T + 2\theta_2\right)\right)}{4c_d\left(4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right) - e^2\rho_1^2 Tw^2\right) - 4e^2 n^2\rho_1^2 Tc_f}, && \text{(A.16)}\\
\frac{ds^*}{dk} &= \frac{Tc_d\left(2c_f\left(b\theta_1 T + 2\theta_2\right) - e\theta_1 m\rho_1 Tw\right)}{c_d\left(8c_f\left(aT + e\rho_2 + e\rho_1 Ty\right) - 2e^2\rho_1^2 Tw^2\right) - 2e^2 n^2\rho_1^2 Tc_f}, \quad \textit{and} && \text{(A.17)}\\
\frac{du(t)^*}{dk} &= \frac{\frac{e\rho_1 Twc_d\left(2c_f(b\theta_1 T + 2\theta_2) - e\theta_1 m\rho_1 Tw\right)}{c_d\left(e^2\rho_1^2 Tw^2 - 4c_f(aT + e\rho_2 + e\rho_1 Ty)\right) + e^2 n^2\rho_1^2 Tc_f} + 2\theta_1 m(T-t)}{4c_f}. && \text{(A.18)}
\end{aligned}$$

From the proof of Lemma 2, in order to have the second-order condition in derivation of s^* to be satisfied, we need to have $c_d\left(e^2\rho_1^2 Tw^2 - 4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right)\right) + e^2 n^2\rho_1^2 Tc_f < 0$. Hence, the sign of denominator of Equation (A.16) is always positive. Now analyzing the inequality in Equation (A.16), we find that when $b < \frac{em\rho_1 w}{2c_f} - \frac{2\theta_2}{\theta_1 T} \equiv \beta_0$, then $\frac{dv(t)^*}{dk} > 0$. Similarly, the denominator of Equation (A.17) is always positive, and by simple algebra we find that when $b < \beta_0$, then $\frac{ds(t)^*}{dk} < 0$. Next, the sign of the denominator of Equation (A.18) is always negative. Analyzing this inequality, we find that $\frac{du(t)^*}{dk} > 0$ holds if and only if $b < \frac{c_d\left(4c_f(2\theta_1 m(t-T)(aT+e\rho_2+e\rho_1 Ty)+e\theta_2\rho_1 Tw)+e^2\theta_1 m\rho_1^2 Tw^2(T-2t)\right)+2e^2\theta_1 mn^2\rho_1^2 Tc_f(T-t)}{2e\theta_1\rho_1 T^2 wc_d c_f} \equiv \beta_1$.

Since $\frac{d^2u(t)^*}{dkdt} = -\frac{\theta_1 m}{2c_f} < 0$, we have $\frac{du(t)^*}{dk}$ strictly decreasing in t. Also observe that

$\frac{c_d\left(4c_f(2\theta_1 m(t-T)(aT+e\rho_2+e\rho_1 Ty)+e\theta_2\rho_1 Tw)+e^2\theta_1 m\rho_1^2 Tw^2(T-2t)\right)+2e^2\theta_1 mn^2\rho_1^2 Tc_f(T-t)}{2e\theta_1\rho_1 T^2 wc_d c_f}$, evaluated at the upper bound of t, i.e., $t = T$ is $\frac{2\theta_2}{\theta_1 T} - \frac{em\rho_1 w}{2c_f}$. Hence, $\beta_0 < \beta_1$ for any $t < T$. Combining this finding with the above threshold values (in addition to ensuring that the second order condition is satisfied), we can write the behavior of the effort levels and the extent of openness with respect to k as presented in the proposition. ■

Proof of Proposition 5

Using the results from Lemma 2 and checking for sensitivity with respect to b and ensuring the second order condition, we obtain:

$$\begin{aligned}
\frac{ds^*}{db} &= \frac{\theta_1 kT^2 c_d c_f}{c_d\left(4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right) - e^2\rho_1^2 Tw^2\right) - e^2 n^2\rho_1^2 Tc_f} > 0,\\
\frac{du^*}{db} &= -\frac{e\theta_1 k\rho_1 T^2 wc_d}{c_d\left(8c_f\left(aT + e\rho_2 + e\rho_1 Ty\right) - 2e^2\rho_1^2 Tw^2\right) - 2e^2 n^2\rho_1^2 Tc_f} < 0,\\
\frac{dv^*}{db} &= -\frac{e\theta_1 kn\rho_1 T^2 c_f}{c_d\left(8c_f\left(aT + e\rho_2 + e\rho_1 Ty\right) - 2e^2\rho_1^2 Tw^2\right) - 2e^2 n^2\rho_1^2 Tc_f} < 0,\\
\frac{d\Pi^*}{db} &= \frac{1}{2}\theta_1 kT^2 s^* > 0.
\end{aligned}$$

Hence, the findings in Proposition 5. ■

Proof of Proposition 6

Using the results from Lemma 2 and checking for sensitivity with respect to y, we obtain:

$$\begin{aligned}
\frac{ds^*}{dy} &= \frac{2\rho_1 Tc_d c_f\left(c_d\left(c_f\left(4ahT - 2ekT\left(b\theta_1 T + 2\theta_2\right) + 4e^2\rho_0\right) + e^2\rho_1 Tw\left(h\rho_1 w + \theta_1 kmT\right)\right) + e^2 hn^2\rho_1^2 Tc_f\right)}{\left(c_d\left(e^2\rho_1^2 Tw^2 - 4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right)\right) + e^2 n^2\rho_1^2 Tc_f\right)^2} \lesseqgtr 0,\\
\frac{du^*}{dy} &= \frac{e\rho_1^2 Twc_d\left(-c_d\left(c_f\left(4ahT - 2ekT\left(b\theta_1 T + 2\theta_2\right) + 4e^2\rho_0\right) + e^2\rho_1 Tw\left(h\rho_1 w + \theta_1 kmT\right)\right) - e^2 hn^2\rho_1^2 Tc_f\right)}{\left(c_d\left(e^2\rho_1^2 Tw^2 - 4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right)\right) + e^2 n^2\rho_1^2 Tc_f\right)^2} \lesseqgtr 0,\\
\frac{dv^*}{dy} &= \frac{en\rho_1^2 Tc_f\left(-c_d\left(c_f\left(4ahT - 2ekT\left(b\theta_1 T + 2\theta_2\right) + 4e^2\rho_0\right) + e^2\rho_1 Tw\left(h\rho_1 w + \theta_1 kmT\right)\right) - e^2 hn^2\rho_1^2 Tc_f\right)}{\left(c_d\left(e^2\rho_1^2 Tw^2 - 4c_f\left(aT + e\rho_2 + e\rho_1 Ty\right)\right) + e^2 n^2\rho_1^2 Tc_f\right)^2} \lesseqgtr 0.
\end{aligned}$$

Further analyzing the above inequalities, we find that when

$$a < -\frac{e\left(c_d\left(e\rho_1 Tw\left(h\rho_1 w + \theta_1 kmT\right) - 2c_f\left(kT\left(b\theta_1 T + 2\theta_2\right) - 2e\rho_0\right)\right) + ehn^2\rho_1^2 Tc_f\right)}{4hTc_dc_f} \equiv \alpha$$

and ensuring that the second order condition is satisfied and the non-negativity conditions of all the variables are satisfied, $\frac{ds^*}{dy} < 0$, $\frac{du^*}{dy} > 0$, and $\frac{dv^*}{dy} > 0$. Hence, the results in the proposition. ■

Proof of Proposition 7

Using the results from Lemma 2 and checking for sensitivity with respect to h and ensuring the second order condition, we obtain:

$$\begin{aligned} \frac{du(t)^*}{dh} > 0 \quad &; \quad \frac{dv^*}{dh} > 0 \\ \frac{ds(t)^*}{dh} \quad &= \quad \frac{\rho_1 T\left(e\rho_1\left(c_f n^2 + c_d w^2\right) - 2c_fc_dy\right) - 2c_fc_d\rho_2}{c_f\left(e^2n^2\rho_1^2 T - 4c_d\left(T\left(a + e\rho_1 y\right) + e\rho_2\right)\right) + c_de^2\rho_1^2 Tw^2} \lesseqgtr 0. \end{aligned} \tag{A.19}$$

Analyzing the inequality in Equation (A.19) (and also ensuring that the second order condition is satisfied along with the non-negativity conditions of all the variables), we find that when $e > \frac{2c_fc_dy}{\rho_1\left(c_fn^2+c_dw^2\right)} \equiv \bar{\mathscr{E}}$ and $T < \frac{2c_fc_d\rho_2}{c_fen^2\rho_1^2+c_de\rho_1^2w^2-2c_fc_d\rho_1 y} \equiv \bar{\mathscr{T}}$, then $\frac{ds^*}{dh} > 0$. ■

Proof of Proposition 8

Using the results from Lemma 4 and checking for sensitivity with respect to k, we get:

$$\frac{dv(t)^*}{dk} \quad = \quad -\frac{nT\left(\theta_1 T\left(amw + bc_fy\right) + 2c_f\theta_2 y + e\theta_1 m\rho_2 w\right)}{4\left(c_f\left(T\left(an^2 + c_dy^2\right) + en^2\rho_2\right) + c_dw^2\left(aT + e\rho_2\right)\right)},$$

From the proof of Lemma 4, we know that the objective function is strictly concave in s. Hence, we can easily derive the result in Proposition 8. ■

Proof of Proposition 9

Using the results from Lemma 4 and checking for sensitivity with respect to h, we get: $\frac{ds^*}{dh} > 0$; $\frac{du^*}{dh} < 0$; $\frac{dv^*}{dh} < 0$. Thus, we have s^* strictly increases with h, $u(t)^*$ and $v(t)^*$ strictly decreases with h. ■

Proof of Proposition 10

Using the results from Lemma 4 and checking for sensitivity with respect to Z, we get:

$$\frac{ds^*}{dZ} = \frac{2yc_dc_f - e\rho_1\left(w^2c_d + n^2c_f\right)}{2c_d\left(w^2\left(aT + e\rho_2\right) + Ty^2c_f\right) + 2n^2c_f\left(aT + e\rho_2\right)} \lesseqgtr 0, \tag{A.20}$$

$$\frac{du^*}{dZ} > 0 \quad ; \quad \frac{dv^*}{dZ} > 0 \tag{A.21}$$

On further examining Equation (A.20), we find that when the condition $e < \frac{2c_fc_dy}{c_fn^2\rho_1 + c_d\rho_1w^2} \equiv \eta$ is satisfied along with the non-negativity conditions for all the variables, we get $\frac{ds^*}{dZ} > 0$, else $\frac{ds^*}{dZ} < 0$. Also, from Equation (A.21), we find that $u(t)^*$ strictly increases with Z and $v(t)^*$ strictly increases with Z. ■

A.3 Tables and Figures

Figure A.1: When to Keep the Technology Proprietary

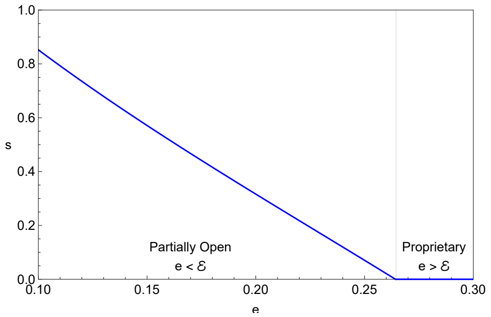


Figure A.2: When to Make the Technology Fully Open

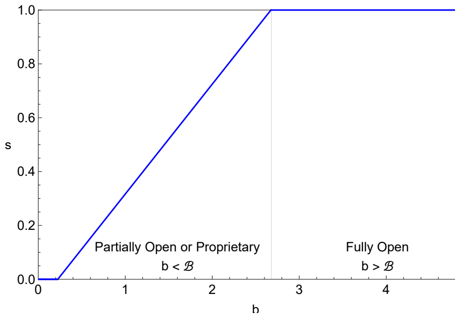
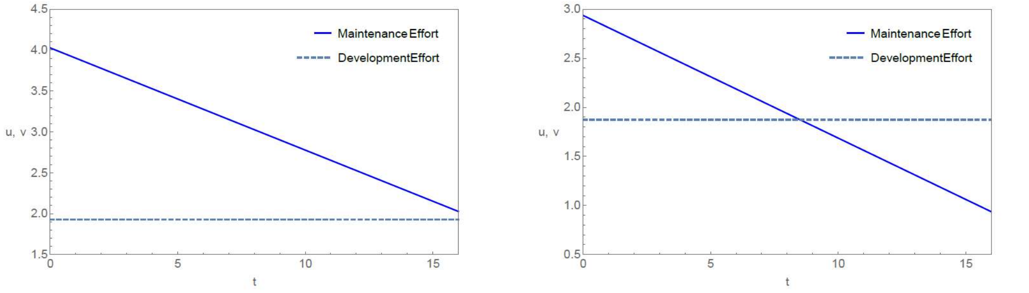


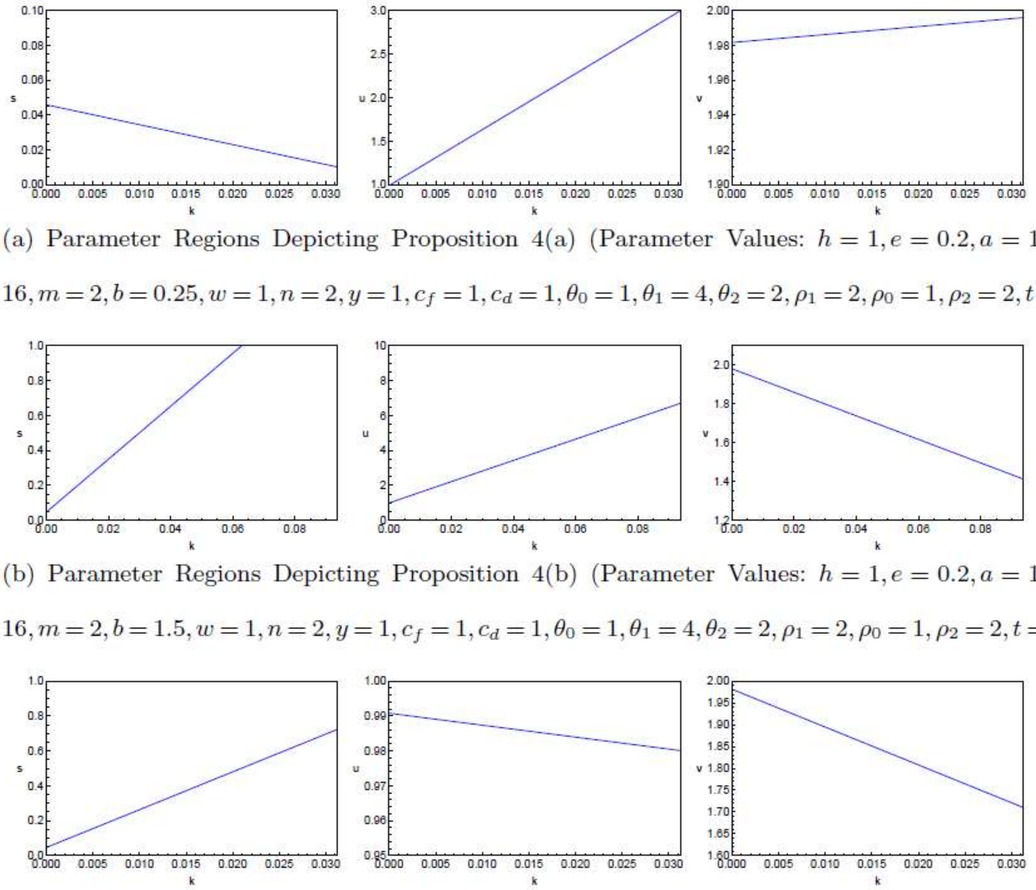
Figure A.3: Should the Firm Focus on Maintenance or Development?



(a) Parameter Regions when $\frac{w}{c_f} > \frac{n}{c_d}$ (Parameter Values: $k = \frac{1}{32}, h = 1, e = 0.12, a = 1, T = 16, m = 2, b = 1, w = 2.1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2$)

(b) Parameter Regions when $\frac{w}{c_f} < \frac{n}{c_d}$ (Parameter Values: $k = \frac{1}{32}, h = 1, e = 0.2, a = 1, T = 16, m = 2, w = 1, b = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2$)

Figure A.4: Impact of the Valuation of Current Version on Firm's Behavior



(a) Parameter Regions Depicting Proposition 4(a) (Parameter Values: $h = 1, e = 0.2, a = 1, T = 16, m = 2, b = 0.25, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2, t = 1$)

(b) Parameter Regions Depicting Proposition 4(b) (Parameter Values: $h = 1, e = 0.2, a = 1, T = 16, m = 2, b = 1.5, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2, t = 1$)

(c) Parameter Regions Depicting Proposition 4(c) (Parameter Values: $h = 1, e = 0.2, a = 1, T = 16, m = 2, b = 2, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2, t = 15$)

Figure A.5: Parameter Regions Depicting the Condition, $b > \beta_1$ in Proposition 4(c) (i.e., β_1)

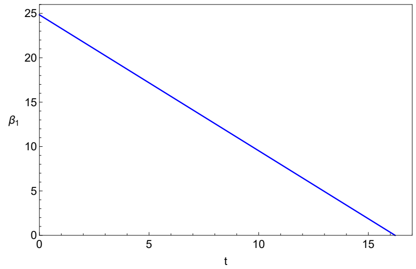
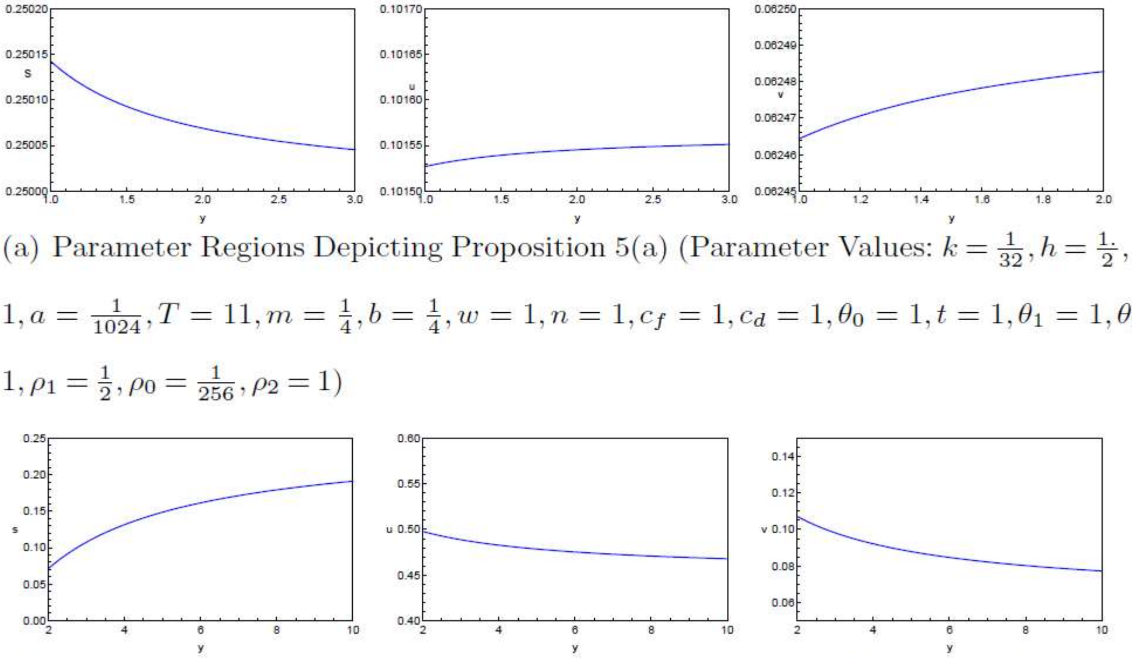


Figure A.6: Impact of Openness Sensitivity Terms on Firm's Behavior



(a) Parameter Regions Depicting Proposition 5(a) (Parameter Values: $k = \frac{1}{32}, h = \frac{1}{2}, e = 1, a = \frac{1}{1024}, T = 11, m = \frac{1}{4}, b = \frac{1}{4}, w = 1, n = 1, c_f = 1, c_d = 1, \theta_0 = 1, t = 1, \theta_1 = 1, \theta_2 = 1, \rho_1 = \frac{1}{2}, \rho_0 = \frac{1}{256}, \rho_2 = 1$)

(b) Parameter Regions Depicting Proposition 5(b) (Parameter Values: $k = \frac{1}{32}, h = \frac{1}{2}, e = 1, a = 1, t = 1, T = 11, m = \frac{5}{2}, b = \frac{1}{4}, w = 1, n = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 1, \theta_2 = \frac{1}{2}, \rho_1 = \frac{1}{2}, \rho_0 = 1, \rho_2 = 1$)

Figure A.7: Impact of the Valuation of Next Version on Firm's Behavior (Parameter Values: $w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \rho_1 = 2, \rho_0 = 1, \rho_2 = 3$. For these parameter values, $e > \bar{\mathscr{E}}$ is always satisfied.)

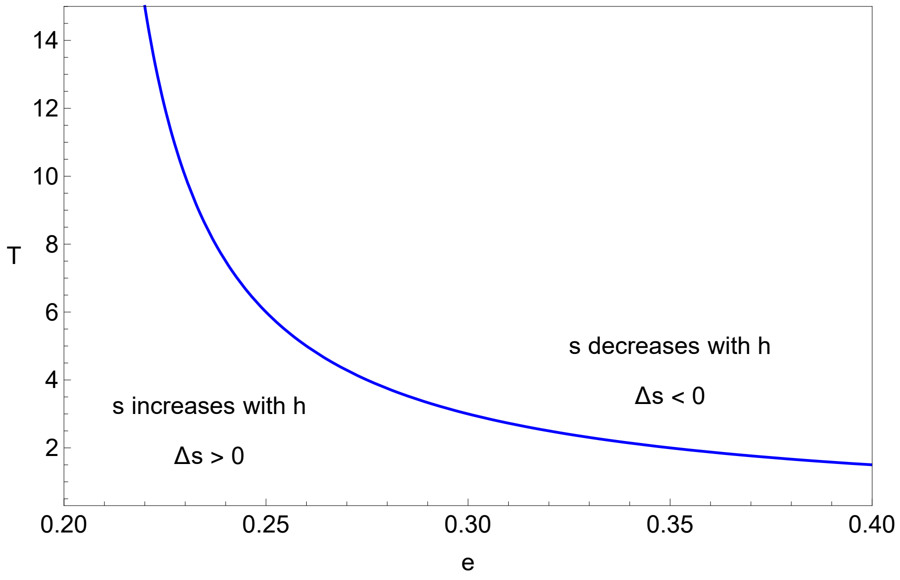
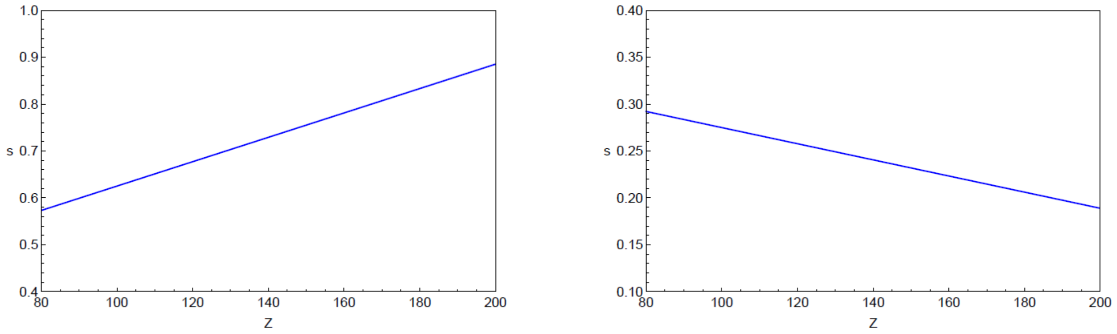


Figure A.8: Impact of Changes in Market Requirements on Firm's Behavior



(a) Parameter Regions when $e < \eta$ (Parameter Values: $k = \frac{1}{32}, h = 1, e = 0.17, a = 0.5, T = 16, m = 2, b = 0.75, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2$)

(b) Parameter Regions when $e > \eta$ (Parameter Values: $k = \frac{1}{32}, h = 1, e = 0.21, a = 0.5, T = 16, m = 2, b = 0.75, w = 1, n = 2, y = 1, c_f = 1, c_d = 1, \theta_0 = 1, \theta_1 = 4, \theta_2 = 2, \rho_1 = 2, \rho_0 = 1, \rho_2 = 2$)

APPENDIX B

APPENDIX FOR CHAPTER 3: A FRAMEWORK FOR ANALYZING INFLUENCER MARKETING IN SOCIAL NETWORKS: SELECTION AND SCHEDULING OF INFLUENCERS

B.1 Proofs

B.1.1 NP-Hard Proofs for Theorems 1, 3, and 4

In this section we provide proofs for Theorems 1, 3, and 4. First, let us start by presenting the proof for Theorem 4.

B.1.1.1 NP Hard Proof for Theorem 4

We use the 3-Satisfiability (3SAT) problem (Garey and Johnson 1979) for our reduction.

<u>3-SATISFIABLILITY (3SAT) PROBLEM:</u>

<u>INSTANCE</u>: A set $X = \{x_1, x_2, \ldots, x_m\}$ of boolean variables and a collection $C = \{C_1 \cap C_2 \cap \ldots \cap C_k\}$ of clauses over X, each of which is a disjunction of literals, $x_1, \bar{x}_1, x_2, \bar{x}_2, \ldots, x_m, \bar{x}_m$ such that $|C_j| = 3$ for $1 \leq j \leq k$ and $\bar{x}_i = 1 - x_i$ for $1 \leq i \leq p$.

<u>SOLUTION</u>: Find an assignment of either a TRUE (1) or a FALSE (0) value to each variable in $\{x_1, x_2, \ldots, x_m\}$, such that the expression C evaluates to TRUE (1).

Given an arbitrary instance of $3SAT$, we construct an instance of the $M1L$ problem.

- Let $m = 3$ and $k = 5$. Thus, $X = \{x_1, x_2, x_3\}$; and $C = \{C_1 \cap C_2 \cap C_3 \cap C_4 \cap C_5\}$, where $C_1 = \bar{x}_1 \cup x_2 \cup \bar{x}_3$, $C_2 = x_1 \cup \bar{x}_2 \cup \bar{x}_3$, $C_3 = \bar{x}_1 \cup x_2 \cup x_3$, and $C_4 = x_1 \cup x_2 \cup x_3$, $C_5 = \bar{x}_1 \cup x_2 \cup \bar{x}_3$.

- The total number of influential users (i.e., size of set W), $|W| = 2m$. Each variable in set $X = \{x_1, \bar{x}_1, x_2, \bar{x}_2, x_3, \bar{x}_3\}$ corresponds to a distinct influential user.

- The cardinality if the number of influential users, $z = p = 3$. Note that the truth assignment for our instance is $\bar{x}_1 = 1$, $\bar{x}_2 = 1$ and $x_3 = 1$.

- Each influential user has direct (or first level) followers defined by the clauses over X, $C = \{C_1 \cap C_2 \cap \ldots \cap C_k\}$. There are a total of k immediate followers ($k = 5$ in the example). Let the benefit to the firm (i.e., b_j) when the message reaches the first level followers C_1 to C_5 be h. Furthermore, let k immediate followers retweet the message of any of the influencer with a probability, $p_{ij} = 1$. Let each of these k followers have same number of followers, i.e., $F_j = f_1$ and the benefit to the firm from reaching each of the followers of followers be equal, $i.e., d_j = s$. Refer Figure B.1 for the instance of $M1L$ problem that we construct.

- A pair of influential users $(x_i, \bar{x}_i)$ have a common identical M followers who retweet with probability, $p_{ij} = 1$ and the benefit to the firm (i.e., b_j) for reaching each of the M followers be equal to g. Further, let each of the M followers have same number of followers i.e., f_2 and let the firm benefit $d_j = r$.

- Retweet Tree of the followers of the influential user x_1 (in our instance) is shown in Figure B.2.

- The Minimum Cardinality, $t = m$ ($t = 3$ in for our instance).

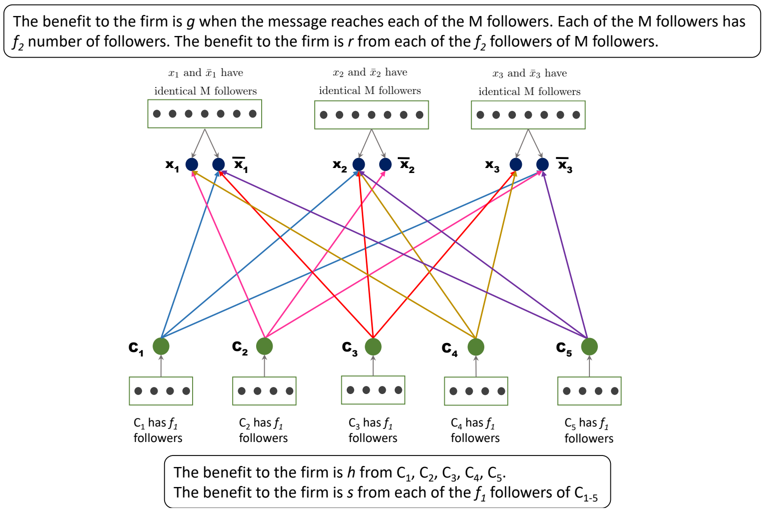
- Let M be a very large number and $r > s, f_2 > f_1$, and $g > h$.

For the above instance of the $M1L$ problem constructed above, we consider the following question:

<ins>Decision Problem</ins>: Does there exist a solution for the $M1L$ problem with the total benefit to the firm (i.e., objective value of problem $M1L$), $\Psi_1 \geq (mMg + mMf_2r) + (kh + kf_1s)$?

The decision problem is NP. Also we can verify that the construction of our decision problem from the $3SAT$ instance can be done in polynomial time. We now show that the decision problem has an answer if and only if the $3SAT$ instance is satisfiable.

Figure B.1: Construction of an Instance of $M1L$ Problem



The benefit to the firm is *g* when the message reaches each of the M followers. Each of the M followers has *f₂* number of followers. The benefit to the firm is *r* from each of the *f₂* followers of M followers.

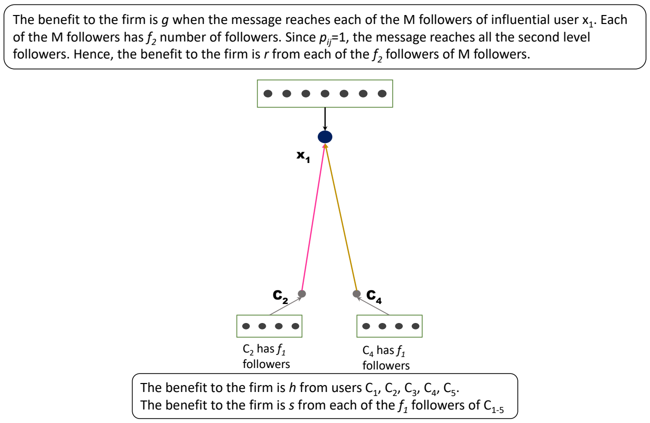
x_1 and $\bar{x}_1$ have identical M followers
x_2 and $\bar{x}_2$ have identical M followers
x_3 and $\bar{x}_3$ have identical M followers
x_1
$\bar{x}_1$
x_2
$\bar{x}_2$
x_3
$\bar{x}_3$
C_1
C_2
C_3
C_4
C_5
C_1 has f_1 followers
C_2 has f_1 followers
C_3 has f_1 followers
C_4 has f_1 followers
C_5 has f_1 followers

The benefit to the firm is *h* from C_1, C_2, C_3, C_4, C_5.
The benefit to the firm is *s* from each of the *f₁* followers of $C_{1\text{-}5}$

If part: Suppose the instance of $3SAT$ is satisfiable, then either x_i or $\bar{x}_i$ has a truth assignment. For each pair of influential users $(x_i, \bar{x}_i)$, we select one user in the pair that has TRUE value 1. Then this selection of m influential users will benefit the firm by $(mgM + mMf_2r)$ via their M followers. Also, since the $3SAT$ instance is satisfiable, each clause $C_i, i = 1, 2, ..., k$ will be reached from the selected m influential users having a TRUE value 1. This in turn further provides a benefit of of $(kh + kf_1s)$ to the firm. Thus, the total benefit that the firm gets is, $\Psi_1 = (mgM + mMf_2r) + (kh + kf_1s)$.

Only If part: Suppose there exists a solution to the decision problem with $\Psi_1 \geq (mgM + mMf_2r) + (kh + kf_1s)$. Since the total benefit to the firm should be at most $(mgM + mMf_2r) + (kh + kf_1s)$, let $\Psi_1 = (mgM + mMf_2r) + (kh + kf_1s)$. Since the cardinality of set W is $t = m$, we can select only m influential users. From above, we know that M is a very large number and $r > s$, $f_2 > f_1$, and $g > h$, hence we can easily show that $(mgM + mMf_2r) > (kh + kf_1s)$. Since $(mgM + mMf_2r) > (kh + kf_1s)$, to obtain the weighted sum equal to Ψ_1, we need to select at least m influential users else $\Psi_1 < (mgM + mMf_2r) + (kh + kf_1s)$. If $m - 1$ influential users are selected, $\Psi_1 = ((m-1)gM + (m-1)Mf_2r) + (kh + kf_1s) < (mgM + mMf_2r) + (kh + kf_1s)$. Further, the m selected influential users need to reach all the followers

Figure B.2: Retweet Tree of Influential user x_1


The benefit to the firm is g when the message reaches each of the M followers of influential user x_1. Each of the M followers has f_2 number of followers. Since p_{ij}=1, the message reaches all the second level followers. Hence, the benefit to the firm is r from each of the f_2 followers of M followers.
x_1
C_2
C_4
C_2 has f_1 followers
C_4 has f_1 followers
The benefit to the firm is h from users C_1, C_2, C_3, C_4, C_5.
The benefit to the firm is s from each of the f_1 followers of $C_{1\text{-}5}$

with weight $(kh+kf_1s)$ (i.e., C_1, C_2, C_3, C_4, C_5). If $m-1$ influencers are selected, we do not reach at least one of the followers with weight $(kh+kf_1s)$. Therefore, by construction only when m influencers are selected, we reach all the followers with weight $(kh+kf_1s)$. Thus, a satisfiable assignment, corresponding to m influential users, for the $3SAT$ instance is now available. This completes the proof. ■

NP Hard Proofs for Theorem 1 and Theorem 3: Since Model $M1L$ is a special case of Model $M2L$ and $M2L^{UB}$. As model $M1L$ is NP-Hard, models $M2L$ and $M2L^{UB}$ are also NP-hard. This completes the proof. ■

B.1.2 Proof of Proposition 11:

Proof: Since any user $j \in V$ follows only one influencer i in set W, then $V_1, V_2, \ldots, V_n$ form a partition of Set V. Thus, the objective function Ψ_2 could be rewritten as $\sum_{i \in W}(\sum_{j \in V_i} b_j y_j + \sum_{j \in V_i} d_j g_j F_j)$.

It is easy to see that $p_{ij} = 0$, if $j \notin V_i$. Thus, Constraint 3.3 can be rewritten as $\delta_j = 1 - p_{ij}x_i, \quad \forall j \in V_i, \forall i \in W$. Then, Constraint 3.4 can be rewritten as $y_j = 1 - \delta_j = p_{ij}x_i, \quad \forall j \in V_i, \forall i \in W$.

Since in V, a user j who is a follower of user i in W, only follows some followers of i, we have $p_{kj} = 0$, if $j \in V_i$ and $k \notin V_i$. Thus, Constraint 3.5 can be rewritten as $\alpha_j = \prod_{k\in V_i}(1-p_{kj}y_k), \quad \forall j \in V_i, \forall i \in W$. For $k \in V_i$, we also have $y_k = p_{ik}x_i$. Then, Constraint 3.6 can be rewritten as $g_j = 1-(1-p_{ij}x_i)\prod_{k\in V_i}(1-p_{kj}y_k) = 1-(1-p_{ij}x_i)\prod_{k\in V_i}(1-p_{kj}p_{ik}x_i), \quad \forall j \in V_i, \forall i \in W$.

We can substitute Constraints 3.4, 3.6 in the objective function, and obtain $\Psi_2 = \sum_{i\in W}(\sum_{j\in V_i} b_j y_j + \sum_{j\in V_i} d_j g_j F_j)$
$= \sum_{i\in W}(\sum_{j\in V_i} b_j p_{ij} x_i + \sum_{j\in V_i} d_j F_j[1-(1-p_{ij}x_i)\prod_{k\in V_i}(1-p_{kj}p_{ik}x_i)])$. We define $f_i(x_i) = \sum_{j\in V_i} b_j p_{ij} x_i + \sum_{j\in V_i} d_j F_j[1-(1-p_{ij}x_i)\prod_{k\in V_i}(1-p_{kj}p_{ik}x_i)]$. When $x_i = 0$, we have $f_i(0) = 0$. When $x_i = 1$, we have $f_i(1) = \sum_{j\in V_i} b_j p_{ij} + \sum_{j\in V_i} d_j F_j[1-(1-p_{ij})\prod_{k\in V_i}(1-p_{kj}p_{ik})] = \Lambda_i$.

For any feasible solution $X = \{x_1, x_2, \ldots, x_n\}$ to Problem $M2L$, we have $\Psi_2 = \sum_{i\in W} f_i(x_i)$. It is easy to see that an optimal solution with constraint $\sum_i x_i \le t$ includes the first t users in Set W after sorting them in the descending order of Λ_i. This completes the proof. ■

B.1.3 Proof of Theorem 2:

For this proof, we will rely on the properties of two logarithmic functions: (i) $\ln(1-py)$, and (ii) $ln(1-p)\times y$. The Taylor's expansion of these two functions are given by:

$$\begin{aligned}\ln(1-py) &= -py - \frac{(py)^2}{2} - \frac{(py)^3}{3} \cdots - \frac{(py)^n}{n} \quad &(B.1)\\ \ln(1-p)y &= -py - \frac{(p)^2 y}{2} - \frac{(p)^3 y}{3} \cdots - \frac{(p)^n y}{n} \quad &(B.2)\end{aligned}$$

When $0 \le y \le 1$, we can easily see that $\frac{(py)^2}{2} < \frac{(p)^2 y}{2}$. Similarly, $\frac{(py)^n}{2} < \frac{(p)^n y}{2}$. Therefore, we have $\ln(1-p)y \le \ln(1-py)$, and $\sum_{k\in V}\ln(1-p_{kj})y_k \le \sum_{k\in V}\ln(1-p_{kj}y_k)$. Exponentiating both sides of the inequality, we have

$$\prod_{k\in V}(1-p_{kj})^{y_k} \le \prod_{k\in V}(1-p_{kj}y_k).$$

Hence, $\alpha_j \ge \bar{\alpha}_j$, and $g_j \le \bar{g}_j$. Closely analyzing the objective functions $\Psi_2^{UB}(x)$ and $\Psi_2(x)$, the first term (i.e., $\sum_{j\in V} b_j y_j$) in both objective functions of $M2L$ and $M2L^{UB}$ are the same. Since $g_j \le \bar{g}_j$, for any given solution set x, $\Psi_2^{UB}(x) \ge \Psi_2(x)$, i.e., the objective value of Model $M2L^{UB}$ is greater than equal to the objective value of Model $M2L$ (the main model).

More formally, if $\overline{\Psi}_2$ is the objective value for the optimal solution for Model $M2L$, as $\Psi_2^{UB}(x) \geq \Psi_2(x)$, we have $\overline{\Psi}_2^{UB} \geq \overline{\Psi}_2$. However, the optimal solution for Model $M2L$ might not be an optimal solution for Model $M2L^{UB}$. That is, there might exist another solution (which is optimal for Model $M2L^{UB}$) such that $\overline{\overline{\Psi}}_2^{UB} \geq \overline{\Psi}_2^{UB}$. Therefore, the objective value of the new model (Model $M2L$) provides an upper bound to the multiplicative two-level influence model (Model $M2L$). ■

B.1.4 Proof of Proposition 12

For this proof, we will use the binomial expansion of $(1-x)^n$. Specifically, $(1-x)^n = \sum_{k=0}^{n} \binom{n}{k} 1^{n-k} (-x)^k$, which can be simplified to $1 - nx + \frac{n(n-1)}{2!} \cdot x^2 - \frac{n(n-1)(n-2)}{3!} \cdot x^3 \dots$. When $x \rightarrow 0$, $(1-x)^n \rightarrow 1 - nx$.

The binomial expansion of $(1-p_{kj})^{y_k} = 1 - p_{kj}y_k + \frac{y_k(y_k-1)}{2!}p_{kj}^2 - \frac{y_k(y_k-1)(y_{kj}-2)}{3!}p_{kj}^3 \dots$. When $p_{kj} \rightarrow 0$, we have $1 - p_{kj}y_k$. Hence, the finding in Proposition 12 holds. ■

B.1.5 Proof of Lemma 5

Compared to Model $M2L$, in Model $M1L$ we ignore the increase in probability of retweet due to influence from users in set V. Therefore for any given solution x, $\Psi_2(x) \geq \Psi_1(x)$, i.e., the objective value of the multiplicative two-level influence model (Model $M2L$) is greater than or equal to the objective value of multiplicative single level influence model (Model $M1L$). Let Ψ_1^o be the objective value for the optimal solution for Model $M1L$. Since $\Psi_2(x) \geq \Psi_1(x)$, we have $\Psi_2^o \geq \Psi_1^o$. However, the optimal solution for Model $M1L$ might not be an optimal solution for Model $M2L$. That is, there exists another solution (which is optimal for Model $M2L$) such that $\Psi_2^* \geq \Psi_2^o$. Thus, the objective value of the multiplicative single level influence model (Model $M1L$) provides an lower bound to the multiplicative two-level influence model (Model $M2L$). ■

B.2 Approximation of Model $M1L$

Clearly, we see that constraint (3.44) makes the $M1L$ a non-linear mixed integer program. Camm et al. (2002) propose a method to linearize this non-linear constraint using the separable programming approach suggested by Saydam and McKnew (1985). Let $W_j^{<1} = \{i|0 < p_{ij} < 1\}$ and $W_j^{=1} = \{i|p_{ij} = 1\}$. We can now replace Equations 3.44 and 3.45 with the following new constraints:

$$h_j = \prod_{i \in W_j^{<1}} (1 - p_{ij})^{x_i}, \quad \forall j \in V \tag{B.3}$$

$$g_j \leq 1 - h_j + \sum_{i \in W_j^{=1}} x_i, \quad \forall j \in V \tag{B.4}$$

$$0 \leq g_j \leq 1; \;\; h_j \geq 0; \;\; x_i, \in \{0, 1\}. \tag{B.5}$$

Further, since h_j is strictly positive, we can now take the log of each side of the non-linear constraints. This results in the following optimization model.

$$\text{Max} \quad \Psi_1 = \sum_j b_j g_j + \sum_j d_j g_j F_j \tag{B.6}$$

subject to:

$$\sum_i x_i \leq t, \tag{B.7}$$

$$\ln(h_j) = \sum_{i \in W_j^{<1}} \ln{(1 - p_{ij})} x_i, \;\; \forall j \in V \tag{B.8}$$

$$g_j \leq 1 - h_j + \sum_{i \in W_j^{=1}} x_i, \quad \forall j \in V \tag{B.9}$$

$$0 \leq g_j \leq 1; \quad h_j \geq 0; \quad x_i \in \{0, 1\}. \tag{B.10}$$

As proposed by Bradley et al. (1977) and Camm et al. (2002), we now approximate the non-linear term $\ln(h_j)$ by taking a convex combination of points, λ, on the curve. Following

Camm et al. (2002), let R_j be the set of break points used to approximate the interval from 0 to 1. Let B_{jt} be the t^{th} breakpoint and λ_{jt} the weighting of the t^{th} breakpoint. We then make the below mentioned linear approximation substitutions:

$$
\begin{aligned}
h_j &= \sum_{t \in R_j} B_{jt}\lambda_{jt}, && \text{(B.11)}\\
\ln(h_j) &= \sum_{t \in R_j} \ln(B_{jt})\lambda_{jt}, && \text{(B.12)}\\
\sum_{t \in R_j} \lambda_{jt} &= 1. && \text{(B.13)}
\end{aligned}
$$

Substituting the above values of h_j and $\ln(h_j)$ into Constraints (B.8) and (B.9), we get the linearized approximation of problem $M1L$ that is presented in model $M1L - AP$.

Mathematical Formulation - $M1L - AP$:

$$\text{Max} \quad \Psi_1^{AP} = \sum_j b_j g_j + \sum_j d_j g_j F_j \tag{B.14}$$

subject to:

$$\sum_i x_i \leq t, \tag{B.15}$$

$$\sum_{t \in R_j} \ln(B_{jt})\lambda_{jt} = \sum_{i \in W_j^{<1}} \ln(1 - p_{ij})x_i, \qquad \forall j \in V \tag{B.16}$$

$$g_j \leq 1 - \sum_{t \in R_j} B_{jt}\lambda_{jt} + \sum_{i \in W_j^{=1}} x_i, \quad \forall j \in V \tag{B.17}$$

$$\sum_{t \in R_j} \lambda_{jt} = 1, \qquad \forall j \in V \tag{B.18}$$

$$0 \leq g_j \leq 1; \quad \lambda_{jt} \geq 0; \quad x_i \in \{0, 1\}. \tag{B.19}$$

B.3 Empirical Analysis

In this section of the Online Appendix, we provide details of the empirical analysis that we conduct to validate the assumptions made to develop a framework to select influencers and to sequence ads by influencers. Although, several recent studies in social media have examined the effects of user-generated content and firm-generated content, no study to the best of our knowledge has systematically analyzed the effect of multiple exposures (i.e., the same content being viewed by the audience multiple times) on engagement, which we operationalize via number of retweets. Against the above background, our first objective is to empirically analyze the relationship between the number of times the same tweet was posted (i.e., number of exposures of a tweet) and number of retweets. The second objective of the empirical study is to examine the impact of the gap between each tweet on the relationship between exposures and engagement levels.

With regards to the first objective, in line with the literature on ads in other contexts such as online and offline ads, we argue that the total number of clicks (or retweets in our context) increases monotonically with number of exposures but with diminishing returns, i.e., the relationship with the number of exposures and number of retweets is non-linear (or Number of Retweets $= f(\text{Exposures})$ is non-linear). Hence, we postulate that the marginal increase in the number of retweets per exposure is decreases with number of exposures (or $\frac{d}{dt} Number\ of\ Retweets = f(\text{Exposures})$), where t denotes the exposure number. In other words, we argue that the number of retweets per exposure decreases with exposures.

B.3.1 Data

To accomplish our objectives, we assemble a novel dataset. Our main data come from Twitter and consists of individual level tweets posted by 37 influencers. The data consists of 33,415 tweets posted by 37 influencers from August 2017 to May 2018. The 37 influencers were randomly picked based on their tweeting behavior. Due to data restrictions from Twitter, we collected tweets posted by these influencers every week, as historical data is not available from Twitter (i.e., we can not get tweets older than a week). Further, since the focus of this empirical study is to examine the effect of multiple exposures, we ignore the tweets that were not repeated by the influencers. Moreover, we winsorize the data at 99.5%, i.e., we drop the tweets over 23 exposures to remove the influence of outliers that can distort the mean values. Thus for our main analysis, the data consists of 18571 tweets. To summarize, our data consists of content of the tweet posted by influencers, number of retweets of the tweet, and the date and time stamp of the tweet. Furthermore, we collect data on personal attributes of an influencer including, the age of his/her twitter account, number of followers, number of friends, and the total number of tweets posted by the influencer. This data helps us control for individual specific heterogeneity of an influencer. The main variables used in this analysis are summarized in Table B.1. In the following subsections, we discuss the operationalization of the variables we use for the analysis.

B.3.2 Dependent Variable

The dependent variable in our study is the number of retweets received by a tweet posted by an influencer. Although an influencer might post multiple tweets with the same content, the followers of influencer may choose to retweet either the first tweet or the subsequent tweets. In this study, our aim is to uncover the relationship between the exposure and number of retweets per exposure. In particular, the dependent variable is *the number of retweets per exposure of an ad* or $\frac{d}{dt}Number\ of\ Retweets$ (denoted by RT in the model). We identify the number of retweets per exposure of each of the 18,571 tweets posted by 37 influencers.

B.3.3 Independent Variables

Our main explanatory variable is the number of exposures of a similar tweet. In order to identify similar tweets, we use an algorithm developed by Raffo (2017) to find similarities between two tweets. This algorithm is necessary because the influencers might slightly change the content of the ad when they repost it. Hence, to identify similar tweets, we employ *Matchit* algorithm. For the main analysis, we set the similarity parameter to 95%. That is, the algorithm will report that two tweets are identical to each other if it finds that at least 95% of the content of those two tweets are similar to each other. Moreover, our results are robust when the similarity parameter is set to 90%. The n^{th} similar tweet represents n

number of exposures (denoted by $Exposures$).

The second explanatory variable of interest is the gap between each tweet of similar content (denoted by Gap), which we measure by calculating the number of days between each tweet (i.e., date difference between the n^{th} and $(n-1)^{th}$ tweet). Since, we cannot calculate the gap for the first tweet, we assign the value of zero to the gap. However, we run re-run our models removing the first tweet as it does not have any gap to ensure robustness.

B.3.4 Control Variables

To control for heterogeneity among influencers we employ several influencer-specific control variables. First, we control for the popularity, which is measured by the number of followers of the influencer (denoted by $FollowerCount$). Next, we control for number of users that an influencer is following (denoted by $FriendsCount$). We also control for total number of tweets that the influencer posted on his twitter account (denoted by $StatusCount$). Finally, we control for the experience of influencer on Twitter, i.e., number of years since the influencer first opened his/her account (denoted by Age). Furthermore, we control for number of hashtags that the tweet consists of (denoted by HashtagsCount). Besides, to control for temporal shocks, we include time fixed effects (i.e., month dummies) and the hour that the tweet was posted (denoted by Hr) in our model. Next, we control for the first tweet (denoted by $First\ Tweet$). Finally, despite federal regulations to use $\#AD$ in the tweet, several influencers fail use it consistently. To control for any heterogeneity when the hashtag is used, we introduce indicator variable denoted by AD.

B.3.5 Model Specification

The dependent variable in our study, RT, is a count variable that takes only non-negative integer variable. Literature suggests that linear regression models are not appropriate for modeling count data and propose count data models such as Poisson and Negative Binomial models. However, we observe that the variance of number of retweets per exposure is nearly four times more than the mean. When the data is over-dispersed, Cameron and Trivedi

(2013) argue that the Poisson model will underestimate the standard errors, which in turn leads to high (and incorrect) levels of significance. Following Cameron and Trivedi (2013), an appropriate model to account for over-dispersion of count data is negative binomial model. We also perform additional analysis to check the significance of the over dispersion parameter. In particular, the p-value for the LR test of $H_o : \alpha = 0$ is < 0.000. Therefore, the estimate of α is greater than zero, which validates our assumption to use negative binomial model. We model number of retweets (RT) of ad i as follows.

$$Pr(RT = y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}(\frac{1}{1+\alpha\mu})^{1/\alpha}(\frac{\alpha\mu}{1+\alpha\mu})^{y} \tag{B.20}$$

$$\begin{aligned} \ln(\mu) &= \beta_0 + \beta_1 Exposures + \beta_2 Gap + \beta_3 Hr + \beta_4 HashtagCount + \beta_5 Ad \\ &\quad + \Delta Influencer\ Controls + \gamma MonthD + \epsilon \end{aligned} \tag{B.21}$$

where $\Gamma(.)$ denotes the gamma distribution, α is the over dispersion parameter, and μ is the mean of RT. The coefficient β_1 is of interest to us. More specifically, if the coefficient is negative, it implies that the marginal number of retweets decreases with number of exposures and vice-versa.

Next, to test the moderating effect of message spacing (i.e., *Gap*) on the relationship between exposures and retweets, we introduce the interaction term ($Exposure \times Gap$). Specifically, we run the following model to test the effect of *Gap*.

$$\begin{aligned} \ln(\mu) &= \beta_0 + \beta_1 Exposures + \beta_2 Gap + \beta_3 Exposure \times Gap + \beta_4 Hr + \beta_5 HashtagCount \\ &\quad + \beta_6 Ad + \Delta Influencer\ Controls + \gamma MonthD + \epsilon \end{aligned} \tag{B.22}$$

In the above model, the coefficient of the interaction effect, β_3, is of interest to us. More specifically, the sign of the coefficient denotes the effect of message spacing.

Table B.1: Variable Operationalization and Summary Statistics

Variable Name	Variable Operationalization	Mean	SD
RT	Number of Retweets per exposure	0.824	4.359
Exposures	Number of times the same content was posted	3.787	4.680
Gap	Number of days between each Tweet	3.404	9.849
First Tweet	Dummy Variable =1 if first exposure, else 0		
Hr	The Time of Tweet (hour)	12.659	6.776
Hashtag Count	Number of Hashtags in the Message	0.206	0.644
log(*Follower Count*)	Number of Followers of Influencer	9.449	0.366
log(*Friends Count*)	Number of Friends of Influencer	8.902	0.551
log(*Status Count*)	Total Number of Tweets Posted by Influencer	11.078	1.101
Age	Age of Influencer's Twitter Account	4.449	0.0933
Month	Month Dummy when the Tweet was Posted		
Ad	Dummy Variable =1 if #AD was used in the Tweet Content, else 0		

B.3.6 Results

Before estimating the results of the proposed model, we estimate the model without the interaction between *Gap* and *Exposures*. The estimates for this model are presented in the first column. Furthermore, to address the concerns about failure to meet the standard regression assumptions (i.e., *i.i.d.* errors), we account for heteroskedastic random errors by using the Whites sandwich covariance matrix, which are robust to heteroscedasticity (Wooldridge 2010). We report the robust standard errors in Table B.2. We present the estimates of the coefficients of parameters of our proposed model (i.e., Equation (B.22)) in Table B.2 (Column 2).

The estimates in the proposed model indicate that the number of *Exposures* has a negative and significant effect on number of retweets. In terms of the magnitude of the effect, we find that on average the number of retweets of a tweet decreases by 5.97% with additional exposure (based on the incident ratio of the negative binomial model specification). This result indicates that the number of retweets tends to increase with number of exposures. We illustrate the relationship between retweets and exposures in Figure B.3. Next, we find that the coefficient of Gap is negative but not statistically significant.

Turning our attention to the moderating role of *Gap*, we find that the coefficient of $Gap \times Exposure$ is significantly negative. To better understand the relationship between the number of exposures and retweets and the moderating effect of days message spacing, we conduct sensitivity analysis. In particular, we plot the relationship between *Exposure* and *Retweet Count* at different values of *Gap*. The results are shown in Figure B.4. This graph accounts for all the effects including the number of exposures, message spacing, influencer individual heterogeneity, and shows the total effect of exposure. As seen in the figure, the non-linear effect of exposures on engagement varies with message spacing (i.e., *Gap*). More specifically, with a higher gap between each tweet, the effect of exposure on the number of retweets diminishes.

B.3.7 Robustness Tests

To further validate our results from the count data model specification, we perform additional analyses with linear models. More specifically, we run the following linear model.

$$\begin{aligned}\ln(RT) &= \beta_0 + \beta_1 Exposure + \beta_2 Gap + \beta_3 Exposure \times Gap + \beta_4 Hr + \beta_5 HashtagCount \\ &\quad + \beta_6 Ad + \Delta Influencer\ Controls + \gamma MonthD + \epsilon\end{aligned}$$

The estimates for the above model are presented in Table B.3. While we find that the main effect of exposures is negative and significant in the model without interaction, the conditional effect of exposures is negative and significant when *Gap* $>=$ 1. Furthermore, we find that the coefficient of the interaction between exposures and message spacing is negative and significant. These results suggest that our findings from the main model are robust to different estimation strategies, providing further empirical support to our results.

Next, we re-estimate our full model with an alternative sample. For our main analysis we winsorized the data at 99.5% (i.e., we dropped tweets with more than 23 exposures) to remove the influence of outliers that can distort the mean values. However, we now re-run the models with full data to ensure robustness. The results of the estimation for the alternative sample are provided in Table B.4. Finally, we run re-run our models removing the first tweet

Table B.2: Regression Results

	(1)	(2)
	RT	*RT*
Exposures	-0.0757***	-0.0616***
	[0.00904]	[0.0119]
Gap	-0.0125***	-0.00281
	[0.00401]	[0.00727]
Exposures × *Gap*		-0.00486*
		[0.00276]
First Tweet	0.137	0.144
	[0.0919]	[0.0916]
Hr	-0.0145***	-0.0149***
	[0.00507]	[0.00508]
log(*Follower Count*)	1.916***	1.928***
	[0.276]	[0.277]
log(*Status Count*)	-2.081***	-2.081***
	[0.0595]	[0.0595]
log(*Friends Count*)	-0.0417	-0.0722
	[0.120]	[0.123]
Age	0.428***	0.423***
	[0.0316]	[0.0317]
Hashtag Count	0.139***	0.139***
	[0.0471]	[0.0471]
Ad	0.348	0.350
	[0.416]	[0.419]
Month Dummy 1	0.161	0.124
	[0.172]	[0.173]
Month Dummy 2	0.301**	0.268*
	[0.152]	[0.152]
Month Dummy 3	0.189	0.164
	[0.169]	[0.170]
Month Dummy 4	0.0674	0.0404
	[0.148]	[0.148]
Month Dummy 5	-0.952***	-0.987***
	[0.274]	[0.274]
Month Dummy 6	-35.20***	-34.27***
	[1.016]	[1.016]
Month Dummy 7	0.685	0.666
	[0.967]	[0.966]
Month Dummy 8	-0.100	-0.119
	[0.616]	[0.613]
Month Dummy 9	-0.610***	-0.623***
	[0.142]	[0.141]
Constant	1.759	1.952
	[1.410]	[1.397]
log(*alpha*)	0.965***	0.958***
	[0.0743]	[0.0750]
Log Likelihood	-13220.71	-13216.55
Pseudo R-Sq	0.2088	0.2091
N	18571	18571

Standard errors in brackets
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

as it does not have any value for the gap to ensure robustness. The results for this alternative specification are presented in Table B.5. The results of the above robustness tests suggest that our findings in the main model are robust to alternative sample and specification.

Table B.3: Results - Robustness Test 1

	(1)	(2)
	log(RT)	log(RT)
Exposures	-0.00163**	0.00224**
	[0.000684]	[0.00101]
Gap	-0.00668***	-0.000811
	[0.000733]	[0.00157]
Exposures × *Gap*		-0.00228***
		[0.000534]
Hr	-0.00689***	-0.00682***
	[0.000536]	[0.000534]
First Tweet	0.00238	0.00470
	[0.00850]	[0.00849]
log(*Follower Count*)	0.602***	0.604***
	[0.0594]	[0.0596]
log(*Status Count*)	-0.398***	-0.402***
	[0.0131]	[0.0133]
log(*Friends Count*)	-0.167***	-0.173***
	[0.0445]	[0.0447]
Age	0.0606***	0.0593***
	[0.00260]	[0.00259]
Hashtag Count	0.0353***	0.0338***
	[0.00848]	[0.00849]
Ad	-0.189**	-0.194**
	[0.0917]	[0.0920]
Constant	0.274	0.286**
	[.]	[0.133]
R-Sq	0.354	0.357
N	18571	18571

Month Dummies were included in all models.
The estimates for Month Dummies are not reported for brevity.
Standard errors in brackets
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B.3.8 Cluster Analysis

To understand the strategies of influencers who post ads on Twitter, we classified influencers using "K-Means" technique. We classify the influencers into four clusters. The mean number of retweets, days between, and number of exposure for each of these clusters is presented in Table B.6. As seen in the table, the strategy of influencers in cluster 1 is to post the same message two times everyday and post the same message an average of 18 times. On the other hand, the strategy of influencers in cluster 2 is to post the same message every

Table B.4: Results - Robustness Test 2

	(1)	(2)	(3)	(4)
	RT	RT	$\log(RT)$	$\log(RT)$
Exposures	-0.00614**	-0.00422*	-0.0000135	0.0000137
	[0.00272]	[0.00226]	[0.0000368]	[0.0000402]
Gap	-0.0107***	0.00944*	-0.00658***	-0.00617***
	[0.00394]	[0.00503]	[0.000720]	[0.000792]
Exposures × *Gap*		-0.0101***		-0.000136*
		[0.00207]		[0.0000751]
First Tweet	0.317***	0.269***	0.0124	0.0118
	[0.0918]	[0.0922]	[0.00803]	[0.00802]
Hr	-0.0145***	-0.0148***	-0.00573***	-0.00570***
	[0.00514]	[0.00509]	[0.000455]	[0.000455]
log(*Follower Count*)	1.969***	1.946***	0.580***	0.579***
	[0.271]	[0.276]	[0.0589]	[0.0589]
log(*Status Count*)	-2.095***	-2.099***	-0.401***	-0.401***
	[0.0635]	[0.0618]	[0.0128]	[0.0128]
log(*Friends Count*)	-0.0180	-0.0905	-0.175***	-0.175***
	[0.125]	[0.127]	[0.0437]	[0.0437]
Age	0.443***	0.428***	0.0637***	0.0638***
	[0.0309]	[0.0308]	[0.00235]	[0.00235]
Hashtag Count	0.192***	0.186***	0.0339***	0.0338***
	[0.0467]	[0.0469]	[0.00720]	[0.00720]
Ad	0.312	0.303	-0.200**	-0.200**
	[0.424]	[0.430]	[0.0910]	[0.0910]
Constant	0.800	1.904	-0.184*	0.480***
	[1.261]	[1.323]	[0.110]	[0.123]
log(*alpha*)	0.997***	0.976***		
	[0.0727]	[0.0739]		
N	21853	21853	21853	21853

Month Dummies were included in all models.
The estimates for Month Dummies are not reported for brevity.
Standard errors in brackets
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

two days on average. The mean number of exposures for the influencers in cluster 2 is 2.4. Likewise, the mean number of exposures for the influencers in cluster 3 is 2.24. However, the strategy of these influencers is to post the message about every 6 days on average. From our data set, the strategy of the influencers in cluster 2 seems to be associated with higher number of retweets.

B.3.9 Summary of Empirical Analysis

In this section, I briefly summarize the main findings from the empirical study.

(a) First, our results suggest a non-linear increasing relationship between exposures and retweets.

Table B.5: Results - Robustness Test 3

	(1)	(2)	(3)	(4)
	RT	*RT*	log(*RT*)	log(*RT*)
Exposures	-0.0786***	-0.0591***	-0.00325***	0.000120
	[0.00886]	[0.0116]	[0.000661]	[0.000955]
Gap	-0.0123***	0.000644	-0.00675***	-0.00167
	[0.00461]	[0.00770]	[0.000753]	[0.00153]
Exposures × *Gap*		-0.00678**		-0.00198***
		[0.00289]		[0.000505]
Hr	-0.0208***	-0.0216***	-0.00686***	-0.00679***
	[0.00604]	[0.00603]	[0.000617]	[0.000614]
log(*Follower Count*)	2.590***	2.628***	0.589***	0.592***
	[0.312]	[0.321]	[0.0648]	[0.0650]
log(*Status Count*)	-2.163***	-2.165***	-0.391***	-0.396***
	[0.0808]	[0.0824]	[0.0146]	[0.0149]
log(*Friends Count*)	-0.404**	-0.481***	-0.152***	-0.162***
	[0.174]	[0.183]	[0.0492]	[0.0495]
Age	0.371***	0.360***	0.0696***	0.0678***
	[0.0415]	[0.0405]	[0.00313]	[0.00311]
Hashtag Count	0.176***	0.174***	0.0354***	0.0336***
	[0.0549]	[0.0547]	[0.00935]	[0.00937]
Ad	0.233	0.235	-0.283***	-0.290***
	[0.608]	[0.611]	[0.101]	[0.102]
Constant	-0.0762	0.365	-0.221	0.207
	[1.469]	[1.451]	[0.263]	[0.149]
log(*alpha*)	0.727***	0.713***		
	[0.0890]	[0.0892]		
N	12552	12552	12552	12552

Month Dummies were included in all models.
The estimates for Month Dummies are not reported for brevity.
Standard errors in brackets
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.6: Cluster Analysis

	Mean	*Mean*	*Mean*
Cluster	*Retweet*	*Days Between*	*Exposure*
1	0.128	1.743	14.406
2	3.246	0.561	1.404
3	1.303	4.009	2.216
4	0.7515	7.228	1.685

(b) Next, the coefficient of the interaction between days between each tweet and exposure is negative and significant. This indicates that days between moderates the relationship between exposure and retweets. We conduct sensitivity analysis to illustrate this moderating effect and the results are shown in Figure B.4. This figure illustrates the interaction effect by plotting predicted retweets corresponding to the frequency of

Figure B.3: Empirical Relationship Between Total Number of Retweets and Exposure

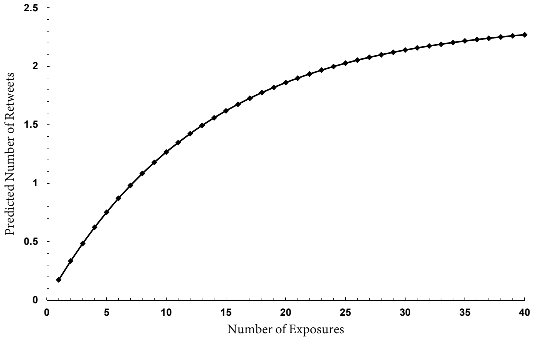
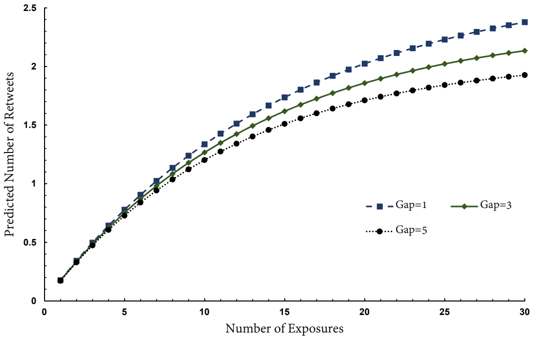


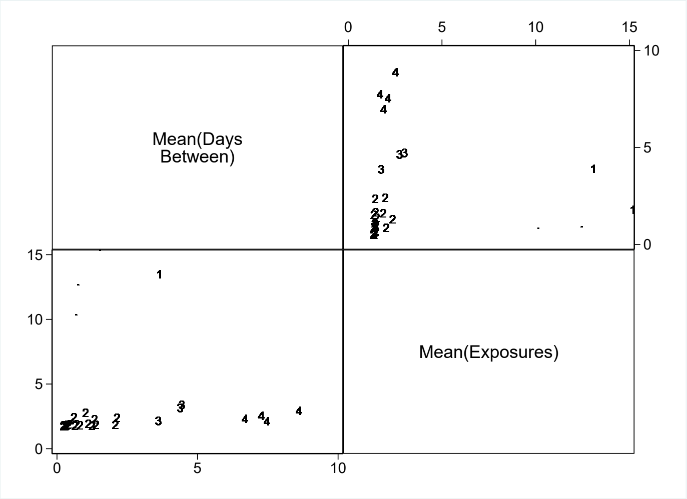
Figure B.4: Empirical Relationship Between Total Number of Retweets, Days Between, and Exposure



exposure at different levels of moderating variables, i.e., days between each tweet.

(c) To glean insights on how different influencers post their ads, we perform exploratory analysis. In particular, we classify influencers into 4 clusters using "K-Means" classification technique. The results are summarized in Figure B.5. The results indicate the strategy of different influencers.

Figure B.5: K-Means Clustering of Influencers



B.4 Experiment Setting for Selection Model

The first level influence was set to $p_{ij} \in$ {Low, High}. Probabilities (i.e., p_{ij}) in *Low* setting were drawn from $Uniform(0,1)/20$ and in High setting were drawn from $Uniform(0,1)/6$ distributions. Similarly, the second level influence was set to $p_{kj} \in$ {Low, High}. Probabilities (i.e., p_{kj}) in *Low* setting were drawn from $Uniform(0,1)/24$ and in High setting were drawn from $Uniform(0,1)/6$ distributions. Next, the number of edges between influencers in set W and followers in set V was set to, a_{ij} ∈{Low Medium, High}. *Low* is drawn from X^5, *Medium* is drawn from X^4, and *High* is drawn from X^3, where $X \sim Uniform[0,1]$. Finally, the number of edges between the followers of influencers is set to a_{kj} ∈{Low Medium, High}. *Low* is drawn from X^6, *Medium* is drawn from X^4, and *High* is drawn from X^2, where $X \sim Uniform[0,1]$. Thus, we generated $3 \times 2 \times 2 \times 3 = 36$ combinations.

B.5 Scheduling Model

In this section, we present the scheduling model that we describe in Section 3.5. Before presenting the model, the list of parameters and decision variables used in the model are presented in Table B.7.

Table B.7: List of Parameters and Variables - Scheduling Model

Symbol		**Description**
Parameters:		
a_{ij}	=	1 if user $j \in V$ is a direct follower of $i \in W$; 0 otherwise.
p_j	=	average probability that user $j \in V$ retweets a message from influencers in W.
b_j	=	total benefit to the firm when the tweet reaches user $j \in V$ through $i \in W$.
d_j	=	benefit to the firm when the tweet reaches one follower of user $j \in V$.
F_j	=	number of followers of user $j \in V$.
B_j	=	the benefit to the firm when user j retweets a message where $B_j = (b_j + d_j F_j)$.
c_i	=	cost to the firm when influencer $i \in W$ posts an ad.
l	=	Number of groups. We assume 5 different groups. The first group posts the ad everyday, the second group posts every other day, the third group posts every three days, the fourth group posts the ad once a week, and the fifth group posts once in 12 days.
m	=	Number of modes. In this problem instance, the number of modes, $m = 24$.
f_{jr}	=	the cumulative probability of user j not retweeting when receiving r number of exposures. $f_{jr} = (1 - p_j)^r$, where r takes the values of $r = 0, 1, 2, \ldots, r_m$.
$\mathscr{E}$	=	Minimum number of Retweets in Time Period t.
M	=	Large Number.
Decision Variables:		
x_{im}	=	1 if $i \in W$ is chosen in mode m; 0 otherwise.
Y_{it}^{l}	=	1 if $i \in W$ in group l posts a message in time period t; 0 otherwise.
K_{jt}	=	number of posts or messages received by user $j \in V$ in time period t.
λ_{jt}	=	1 if $K_{jt} > 0$; 0 otherwise.
L_{jt}	=	number of posts or messages received by user $j \in V$ until time period t.
θ_{jt}	=	number consecutive times user $j \in V$ did not receive a message in time period t.
δ_{jt}	=	number days after which a user $j \in V$ receives a message.
α_{jt}^{n}	=	1 if a user gets a message after not receiving a message for n previous consecutive periods; 0 otherwise.
U_{rjt}	=	dummy variable that denotes total number of received by user $j \in V$ until time period t. E.g., $U_{rjt} = 1$ if user j receives a total of r exposures until time period t.
h_{jt}	=	the cumulative probability of user $j \in V$ not retweeting a message in time period t.
g_{jt}	=	the cumulative probability of user $j \in V$ retweeting a message in time period t.
D_{jt}	=	Discount factor for cumulative probability of user $j \in V$ retweeting a message in time period t.
γ_{jt}	=	marginal increase in cumulative probability (without discount factor) of user $j \in V$ retweeting a message in time period t.
e_{jt}	=	marginal increase in cumulative probability (with discount factor) of user $j \in V$ retweeting a message in time period t.

The mathematical formulation is formally presented below.

Mathematical Formulation - Scheduling Model:

$$\text{Max} \;\; \Pi = \sum_{t=1}^{T} \sum_{j \in V} e_{jt} B_j - \sum_{i \in W} \sum_{l=1}^{5} \sum_{t=1}^{T} c_i Y_{it}^{l} \tag{B.23}$$

Subject to:

$$\sum_{m=1}^{12} x_{im} \leq 1, \qquad \forall i \in W \tag{B.24}$$

$$\sum_{t=1}^{T} Y_{it}^{1} = T x_{i1}, \qquad \forall i \in W \tag{B.25}$$

$$Y_{i1}^{2} + Y_{i3}^{2} + Y_{i5}^{2} + Y_{i7}^{2} + Y_{i9}^{2} + Y_{i11}^{2} = 6x_{i2}, \qquad \forall i \in W \tag{B.26}$$

$$Y_{i2}^{2} + Y_{i4}^{2} + Y_{i6}^{2} + Y_{i8}^{2} + Y_{i10}^{2} + Y_{i12}^{2} = 6x_{i3}, \qquad \forall i \in W \tag{B.27}$$

$$Y_{i1}^{3} + Y_{i4}^{3} + Y_{i7}^{3} + Y_{i10}^{3} = 4x_{i4}, \qquad \forall i \in W \tag{B.28}$$

$$Y_{i2}^{3} + Y_{i5}^{3} + Y_{i8}^{3} + Y_{i11}^{3} = 4x_{i5}, \qquad \forall i \in W \tag{B.29}$$

$$Y_{i3}^{3} + Y_{i6}^{3} + Y_{i9}^{3} + Y_{i12}^{3} = 4x_{i6}, \qquad \forall i \in W \tag{B.30}$$

$$Y_{i1}^{4} + Y_{i7}^{4} = 2x_{i7}, \qquad i \in W \tag{B.31}$$

$$Y_{i2}^{4} + Y_{i8}^{4} = 2x_{i8}, \qquad i \in W \tag{B.32}$$

$$Y_{i3}^{4} + Y_{i9}^{4} = 2x_{i9}, \qquad i \in W \tag{B.33}$$

$$Y_{i4}^{4} + Y_{i10}^{4} = 2x_{i10}, \qquad i \in W \tag{B.34}$$

$$Y_{i5}^{4} + Y_{i11}^{4} = 2x_{i11}, \qquad i \in W \tag{B.35}$$

$$Y_{i6}^{4} + Y_{i12}^{4} = 2x_{i12}, \qquad i \in W \tag{B.36}$$

$$Y_{i1}^{5} = x_{i13}, \qquad i \in W \tag{B.37}$$

$$Y_{i2}^{5} = x_{i14}, \qquad i \in W \tag{B.38}$$

$$Y_{i3}^{5} = x_{i15}, \qquad i \in W \tag{B.39}$$

$$Y_{i4}^{5} = x_{i16}, \qquad i \in W \tag{B.40}$$

$$Y_{i5}^{5} = x_{i17}, \qquad i \in W \tag{B.41}$$

$$Y_{i6}^{5} = x_{i18}, \qquad i \in W \tag{B.42}$$

$$Y_{i7}^{5} = x_{i19}, \qquad i \in W \tag{B.43}$$

$$Y_{i8}^{5} = x_{i20}, \qquad i \in W \tag{B.44}$$

$$Y_{i9}^{5} = x_{i21}, \qquad i \in W \tag{B.45}$$

$$Y_{i10}^{5} = x_{i22}, \qquad i \in W \tag{B.46}$$

$$Y_{i11}^{5} = x_{i23}, \qquad i \in W \tag{B.47}$$

$$Y_{i12}^{5} = x_{i24}, \qquad i \in W \tag{B.48}$$

$$K_{jt} = \sum_{i \in W} \sum_{l=1}^{5} a_{ij} Y_{it}^{l}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.49}$$

$$\lambda_{jt} \leq M K_{jt}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.50}$$

$$K_{jt} \leq M \lambda_{jt}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.51}$$

$$\theta_{j0} = 0, \qquad \forall j \in V \tag{B.52}$$

$$\theta_{jt} = (\theta_{jt-1} + 1)(1 - \lambda_{jt}), \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.53}$$

$$\delta_{jt} = (\theta_{jt-1})(\lambda_{jt}), \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.54}$$

$$\delta_{jt} = \sum_{n=1}^{12} n \alpha_{jt}^{n}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.55}$$

$$\sum_{n=1}^{12} \alpha_{jt}^{n} \leq 1, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.56}$$

$$\begin{aligned} D_{jt} = .05\alpha_{jt}^{1} + .10\alpha_{jt}^{2} + 0.15\alpha_{jt}^{3} + .20\alpha_{jt}^{4} \\ + .25\alpha_{jt}^{5} + .30\alpha_{jt}^{6}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \end{aligned} \tag{B.57}$$

$$L_{jt} = \sum_{\tau=1}^{t} K_{j\tau} \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.58}$$

$$L_{jt} = \sum_{r=0}^{r_m} r U_{rjt}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.59}$$

$$\sum_{k=0}^{r_m} U_{kjt} = 1, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.60}$$

$$h_{jt} = \sum_{r=0}^{r_m} f_{jr} U_{rjt}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.61}$$

$$g_{jt} = 1 - h_{jt}, \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.62}$$

$$g_{j0} = 0, \qquad \forall j \in V \tag{B.63}$$

$$\gamma_{jt} = (g_{jt} - g_{j,t-1}), \qquad \forall j \in V, \;\; t = 1, 2, \ldots, T \tag{B.64}$$

$$\sum_{j \in V} \sum_{t=1}^{3} \gamma_{jt} \geq \mathscr{E}_1, \tag{B.65}$$

$$\sum_{j \in V} \sum_{t=4}^{6} \gamma_{jt} \geq \mathscr{E}_2, \tag{B.66}$$

$$\sum_{j \in V} \sum_{t=7}^{9} \gamma_{jt} \geq \mathscr{E}_3, \tag{B.67}$$

$$\sum_{j \in V} \sum_{t=10}^{1} 2\gamma_{jt} \geq \mathscr{E}_4, \tag{B.68}$$

$$e_{jt} = \gamma_{jt}(1 - D_{jt}), \qquad \forall j \in V,\ \ t = 1, 2, \ldots, T \tag{B.69}$$

$$x_{im}, Y_{it}, U_{rjt} \in \{0, 1\}, \qquad \forall i \in W, t = 1, 2, \ldots, T. \tag{B.70}$$

$$\alpha_{jt}, \quad \lambda_{jt} \in \{0, 1\}, \qquad \forall j \in V, t = 1, 2, \ldots, T. \tag{B.71}$$

$$l = 1, 2, 3, 4; \ \ m = 1, 2, \ldots, 12. \tag{B.72}$$

In the above model, the objective function, Π (i.e., Equation B.23), maximizes the total benefit to the firm minus the cost of hiring influencers. Constraint (B.24) ensures that an influencer, if selected, is assigned to only one mode. Constraints (B.25)-(B.48) keep track of the posts by influencers depending on which group (and mode) they are assigned to. Constraint (B.49) represents the number of exposures received by user $j \in V$ in time period t. Constraints (B.50)-(B.54) checks whether user j receives an ad on consecutive days, if not, these constraints keep track of number of days that the user did not receive a message from the influencers. Constraints (B.55)-(B.57) generate the required discount factors for not sending an ad to the users on consecutive days. Constraint (B.58) represents the total number of exposures received by user $j \in V$ till time period t. Constraints (B.59) and (B.60) are utilized to generate dummy variables U_{kjt}. These constraints allow us to linear the non-linear relationship between number of exposures and the probability of retweet. Constraint (B.61) calculates the cumulative probability of user j not retweeting till time period t. Constraint (B.62) tracks the cumulative probability that user $j \in V$ retweets a message till time period t and Constraint (B.64) tracks the marginal increase in cumulative probability in time period t. Constraints (B.65)-(B.68) ensure that the firm gets a minimum of $\mathscr{E}$ number of retweets in the n^{th} time period. Constraint (B.69) discounts the marginal benefit, i.e., penalizes for not reaching the users on consecutive days.

Explanation of Constraints (B.59)-(B.61): We assume $f_{jr} = (1 - p_j)^r$ and r takes the values of $r = 0, 1, 2, \ldots, r_m$. Assume that values of f_{jr} can be found a prior for all possible values of r, where r_m is the maximum possible values for L_{jt}'s.

$h_{jt} = \sum_{r=0}^{r_m} f_{jr} U_{rjt}$

$L_{jt} = \sum_{r=0}^{r_m} r U_{rjt}$

$\sum_{k=0}^{r_m} U_{kjt} = 1$

B.5.1 Alternative Formulations to Linearize Non-Linear Constraints

Non-linear Constraints: In the above formulation, there are three non-linear constraints namely, Constraints (B.53), (B.54), and (B.69). We now provide discussion on how to convert these non-linear equations into linear form, which will allows us to employ a linear optimization solver (i.e., *CPLEX*) rather than a non-linear solver.

Non-linear constraint (B.53):

$\theta_{jt} = (\theta_{jt-1} + 1)(1 - \lambda_{jt})$.

We can linearize this constraint using the below set of constraints.

$\theta_{j1} = (1 - \lambda_{jt})$

$\theta_{jt} \leq M(1 - \lambda_{jt}), \quad \forall j \in V, \; t = 2...T$

$\theta_{jt} \leq \theta_{jt-1} + 1, \quad \forall j \in V, \; t = 2...T$

$\theta_{jt} \geq \theta_{jt-1} + 1 - M\lambda_{jt}, \; \forall j \in V, \; t = 2...T$

$\theta_{jt} \geq 0, \quad \forall j \in V, \; t = 2...T$

Non-linear constraint (B.54):

$\delta_{jt} = (\theta_{jt-1})(\lambda_{jt})$.

We can linearize this constraint using the below set of constraints.

$\delta_{j1} = 0$

$\delta_{jt} \leq M(\lambda_{jt}), \qquad \forall j \in V, \;\; t = 2...T$

$\delta_{jt} \leq \theta_{jt-1}, \qquad \forall j \in V, \;\; t = 2...T$

$\delta_{jt} \geq \theta_{jt-1} - M(1 - \lambda_{jt}), \qquad \forall j \in V, \;\; t = 2...T$

$\delta_{jt} \geq 0, \qquad \forall j \in V, \;\; t = 2...T.$

Non-linear constraint (B.69):

$e_{jt} = \gamma_{jt}(1 - D_{jt}).$

In this constraint, we first substitute D_{jt} with the right hand side of constraint (eq6.6).

$e_{jt} = \gamma_{jt}(1 - .05\alpha^1_{jt} + .10\alpha^2_{jt} + 0.15\alpha^3_{jt} + .20\alpha^4_{jt} + .25\alpha^5_{jt} + .30\alpha^6_{jt}).$

In the above equation, we have six non-linear functions (i.e., $\gamma_{jt} \times \alpha^1_{jt}, \gamma_{jt} \times \alpha^2_{jt}, \ldots \gamma_{jt} \times \alpha^6_{jt},$). Let $e^n_{jt} = \gamma_{jt} \times \alpha^n_{jt}$, where $n = 1, 2, \ldots, 6$). Thus, $e_{jt} = \gamma_{jt} - (.05e^1_{jt} + .10e^2_{jt} + 0.15e^3_{jt} + .20e^4_{jt} + .25e^5_{jt} + .30e^6_{jt})$.

We can linearize the six non-linear terms using the below set of constraints. For brevity, we only present the linearization method only for one of the non-linear terms (i.e., e^1_{jt}).

$e^1_{jt} \leq \alpha^1_{jt}, \qquad \forall j \in V, \;\; t = 2...T$

$e^1_{jt} \leq \gamma_{jt}, \qquad \forall j \in V, \;\; t = 2...T$

$e^1_{jt} \geq \gamma_{jt} + \alpha^1_{jt} - 1, \qquad \forall j \in V, \;\; t = 2...T \; e^1_{jt} \geq 0, \qquad \forall j \in V, \;\; t = 2...T.$

B.5.2 Experiment Setting - Without Minimum Engagement Level Constraint:

We begin by specifying the default setting of the test bed. The number of influencers is 8 (i.e., $|W| = 8$). Size of set $|V|$ is 200.[1]

- % First level followers, $a_{ij} \in$ {Low, High}.
- Probability of retweet $p_j \in$ {Low, High}.
- Cost of hiring an influencer $C_i \in$ {Low, High}.
- % Overlap of followers among influencers $\in$ {Low, High}.

[1] This represents a relatively small problem size. This assumption is required to ensure that the solver is able to provide us with an *optimal* solution for MIP. Furthermore, we assume only 4 groups (and 12 modes) instead of 5 groups, i.e., we ignore modes 13-24).

For each of the instances created using above criteria, we generate unique instances by drawing the values of P_j, a_{ij}, and expected benefit B_j which is drawn from a continuous uniform distribution ($Uniform(100, 200)$). To summarize, we simulated a set of 96 instances to ensure that our test bed covers different scenarios. Table B.14 summarizes the scenario where the overlap of followers is high among all the influencers (i.e., the influencers are likely to have several similar followers) and all the influencers have either low or high levels of followers. Table B.15 summarizes the scenario where there are two categories of influencers. The influencers are categorized as Set 1 and Set 2. In the scenarios tested in Table B.15, the overlap of followers is high among all the influencers (i.e., the influencers are likely to have several similar followers) and all the influencers have either low or high levels of followers as noted in the table. Finally, the scenarios considered in Table B.16, in addition to varying the levels of followers and costs of hiring influencers among influencers in set 1 and set 2, we also vary the %overlap of followers for influencers. Specifically, high overlap denotes the scenario where influencers are likely to have the same set of followers. Specifically, in these scenarios, influencers 1 to 4 have a high overlap of followers among themselves, while influencers 5 and 6 have a high overlap of followers among themselves, and influencers 7 and 8 have a high overlap of followers among themselves. Furthermore, there is no overlap of followers between the set 1 (i.e., influencers 1 to 4) and set 2 (i.e., influencers 5 to 8).

B.5.2.1 Computational Experiment - With Minimum Engagement Level Constraint

We now perform computational experiments by including the constraints pertaining to the minimum engagement levels. The default setting of the test bed is as follows. The number of influencers is 8 (i.e., $|W| = 8$). Size of set $|V|$ is 200 and $T = 6$. We set the number of first level followers high for all influencers and the probability of retweet p_j is also set to high values. Whereas the cost of hiring an influencer c_i is Uniform (125,200). To generate insights, we change the network type, value of $\mathscr{E}$, i.e., minimum number of retweets required by the firm, and the frequency, i.e., how often do we need to reach the target engagement levels. With respect to network type, we assume three different networks (namely, Type 1, Type 2, and Type 3) and they are depicted in Figures B.6-B.6. We generate unique instances by drawing the values of p_j, a_{ij}, and expected benefit B_j is drawn from continuous uniform distribution ($Uniform(60, 80)$).

Figure B.6: Computational Experiment - With Minimum Engagement Level Constraint: Type 1

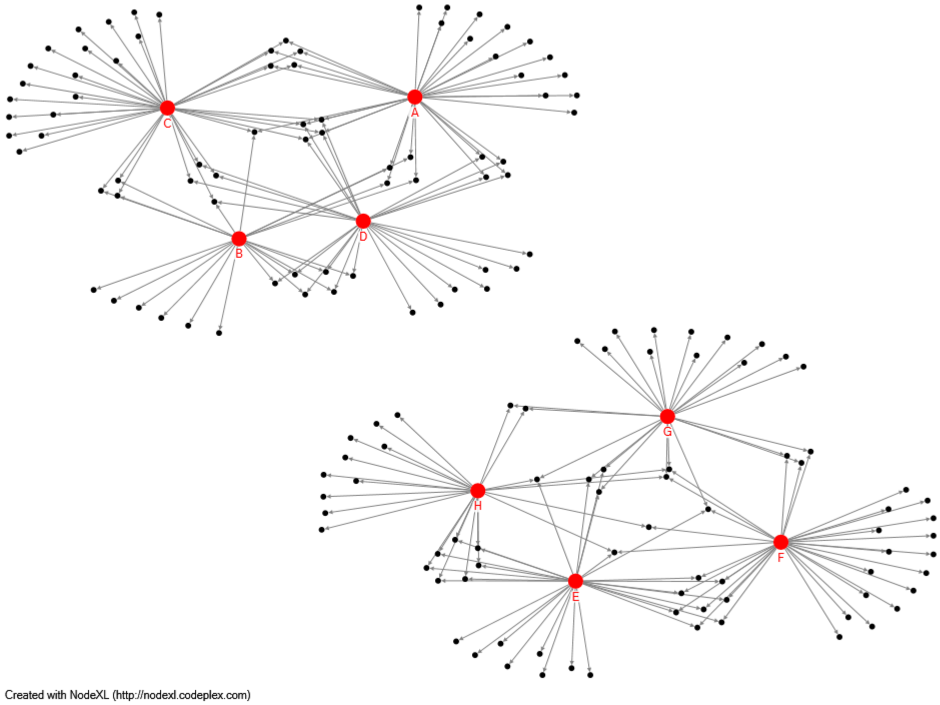


Figure B.7: Computational Experiment - With Minimum Engagement Level Constraint: Type 2

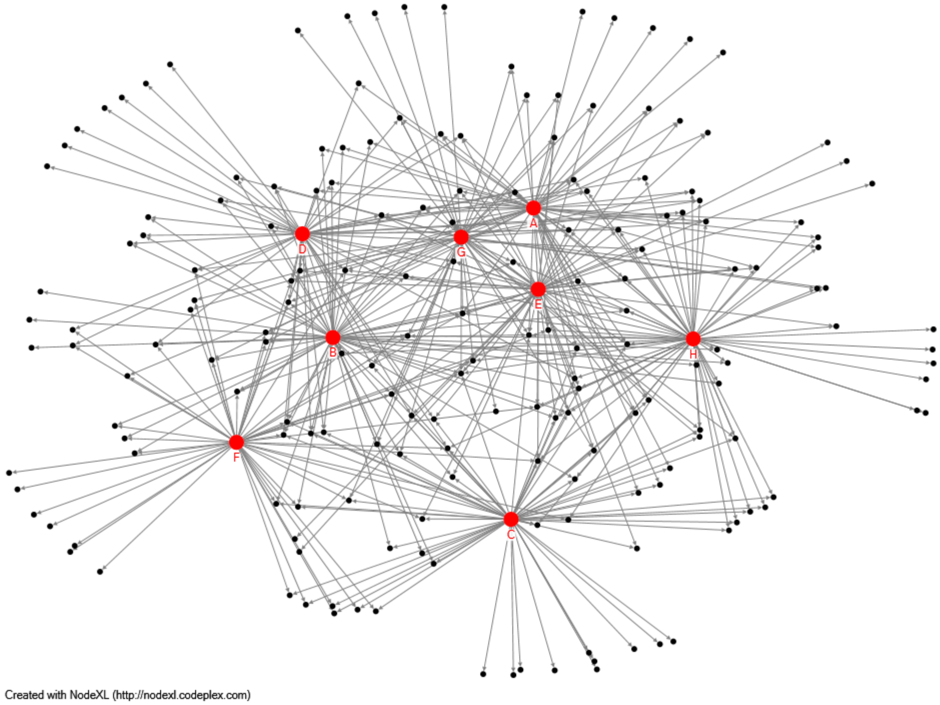
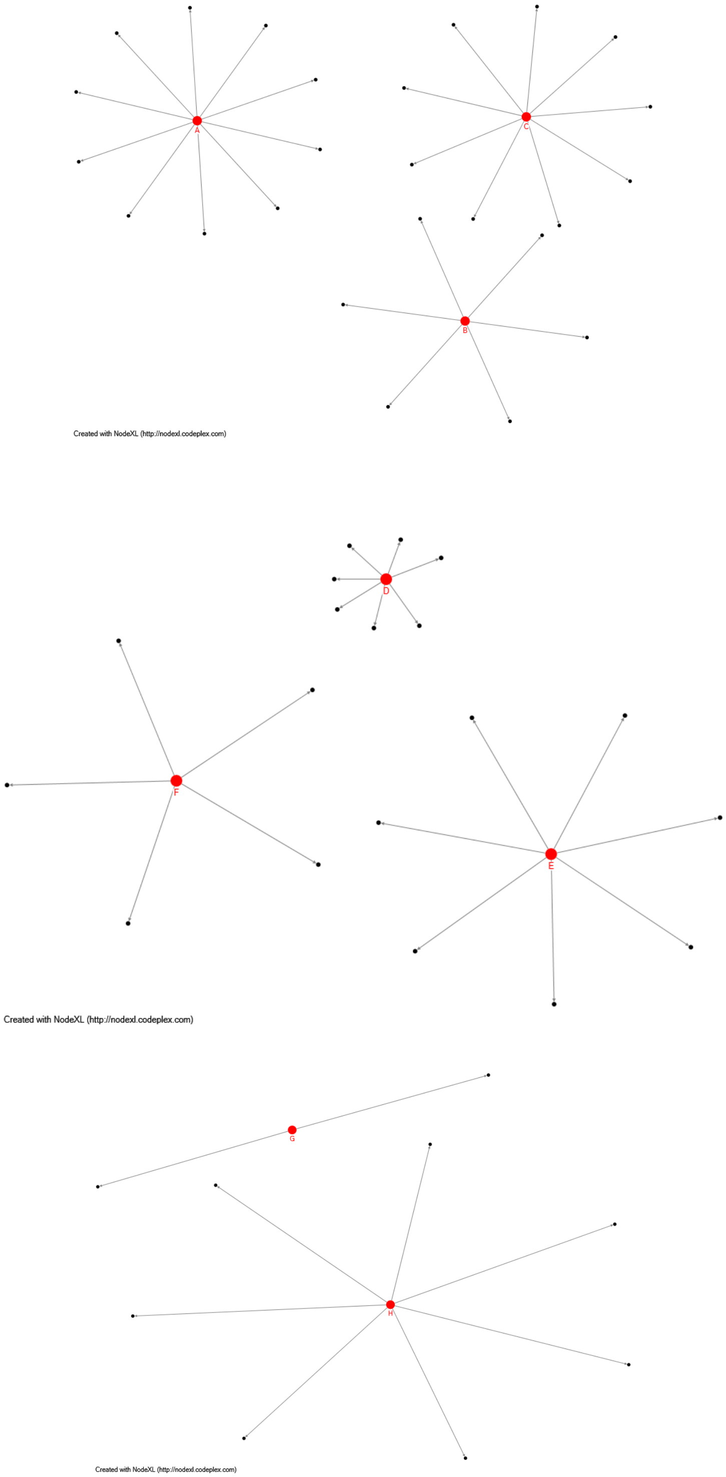


Figure B.8: Computational Experiment - With Minimum Engagement Level Constraint: Type 3


A
C
B
Created with NodeXL (http://nodexl.codeplex.com)
D
F
E
Created with NodeXL (http://nodexl.codeplex.com)
G
H
Created with NodeXL (http://nodexl.codeplex.com)

B.6 Tables

Table B.8: Computational Evaluation of Upper Bound Model ($M2L^{UB}$)

- % Optimality gap is calculated using the following formula. $\%Gap = \frac{\Psi_2^{UB} - \Psi_2}{\Psi_2}$, where Ψ_2^{UB} the objective function value of the upper bound problem and Ψ_2 is the objective function value of the main model.
- *Average optimality gap* is the average over 30 instances and over $t = 1..10$, i.e., average over 300 random instances.
- *Maximum optimality gap* denotes the maximum gap over 30 instances and over $t = 1..10$, i.e., maximum gap over 300 randomly generated instances.
- *% Same Solution* represents the percentage of runs where the solution set (i.e., optimal influencer set) obtained from the upper bound model is same as the solution set of the Model $M2L$.

Instances				**%Optimality Gap**		**% Same Solution**
p_{kj}	a_{ij}	p_{ij}	a_{kj}	Average % Optimality Gap	Maximum % Optimality Gap	
High	High	High	High	3.47%	4.49%	94.33%
			Medium	2.44%	3.16%	95.67%
			Low	1.07%	1.39%	98.67%
		Low	High	4.52%	5.07%	96.67%
			Medium	2.95%	3.53%	97.67%
			Low	1.29%	1.54%	96.00%
Low	High	High	High	0.61%	0.78%	94.67%
			Medium	0.24%	0.30%	99.33%
			Low	0.08%	0.10%	100.00%
		Low	High	0.72%	0.86%	99.00%
			Medium	0.29%	0.34%	100.00%
			Low	0.09%	0.11%	100.00%
High	High | Low	High | Low	High	3.45%	4.42%	98.67%
			Medium	2.57%	3.14%	99.33%
			Low	1.28%	1.96%	98.67%
		Low | High	High	2.73%	4.19%	97.33%
			Medium	2.25%	3.19%	98.67%
			Low	1.10%	1.53%	97.33%
Low	High | Low	High | Low	High	0.61%	0.72%	99.67%
			Medium	0.26%	0.34%	98.67%
			Low	0.09%	0.52%	99.33%
		Low | High	High	0.53%	0.69%	98.67%
			Medium	0.24%	0.33%	96.67%
			Low	0.07%	0.29%	97.67%
High	High|Medium|Low	High|Medium|Low	High	2.03%	3.71%	98.00%
			Medium	1.93%	3.90%	98.67%
			Low	0.89%	1.30%	98.67%
		Low|Medium|High	High	4.02%	4.84%	99.67%
			Medium	2.95%	3.41%	99.67%
			Low	1.29%	1.52%	99.33%
Low	High|Medium |Low	High|Medium|Low	High	0.45%	0.65%	98.00%
			Medium	0.21%	0.30%	99.33%
			Low	0.07%	0.98%	99.33%
		Low|Medium|High	High	0.68%	0.78%	100.00%
			Medium	0.30%	0.35%	99.33%
			Low	0.09%	0.14%	98.00%

Table B.9: Computational Evaluation of Lower Bound Model ($M1L$)

- % Optimality gap is calculated using the following formula. $\%Gap = \frac{\Psi_2 - \Psi_1}{\Psi_2}$, where Ψ_1 denotes the objective function value of the lower bound problem and Ψ_2 is the objective function value of the main model.
- *Average optimality gap* is the average over 30 instances and over $t = 1..10$, i.e., average over 300 instances.
- *Maximum optimality gap* denotes the maximum gap over 30 instances and over $t = 1..10$, i.e., maximum gap over 300 instances.
- *% Same Solution* represents the percentage of runs where the solution set (i.e., optimal influencer set) obtained from the lower bound model is same as the solution set of Model $M2L$.

Instances				**%Optimality Gap**		**% Same Solution**
p_{kj} (i.e., second level influence)	a_{ij} (i.e.,number of first level followers)	p_{ij} (i.e., first level influence)	a_{kj} (i.e., number of edges within V)	Average % Gap	Maximum % Gap	
High	High	High	High	79.28%	85.98%	13.67%
			Medium	50.65%	56.73%	35.33%
			Low	21.50%	27.00%	52.67%
		Low	High	81.74%	86.39%	13.67%
			Medium	52.58%	57.50%	37.67%
			Low	22.53%	27.25%	52.67%
Low	High	High	High	51.60%	60.94%	25.67%
			Medium	20.76%	24.94%	65.67%
			Low	6.41%	8.47%	84.33%
		Low	High	53.42%	61.28%	26.33%
			Medium	21.68%	25.35%	66.33%
			Low	6.74%	8.57%	79.67%
High	High | Low	High | Low	High	79.82%	83.60%	37.00%
			Medium	52.60%	58.62%	58.00%
			Low	24.82%	29.92%	69.33%
		Low | High	High	76.67%	84.59%	25.67%
			Medium	49.92%	59.92%	39.00%
			Low	23.10%	28.16%	65.67%
Low	High | Low	High | Low	High	52.48%	56.63%	53.67%
			Medium	22.11%	26.25%	79.33%
			Low	7.62%	9.66%	90.00%
		Low | High	High	49.70%	56.09%	40.33%
			Medium	20.97%	24.54%	61.67%
			Low	5.69%	6.99%	88.33%
High	High | Medium | Low	High | Medium | Low	High	72.51%	81.43%	66.67%
			Medium	47.68%	55.71%	78.67%
			Low	20.19%	24.94%	85.67%
		Low | Medium | High	High	80.71%	83.56%	56.67%
			Medium	53.36%	57.41%	67.33%
			Low	22.97%	25.70%	76.00%
Low	High | Medium | Low	High | Medium | Low	High	47.53%	54.76%	77.00%
			Medium	19.85%	24.39%	89.00%
			Low	6.06%	8.70%	95.00%
		Low | Medium | High	High	53.30%	56.51%	67.67%
			Medium	22.58%	25.30%	83.67%
			Low	6.96%	7.97%	91.33%

Table B.10: Twitter Case Study Problem-I: Results

- Set t represents the number of influencers to be selected from set W.
- The objective value for the main model (Ψ_2) is calculated by using the solution set obtained from upper bound model.

a_{kj}	Number of Influencers (t)	Objective Value (Ψ_2)	Benefit per Influencer ($\frac{\Psi_2}{t}$)	% Improvement with an Additional Influencer
Low number of edges within set V	1	552.34	552.34	-
	2	1061.19	530.60	92.13%
	3	1516.27	505.42	42.88%
	4	1687.15	421.79	11.27%
	5	1788.66	357.73	6.02%
	6	1857.88	309.65	3.87%
	7	1907.77	272.54	2.69%
	8	1943.53	242.94	1.87%
	9	1966.17	218.46	1.17%
	10	1977.03	197.70	0.55%
	11	1987.14	180.65	0.51%
	12	1994.53	166.21	0.37%
	13	1999.29	153.79	0.24%
	14	2002.69	143.05	0.17%
	15	2006.06	133.74	0.17%
	16	2009.53	125.60	0.17%
	17	2012.68	118.39	0.16%
	18	2013.56	111.86	0.04%
High number of edges within set V	1	2775.67	2775.67	-
	2	5088.96	2544.48	83.34%
	3	7102.64	2367.55	39.57%
	4	7900.62	1975.15	11.24%
	5	8326.10	1665.22	5.39%
	6	8653.87	1442.31	3.94%
	7	8843.61	1263.37	2.19%
	8	9000.86	1125.11	1.78%
	9	9103.25	1011.47	1.14%
	10	9148.97	914.90	0.50%
	11	9192.51	835.68	0.48%
	12	9226.23	768.85	0.37%
	13	9246.68	711.28	0.22%
	14	9261.37	661.53	0.16%
	15	9275.14	618.34	0.15%
	16	9288.65	580.54	0.15%
	17	9301.06	547.12	0.13%
	18	9305.90	516.99	0.05%

Table B.11: Twitter Case Study Problem-II: Results

- Set t represents the number of influencers to be selected from set W.
- The objective value for the main model (Ψ_2) is calculated by using the solution set obtained from the upper bound model.

	Low number of edges within set V, i.e., low a_{kj}			High number of edges within set V, i.e., high a_{kj}		
Number of Influencers (t)	Objective Value (Ψ_2)	Benefit per Influencer ($\frac{\Psi_2}{t}$)	% Improvement with an Additional Influencer	Objective Value (Ψ_2)	Benefit per Influencer ($\frac{\Psi_2}{t}$)	% Improvement with an Additional Influencer
1	1342.02	1342.02	-	4056.99	4056.99	-
2	2246.86	1123.43	67.42%	6620.5	3310.25	63.19%
3	2584	861.33	15.00%	7507.01	2502.34	13.39%
4	2898.61	724.65	12.18%	8292.52	2073.13	10.46%
5	3167.47	633.49	9.28%	8939.36	1787.87	7.80%
6	3368.06	561.34	6.33%	9371.1	1561.85	4.83%
7	3536.6	505.23	5.00%	9753.66	1393.38	4.08%
8	3647.82	455.98	3.14%	10116.94	1264.62	3.72%
9	3755.49	417.28	2.95%	10424.7	1158.30	3.04%
10	3860.21	386.02	2.79%	10709.1	1070.91	2.73%
11	3955.64	359.60	2.47%	10903.21	991.20	1.81%
12	4046.89	337.24	2.31%	11095.84	924.65	1.77%
13	4136.17	318.17	2.21%	11273.89	867.22	1.60%
14	4206.15	300.44	1.69%	11441.01	817.22	1.48%
15	4275.46	285.03	1.65%	11576.54	771.77	1.18%
16	4341.91	271.37	1.55%	11708.01	731.75	1.14%
17	4399.27	258.78	1.32%	11831.52	695.97	1.05%
18	4451.4	247.30	1.18%	11938.03	663.22	0.90%
19	4497.23	236.70	1.03%	12034.31	633.38	0.81%
20	4541.36	227.07	0.98%	12113.22	605.66	0.66%
21	4582.03	218.19	0.90%	12190.49	580.50	0.64%
22	4615.15	209.78	0.72%	12265.75	557.53	0.62%
23	4645.64	201.98	0.66%	12336.71	536.38	0.58%
24	4675.89	194.83	0.65%	12389.34	516.22	0.43%
25	4694.44	187.78	0.40%	12426.76	497.07	0.30%
26	4708.59	181.10	0.30%	12459.57	479.21	0.26%
27	4722.42	174.90	0.29%	12488.8	462.55	0.23%
28	4735.56	169.13	0.28%	12514.1	446.93	0.20%
29	4747.49	163.71	0.25%	12538.72	432.37	0.20%
30	4758.59	158.62	0.23%	12562.79	418.76	0.19%
31	4766.06	153.74	0.16%	12579.02	405.77	0.13%
32	4770.65	149.08	0.10%	12589.97	393.44	0.09%
33	4773.8	144.66	0.07%	12596.77	381.72	0.05%
34	4774.33	140.42	0.01%	12598.17	370.53	0.01%
35	4774.73	136.42	0.01%	12599.25	359.98	0.01%
36	4775.12	132.64	0.01%	12600.2	350.01	0.01%
37	4775.12	129.06	0.00%	12600.2	340.55	0.00%

Table B.12: Case Study Problem-I: Performance of our Model-based Solution vs Current Industry Practices

- For Benchmark 1, influencers were selected based on number of followers of each influencer. The data on number of followers of each influencer is presented in Table3.4. For example, in problem instance where $t = 2$, influencer E and L had the largest number of followers.
- In case of Benchmark 2, influencers were selected based on number of active followers (i.e., number of followers who retweeted at least once) of each influencer. The data on % number of active followers of each influencer is presented in Table3.4, see column - % of followers who retweeted. For example, in problem instance where $t = 2$, influencer K and R had the largest number of active followers.
- For Benchmark 3, influencers were selected based on number of active followers (i.e., number of followers who retweeted at least once) of each influencer. The data on number of active followers of each influencer is presented in Table3.4, see column - Number of retweeters. For example, in problem instance where $t = 2$, influencer E and L had the largest number of active followers.
- Gap represents the % difference in objective value obtained from the model based solution and the solution obtained for Benchmarks 1, 2, and 3 respectively.

a_{kj}	Number of Influencers (t)	**Model-based Solution**	**Benchmark 1**		**Benchmark 2**		**Benchmark 3**	
		Objective Value (Ψ_2)	Objective Value	%Optimality Gap	Objective Value	%Optimality Gap	Objective Value	%Optimality Gap
Low number of edges within set V	1	552.34	482.31	12.68%	0.90	99.84%	482.31	12.68%
	2	1061.19	996.57	6.09%	553.24	47.87%	996.57	6.09%
	3	1516.27	999.79	34.06%	1021.49	32.63%	1516.27	0.00%
	4	1687.15	1519.48	9.94%	1517.16	10.08%	1618.10	4.09%
	5	1788.66	1596.14	10.76%	1618.99	9.49%	1654.07	7.52%
	6	1857.88	1606.27	13.54%	1654.95	10.92%	1661.55	10.57%
	7	1907.77	1656.44	13.17%	1662.44	12.86%	1738.04	8.90%
	8	1943.53	1667.31	14.21%	1665.80	14.29%	1748.17	10.05%
	9	1966.17	1768.80	10.04%	1742.29	11.39%	1752.97	10.84%
	10	1977.03	1773.61	10.29%	1905.31	3.63%	1802.90	8.81%
	11	1987.14	1809.49	8.94%	1910.08	3.88%	1965.79	1.07%
	12	1994.53	1972.41	1.11%	1920.19	3.73%	1976.65	0.90%
	13	1999.29	1979.80	0.98%	1942.84	2.82%	1979.80	0.97%
	14	2002.69	2002.44	0.01%	1992.68	0.50%	2002.44	0.01%
	15	2006.06	2005.91	0.01%	1996.09	0.50%	2003.32	0.14%
	16	2009.53	2009.32	0.01%	2006.94	0.13%	2006.80	0.14%
	17	2012.68	2012.68	0.00%	2010.42	0.11%	2010.20	0.12%
	18	2013.56	2013.56	0.00%	2013.56	0.00%	2013.56	0.00%
High number of edges within set V	1	2775.67	2403.35	13.41%	6.31	99.77%	2403.35	13.41%
	2	5088.96	4760.99	6.44%	2781.54	45.34%	4760.99	6.44%
	3	7102.64	4777.99	32.73%	4970.98	30.01%	7102.64	0.00%
	4	7900.62	7118.52	9.90%	7107.84	10.03%	7539.32	4.57%
	5	8326.10	7494.00	9.99%	7544.46	9.39%	7703.52	7.48%
	6	8653.87	7541.95	12.85%	7708.63	10.92%	7739.20	10.57%
	7	8843.61	7738.56	12.50%	7744.31	12.43%	8107.33	8.33%
	8	9000.86	7783.96	13.52%	7758.47	13.80%	8154.42	9.40%
	9	9103.25	8211.09	9.80%	8126.40	10.73%	8175.65	10.19%
	10	9148.97	8232.30	10.02%	8864.79	3.11%	8368.04	8.54%
	11	9192.51	8392.96	8.70%	8885.46	3.34%	9101.13	0.99%
	12	9226.23	9125.68	1.09%	8931.49	3.19%	9144.72	0.88%
	13	9246.68	9159.46	0.94%	9034.10	2.30%	9159.46	0.94%
	14	9261.37	9261.37	0.00%	9221.61	0.43%	9261.37	0.00%
	15	9275.14	9273.79	0.01%	9235.39	0.43%	9266.23	0.10%
	16	9288.65	9287.55	0.01%	9278.82	0.11%	9278.64	0.11%
	17	9301.06	9301.06	0.00%	9291.23	0.11%	9292.40	0.09%
	18	9305.90	9305.90	0.00%	9305.90	0.00%	9305.90	0.00%

Table B.13: Case Study Problem-II: Performance of our Model-based Solution vs Current Industry Practices

a_{kj}	Number of Influencers (t)	Model-based Solution	Benchmark 1		Benchmark 2		Benchmark 3	
		Objective Value (Ψ_2)	Objective Value	%Optimality Gap	Objective Value	%Optimality Gap	Objective Value	%Optimality Gap
Low number of edges within set V	1	1342.02	364.78	72.82%	402.72	69.99%	1039.72	22.53%
	2	2246.86	1363.8	39.30%	517.16	76.98%	1404.29	37.50%
	3	2584.00	1477.34	42.83%	1517.65	41.27%	1483.31	42.60%
	4	2898.61	1477.34	49.03%	1739.11	40.00%	1682.39	41.96%
	5	3167.47	1509.24	52.35%	1755.43	44.58%	1719.44	45.72%
	6	3368.06	1524.19	54.75%	1831.21	45.63%	1734.21	48.51%
	7	3536.60	1597.81	54.82%	1908.3	46.04%	1847.26	47.77%
	8	3647.82	1704.63	53.27%	1978.66	45.76%	1899.79	47.92%
	9	3755.49	1709.41	54.48%	2031.4	45.91%	3021.84	19.54%
	10	3860.21	1756.5	54.50%	2223.44	42.40%	3226.3	16.42%
	11	3955.64	1948.57	50.74%	2259.94	42.87%	3325.01	15.94%
	12	4046.89	1962.05	51.52%	2615.8	35.36%	3396.71	16.07%
	13	4136.17	1962.45	52.55%	3674.17	11.17%	3457.6	16.41%
	14	4206.15	1999.29	52.47%	3688.26	12.31%	3561.28	15.33%
	15	4275.46	2087.24	51.18%	3779.01	11.61%	3873.17	9.41%
	16	4341.91	3199.58	26.31%	3882.42	10.58%	3963.54	8.71%
	17	4399.27	3211.19	27.01%	3977.84	9.58%	3977.84	9.58%
	18	4451.40	3285.01	26.20%	4008.42	9.95%	4229.27	4.99%
	19	4497.23	3327.26	26.02%	4095.1	8.94%	4259.84	5.28%
	20	4541.36	3434.02	24.38%	4169.12	8.20%	4346.18	4.30%
	21	4582.03	3480.92	24.03%	4190.63	8.54%	4417.05	3.60%
	22	4615.15	3493.97	24.29%	4198.14	9.04%	4457.78	3.41%
	23	4645.64	3587.62	22.77%	4225.76	9.04%	4457.78	4.04%
	24	4675.89	3606.45	22.87%	4473.66	4.32%	4476.4	4.27%
	25	4694.44	3904.6	16.83%	4514.39	3.84%	4487.52	4.41%
	26	4708.59	3905.02	17.07%	4514.91	4.11%	4518.03	4.05%
	27	4722.42	3975.8	15.81%	4599.04	2.61%	4522.64	4.23%
	28	4735.56	4032.76	14.84%	4610.16	2.65%	4608.27	2.69%
	29	4747.49	4332.13	8.75%	4622.15	2.64%	4654.08	1.97%
	30	4758.59	4335.31	8.90%	4622.15	2.87%	4666.01	1.95%
	31	4766.06	4419.57	7.27%	4626.76	2.92%	4673.48	1.94%
	32	4770.65	4420.1	7.35%	4672.6	2.06%	4674.01	2.03%
	33	4773.80	4427.61	7.25%	4758.05	0.33%	4758.05	0.33%
	34	4774.33	4458.13	6.62%	4758.46	0.33%	4758.44	0.33%
	35	4774.73	4650.6	2.60%	4758.85	0.33%	4758.85	0.33%
	36	4775.12	4761	0.30%	4761.99	0.27%	4771.97	0.07%
	37	4775.12	4775.12	0.00%	4775.12	0.00%	4775.12	0.00%
High number of edges within set V	1	4056.99	1227.89	69.73%	1335.1	67.09%	3382.85	16.62%
	2	6620.50	4348.14	34.32%	1877.63	71.64%	4456.09	32.69%
	3	7507.01	4648.58	38.08%	4929.92	34.33%	4694.23	37.47%
	4	8292.52	4648.58	43.94%	5495.85	33.73%	5247.82	36.72%
	5	8939.36	4726.79	47.12%	5556.27	37.84%	5365.16	39.98%
	6	9371.10	4773.98	49.06%	5752.32	38.62%	5410.24	42.27%
	7	9753.66	4989.96	48.84%	5993.01	38.56%	5863.75	39.88%
	8	10116.94	5276.32	47.85%	6187.66	38.84%	6019.2	40.50%
	9	10424.70	5292.74	49.23%	6341.29	39.17%	8719.68	16.36%
	10	10709.10	5406.52	49.51%	6836.58	36.16%	9161.32	14.45%
	11	10903.21	5923.47	45.67%	6944.25	36.31%	9379.06	13.98%
	12	11095.84	5959.28	46.29%	7860.78	29.16%	9532.3	14.09%
	13	11273.89	5960.83	47.13%	10212.75	9.41%	9690.33	14.05%
	14	11441.01	6074.18	46.91%	10246.56	10.44%	9978.16	12.79%
	15	11576.54	6404.87	44.67%	10435.84	9.85%	10679.67	7.75%
	16	11708.01	9036.32	22.82%	10711.34	8.51%	10863.7	7.21%
	17	11831.52	9067.36	23.36%	10904.96	7.83%	10904.96	7.83%
	18	11938.03	9220.26	22.77%	10987.82	7.96%	11436.78	4.20%
	19	12034.31	9308.28	22.65%	11241.89	6.58%	11492.04	4.51%
	20	12113.22	9607.13	20.69%	11377.72	6.07%	11738.18	3.10%
	21	12190.49	9722.89	20.24%	11417.49	6.34%	11864.08	2.68%
	22	12265.75	9753.17	20.48%	11434.88	6.77%	11937.2	2.68%
	23	12336.71	9951.44	19.33%	11485.73	6.90%	11937.2	3.24%
	24	12389.34	9989.4	19.37%	11992.99	3.20%	11970.88	3.38%
	25	12426.76	10646.95	14.32%	12065.52	2.91%	11996.33	3.46%
	26	12459.57	10648.02	14.54%	12066.97	3.15%	12074.11	3.09%
	27	12488.80	10788.26	13.62%	12282.71	1.65%	12085.43	3.23%
	28	12514.10	10926.85	12.68%	12307.67	1.65%	12233.93	2.24%
	29	12538.72	11567.62	7.74%	12332.2	1.65%	12311.96	1.81%
	30	12562.79	11574.83	7.86%	12332.2	1.84%	12336.41	1.80%
	31	12579.02	11796.4	6.22%	12343.33	1.87%	12352.88	1.80%
	32	12589.97	11797.88	6.29%	12420.76	1.34%	12354.3	1.87%
	33	12596.77	11814.87	6.21%	12566.18	0.24%	12566.18	0.24%
	34	12598.17	11893.48	5.59%	12567.13	0.25%	12567.26	0.25%
	35	12599.25	12249.77	2.77%	12568.22	0.25%	12568.22	0.25%
	36	12600.20	12563.22	0.29%	12575.02	0.20%	12593.4	0.05%
	37	12600.20	12600.2	0.00%	12600.2	0.00%	12600.2	0.00%

Table B.14: Computational Evaluation-I of Scheduling Model

			EXPERIMENT 1								EXPERIMENT 2								EXPERIMENT 3							
	Number of followers	Cost of hiring an influencer	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Low p_j	High	High | Low	5	2	0	0	1	1	3	1	2	10	0	0	1	1	1	1	3	2	7	0	1	1	1	1
	High	Low | High	1	1	1	1	0	8	0	0	1	1	1	1	0	8	0	0	1	1	1	1	0	8	0	0
	Low	High | Low	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	1	0	0
	Low	Low |High	1	0	0	2	0	0	0	0	1	3	0	2	0	0	0	0	1	2	0	2	0	0	0	0
High p_j	High	High | Low	4	6	7	4	3	2	2	3	4	8	5	4	3	2	2	4	5	7	7	4	2	1	2	3
	High	Low | High	1	1	2	1	0	7	7	7	1	1	2	1	0	7	7	7	1	1	2	1	0	4	4	0
	Low	High | Low	4	7	0	7	4	2	7	2	4	7	0	8	4	2	8	2	4	7	0	8	4	2	4	4
	Low	Low | High	2	2	4	2	7	7	0	7	2	2	5	2	7	4	0	7	2	2	4	4	7	4	0	7

Notes: In the above table, x_1=5 implies x_{15}=1, that is influencer 1 was selected and assigned to mode 5. Mode 5 implies that this influencer was assigned to tweet the ad once every three days.

Mode 1 := Influencer is assigned to tweet everyday for two weeks.
Modes 2 and 3 := Influencer is assigned to tweet every two days for two weeks.
Modes 4, 5, and 6 := Influencer is assigned to tweet once every three days for two weeks.
Modes 7, 8, 9, 10, 11, and 12 := Influencer is assigned to tweet once every six days for two weeks.

Table B.15: Computational Evaluation-II of Scheduling Model

	Number of followers		Cost of hiring an influencer		EXPERIMENT 1								EXPERIMENT 2								EXPERIMENT 3							
					SET 1				SET 2				SET 1				SET 2				SET 1				SET 2			
	Set 1	Set 2	Set 1	Set 2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
High p_j (i.e., probability of retweet)	High	High	Low	High	1	1	2	1	0	8	7	10	1	1	1	2	7	2	0	0	2	1	1	2	4	7	0	7
	High	High	High	Low	0	7	0	9	2	1	1	2	0	0	0	0	1	1	2	2	0	0	0	0	3	1	1	2
	Low	High	High	Low	0	7	0	9	2	1	1	2	0	0	0	0	1	1	2	2	0	0	0	0	3	1	1	2
	Low	High	Low	High	4	4	7	5	4	2	4	5	6	0	4	5	2	2	4	4	4	7	7	7	4	2	2	4
	High	Low	Low	High	1	1	1	1	0	0	0	0	1	1	4	2	0	0	0	0	1	1	3	2	0	0	0	0
	High	Low	High	Low	2	2	4	2	0	7	7	0	3	2	7	4	4	4	2	4	3	2	4	4	0	7	11	7
	Low	Low	High	Low	4	7	0	7	4	2	7	2	7	4	4	0	2	2	2	7	7	4	4	0	2	6	4	2
	Low	Low	Low	High	2	2	4	2	7	7	0	7	2	2	2	4	7	7	7	0	2	2	2	8	7	0	0	4
Low p_j (i.e., probability of retweet)	High	High	Low	High	1	1	1	1	0	8	0	0	1	1	1	1	0	2	0	0	1	1	1	1	0	0	0	0
	Low	Low	Low	High	1	0	0	2	0	0	0	0	0	1	1	0	0	0	0	0	2	1	2	0	0	0	0	0
	High	Low	Low	High	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
	Low	High	Low	High	3	0	0	4	0	1	2	4	0	0	0	0	2	1	4	0	0	0	0	0	0	2	2	3
	Low	High	High	Low	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
	High	Low	High	Low	1	1	0	2	0	0	0	0	1	1	0	7	0	3	0	0	1	1	1	0	0	0	0	7
	High	High	High	Low	5	2	0	0	1	1	3	1	6	5	7	0	1	1	1	1	0	2	0	0	1	1	3	1
	Low	Low	High	Low	0	0	0	0	2	0	0	0	0	0	0	0	0	2	7	0	0	0	0	0	0	0	0	1

Notes: In the above table, x_1=1 implies x_{11}=1, that is influencer 1 was selected and assigned to mode 1. Mode 1 implies that this influencer was assigned to tweet the ad everyday.

Mode 1 := Influencer is assigned to tweet everyday for two weeks.
Modes 2 and 3 := Influencer is assigned to tweet every two days for two weeks.
Modes 4, 5, and 6 := Influencer is assigned to tweet once every three days for two weeks.
Modes 7, 8, 9, 10, 11, and 12 := Influencer is assigned to tweet once every six days for two weeks.

Table B.16: Computational Evaluation-III of Scheduling Model

	Overlap of followers		Number of followers		Cost of hiring an influencer		EXPERIMENT 1								EXPERIMENT 2								EXPERIMENT 3							
							SET 1				SET 2				SET 1				SET 2				SET 1				SET 2			
	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
High p_j (i.e., probability of retweet)	High	Low	Low	High	High	Low	7	7	0	0	4	1	2	7	0	7	7	0	2	1	2	4	7	7	7	0	2	2	2	1
	High	Low	Low	High	Low	High	2	2	7	4	8	4	7	8	7	1	3	4	4	2	7	7	2	4	4	4	7	5	4	0
	High	Low	High	High	Low	High	1	2	1	2	7	4	8	4	1	2	2	2	10	8	8	4	1	1	1	3	7	2	4	5
	High	Low	High	High	High	Low	4	4	2	4	2	2	2	1	2	4	5	4	2	3	3	1	2	4	2	4	4	2	2	1
	High	Low	High	Low	High	Low	2	7	3	4	4	0	7	0	4	2	4	8	7	4	7	0	2	4	2	4	0	7	7	0
	High	Low	High	Low	Low	High	1	3	1	2	7	0	0	0	2	1	3	2	0	1	0	0	1	2	1	2	0	0	0	0
	High	Low	Low	Low	Low	High	2	2	4	0	7	0	0	0	5	2	4	4	0	0	0	0	2	2	7	4	0	0	0	0
	High	Low	Low	Low	High	Low	7	7	7	0	4	0	7	0	7	7	0	0	0	0	0	0	7	4	0	7	7	7	7	0
Low p_j (i.e., probability of retweet)	High	Low	Low	High	High	Low	0	0	0	0	0	1	2	0	0	0	0	0	1	1	0	0	0	0	0	0	2	3	1	0
	High	Low	Low	High	Low	High	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	High	Low	High	High	Low	High	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
	High	Low	High	High	High	Low	2	0	2	0	4	2	0	1	2	0	0	0	0	2	2	1	8	0	2	4	0	1	3	2
	High	Low	High	Low	High	Low	2	0	3	0	0	0	0	0	3	2	0	0	0	0	0	0	8	0	2	7	0	0	0	0
	High	Low	High	Low	Low	High	1	1	1	1	0	0	0	0	1	1	1	2	0	0	0	0	1	1	1	1	0	0	0	0
	High	Low	Low	Low	Low	High	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	High	Low	Low	Low	High	Low	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Notes: In the above table, x_1=7 implies x_{17}=1, that is influencer 1 was selected and assigned to mode 7. Mode 7 implies that this influencer was assigned to tweet the ad once a week.

Mode 1 denotest that an influencer is assigned to tweet everyday for two weeks.
Modes 2 and 3 denotest that an influencer is assigned to tweet every two days for two weeks.
Modes 4, 5, and 6 denotest that an influencer is assigned to tweet once every three days for two weeks.
Modes 7, 8, 9, 10, 11, and 12 denotest that an influencer is assigned to tweet once every six days for two weeks.

Table B.17: Computational Evaluation-IV of Scheduling Model

Network Type	**Frequency of E**	**E**	x_A	x_B	x_C	x_D	x_E	x_F	x_G	x_H
Type 1	Weekly	Low	4	2	2	7	4	2	2	7
	Weekly	Medium	2	4	2	4	4	2	4	4
	Weekly	High	2	4	2	4	4	1	7	5
	Weekly	Very High	2	4	2	4	4	1	5	2
	Semi-weekly	Low	2	4	2	7	4	2	7	4
	Semi-weekly	Medium	2	4	3	7	4	2	7	5
	Semi-weekly	High	2	4	3	7	4	3	7	5
	Semi-weekly	Very High	3	4	3	7	4	3	10	4
Type 2	Weekly	Low	2	3	3	4	7	2	5	2
	Weekly	Medium	2	3	1	4	5	2	4	2
	Weekly	High	2	3	1	2	3	2	2	2
	Weekly	Very High	2	1	1	2	3	1	2	3
	Semi-weekly	Low	3	2	2	5	7	2	4	4
	Semi-weekly	Medium	2	2	2	4	7	3	4	5
	Semi-weekly	High	3	2	3	5	7	2	4	4
	Semi-weekly	Very High	3	3	3	10	7	2	4	2
Type 3	Weekly	Low	7	7	4	7	0	7	0	0
	Weekly	Medium	4	7	4	7	0	7	0	0
	Weekly	High	4	7	4	7	0	7	0	7
	Weekly	Very High	4	7	4	4	7	7	0	7
	Semi-weekly	Low	10	7	4	10	0	0	0	7
	Semi-weekly	Medium	4	10	4	7	0	10	0	0
	Semi-weekly	High	4	10	4	10	0	7	0	7
	Semi-weekly	Very High	3	4	3	10	0	7	0	7

Notes: In the above table, x_1=7 implies x_{17}=1, that is influencer 1 was selected and assigned to mode 7. Mode 7 implies that this influencer was assigned to tweet the ad once a week.

The first column denotes "Network Type." Type 1 network is presented in Figure 26, Type 2 network is illustrated in Figure 27, and Type 3 network is depicted in Figure 28.

The second column, Frequency of E, denotes how often a firm has a minimum number of retweets requirement. Weekly denotes, that the firm needs E number of retweets every week. Semi-week denotes that the firms needs atleast E number of retweets every 3 days.

The third column, E, denotes what is minimum number of retweets that are required by the firm. Low indicates, the number of retweets required during the time period is low, whereas, Very High denotes that the number of retweets required during the time periof is very high.

For this experiment, T= 6 days.
Mode 1 denotes that an influencer is assigned to tweet everyday.
Modes 2 and 3 denotest that an influencer is assigned to tweet every two days.
Modes 4, 5, and 6 denotest that an influencer is assigned to tweet once every three days.
Modes 7, 8, 9, 10, 11, and 12 denotest that an influencer is assigned to tweet once during the week.

B.7 Extensions

B.7.1 Model with Budget Constraint

We now present the our main model, i.e., the multiplicative two-level influence problem (Model $M2L$). In particular, in this section we focus on multiplicative framework for calculating the cumulative probability of retweet. Furthermore, as suggested by our data, we assume that the users in set V may be influenced by users in set W and V. Before presenting Model $M2L$, we first define our parameters and decision variables.

Parameters:

$$a_{ij} = \begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } i \in W, \\ 0, & \text{otherwise.} \end{cases}$$

$$a_{kj} = \begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } k \in V, \\ 0, & \text{otherwise.} \end{cases}$$

b_j = total benefit to the firm when the tweet reaches user $j \in V$ through all $i \in W$.

c_i = cost to the firm for hiring influencer $i \in W$.

d_j = benefit to the firm when the tweet reaches one follower of user $j \in V$.

F_j = number of followers of user $j \in V$.

p_{ij} = probability of user $j \in V$ retweeting the message of influencer $i \in W$; $p_{ij} = 0$ if $a_{ij} = 0$.

p_{kj} = probability of user $j \in V$ retweeting the message of user $k \in V$; $p_{ij} = 0$ if $a_{jk} = 0$.

Decision Variables:

α_j = the cumulative probability of user $j \in V$ not retweeting a message from users in V.

δ_j = the cumulative probability of user $j \in V$ not retweeting a message from users in W.

g_j = the cumulative probability of user $j \in V$ retweeting a message.

$$x_i = \begin{cases} 1, & \text{if } i \in W \text{ is chosen as a seeder,} \\ 0, & \text{otherwise.} \end{cases}$$

Mathematical Formulation - Model M2L:

$$\text{Max} \quad \Psi_2 = \sum_{j \in V} b_j(1-\delta_j) + \sum_{j \in V} d_j g_j F_j \tag{B.73}$$

subject to:

$$\sum_i c_i x_i \leq \mathscr{B}, \tag{B.74}$$

$$\delta_j = \prod_{i \in W} (1 - p_{ij} x_i), \qquad \forall j \in V \tag{B.75}$$

$$\alpha_j = \prod_{k \in V} (1 - p_{kj}(1-\delta_k)), \forall j \in V \tag{B.76}$$

$$g_j = 1 - \alpha_j \delta_j, \qquad \forall j \in V \tag{B.77}$$

$$x_i \in \{0,1\} \qquad . \tag{B.78}$$

In the above model, the objective function, Ψ_2 (i.e., Equation B.73), maximizes the total benefit to the firm when tweets reach the users on Twitter. Constraint (B.74) ensures that the cost of hiring the influencers is below the firm's budget level. Constraint (B.75) denotes the expected probability of user $j \in V$ not retweeting message from user $k \in V$. Constraint (B.76) represents the expected probability that follower $j \in V$ does not retweets the message from influencers $i \in W$ and $j \in V$. Constraint (B.77) is the expected probability of user $j \in V$ retweeting a message.

B.7.2 Model with Coverage Constraints

We now present the our main model, i.e., the multiplicative two level influence problem (Model $M2L$). In particular, in this section we focus on multiplicative framework for calculating the cumulative probability of retweet. Furthermore, as suggested by our data, we assume that the users in set V may be influenced by users in set W and V. Before presenting Model $M2L$, we first define our parameters and decision variables. There are R set of regions. We need to cover certain number of users in region $r \in R$ (N_r^1) in set V and followers of V (N_r^2) in region $r \in R$.

Parameters:

$$a_{ij} = \begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } i \in W, \\ 0, & \text{otherwise.} \end{cases}$$

$$a_{kj} = \begin{cases} 1, & \text{if user } j \in V \text{ is a direct follower of } k \in V, \\ 0, & \text{otherwise.} \end{cases}$$

p_{ij} = probability of user $j \in V$ retweeting the message of influencer $i \in W$; $p_{ij} = 0$ if $a_{ij} = 0$.

p_{kj} = probability of user $j \in V$ retweeting the message of user $k \in V$; $p_{ij} = 0$ if $a_{jk} = 0$.

b_j = benefit to the firm when the tweet reaches user $j \in V$ through all $i \in W$.

d_j = benefit to the firm when the tweet reaches one follower of user $j \in V$.

F_j = number of followers of user $j \in V$.

$$e_{jr} = \begin{cases} 1, & \text{if user } j \in V \text{ belongs to region } r \in R, \\ 0, & \text{otherwise.} \end{cases}$$

f_{jr} = number of followers of user $j \in V$ belonging to region $r \in R$. Note that $F_j = \sum_{r \in R} f_{jr}$.

N_r^1 = minimum number of users in V belonging to region $r \in R$ required to be covered.

N_r^2 = minimum number of followers of V belonging to region $r \in R$ required to be covered.

M = large number.

Decision Variables:

$$x_i = \begin{cases} 1, & \text{if } i \in W \text{ is chosen as a seeder,} \\ 0, & \text{otherwise.} \end{cases}$$

g_j = the cumulative probability of user $j \in V$ retweeting a message.

δ_j = the cumulative probability of user $j \in V$ not retweeting a message from users in W.

α_j = the cumulative probability of user $j \in V$ not retweeting a message from users in V.

Mathematical Formulation - Model M2L:

$$\text{Max} \quad \Psi_2 = \sum_{j \in V} b_j(1-\delta_j) + \sum_{j \in V} d_j g_j F_j \tag{B.79}$$

subject to:

$$\sum_i x_i \le t, \tag{B.80}$$

$$\delta_j = \prod_{i \in W} (1 - p_{ij} x_i), \qquad \forall j \in V \tag{B.81}$$

$$\alpha_j = \prod_{k \in V} (1 - p_{kj}(1-\delta_k)), \forall j \in V \tag{B.82}$$

$$g_j = 1 - \alpha_j \delta_j, \qquad \forall j \in V \tag{B.83}$$

$$y_j \ge 1 - \delta_j, \qquad \forall j \in V \tag{B.84}$$

$$y_j \le M(1-\delta_j), \qquad \forall j \in V \tag{B.85}$$

$$\sum_{j \in V} e_{jr} y_j \ge N_r^1, \qquad \forall r \in R \tag{B.86}$$

$$\sum_{j \in V} f_{jr} y_j \ge N_r^2, \qquad \forall r \in R \tag{B.87}$$

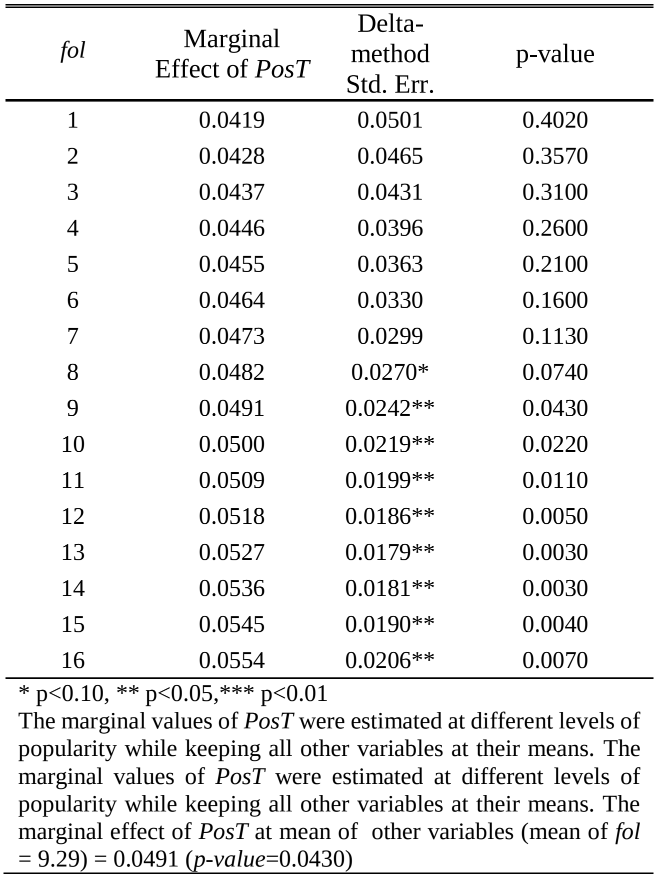
$$y_j, x_i \in \{0, 1\} \tag{B.88}$$

APPENDIX C

APPENDIX FOR CHAPTER 4: THE EFFECTS OF SOCIAL MEDIA CONTENT ON ENGAGEMENT

C.1 Additional Tables

Table C.1: Conditional Marginal Effects of $PosT$



fol	Marginal Effect of *PosT*	Delta-method Std. Err.	p-value
1	0.0419	0.0501	0.4020
2	0.0428	0.0465	0.3570
3	0.0437	0.0431	0.3100
4	0.0446	0.0396	0.2600
5	0.0455	0.0363	0.2100
6	0.0464	0.0330	0.1600
7	0.0473	0.0299	0.1130
8	0.0482	0.0270*	0.0740
9	0.0491	0.0242**	0.0430
10	0.0500	0.0219**	0.0220
11	0.0509	0.0199**	0.0110
12	0.0518	0.0186**	0.0050
13	0.0527	0.0179**	0.0030
14	0.0536	0.0181**	0.0030
15	0.0545	0.0190**	0.0040
16	0.0554	0.0206**	0.0070

* p<0.10, ** p<0.05,*** p<0.01

The marginal values of *PosT* were estimated at different levels of popularity while keeping all other variables at their means. The marginal values of *PosT* were estimated at different levels of popularity while keeping all other variables at their means. The marginal effect of *PosT* at mean of other variables (mean of *fol* = 9.29) = 0.0491 (*p-value*=0.0430)

Table C.2: Robustness Check: Alternative Models

		Model R1	Model R2	Model R3
	DV	***RT***	***log(RT)***	***log(RT)***
		(SE)	(SE)	(SE)
Main Effects	*PosT*	-0.00563	0.0107	0.0170
		(0.0355)	(0.0419)	(0.0374)
	NegT	0.334***	0.407***	0.407**
		(0.0490)	(0.152)	(0.141)
	MixT	0.0930***	0.140***	0.145**
		(0.0106)	(0.0456)	(0.0461)
	fol	0.148***	0.302***	0.111*
		(0.0188)	(0.0230)	(0.0517)
	PartyPosT	0.000705**	0.00142**	0.00130*
		(0.000352)	(0.000582)	(0.000605)
	PartyNegT	-0.00106^{+}	-0.00140***	-0.000976^{+}
		(0.000706)	(0.000483)	(0.000537)
	PartyMixT	0.0000137	-0.0000231	-0.0000216
		(0.0000783)	(0.000119)	(0.000123)
Interaction Effects	*fol* x *PosT*	0.00145	0.00570**	0.00637*
		(0.00238)	(0.00281)	(0.00266)
	fol x *NegT*	-0.0226***	-0.0310***	-0.0311**
		(0.00341)	(0.0104)	(0.00905)
	fol x *MixT*	-0.00558***	-0.00805**	-0.00837*
		(0.000712)	(0.00346)	(0.00353)
	PosT x *PartyPosT*	0.0000550	0.0000195	0.0000317
		(0.0000372)	(0.0000652)	(0.0000780)
	PosT x *PartyNegT*	0.0000214	-0.0000595	-0.000106
		(0.0000791)	(0.000163)	(0.000193)
	NegT x *PartyPosT*	-0.000197***	-0.000339***	-0.000373**
		(0.0000602)	(0.000116)	(0.000113)
	NegT x *PartyNegT*	0.000251*	0.000638***	0.000691**
		(0.000129)	(0.000205)	(0.000177)
Controls	*partyMentions(lag)*	-0.318***	-0.506***	-0.531***
		(0.105)	(0.148)	(0.115)
	partyRetweets(lag)	0.221**	0.260***	0.305***
		(0.104)	(0.0330)	(0.0222)
	mentions(lag)	0.0000210***	0.272***	0.219***
		(0.00000548)	(0.0581)	(0.0500)
	Intercept	-1.428***	-8.520	2.881**
		(0.430)	(6.470)	(0.738)
	Region Controls	[No]	[Yes]	[No]
	Education Controls	[No]	[Yes]	[No]
	Time Controls	[Yes]	[Yes]	[Yes]
	Time invariant Controls	[No]	[Yes]	[No]
	Opp HGC Tone(lag)	[Yes]	[Yes]	[Yes]
	N	1313	1313	1313
	Log-likelihood/R-sq	-5420.01	0.671	0.494

Robust standard errors clustered at party level are in parentheses.
* p<0.10, ** p<0.05,*** p<0.01, $^{+}$ p<0.13.
Model R1 is FE-NB; Model R2 is random effects; Model R3 is fixed effects. Time invariant controls are *Age*, *assets*, and *Male.*

Table C.3: Robustness Check: Alternate Sample

		Model R4A	Model R4B	Model R4C	Model R4D
	DV	***$log(RT_{ipt})$***	***RT_{ipt}***	***$log(RT_{ipt})$***	***$log(RT_{ipt})$***
Main Effects of HGC	*PosT*	0.0137	-0.0155	0.0117	0.027
		(0.0475)	(0.0356)	(0.0354)	(0.0344)
	NegT	0.489***	0.355***	0.404***	0.377**
		(0.128)	(0.0488)	(0.134)	(0.134)
	MixT	0.135***	0.0957***	0.149***	0.150***
		(0.0415)	(0.0105)	(0.0395)	(0.0412)
	Fol	0.181***	0.141***	0.271***	0.164**
		(0.0295)	(0.0176)	(0.0221)	(0.0698)
	PartyPosT	0.00149***	0.000648*	0.00105***	0.000962**
		(0.000274)	(0.000333)	(0.000365)	(0.000355)
	PartyNegT	-0.00106*	-0.00135**	-0.000763**	-0.000442
		(0.000637)	(0.000667)	(0.000379)	(0.000442)
	PartyMixT	-0.0000841	0.00000823	-0.0000175	-0.00000483
		(0.0000749)	(0.0000745)	(0.0000719)	(0.0000797)
Interaction Effects	*fol* x *PosT*	0.00407	0.00215	0.00667**	0.00626*
		(0.00267)	(0.00238)	(0.00270)	(0.00324)
	fol x *NegT*	-0.0365***	-0.0242***	-0.0321***	-0.0305***
		(0.00908)	(0.00342)	(0.00919)	(0.00897)
	fol x *MixT*	-0.00763**	-0.00571***	-0.00857***	-0.00872**
		(0.00316)	(0.000710)	(0.00309)	(0.00323)
	PosT x *PartyPosT*	0.0000468	0.0000638*	0.0000383	0.0000422
		(0.0000287)	(0.0000375)	(0.0000529)	(0.0000611)
	PosT x *PartyNegT*	-0.0000442	0.0000239	-0.000143	-0.000185
		(0.0000930)	(0.0000797)	(0.000146)	(0.000166)
	NegT x *PartyPosT*	-0.000359**	-0.000216***	-0.000361***	-0.000364**
		(0.000148)	(0.0000613)	(0.000120)	(0.000119)
	NegT x *PartyNegT*	0.000678**	0.000280**	0.000753***	0.000777***
		(0.000343)	(0.000130)	(0.000205)	(0.000183)
Controls	*mentions(lag)*	0.432***	0.0000224***	0.279***	0.200***
		(0.0649)	(0.00000537)	(0.0553)	(0.0494)
	partyMentions(lag)	-0.045	-0.300***	-0.415***	-0.451***
		(0.0450)	(0.103)	(0.0873)	(0.0745)
	partyRetweets(lag)	0.0734	0.222**	0.236***	0.295***
		(0.0526)	(0.104)	(0.0410)	(0.0735)
	partyExpenditure(lag)	-0.135**		0.585*	
		(0.0577)		(0.299)	
	Region/Education Controls	[Yes]	[NO]	[Yes]	[NO]
	Time Controls	[Yes]	[Yes]	[Yes]	[Yes]
	Opp HGC Tone	[Yes]	[Yes]	[Yes]	[Yes]
	Intercept	1.834	-1.640***	-10.73*	1.303
		(1.943)	(0.357)	(6.016)	(0.846)
	Observations	1841	1799	1841	1841
	Log Likelihood/*R-sq*	-3156.019	-5997.38	0.699	0.589

Robust standard errors clustered at party level are in parentheses. The coefficients of Model R4A were estimated using linear mixed effects model. The coefficients of Model R4B were estimated using negative binomial fixed effects model. The coefficients of Model R4C were estimated using random effects model (Party dummy variables are included). The coefficients of Model R4D were estimated using fixed effects model. Time invariant variable (*Age*, *assets*, and *Male*) were included for mixed and random effects model (i.e., Models R4A and R4C).
* p<0.10, ** p<0.05,*** p<0.01

Table C.4: Robustness Check: Accounting for Selection Bias

		Model R5
	DV	***log(RT)***
		(Robust SE)
Main Effects	PosT	0.0462
		(0.0468)
	NegT	0.421***
		(0.0920)
	MixT	0.130***
		(0.0410)
	Fol	0.0892***
		(0.0344)
	PartyPosT	0.000977***
		(0.000161)
	PartyNegT	-0.000821***
		(0.000271)
	PartyMixT	-0.0000753
		(0.0000604)
Interaction Effects	*fol* x *PosT*	0.00198
		(0.00253)
	fol x *NegT*	-0.0328***
		(0.00754)
	fol x *MixT*	-0.00728**
		(0.00316)
	PosT x *PartyPosT*	0.0000615**
		(0.0000302)
	PosT x *PartyNegT*	-0.000108
		(0.0000982)
	NegT x *PartyPosT*	-0.000348**
		(0.000136)
	NegT x *PartyNegT*	0.000741**
		(0.000336)
Controls	*mentions(lag)*	0.436***
		(0.0553)
	partyMentions(lag)	-0.180***
		(0.0401)
	partyRetweets(lag)	0.114**
		(0.0444)
	Inverse Mills	**-0.146*** **
		(0.0282)
	Intercept	0.682
	Region / Education Controls	[Yes]
	Opp HGC Tone Controls	[Yes]
	Time invariant controls	[Yes]
	Time Controls	[Yes]
	N	2310
	Log-likelihood	-3805.206

Robust standard errors clustered at party level are in parentheses. The coefficients are estimated as mixed effects model using full maximum likelihood model. Time invariant controls are *Age*, *assets*, and *Male*.
* p<0.10, ** p<0.05,*** p<0.01

Table C.5: Alternate Measure for Engagement

DV	Model R6 ***log(Mentions$_{ipt}$+RT$_{ipt}$)*** (Robust SE)
PosT	0.0815
	(0.0806)
NegT	0.814***
	(0.200)
MixT	0.247***
	(0.0934)
log(Fol)	0.840***
	(0.198)
PartyPosT	0.00211***
	(0.000491)
PartyNegT	-0.00317***
	(0.000882)
PartyMixT	0.000148
	(0.000150)
fol x *PosT*	0.00220
	(0.00466)
fol x *NegT*	-0.0605***
	(0.0147)
fol x *MixT*	-0.0145**
	(0.00681)
PosT x *PartyPosT*	0.0000447
	(0.0000580)
PosT x *PartyNegT*	-0.000130
	(0.000199)
NegT x *PartyPosT*	-0.000538*
	(0.000289)
NegT x *PartyNegT*	0.00115*
	(0.000609)
log(Assets)	0.00186
	(0.120)
Age	0.0351***
	(0.0126)
Male	1.503*
	(0.854)
partyMentions(lag)	-0.898***
	(0.157)
partyRetweets(lag)	0.513***
	(0.112)
partyExpenditure	-0.641***
	(0.0521)
Intercept	5.417**
	(2.573)
Region & Education Controls	[Yes]
Opp HGC Tone	[Yes]
Time Controls	[Yes]
N	1313
Log-likelihood	-3029.382

Robust standard errors clustered at party level are in parentheses.
* p<0.10, ** p<0.05,*** p<0.01