Essays on Social Network Propagation, Online Privacy, and Security

by

Hooman Hidaji

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Operations and Information Systems

Faculty of Business
University of Alberta

ABSTRACT

This thesis is composed of four essays addressing problems in the domain of information systems (IS). In the first essay, we study methods for improving propagation of messages in both consumer and enterprise social networks. We present the formal definition and analysis of the problem, and use the hop-constrained minimum spanning tree (HMST) model to find cost-effective seeds and possible new connections that result in networks with improved propagation properties. Moreover, we present new heuristic algorithms that substantially improve the solution quality for the HMST problem, as tested on both random and real-world networks.

In the second essay, we study the decision making of publisher websites in using third parties. We propose a two-sided economic model that captures the interaction between users, publisher websites, and third parties. Specifically, we focus on the effect of user privacy concerns on information sharing behavior of publisher websites. We then analyze welfare aspects and provide insights on the impact of industry regulations on stakeholders. The model is validated using an exploratory empirical analysis of publisher websites' third party sharing. Following this topic, in the third essay, we examine the impact of user privacy concerns as the self-regulatory mechanism that induces the website publisher to respect user privacy concerns. We conduct experiments designed to test the impact of users' privacy concerns, and find that the privacy concerns do affect the sharing intensity of user information by the websites. We analyze the effectiveness of passive "Do Not Track" and active "AdBlock Plus" privacy tools in a self-regulated environment. Interestingly, we find that the "Do Not Track"

request does not always serve its intended purpose, but is actually being used by many websites as a signal to substantially increase the user information sharing intensity.

Finally, in the fourth essay, we examine a firm's choice of information technology supplier, where customers' demand changes in response to adverse events or incidents that occur at the firms. We specifically model the strategic choice of firms choosing between either a shared supplier versus an independent supplier. In a symmetric duopoly setting, we show that this choice depends on the customer demand reactions to adverse events as well as relative risks of the suppliers. We also analyze the effectiveness of regulation and cooperation in improving firms' profit.

PREFACE

Some of the research projects that I report in this thesis are part of research collaborations. Chapter 2 has been published as Gopal, R., Hidaji, H., Patterson, R. A., Rolland, E., & Zhdanov, D. (2016). "Design Improvements for Message Propagation in Malleable Social Networks." *Production and Operations Management* 25(6): 993-1005. In this paper, I was responsible for the data collection and analysis as well as the manuscript composition. My supervisor, Dr. Raymond Patterson, assisted me in doing the analysis, and my co-authors helped me by providing feedback and suggestions, as well as in editing the manuscript. Chapter 3 has been accepted for publication at *MIS Quarterly* as Gopal, R., Hidaji, H., Patterson, R. A., Rolland, E., & Zhdanov, D. "How Much to Share with Third Parties? Users' Privacy Concerns and Website's Dilemma." For this paper, I was responsible for data collection and analysis with assistance from my supervisor, and I developed the model using suggestions and feedback from my co-authors. They also helped me in editing the manuscript. Chapter 4 is a continuation of this paper, in which I was responsible for the data gathering and analysis, with guidance from the same co-authors. Finally, Chapter 5 is result of a collaboration with Dr. Bora Kolfal, Dr. Raymond Patterson, Dr. Erik Rolland, and Dr. Lisa Yeo. In this project, I was responsible for developing the model and its analysis, for which I received feedback and guidance from my co-authors, as well as in editing the manuscript.

ACKNOWLEDGMENETS

First, I would like to thank my supervisor, Dr. Raymond Patterson, for his unreserved and patient guidance, encouragement, and advice he has provided throughout my Ph.D. program. I would also like to thank my supervisory committee, Dr. Armann Ingolfsson and Dr. Bora Kolfal for their guidance and support, as well as their teachings. I would like to thank my co-authors Dr. Ram Gopal, Dr. Erik Rolland, Dr. Lisa Yeo, and Dr. Dmitry Zhdanov for their guidance and support. I would also like to thank Dr. Yonghua Ji for his teachings.

My sincere thanks go to Dr. David Deephouse, Business PhD Program Associate Dean, and Dr. Karim Jamal, Chair of Department of Accounting, Operations and Information Systems as well as the staff at Alberta School of Business, especially Debbie Giesbrecht and Jeanette Gosine at the Ph.D. Office and Deb Picken, Sharon Luyendyk, and Karmeni Govender at the Department of Accounting, Operations and Information Systems for their support.

I would like to give special thanks to my amazing wife, Elshan, my dedicated parents, Farkhondeh and Mahmood, and my kind and caring brothers, Hajir and Hatef for their unconditional love, support, and encouragements. Finally, many thanks go to my fellows and friends, Dr. Mohammad Delasay Sorkhab, Dr. Amir Rastpour, Mohamad Soltani, Can Sun, and Mostafa Rezaei.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

The twenty-first century marks the dominance of a digital world where individuals connect, socialize, and consume content using online social networks, and go online to get news and shop among other activities. This online trend has summoned businesses and firms to propagate their marketing messages and advertise in this remarkable marketplace. Firms and organizations rely heavily on internet- and cloud-based infrastructure for their operations, making the stability and security of these services vital to their success. It is now more important than ever to study the different aspects of this online industry. This thesis is composed of essays addressing four problems in the domain known as the information systems (IS) literature. In the first essay, we study methods for improving propagation of messages in online social networks. In the second essay, we study the decision making of publisher websites, and how it is impacted by privacy concerns of users. The third essay extends this analysis by studying the impact of privacy tools on third party usage in a self-regulated setting. Finally, in the fourth essay, we propose a modelling approach for strategizing of firms who may want to jointly invest in security practices of cloud-based infrastructures.

This thesis is organized in a paper-based format, with four essays provided in Chapters 2 to 5. In Chapter 2, we introduce the problem of social network propagation optimization, where both seeds and connections in the network can be altered in order to improve propagation. The problem is modeled as a mixed-integer linear programming (MILP) based on the hop-constrained minimum spanning tree (HMST) problem. We provide effective and efficient

heuristics for solving this problem, and apply these methods to both random and real-world networks. We posit the importance of manipulations in networks, and how they can affect the performance of the social networks in terms of propagation of messages. This approach is a result of online social networks trying to acquire strategic advantage through making the network more useful for users, but also creating more monetization opportunities, usually through advertising campaigns.

While Chapter 2 looks at the problem from the point of view of a social network and advertisers who want to maximize performance of the network, in Chapter 3 we study how the preferences of the users affect the decision making of online businesses. More specifically, we investigate the effect of user privacy concerns on a publisher website's decision of sharing user information with third parties. We provide insights on how user privacy concerns drive the website and third party industry. We further explore the social welfare implications of the different parameters, and develop policy implications. The model is verified through an exploratory analysis using real-world data and robustness check. Chapter 4 extends this analysis by examining the impact of user privacy concerns as the self-regulatory mechanism, or "invisible hand", that induces the website publisher to respect user privacy concerns. We conduct experiments designed to test the impact of users' privacy concerns, and find that the privacy concerns do affect the sharing intensity of user information by the websites. We analyze the effectiveness of passive "Do Not Track" and active "AdBlock Plus" privacy tools in a self-regulated environment. Interestingly, we find that the "Do Not Track" request does not always serve its intended purpose, but is actually being used by many websites as a signal to substantially increase the sharing intensity. Our findings provide important directions for shaping policy and future research in the domain of online privacy.

In Chapter 5 we explore the problem of information technology (IT) security investment for firms in a supply chain setting. We consider two firms in a duopoly, where IT security breaches affect the demands for both firms. Firms can choose to independently invest in different security suppliers, or to share their investments in a single security supplier. We provide several insights regarding firm strategies in different settings. Most importantly, we find that the firms are better off sharing their resources when the competition is high among the firms. Finally, Chapter 6 summarizes the findings of this thesis.

CHAPTER 2

Design Improvements for Message Propagation in Malleable Social Networks

2.1 Introduction

The use of enterprise social networks (ESN) (e.g., Tibbr, SocialCast, Jive, IBM Connections, Socialtext, etc.) is becoming widespread among large and successful organizations (Deloitte 2013). ESNs not only include the network of a firm's employees, but also the partners from the firm's supply chain. Additionally, consumer social networks (CSN) (e.g., Facebook, Twitter, Google+, etc.) are the external equivalent to ESNs, allowing the firm to connect with customers as well. ESNs support the transfer of innovative ideas and messages throughout an organization's social network. ESNs enable network members to interact with each other and to propagate knowledge and information. Employees, supply chain partners, and customers represent extended human resources of the firm, and social networks allow its users to learn about new information, ideas, and innovation from their peers, which can augment traditional communication methods.

One of the crucial properties of social networks is their ability to propagate information, ideas, and messages. Propagation starts with one or several "seed" users and spreads through the network. Propagation in a network is realized because of the users' willingness to transfer the information to other users. Timely and reliable propagation is not only desirable, but is often the main role of a network. Therefore, it is important to understand how the propagation of a network is affected by different design improvements. This chapter specifically focuses on the context of messages that are transferred in a cascade-style propagation in social networks.

Cascade-style propagation is the type of propagation where ideas and information propagate from person to person through social network connections in an all-or-nothing fashion. The propagation of ideas and information, also known as messages, does not depend on the collective influence from many users in the network, but a single user alone may transfer the message to her/his connections.

While some networks have pre-defined and fixed architectures (e.g., computer networks and cell phone towers), and others are purely self-organized and not subject to external design intervention (e.g., biological networks), we define ***malleable networks*** as networks in which new connections can be intentionally formed. Note that while intentional interventions in the networks may not be totally under control of the network administrators, these administrators still have influence on alteration of network topology. Valente (2012) describes the four different strategies for network intervention as (i) identifying the influential individuals (seeds), (ii) segmentation, (iii) induction or excitation of network, and (iv) alteration of the network. We are specifically interested in intervention types (i) and (iv). We posit that ESNs are especially malleable, in that self-organized network structures and the propagation properties can be augmented, or at least influenced, by managers through job assignments, corporate social engagements, educational workshops, and other means. Moreover, in CSNs (e.g., Facebook, Twitter), the network administrators can cooperate with firms (e.g., advertising campaigners) in order to find improvements in the network that are beneficial to the cascade-style propagation of the message. CSNs also take on these efforts so as to create groups of people with similar interests that can be used as an effective audience for marketing purposes. This chapter adopts the point of view of the network administrator of either malleable ESNs (an agent of the firm) or malleable CSNs (e.g., Facebook, Twitter).

In this chapter, we are interested in the first and last of these network interventions as categorized above (Valente 2012). In other words, in order to enhance network propagation, we focus on the seeding of the message, and alterations in the structure of the network, mainly by creation of new connections. While prior literature focuses mainly on message seeding to improve propagation (e.g., Domingos and Richardson 2001; Kempe, Kleinberg, and Tardos 2003; and Ni, Xie, and Liu 2010), we argue that better results can be expected if seeding and alterations are considered jointly. Several approaches have been proposed to make alterations in topologies of the networks. In the ESN context; Cross, Borgatti, and Parker (2002) study informal social networks in organizations, and provide examples of how managers can influence these informal networks at individual and whole network levels. They use social network analysis to find and promote effective collaborations within important individuals and groups. In a real-world experiment, practical tools such as joint staffing of projects and mixed revenue incentives are used to alter the network. The authors find that these actions not only alter networks, but that they also have significant impact on the performance of the collaboration network as a whole. In another study, Cross, Martin, and Weiss (2006) provide how network analysis can help companies understand the potential points for improvement in their employee collaboration network and act on it. On the other hand, in the CSN setting, network administrators utilize different friend recommendation systems to find proper connections. These systems aim at "suggesting suitable matches to people in a way that increases the likelihood of a positive interaction" (Kim et al. 2012). Consumer social networks often utilize proprietary recommendation systems, actively trying to alter and enhance their networks (Moricz et al. 2010).

Network intervention through seeding and alterations is costly. In the ESN context, the seeding cost is due to finding the influential seeds and inducing them with a message or idea. The new connection cost is the effort needed to develop a relationship between two employees or sets of employees, such as in the cases provided by Cross et al. (2002) and Cross et al. (2006) above. In CSNs, seeding cost includes the effort needed to identify the proper users for starting the message in the network, and directly sending the message to the users. The new connection creation cost includes search and recommendation effort to identify potential connections and encourage them. These costs are further discussed in Section 2.2.

The problem of network intervention and improvement is important to operations managers from a strategic point of view. Boyer, Swink, and Rosenzweig (2005) suggest several theoretical perspectives to ground operations management research into firm strategy. Among them is the resource based view (RBV) of the firm (Barney 1991), in which both supply chain partners and customers are often considered to be extended resources of the firm (Ward, Rolland, and Patterson 2005; Rolland, Patterson, and Ward 2009). This is especially true for customers who are highly engaged with the firm's delivery of products or services (e.g., healthcare and education), perhaps because they are substantially involved with product or service co-creation (Rolland, Patterson, and Ward 2010). From the RBV perspective, the ability to effectively and efficiently propagate ideas and knowledge to key human resources with ESNs and CSNs could very well create a sustainable competitive advantage for the firm (Turban, Bolloju, and Liang 2011). Improved malleable social networks represent tangible infrastructure to propagate ideas and messages not only within the firm, but also to supply chain partners and customers, thus creating a sustainable competitive advantage, because it is valuable, rare, and difficult to imitate.

In this chapter, we specifically consider the problem of improving the message propagation to a specific threshold at the minimum cost. In other words, the objective is to enhance an existing network for propagation of a message to all nodes within a certain propagation performance threshold at the lowest cost. Note that this is one possible formulation of propagation improvement problem, and there are many other possible settings that can be addressed, and fall outside of the domain of this chapter. Some alternative problem formulations are discussed in the conclusions section.

We adopt the hop-constrained minimum spanning tree (HMST) MILP, and use the CPLEX solver and meta-heuristic algorithms (which we collectively refer to as the HMST model) to solve this problem. The HMST model allows us to find low-cost seeds and new connections in the network, and at the same time, provide acceptable propagation performance. The benefit of using the HMST in this context is that it can jointly find good seeds and new connections for improving the propagation. We discuss in Section 2.3 that there is some evidence for how propagation in social networks is constrained with hops, thus making the HMST an appropriate model in this context. We argue that creating a hop-constrained spanning tree structure in the network can guarantee a certain propagation performance. We further demonstrate this through simulation analysis of network propagations.

The contributions of this chapter are twofold. First, we introduce the propagation improvement problem, where the malleable social networks can be altered at a cost in order to improve message propagation performance. We provide a framework for jointly finding effective seeds, as well as identifying network alterations using the HMST model. Through numerical experiments and simulation analysis, we demonstrate how creation of these new connections can improve propagation in the network at a low cost. Moreover, we provide some

theoretical analysis on the tradeoff between seeding and connection cost, and how that affects the structure of the solution. Second, from a methodological perspective, we provide new and effective solution methods for the HMST problem. We demonstrate the efficiency of the proposed methods through extensive computational experiments, with randomly generated data as well as data based on real-world social network topologies.

The remainder of this chapter is organized as follows: In Section 2.2, we provide a literature review of the message propagation problem, focusing on social network context. Some concepts of the problem are also introduced in this section. In Section 2.3 we propose the model for finding the seeds and alterations in malleable social networks that will improve propagation at low cost. In Section 2.4 we propose efficient and effective heuristics for solving the model. Section 2.5 provides the details of the propagation model that is used to study the improvement of propagation performance in a network. Section 2.6 presents computational results, and Section 2.7 concludes the chapter.

2.2 Message Propagation in Social Networks

There exists significant literature for modeling the propagation and diffusion of information in social networks. Two of the most prominent model classes in this area are threshold (Kempe et al. 2003) and cascade (Goldenberg, Libai and Muller 2001) propagation models. In the threshold models, propagation depends on the collective influence of people on each other. In other words, for a user to become active, the summation of influence from her/his connections should be higher than a given threshold. This model is useful in propagation of information artifacts that are dependent on network externalities. In contrast, cascade model message propagation does not rely on the collective influence of network connections, and each person can propagate messages to his/her connections with a given probability. In other words, the information is not spread by

the summation of connections, but by a single connection that may be sufficiently powerful, trustworthy, or influential. The cascade model is useful for propagation of news, information, ideas, viral messages, and rumors. In graph terms, propagation of information starts from one or several "activated nodes" or "seeds". At each stage, every activated node can in turn activate its non-active neighbors with a certain probability. In this study, we consider a special directed cascade propagation model where the propagation probabilities are determined by the type of connection between the nodes. The propagation model will be described in detail in Section 2.5.

In this chapter, we consider two ways to improve propagation performance in networks: identifying influential individuals, often referred to as seeding (Domingos and Richardson 2001), and performing network alterations. Both seeding and alteration (creating new connections) efforts are typically costly. In ESN context, seeding cost can be seen as the cost of conveying a message to the influential employees, possibly through some form of training or education. In our approach, we model seeding as connecting a "source" (or a root node) to another node in order to enable efficient information propagation. In CSN context, the seeding is usually in the form of advertising, which is typically costly. Moreover, the amount of direct advertising that can be made is limited, as user attention will wane if too much advertising is presented, or worse, the user might stop using the product entirely. Advertisers may have the option to specifically choose the seeds for their advertising message (Facebook 2015a). There is a stream of literature on seeding of messages in order to improve propagation, such as Domingos and Richardson (2001), Nguyen and Zheng (2013), and Ni et al. (2010).

While seeding has an impact on message propagation, propagation may also be improved by altering the network itself. While Valente (2012) suggested three different tactics that might be considered for alterations (adding/deleting nodes, adding/deleting links, or rewiring existing

links), our focus is on adding links. However, such alterations face two challenges. First, identifying and executing an alteration is difficult and expensive. Second, there may be limits to the number of meaningful connections that each user can have (Gonçalves, Perra and Vespignani 2011). Utilizing Valente's generic strategies, this chapter focuses on finding alterations in the form of adding links, and at the same time locating the best seed nodes. Our approach is motivated by practice. Both Facebook and LinkedIn engage in network topology alteration strategies by suggesting new connections (links) in the networks. In fact, Facebook uses a number of methods to create new connections (Udemy 2014). These include suggesting friends of friends, suggesting users who search for a person to that same person, mining potential friends from school or work, and analyzing online activities for commonalities such as tagged photos, wall posts, likes, and comments. To determine the strength of connection, Facebook also allows users to designate friends as "acquaintances" in order to better estimate the connection strength (Udemy 2014). Facebook is also trying to connect people by showing content that users have liked or not liked (Kosner 2013).

In order to illustrate our problem context, we provide an example of how designed network alteration activities can take place in CSNs. Consider a professional software development company that is interested in promoting their new business analytics software through word-of-mouth marketing. They want to inform startup companies in the technology sector about their product. A professional social networking website such as LinkedIn has a base network that includes high-level managers in such companies, and is interested in business-to-business advertising (Carter 2012). Users can be reached both by directly seeding the advertising message to an individual, or by indirectly transmitting the message through word-of-mouth in a user-to-user manner. For example, some advertising campaigns in LinkedIn advertise live

discussions or messages that can attract a very special audience (Carter 2012). Facebook and LinkedIn have recently started targeting users based on their network of friends (Facebook 2015b; LinkedIn 2015). Some other examples of potential applications include healthcare education propagation (Amirkhanian et al. 2003), smoking cessation influence propagation (Christakis and Fowler 2008), and trust and distrust propagation (Guha et al. 2004).

This study is related to the stream of literature on network propagation, especially influence maximization, first introduced by Domingos and Richardson (2001). In the influence maximization problem, the goal is to find the set of initial nodes that, by seeding a message to them, will maximize the total number or proportion of influenced nodes in the network. Usually, the total number of available seeds is given, which can be interpreted as a budget constraint. Nguyen and Zheng (2013) provide a generalized version of this problem as the budgeted influence maximization problem. Ni et al. (2010) have taken a slightly different approach where they find the initial seed nodes that minimize expected complete influence time (CIT), and consider the case where the number of initial nodes is given. In the Ni et al. (2010) approach, CIT is defined as a measure of how long it takes to propagate the message to 100% of the population. In this chapter, we use the CIT to measure network propagation performance.

Our study is also remotely related to the literature on social contagion (Burt 1987) and influence networks (Friedkin and Johnsen 1990). The propagation mechanisms in influence networks are different from what we study in this chapter. Here, we study propagation of messages, that are not affected by externalities of the network or accumulation of influence.

The general propagation problem can be stated as follows: minimize the CIT of stochastic message propagation in a malleable network while simultaneously minimizing the cost of both

seeding and network modifications. In this chapter, instead of minimizing both CIT and costs, we posit a constraint on CIT, and minimize the costs. This is different from the prior literature in two key ways. First, we consider alterations in network connections in addition to seeding, whereas in prior research, networks are considered static. Second, we do not constrain the budget or the available number of seeds, but consider the cost to be a variable in the model. The HMST problem, as will be shown in Section 2.3, is a simplification of this propagation problem.

2.3 Problem Formulation

We first provide theoretical analysis for the problem of reaching all nodes in a linear network within a given hop constraint. We analyze networks with both unidirectional and bidirectional propagation flows, and provide the optimal solution structures based on the costs of adding seeds and new connections to the network. The details of the model are provided in Appendix 2.1. As shown in the analysis, closed-form results can be obtained for specific cases. We find that for a given set of non-dominated solutions with equivalent propagation performance, the optimal solution in terms of cost of seeds and new connections, depends solely on the ratio of seed to new connection cost. Moreover, we find that the structure of the optimal solution can contain all seeds, one seed and all connections, or a combination of many seeds and many new connections.

For the bidirectional propagation flow problem of a linear network with N nodes and hop limit of H, let C_S and C_C be the cost of seeds and new connections, respectively, utilizing N_S seeds and N_C new connections. In order to find the lowest cost solution, we need to solve the following problem (see Appendix 2.1 for how these equations are derived):

$$(N_S^*, N_C^*) = ArgMin[Cost(N_S, N_C) = N_S C_S + N_C C_C] \tag{2.1}$$

$$s.t. \quad N_S \in \{1, \ldots, \left\lceil \frac{N}{1+2(H-1)} \right\rceil\} \tag{2.2}$$ and $$N_C = \left\lceil \frac{Max\{0, N - N_S(1+2(H-1))\}}{1+2(H-2)} \right\rceil \tag{2.3}$$

For the unidirectional propagation flow problem, the only difference is that the constraints for N_S and N_C are altered as $N_S \in \left\{1, \ldots, \left\lceil \frac{N}{1+(H-1)} \right\rceil \right\}$, and $N_C = \left\lceil \frac{Max\{0, N-N_S(1+(H-1))\}}{1+(H-2)} \right\rceil$. Solving these problems gives the number of seeds and new connections for a given seeding and connection cost. Finding all non-dominated solutions of the problem, the optimal solution depends on the ratio of the costs (C_S/C_C). In our setting where the possible interventions in the network are individual seeding and network alterations, this is an intuitive result. The reason is that if the network administrator can choose between seeding and new connections, her/his decision will depend on the relative costs of each of these actions. We find that the optimal solution can have some combination of seeds and new connections, depending on the relative costs. See Appendix 2.1 for complete analysis of these particular cases.

The closed-form solutions presented in (2.1), (2.2), and (2.3) are applicable only to two very specific cases of linear networks. While illustrating that the solutions are dependent on relative costs of seeding and new connections, as well as the network structure, closed form solutions that apply to general networks are not possible. Next we seek to create a mixed-integer linear programming (MILP) approach to create a solution methodology that applies to any network structure. In the remainder of this section, we formulate the HMST problem for malleable networks.

Assuming that new connections can be created in the network at a cost, the problem at hand is to find the seed nodes and new connections to be created so that the expected CIT is minimized at lowest cost. Ideally one would want to minimize both expected CIT and cost, but this bi-objective problem is not easily solvable and might not have a single optimal solution. So we propose a simplification to this problem. To obtain the HMST simplification, the CIT is

treated as a constraint. The goal of the HMST is to create a minimum spanning tree network where no path exceeds a pre-specified number of links, known as hops. We argue that the HMST problem provides a suitable framework for finding the lowest cost seeds and connection improvements that will result in a network with acceptable CIT. The hop constraint in the HMST problem, which requires each node to be within *H* hops from the root node, limits the expected CIT in the network. The HMST problem in graph-theoretic notation is defined as follows:

> Let $G = (V, A)$ be a directed network with node set $V = \{1,2,\ldots,N\}$ where node 1 is defined as the root node, and *A* is the set of directed arcs (i, j) connecting nodes in V. A positive arc cost C_{ij} is associated with each directed arc. *The HMST problem is to find a minimum cost spanning tree (MST) of G, subject to a hop constraint.* The MST is a directed tree connecting every node such that every node other than the root node has exactly one incoming arc, and the root node has no incoming arcs. The hop constraint in the spanning tree limits the number of arcs on the unique path from the root node to any other node to be no more than the given number H.

Note that while the problem formulation is given for a directed network, the methodology can be used for networks with bidirectional connections as well. The important factor is that while the network can be bidirectional, the propagation flows is in a single direction. The root node in this problem represents the source of the information, and the nodes with direct connections to the root node represent the seed nodes where the message is seeded. The hops in a network represent the maximum allowable information degradation and/or transmission delay, i.e., CIT.

The HMST problem was first introduced by Gouveia (1995) to describe telecommunication networks with a guaranteed performance measure as dictated by the limited number of hops between the root node and any other node in the network. In such networks, delay and reliability are directly related to the number of hops that a message should travel until it reaches the destination. This problem has found applications in multicommodity flow (Gouveia 1996) and wireless network location (Clementi et al. 2005) problems. For a review on HMST problem formulations and methods, refer to Dahl, Gouveia and Requejo (2006), and more recently to Akgun (2011). The HMST problem creates a suitable framework for minimizing the cost while constraining the propagation quality to an acceptable limit. Figure 2.1 is a graphical representation of a 20 node HMST with a maximum of 3 hops in 100 by 100 Euclidean space. Node 1 (root node) is located in the middle of the space and represents the information source. Nodes directly connected to the root node are known as seeds. In the example given in Figure 2.1, there is only one seed node.

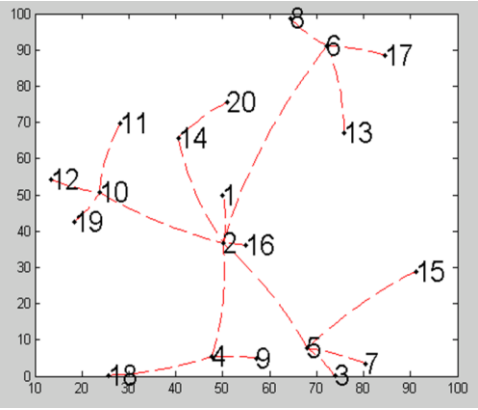


Figure 2.1. Graphic Representation of a 20 Node, 3 Hop HMST

To the best of our knowledge, the HMST problem has not been considered in the social network context. However, there is evidence showing that hop constraint also exists in social networks. Research suggests that "under 5" is a realistic hop length in many social settings.

Watts and Peretti (2007) provide several marketing examples for how many hops a marketing message can travel in a network, and show how the message intensity is diminished as it reaches outer nodes after hopping from node to node. Leskovec, Adamic and Huberman (2007) and Leskovec et al. (2008) study the propagation of viral marketing messages and evolution of social networks, respectively, to find that the majority of transmissions and connections happen within a limited number of hops. Cha, Mislove and Gummadi (2009) find that viral messages are not generally transmitted to nodes more than a few hops apart. Lv and Pan (2014) propose a special cascade propagation model that is modified to take into account that messages do not travel more than a given number of hops in order to represent word-of-mouth marketing. In the next section, we propose an improvement heuristic method for finding less costly HMST solutions, compared to methods that have previously been developed in the literature.

2.4 Solution Methods

The HMST problem is proven to be NP-hard (Gouveia 1995). Several heuristic algorithms have been proposed for HMST that use solution techniques originally developed for the minimum spanning tree (MST) problem, such as Prim's (1957), Kruskal's (1956) and Esau and William's (1966) algorithms. Clementi et al. (2005) modifies the Prim and Kruskal algorithms for the HMST. The solution quality of these algorithms is not satisfactory when there are some non-Euclidean perturbations in the data. Fernandes, Gouveia and Voß (2007) propose meta-heuristic solution methods using "repetitive" and "pilot" heuristics with multiple starting-point solutions. They use the Esau-Williams (EW) improvement algorithm in a repetitive algorithm to search for better solutions. In section 2.6, we find that these methods perform poorly in terms of solution quality and runtime for larger social network problems with perturbed and non-Euclidean costs. Among exact methods for the problem, Akgun (2011) presents the most efficient mathematical

programming formulation for the HMST to date. His novel formulation for the HMST problem uses Miller–Tucker–Zemlin constraints and outperforms flow-based (Gouveia 1996) and hop-indexed (Pirkul and Soni 2003) formulations both in terms of linear programming bounds and runtime. However, we find that exact procedures, including Akgun (2011), are not computationally efficient for solving large problem instances, especially for non-Euclidean problems.

Our goal in this section is to present a computationally efficient heuristic to solve the HMST problem. Here, we propose an improvement meta-heuristic technique that builds upon methods from prior literature as starting point solutions. We have taken a hybrid approach to solving this problem. As starting points, we use slightly modified forms of three techniques:

1. Prim: Prim's (1957) technique subject to the hop constraint;

2. FGV: Heuristic proposed by Fernandes et al. (2007) called ILA; and

3. Akgun: Akgun's (2011) formulation using CPLEX with time limited to 3600 seconds.

These algorithms are further described in Appendix 2.2. We next apply a series of heuristic steps to each starting point solution. The first step is a 2-opt swap (labeled *Swap)* where we search possible connections between any two nodes for a lower cost spanning tree. The second step is local branch optimization and aggregation. We create branches based on the nodes' number of hops in solution from the first step, optimize them using Akgun's (2011) formulation (using limited CPU time), and then aggregate the solutions into a single HMST. The third step is a one-opt heuristic expansion, a simplified version compared to the original heuristic expansion technique presented in Jayaraman, Patterson, and Rolland (2003). This one-opt heuristic expansion technique creates sub-problems by additionally considering all arcs entering or exiting

each node, one node at a time, and looks for improvements in the solution quality. Only one pass of all nodes is performed and the best solution is retained. We use the abbreviation *SLE* for the entire three-step improvement heuristic. *Swap* and *SLE* are both local optimization techniques that consider many simultaneous changes to the current solution. This local optimization process effectively drives towards improved solutions, all with reasonable computational effort. The algorithm implementation interactively uses Akgun's formulation to generate sub-problem solutions in steps two and three of the algorithm. In the next section, we discuss the propagation model that is used to calculate the simulated CIT in the networks.

2.5 Propagation Model

In this thesis chapter, we introduce the HMST problem as a method for finding seeds and connections to enhance propagation in malleable networks with the goal of total penetration. In order to analyze the propagation performance of the networks, we provide the details of our propagation model in this section. This model will serve as the foundation for the computational experiments, and enables us to analyze how the network alterations suggested by the HMST solutions can affect the expected propagation performance of information in the network in terms of CIT. The model is based on the cascade propagation model, as proposed by Kempe et al. (2003). In the cascade model, propagation of information starts from one or several activated nodes or seeds. The propagation occurs stage by stage, and continues until all nodes in the network are activated. At each stage, every activated node can activate its non-active neighbors with a certain probability. In our model, the probability is determined by the type of connection between the two neighboring nodes. In this study, we consider a special directed cascade propagation model where the propagation probabilities are classified into three sets as follows: existing connections in the network (Q), altered or newly created connections (M), and weak

connections (L). These probabilities are random variables in our simulation model. Let the expected probability value be as follows: $E(P(Q)) \geq E(P(M)) >> E(P(L)) > 0$. These sets correspond to connections that exist in the current network. Altered or newly created connections are alterations that are suggested by HMST solution. We consider a very small propagation probability for weak connections. As explained in Section 2.2, this is due to the fact that social network users may be influenced by users who they are not directly connected to. In ESN and CSN settings, this may happen through daily interactions or through internal or external newsletters for example. Because of this, the propagation model is devised in a way that every active node can potentially propagate to any non-active node at each stage, and thus complete influence will eventually happen in the network. We use expected complete influence time (CIT) as our propagation performance measure. The details of the propagation probability calculations are given in Appendix 2.3.

In order to compare the propagation performance of the altered networks with non-altered networks, we calculate CIT for the non-enhanced networks by employing both a random seeding and a greedy seeding algorithm. The greedy seeding algorithm finds the initial seed set by adding nodes that minimize the simulated CIT, and these nodes are added sequentially until the desired quantity of nodes to be seeded are in the set. This algorithm is equivalent to the algorithm provided in Ni et al. (2010) for the case where the target set includes all nodes. Refer to Appendix 2.4 for a detailed description of the greedy algorithm. The random seeding follows a similar algorithm, except that the nodes are added to the seed set randomly, without any preferences based on CIT improvement.

2.6 Computational Results

In order to test the impact of the HMST improvement heuristics, we study problems typical of both real-world and randomly generated networks. We test both ESNs and CSNs using real-world data. For ESN, we experiment on the Enron e-mail network. While the Enron data is now more than a decade old, it provides a network structure that is similar to the organizations of today. Moreover, while the organizational structures may evolve over time, the basic problem that our modeling approach captures has not changed, and our model of message propagation works for any network structure. For CSNs we use ego networks of Twitter and Facebook users, and also two sets of randomly generated problems of the type that are usually used in HMST literature in order to verify the efficiency of the proposed heuristic. Problem sizes included in the tests range from small (10 nodes) to extremely large (up to 1,500 nodes; see appendices 2.7 and 2.8). These problem sizes are larger than prior HMST literature which is typically limited to under 100 nodes. Moreover, these problem sizes are practical, as the networks that are being analyzed in this study are created by a pre-processing step or involve work networks or campaigns. We believe that the robustness of results from applying the methodology to different types of networks verifies the HMST approach, and the proposed heuristic. The basic problem that our modeling approach captures does not change over time or for different types of networks. We test our improvement algorithm with the three initial HMST solution algorithms discussed in Section 2.3. Prim, FGV, and Akgun represent the initial HMST solution algorithms and Prim+SLE, FGV+SLE and Akgun+SLE are solutions after the proposed improvement heuristic is applied to the three initial solution techniques. The hop constraints used in computational results are $H = 3, 4$ and 5. These parameters are based on prior research related to message propagation in various social networks. Specifically, Cha, Mislove and Gummadi

(2009) found that for Flickr photos, 4 hops covered 36% of the entire network. Ye and Wu (2010) found that for Twitter message propagation, 62.9% of the (re-tweeted) messages propagated 3 or less hops, and 37.1% propagated four or more hops. As such, we conclude that the range 3 to 5 hops is sufficiently appropriate and interesting. We can measure the propagation performance of the network with different number of seeds using the propagation setting described in Section 2.5. A random seeding (Random) and the greedy seeding method (Greedy) from Ni et al. (2010) are used to measure the average CIT with different number of seeds in the network without any network interventions. For a detailed description of problem generation procedures refer to Appendix 2.5. Experiments are run on an IBM X Series 3550 machine using 1 of 8 Intel Xeon CPU X5460 @ 3.16 GHz processors with 32 GB RAM, running Matlab 7.8.0 and AMPL with CPLEX 11.0.1.

2.6.1 Enterprise Social Network Problems

In this section we provide the results of using the proposed methodology on the Enron e-mail network released by the Federal Energy Regulatory Commission during an investigation on the firm. The Enron e-mail corpus has been extensively studied in social and communication network literature (Diesner, Frantz, and Carley 2005). Using the network implied by Enron's e-mail activity, we study how the proposed methodology can improve upon an existing ESN to propagate cascade-style messages that may help improve collaboration, efficiency, and innovation in the organization. The ESN representation is created using the internal e-mail data, having number of e-mails communicated between each pair of employees as the tie strengths. Multiple data sets are obtained by observing the cumulative e-mail history at different points in time. Figure 2.2 provides the average CIT-Cost figure for an Enron network. The HMST-altered solutions are provided for the three starting solutions (Prim, Akgun, and FGV) and the improved

solutions by the proposed improvement meta-heuristic (SLE), for the three different number of hops (*H=3,4,*5). For non-altered networks, Random and Greedy seeding methods are provided for comparison. CIT is calculated using the propagation model from Section 2.5. Because the CIT is an estimate of the expected value based on random simulations, we run 5 simulations for each problem set to find the average CIT for each problem set.

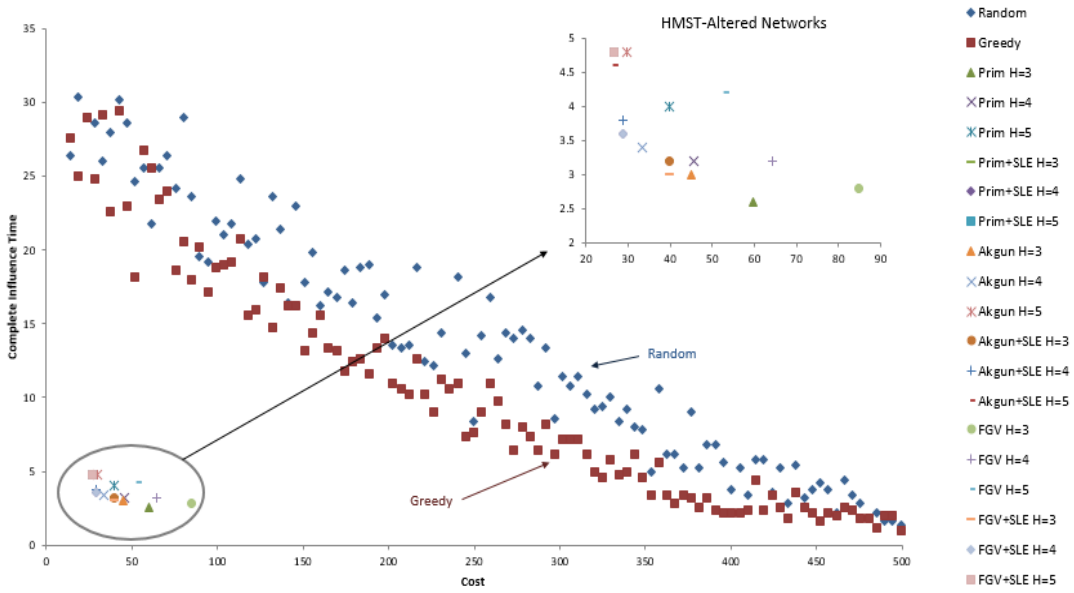


Figure 2.2. Average CIT-Cost for an Enron Network of Size *N=111*

By comparing the HMST-altered solutions to non-altered solutions of Random and Greedy, it can be seen that the improved solutions provide significantly better CIT at lower costs. While the HMST-improved methods for a given hop constraint have comparable CITs, the solutions that are improved by the SLE heuristic generally have lower costs. We provide further analysis of the Enron ESN problems in Appendix 2.6.

While the above analysis partially illustrates the performance of the proposed solution methodology to improve CIT for ESNs at low cost, in the next two subsections we provide

extensive experiments that will further validate the effectiveness and efficiency of the HMST approach.

2.6.2 Consumer Social Network Problems

In this section we apply the proposed methodology to CSNs. Here we analyze Twitter and Facebook ego networks (EN) from the dataset used in McAuley and Leskovec (2012). An EN is the network of all individuals connected to a single person, known as the ego node. The ego node itself is removed from the network for our analysis. We believe that, for the purposes of testing propagation performance, an EN is a good representative of the structure of a network of people with similar interests. The ego networks were collected from users in 2012, and represent established connections of a user, as is shown by relatively large number of degrees in these networks (Jure Leskovec, personal communication, June 11, 2015). The network calculations are explained in detail in Appendix 2.5.

We analyze Twitter and Facebook networks with a wide variety of sizes. In this section, other than the HMST-altered methods from Section 2.6.1, we also report the Prim+Swap solution created by performing only the two-opt swap heuristic improvement (Swap) on the initial Prim solution. This is the fastest of our HMST-altered solutions and can be calculated in reasonable time even for extremely large problems. The HMST solution costs and computational times are presented in Table 2.1. Solution values are presented as average percentage gap from the best solution among all available solution techniques. We have aggregated the problems to 2 groups of small and large for presentation purposes, the complete tables can be found in Appendix 2.7.

Table 2.1. Average Gaps from Best Known Solutions and Times for EN Problems

Problem Set		Average Percentage Gap from Best Known Solution							Solution Times in Seconds						
Network Size	Number of Problems	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun +SLE	FGV+ SLE	Prim	Akgun	FGV	Prim+Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10 to 45	30	9.2	0.4	50.0	4.6	0.2	0.1	0.4	0.0	2,776.4	53.3	0.1	10,093.4	12,288.1	9,882.3
82 to 124	6	75.4	23.2	182.6	31.9	3.1	1.7	1.0	1.2	3,601.1	114,036.9	6.7	56,956.0	47,980.3	161,029.6

It can be seen in Table 2.1 that Prim and FGV do not find any good solutions to the EN problems, and while Akgun performs generally well for problems of sizes smaller than 45, its performance deteriorates for larger problems. The SLE heuristic is able to improve all three initial solutions close to the best known solution. The proposed heuristic improvement results in substantial cost savings for EN problems, and the savings are increased as the problem size increases. It can be seen that all three initial algorithms perform poorly overall, but that our heuristic can greatly improve the solution quality. As the problem size increases, the computational time becomes increasingly burdensome for all methods except Prim+Swap.

We next analyze the performance of the HMST solutions in terms of propagation efficiency using CIT measure. Table 2.2 provides the summary of propagation results for the EN problems. The complete table is provided in Appendix 2.7.

While there are some variations in the CITs for different methods, all of them are within an acceptable range of CIT performance as compared to the respective hop constraints (H). These results confirm that the HMST structure in the network has succeeded in effectively propagating the message to all users within the hop constraint.

Table 2.2. Average CIT for EN Problems

Problem Set			Average Percentage Gap from Best Known Solution						
Size Range	Number of Problems	H	Prim	Akgun	FGV	Prim+Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10 to 45	30	3	2.4	2.3	2.1	2.3	2.3	2.3	2.3
		4	2.8	2.5	2.7	2.8	2.8	2.5	2.9
		5	2.9	2.8	2.8	2.8	2.8	2.8	3.0
82 to 124	6	3	3.0	2.8	2.5	2.9	3.0	2.8	2.6
		4	3.6	3.4	3.0	3.8	3.5	3.4	3.5
		5	3.9	3.6	3.7	4.0	3.7	3.9	4.0

Using CIT-cost graphs, Figure 2.3 illustrates the tradeoff between CIT and cost of the solution for a 96 node Twitter EN problem. The graphs for other problems exhibit roughly the same behavior, so we present only one problem here.

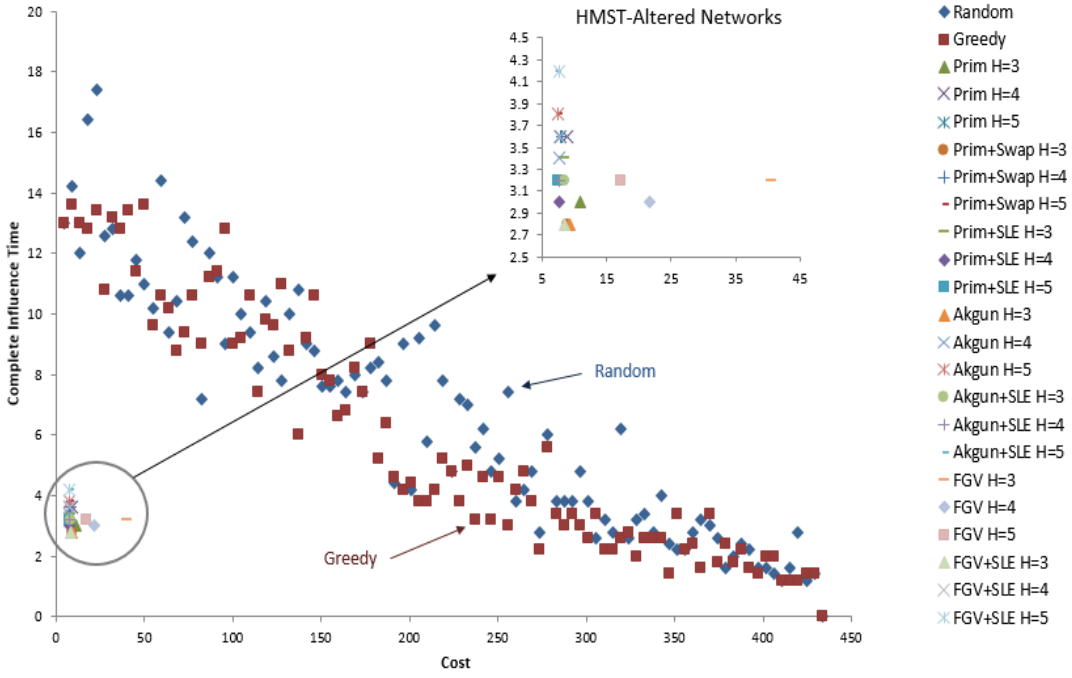


Figure 2.3. Average CIT-Cost for a Twitter EN Problem of Size N=96

Per visual inspection of Figure 2.3, we are able to confirm that while the Greedy seeding performs better than the Random seeding method; the HMST-altered networks again dominate

seeding-only methods in terms of both cost and CIT. Additionally, while all of the HMST-altered solutions are within an acceptable CIT range, it is seen that there are big differences among them in terms of cost. Further analysis of the ENs are provided in Appendix 2.7.

2.6.3 Randomly-Generated Social Network Problems

We have also studied randomly-generated problems. Two sets of problems were generated. In the first set, the connection cost between a pair of nodes is the Euclidean distance between the nodes that are randomly located on a square space, with root node (node 1) located in the center (EC problem set). This type of problem is common in HMST literature. For the EC problems, the previously available HMST solution techniques of Akgun and FGV perform well in finding low cost solutions in reasonable time. In terms of propagation, as expected, all of the enhanced networks have much better CITs compared to random and greedy seeding-only propagation. In the second set, some connection costs of the EC problems are randomly set to zero to indicate existing connections, and the costs of connections between disconnected nodes are recalculated based on the number of their mutual connections (perturbed-EC problem set). In perturbed-EC problems, the previously available HMST solution methods do not do so well in terms of solution quality, but the proposed SLE heuristic can greatly improve the solutions. Because the analysis on real-world ego network problems provides all the main results and discussions, for brevity, we present the computational results of randomly-generated problems in Appendix 2.8.

2.6.4 Discussion of Computational Results

In summary, we illustrate that the use of HMST-based formulation and additional heuristics of SLE and Swap have substantial benefits for improving propagation in malleable social networks. Specifically, we find that the HMST structure in the network helps to propagate a message to all users within the hop constraint. HMST-altered networks dominate seeding-only methods in

terms of both cost and CIT, and they require much smaller number of seeds to propagate effectively. In terms of the heuristic design, we have shown that our heuristic method, SLE, is able to identify good solutions for larger networks where previously proposed algorithms do not perform well. The SLE heuristic is able to improve on initial solutions in a variety of settings, including Enron ESN, and Facebook and Twitter CSNs, and randomly generated networks.

2.7 Discussion and Conclusions

A common assumption is that networks, and especially social network structures, are evolving naturally and cannot be manipulated by a network manager. However, we observe that whether the social network is face-to-face or electronic, there is a rather long history of attempts to modify the malleable network structure. Moreover, in organizational contexts we observe that companies do try to create new connections among their employees and their supply chain partners through ESNs, and with their customers through CSNs. In this chapter, we study how these malleable networks can be intervened in order to enhance propagation performance.

The contributions of our study are two-fold. First, we propose the HMST problem as a framework for finding low cost network enhancements having acceptable CIT within a hop constraint. This framework can jointly determine which nodes should be targeted for seeding and which connections should be created in order to improve propagation performance. Through simulation of a cascade propagation model, we have shown how creating the HMST solution for the network will improve the propagation as measured by CIT. Using the HMST approach on the network greatly reduces CIT at a reasonable cost, as compared to seed-only methods in randomly generated and real-world networks. Second, we show through computational experiments that creating any HMST in the network can greatly improve the CIT, but a good HMST solution is also more cost-effective. We propose the SLE improvement heuristic

procedure that greatly reduces the costs of the HMST solutions. From a Design Science research point of view, we advance the field through heuristic design improvements for the HMST. We illustrate how the HMST problem can be applied to the cascade propagation problem. Thus, our approach contributes to the theoretical development of Design Science research.

The clear (and most important) practical implication of our work is that social networks can become more effective conduits for message propagation by simultaneously considering network design changes and potential seeding points. Our work contributes theoretically by combining two streams of literature on seeding and network alterations. We demonstrate that combining these two components and considering them simultaneously can substantially improve message propagation.

This study has important managerial implications for different types of networks, such as social, collaboration, campaigns, and work networks. We empirically show that deliberate network manipulation is a possible tool for improvement of propagation in malleable networks, and can be possibly done using efficient heuristics. We demonstrate that use of active network alterations through creation of new connections helps to improve reach at a lower cost as compared to the seeding-only approaches. Such improvements can have substantial impact on the effectiveness of social networks. In CSNs, improvements can result in groups of users with similar interests, and this can improve marketing campaigns. In ESNs, the benefit is manifested through higher efficiency in execution of tasks and improved collaboration and innovation.

It should be noted however, that network administrators should perform network interventions with great care. If carried out poorly, user resistance due to improper or excessive

manipulation efforts can occur. Moreover, attempts to alter the network structure may be risky in terms of propagation improvement and costly.

Extending our work to consider threshold-style propagation is a viable and practical avenue for future research. Other possible formulations of this problem can consider the nodes or users as intelligent agents with utilities that depend on the amount of enhancement recommendations and the usefulness of such efforts. The question in such a game-theoretic research would be to examine system equilibrium and the impact of changes such as propagation probabilities on a possible equilibrium.

CHAPTER 3

How Much to Share with Third Parties? Users' Privacy Concerns and Publisher Website's Dilemma

3.1 Introduction

Publisher websites enable users to obtain various services and information. These websites outsource components of their websites, and present content and services provided by third party providers on their pages. Thus, the user experience of visiting a publisher website involves interactions with many third parties. Use of third parties is pervasive among top publisher websites. Gopal et al. (2014) investigate 700 popular websites, showing that these publisher websites utilize an average of 13.5 (and up to 70 in some cases) third parties. Since third parties can and do obtain user information from publisher websites, it creates natural tension between the use of third party components and user privacy. A U.S. Senate report states:

> "A visit to an online news site may trigger interactions with hundreds of other parties that may be collecting information on the consumer as he travels the web. The Subcommittee found, for example, a trip to a popular tabloid news website triggered a user interaction with some 352 other web servers as well. Many of those interactions were benign; some of those third-parties, however, may have been using cookies or other technology to compile data on the consumer. The sheer volume of such activity makes it difficult for even the most vigilant consumer to control the data being collected or protect against its malicious use." (United States Senate 2014)

The most familiar sharing mechanism involves the use of cookies, but other more sophisticated approaches exist as well. This sharing of user information with third parties is typically done without explicit user consent or appropriate disclosure mechanisms. For example, policies regarding shared information are often hidden deep within complex privacy statements, and users share readily via convenient one-click sign-up mechanisms such as "Sign up with your Facebook, Google, or Twitter account".

Reduction in cost of information storage and processing has enabled firms to collect and utilize large amounts of user information. It has become extremely difficult, if not impossible, to know who is tracking users online (Schoen 2009). Interestingly, the extent of collected information is not limited to browsing data, and the data can be used for identification or re-identification of individuals when used alongside other sources (Krishnamurthy and Wills 2009a). Privacy issues that arise from the increased usage of third parties and cookies is a public concern, and is being investigated by authorities and policy makers such as the Federal Trade Commission and the European Union (Mayer and Mitchell 2012). Turow et al. (2009) surveys users in the United States to find that between 68 to 87 percent of them do not want to be tracked for advertising purposes. McDonald and Cranor (2010) also find that only 20% of users prefer targeted online advertising over random advertising. Mayer and Mitchell (2012) provide a review of the policies and technologies surrounding web tracking. They note the fact that regulation is lacking behind the fast-growing industry, and emphasize the importance of discussions and debates on the topic.

Publisher websites can have a variety of revenue streams. Two main monetization approaches are 1) subscription services, and 2) selling of user information for purposes such as affiliate marketing (e.g., lead generation), targeting, and customization. Publisher websites

utilize one or a combination of these approaches. For example, consider the news websites for *Financial Times* (2015) and *Washington Times* (2015) depicted in Figure 3.1, highlighting various visible third party components. While *Financial Times* requires readers to subscribe in order to read articles, reading articles in *Washington Times* is free. Thus the *Washington Times* website operation depends entirely on income from third parties that pay the publisher website to obtain user information. We find that *Financial Times* shares with fewer third parties than *Washington Times* (22 compared to 36). In some cases, sharing of user information can be quite disconcerting.

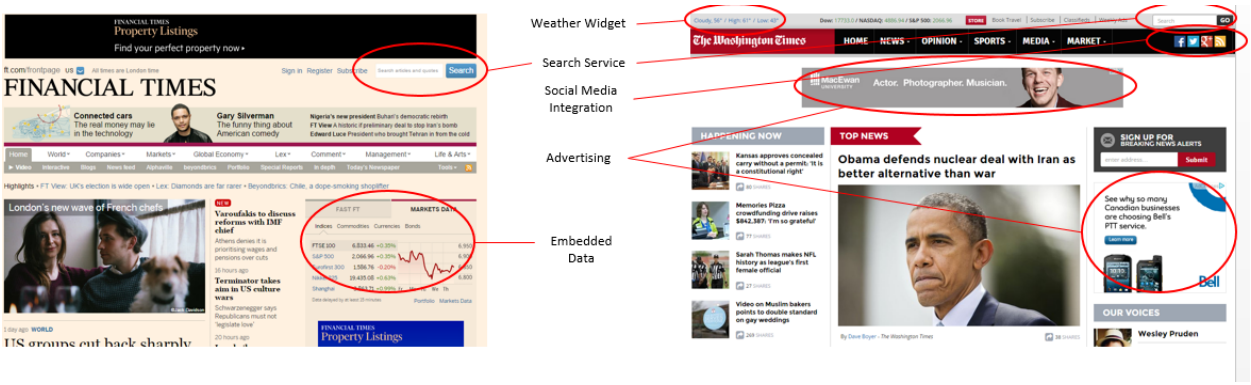


Figure 3.1. Two Sample Publisher Websites with Selected Third Party Content Highlighted

Most organizations involved with web publishing face a choice of using a variety of financial models to derive revenue. For example, the New York Times simultaneously employs both advertising and subscription revenues (New York Times 2016), including providing ads to users with subscriptions (Singleton 2016). Somaiya (2015a, b) describes this dual strategic approach by the New York Times. A mix of subscription and advertising revenues, which is consistent with the New York Times strategy, is often observed. As a second anecdotal observation, ads are presented in conjunction with mobile subscriptions by the Los Angeles

Times. The advertising may possibly be altered with a subscription, but clearly a blending of both revenue generation strategies is observed.

In the literature, some researchers have attested to the presence of several sources of monetization. For example, Kumar and Sethi (2009) note "The accumulated evidence indicates that pure revenue models, such as free-access models and pure subscription fee-based models, are not sufficient to support the survival of online information sellers. Hence, hybrid models based on a combination of subscription fees and advertising revenues are replacing the pure revenue models." Casadesus-Masanell and Hervas-Drane (2015) note "Firms compete for consumer information and derive revenues both from consumer purchases as well as from disclosing consumer information in a secondary market." One important way for websites to make revenue from third parties is through lead generation. Third parties involved in the collection of user information for lead generation include advertisers and advertising agencies, content providers, and data aggregators, among others. These third parties provide support and information required for improved targeting of website ads and improved sales opportunities. Similarly, behavioral targeting, which utilizes information shared with the many third parties, tracks user behavior within and across websites (Helft and Vega 2010) to combine user data to present the most relevant ads. We specifically consider the impact of user privacy concerns resulting from the sale of user information to generate additional revenues for the publisher website.

While the problem of information privacy in publisher websites is faced by many, it has not received much attention in the academic literature. An extensive body of work has addressed information privacy in the context of e-commerce where users willingly provide their information to companies (Li 2012), but there is a gap in the literature concerning the use of third

parties by publisher websites and the disclosure of personal information, along with publisher website's decision-making on deriving revenue from both users and third parties. In this chapter, we address this gap by considering the issue of web traffic monetization versus information privacy from an economic perspective. Using a two-sided stylized economic model, we analyze how the publisher websites control monetization of both users and third parties through setting user subscription prices and third party royalties. The analysis is provided for duopolistic publisher websites, for many users and third parties. We find that user privacy concerns can impact publisher websites' monetization decisions. In asymmetric settings, we demonstrate that publisher websites may choose drastically different business models, ranging from a focus on privacy-sensitive users to price-sensitive users. We also provide welfare analysis, and analysis of the impact of several practical regulatory tools that can be used to help improve the surplus of the users, publisher websites, and/or third parties.

The contributions of this chapter are as follows. First, we provide a two-sided economic model that describes the decision-making process of the publisher websites based on the privacy concerns of the user, participation incentives of third party service providers, and the publisher website's own incentives to maximize profits. Second, we discuss the effects of privacy concerns on the stakeholders, the impact on third party industry concentration, and implications for policymaking. Third, we contribute to the two-sided market literature, and discuss the problem where the two sides affect each other both positively and negatively. Finally, we provide an empirical validation and partial support for several important aspects of the model. This empirical validation also serves to illustrate the problems surrounding information privacy versus publisher website monetization that publisher websites, users, third parties, and policymakers face. The proposed model explains differences in third party sharing by publisher websites, and

provides managerial and policy insights for publisher websites, policymaking organizations, and governments.

The remainder of the chapter is organized as follows. Literature review is provided in next section. We then present the analytical model and an extension dealing with asymmetry in user privacy concerns. Subsequent sections present discussions on effects of privacy concerns on market concentration, analysis of the third party market structure, and implications for public policy and regulatory considerations. We conclude with the model robustness check, empirical analysis and discussion of our findings.

3.2 Literature Review

This chapter is related to several streams of literature. First, it is relevant to the literature on third party sharing in publisher websites. Second, there is the literature on online privacy and its implications. Lastly, we contribute to the literature in two-sided markets.

Despite the omnipresent use of third parties in publisher websites, work that addresses such third parties is sparse. Third parties do provide some benefit by providing additional services to websites which can be inferred by the pandemic use of third parties in websites (Mayer and Mitchell 2012). Adler et al. (2002) among others, have considered scheduling of online advertising. This stream of literature considers more technical aspects of resource management in publisher websites, but does not provide insights on the larger picture of publisher website decision-making in presence of revenues from both users and third parties, and when users have privacy concerns. Chen and Stallaert (2014) provide an economic analysis of online behavioral advertising. They find the conditions for which the use of behavioral advertising is better than traditional advertising for the publisher website. While their model can

incorporate the privacy concerns of users in the form of opting out of the service, the authors only consider the problem of choosing between traditional and behavioral advertising, and do not consider the problem from a privacy point of view. Kumar and Sethi (2009) also consider the problem of "online information sellers" and dynamic pricing in this context. Their study is one of the few that consider both subscription and advertising revenues simultaneously, and they use optimal control theory to dynamically price advertising and subscriptions.

There are many papers that study the effect of third parties on user information diffusion. Krishnamurthy and Wills (2006) find that "the size of the privacy footprint is a legitimate cause for concern". They also find significant increase in privacy footprint over a six month period. Krishnamurthy and Wills (2009b) show in a longitudinal study that the sharing and aggregation of user information has been increasing, while the number of entities involved has been decreasing as a result of acquisitions. One of the issues that arise as a result of privacy leakage is discrimination among different users. Krishnamurthy et al. (2011) study websites that require users to register and provide personal information, and find that 75% of the popular websites studied leak sensitive user information to third parties. Mikians et al. (2012) and Valentino-DeVries et al. (2012) among others, provide evidence for price and search discrimination in e-commerce setting, which is based on user information on the web. Recently, there has been even more concern about the implications of information sharing with third parties. Quintin (2015) provides evidence for sharing of users' health related and other information. This stream of literature magnifies the importance of understanding how publisher websites operate and their incentives in sharing user information. Malandrino and Scarano (2013) study how third party sites collect and aggregate data, and build personal profiles of users. They provide an empirical study on how user's privacy can be undermined because of such privacy violations, and

experiment with tools that can inform users and give them control over such activities. However, these tools are only used by tech savvy users, and not by the majority of users.

Online information privacy has been studied by several researchers. While this literature does not directly focus on third parties, many of the principles are applicable to information privacy in third party sharing. Using an economic model, Chellappa and Shivendu (2007) consider the personalization versus privacy tradeoff that users make when they reveal their personal information online for more personalization. Smith et al. (2011) provides a review of studies on information privacy. Li (2012) provides a comprehensive review of the extensive online information privacy literature, and provides a framework for theoretical research on the user's privacy decision-making. They provide that user information disclosure in form of third parties usage can be explained by theories such as privacy calculus theory, risk calculus theory, and dual-calculus theory among many foundational theories. In this chapter, we argue that the publisher website behavior cannot be viewed in isolation, as it is affected by both users and third parties. The publisher website realizes that users factor privacy in their economic evaluation of transacting with the publisher website. According to agency theory and utility maximization theory and their application in information privacy (Li 2012), a publisher website sets decision variables to maximize total profit from users and third parties collectively.

One of the more relevant studies to ours is Casadesus-Masanell and Hervas-Drane (2015), where authors study how the competition between online firms is affected by the user privacy concerns. This study is also one of the few considering the effect of privacy on publisher website decision-making. Similar to our study, they consider online firms who derive revenue from users via subscriptions and by "disclosing consumer information in a secondary market", where positive and negative cross-side network effects exist among users and third parties in the

secondary market. They find that in competition, firms differentiate among themselves and focus on one of the revenue sources. Our study is different from theirs in that we focus on the internal decision-making of a publisher website. We consider the factors that affect the balance that publisher websites must achieve between users' desire for privacy and monetization of user information, and the implicit privacy violations that monetization entails. Moreover, we focus on the participation of users and third parties, as well as the impact of user privacy concerns on third party industry.

This chapter is related to the significant literature on two-sided markets, where a platform provider is affected by two markets that interact and create network effects. Rochet and Tirole (2003) and Parker and Van Alstyne (2005) study the pricing strategies in such markets. Anderson et al. (2013) consider the platform investment in quality in two-sided networks. Most of the studies in this stream consider markets in which positive indirect network effects are present among the two markets. Casadesus-Masanell and Hervas-Drane (2015) is one of the few that consider both positive and negative cross-sided network effects. In this chapter, we model the problem as a two-sided market, having users on one side and third parties on the other, both contributing to the profit making of a publisher website. We consider both positive and negative network effects among the users and third parties. While third parties enjoy having more users on the publisher website, users do not appreciate third parties due to privacy concerns. This will be discussed further in the Section 3.3 below.

3.3 Model

In this section, we propose an economic model in order to describe and analyze the problem of third party usage in publisher websites. The notations are provided in tables 3.1 and 3.2.

Table 3.1. Model Parameters and Variables

Notation	Definition
y	Location of a user in Hotelling's model, $0 \leq y \leq 1$
t	Hotelling's fit cost or publisher website differentiation, $0 \leq t$
X	Intrinsic value of the publisher website for users, $X > 0$
$U_i(y)$	Utility of a user at location y for publisher website i. A user will use the publisher website with higher utility when, $Max\{U_1(y), U_2(y)\} \geq 0 \quad \forall\, y \in [0,1]$
N_{U_i}	Number of users for publisher website i, $N_{U_i} > 0 \quad \forall i = 1,2$
v	User's perceived disutility from each third party or user privacy concerns, $v \geq 0$
M_U	Total number of potential users in the market, $M_U \geq N_{U_i} \quad \forall i = 1,2, \;\; M_U > 0$
Π_{D_i}	Third party profit from publisher website i. A third party will join the publisher website i if $\Pi_{D_i} \geq 0$.
φ	Fixed cost of a third party, $\varphi \geq 0$
Φ	Maximum third party fixed cost, $\Phi > 0$
N_{D_i}	Number of third parties on publisher website i, $N_{D_i} \geq 0$
M_D	Total number of potential third parties in the market, $M_D \geq N_{D_i} \;\; \forall i = 1,2 \quad M_D > 0$
R_D	Third party's net revenue from each user's information, $R_D \geq 0$
Π_{W_i}	Publisher website i's profit, $\Pi_{W_i} \geq 0 \quad \forall i = 1,2$
Z_U	Total user surplus, total utility surplus for all users from both publisher websites
Z_D	Total third party surplus, total profit of all third parties from both publisher websites

Table 3.2. Model Decision Variables

Notation	Definition
R_{W_i}	Publisher website i's per user royalty (paid to the publisher websites by third party), $0 \leq R_{W_i} \leq \infty$
P_{W_i}	Publisher website i's price per user (paid to the publisher website by user), $0 \leq P_{W_i} \leq \infty$

3.3.1 Base Model

We consider a two-sided market model involving three sets of players: two publisher websites, publisher website users, and third parties. The publisher websites are seen as the platforms where users participate to get a certain utility, and third parties participate to get access to user information. The analysis is provided for two publisher websites, and for multiple third parties and users. The model employs a duopolistic price-maker with royalties and subscription prices as publisher websites' decision variables, and it incorporates the two-sided network effects of users and third parties.

In our model, the third parties can make revenue from the users on the publisher website. It is assumed that more users result in more information for the third party to collect. On the other hand, we assume that third parties provide no additional benefit to the user. While in many instances third parties do provide some utility to the user, here we consider the case in which the publisher website is considering whether or not to outsource a service on the publisher website, where the third party is a substitute for the publisher website's own service. In this setting the third party brings only disutility to the user with no additional benefit to the user than what they already receive from the publisher website in terms of an intrinsic value X. There are many studies that provide evidence for the negative utility of third parties for users (Krishnamurthy and Wills 2006; Turow et al. 2009). Moreover, Krishnamurthy et al. (2007) found that blocking of third parties does not significantly affect the usability of publisher websites.

In the duopoly setting, the two publisher websites compete for users. Users will choose exactly one of the two publisher websites (the user market is covered by two publisher websites), but third parties can participate in either of the publisher websites, both, or not participate at all. The publisher websites are symmetric in terms of the users' intrinsic valuation for the publisher

website, user privacy concerns, and the revenue that third party makes from user information. In Section 3.4.2, we relax the user privacy concern symmetry assumption.

A Hotelling model is used to differentiate users' utility from either publisher website. Publisher website one is located at location 0, and publisher website two is located at 1, with fit cost of t. Users are uniformly located between locations 0 and 1. If a user at location y decides to go with publisher website 1, her utility is modeled as:

$$U_1(y) = u_1 - ty, \qquad u_1 = X - N_{D_1}v - P_{W_1} \tag{3.1}$$

where N_{D_1} is the number of third parties on the publisher website 1, P_{W_1} is the price of using publisher website 1. v is user's disutility from each third party, or user's sensitivity to privacy violations. From now on, we call this parameter user privacy concerns. The argument ty is the user fit cost to use publisher website 1. Similarly, utility of user for the second publisher website is:

$$U_2(y) = u_2 - t(1-y), \qquad u_2 = X - N_{D_2}v - P_{W_2} \tag{3.2}$$

where N_{D_2} and P_{W_2} are the number of third parties on the publisher website 2 and the price of using publisher website 2, respectively. Users will choose the publisher website that yields higher utility. Thus, using the two utility functions, the location of the indifferent user between two publisher websites, $\hat{y}$ can be calculated as:

$$u_1 - t\hat{y} = u_2 - t(1-\hat{y}) \Rightarrow \hat{y} = \frac{t+(N_{D_2}-N_{D_1})\,v + (P_{W_2}-P_{W_1})}{2t} \tag{3.3}$$

Assuming that the total number of potential users in the market is M_U, the number of users for each publisher website i, N_{U_i} is calculated as:

$$N_{U_i} = M_U \frac{t+(N_{D_{-i}}-N_{D_i})\,v + (P_{W_{-i}}-P_{W_i})}{2t} \geq 0 \quad (3.4)$$

We assume that the third party pays per user royalties R_{W_i} to each publisher website i that they participate in. This is especially the case for the advertising and lead generation third parties, which pay the publisher website per impression or per click on ads, and this is directly correlated with the number of users on the publisher website. The business model of the third party is a form of revenue sharing, where the third party generates revenue from the service and/or lead generation on the website, and then shares some of their profit with the website. Per user royalty, or simply royalty, is set by the publisher website as a decision variable. While in practice royalties may be set by the third party, here we consider the case where the website sets the royalty. Doing so enables us to analyze and provide insights on how the publisher website balances the needs of users and third parties. The publisher website controls the number of third parties, and thus the total effect of user privacy concerns in their utility, by setting royalties. For a third party with the fixed cost of φ, the profit from each publisher website i is calculated as:

$$\Pi_{D_i}(\varphi) = N_{U_i} R_D - N_{U_i} R_{W_i} - \varphi = N_{U_i} (R_D - R_{W_i}) - \varphi \quad (3.5)$$

where R_D is the net revenue that the third parties can obtain from each user's information, or simply third party revenue from user information. Third parties have a fixed cost for their operations, which is assumed to be uniformly distributed over $[0, \Phi]$. The fixed cost of a third party that is indifferent between joining or not joining a publisher website i is characterized by $\widehat{\varphi_i} = N_{U_i} (R_D - R_{W_i})$. The third parties having $\varphi < \widehat{\varphi_i}$ will join the publisher website i. The ratio of third parties that will provide the service to publisher website i is calculated as $\frac{\widehat{\varphi_i}}{\Phi}$.

Assuming that there are overall M_D number of potential third parties in the market, the number of third parties that will join the publisher website i, N_{D_i} is calculated as:

$$N_{D_i} = M_D \frac{N_{U_i}(R_D - R_{W_i})}{\Phi} \geq 0 \tag{3.6}$$

For the number of third parties to be positive, we need to have $R_D - R_{W_i} \geq 0\,, \ \forall i = 1,2$. This is the third party participation constraint.

It can be seen that cross-sided network effects are present among the number of users and number of third parties. However, the network effects are not positive in both ways as we see in the majority of two-sided market literature. Instead, the externalities are positive in one direction (users to third parties) and negative in the other (third parties to users). In other words, third parties prefer a higher number of users, but users prefer a lower number of third parties. The duopolistic publisher website benefits from both, as they provide revenue for the publisher website.

Solving for N_{U_i} and N_{D_i} in (3.4) and (3.6), we can calculate the number of users and third parties for each publisher website i as:

$$N_{U_i} = M_U \frac{\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D v\,(R_D - R_{W_{-i}})}{2\Phi t + M_U M_D v((R_D - R_{W_i}) + (R_D - R_{W_{-i}}))} \qquad \forall i = 1,2 \tag{3.7}$$

$$N_{D_i} = M_D \frac{M_U(R_D - R_{W_i})(\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D v\,(R_D - R_{W_{-i}}))}{\Phi(2\Phi t + M_U M_D v((R_D - R_{W_i}) + (R_D - R_{W_{-i}})))} \qquad \forall i = 1,2 \tag{3.8}$$

As stated earlier, the publisher websites decide on subscription price and royalties. Profit for each publisher website i is calculated as:

$$\Pi_{W_i} = N_{U_i} P_{W_i} + N_{D_i}\, N_{U_i}\, R_{W_i} \qquad \forall i = 1,2 \tag{3.9}$$

Note that in this setting, the publisher website generates revenue from only two sources: users paying a price in exchange for access to the website, and third parties paying a royalty in return for user information. In reality, the website can also generate revenue from third parties without providing them with user information. An example for this is advertising where third parties only have access to information regarding the content of the publisher website, and not the users themselves. While this additional revenue source can be easily added to the model, we do not consider this for three reasons. First, the focus of this chapter is mainly on the privacy tradeoff that the users make by going to websites with third parties who collect information about users. Second, our experiments on the publisher websites (in Section 3.6) show that publisher websites predominantly provide third parties with user information. Third, a model with added revenue from third parties that do not collect user information (e.g., traditional advertising) provides the same major insights as the model without such added revenue, and therefore we omit these third parties for parsimony.

By substituting for N_{U_i} and N_{D_i} from (3.7) and (3.8) in the publisher website profit equation (3.9) we obtain the formula for publisher website profit. Each publisher website decides on its royalty and price, independently of the other publisher website. Using the first and second order conditions, we can calculate the optimal royalties and price of each publisher website, as is given in Lemma 3.1. In this case, the two firms set symmetric prices and royalties. The proofs for lemmas and propositions are provided in Appendix 3.1. We note from the discussion above and the discussion in Appendix 3.1 that our assumptions include that the publisher website profit is continuous and twice differentiable with respect to website price and royalties, there is a maximum profit (the profit function is concave with respect to both subscription price and royalties), and there exists a positive number of users and third parties.

Lemma 3.1 *In equilibrium, the duopolistic publisher websites choose the following symmetric royalties and publisher website prices*

$$R_{W_i}{}^* = R_W{}^* = \frac{R_D+v}{2} \qquad \forall i = 1,2 \qquad (3.10)$$

$$P_{W_i}{}^* = P_W{}^* = \frac{4\,\Phi t - M_D M_U (R_D - v)^2}{4\,\Phi} \geq 0 \qquad \forall i = 1,2 \qquad (3.11)$$

In order to have positive prices, we need to have $4\Phi\, t - M_D M_U (R_D - v)^2 \geq 0$. Using Lemma 3.1, Proposition 3.1 provides the effect of model parameters on the decision variables of the publisher websites.

Proposition 3.1 *The optimal publisher website royalty ($R_W{}^*$) and optimal publisher website price ($P_W{}^*$) in equilibrium satisfy the following:*

(i) $R_W{}^*$ increases with user privacy concerns (v) and third party revenue from user information (R_D).

(ii) $P_W{}^*$ increases with user privacy concerns (v) and publisher website differentiation (t). $P_W{}^*$ decreases with third party revenue from user information (R_D), total number of potential users (M_U), and total number of potential third parties (M_D).

Proposition 3.1 provides several important insights. As seen in part *(i)*, when users' privacy concerns are high, publisher websites set a high royalty price resulting in decreased third party participation and user information sharing. The royalty price, in this sense, is a lever for the publisher websites to manage the number of third parties. For a publisher website whose users are more concerned about their privacy, or who have sensitive information, the publisher website reduces the amount of privacy violation through increase in royalties. This demand

control mechanism can be observed when publisher websites charge higher prices for presenting fewer ads (Moss 2014).

On the other hand, when the revenue that the third party can make from user information increases, the third parties will be willing to pay more royalty to participate on the website. For example, the expected value of a purchase referral for an automobile sale increases when the purchase intent certainty is higher. When a user searches on publisher websites such as Edmunds.com, it is a very good indication that the user is on the market for an automobile, and the publisher website can charge the third party a high price for this lead generation. This lead generation phenomenon can also be observed in advertising keyword pricing such as Google's AdWords, where insurance, loans, and mortgage keyword searches demand high prices (Wordstream 2011).

From part *(ii)*, we observe that the publisher websites increase the price as users' privacy concerns increase. The reason for this is that as we see in part *(i)*, an increase in user privacy concerns will cause the publisher website to reduce third party usage through increased royalties. This causes fewer third parties to participate on the website (as is shown in Proposition 3.2). On the other hand, because of the lower number of third parties participating on the publisher website, users enjoy higher utility and thus have a higher willingness to pay. So the publisher website can increase its subscription price. This describes a natural phenomenon wherein if the publisher website cannot make profit from third parties, it needs to increase the user subscription price, which reduces the publisher website's user base. This can be seen with publisher websites that require a payment in order to remove advertisement from their page. For example, The Washington Post, Forbes, and Wired Magazine ask users who use ad blockers to either pay a certain fee or subscribe in order to be able to use the websites' services (Barr 2016). This can

also be construed as the fee for not having user's information shared with advertising third parties.

The price also increases with publisher website differentiation (measure by Hotelling's fit parameter, t). This is intuitive, because as the differentiation increases, publisher websites move towards monopolies, and can charge higher prices from the users. The implication of this is that when publisher websites operate in markets without real competitors, they can set high user subscription prices. As the market expands and new competitors enter the market, the business model of the publisher website transforms to no-fee subscription, and it focuses on revenue from the third parties.

The publisher website's optimal user subscription price will decrease if the third party's revenue from users increases. In this case, the third party can obtain higher revenue from user information, thus the third party is willing to pay higher royalties to the publisher website. At the same time, more users are willing to use the publisher website if the price is lower. Essentially, the publisher website increases user participation by dropping its prices, and monetizes the users through third parties. The example of valuable advertising keywords (Wordstream 2011) applies to this case as well, where high expected value of user information, may enable publisher website to forgo the subscription fee, for higher returns from the third party.

The effect of dropping the price of one side to monetize the other side is previously seen in two-sided market models where positive network effects are present among both sides, and is known as cross-sided network effect (Eisenmann et al. 2006). Here, we find that the effect is

also present in a two-sided market model with both positive and negative cross-sided network effects.

The publisher website price decreases with the total number of potential users in the market. Based on this, we expect that publisher websites with a specialized and small user base would set high subscription prices. In contrast, publisher websites with more generalized appeal and larger audiences would set low (or zero) subscription price, and rely on lead generation and revenue from third parties instead. Figure 3.2 summarizes the findings in Proposition 3.1.

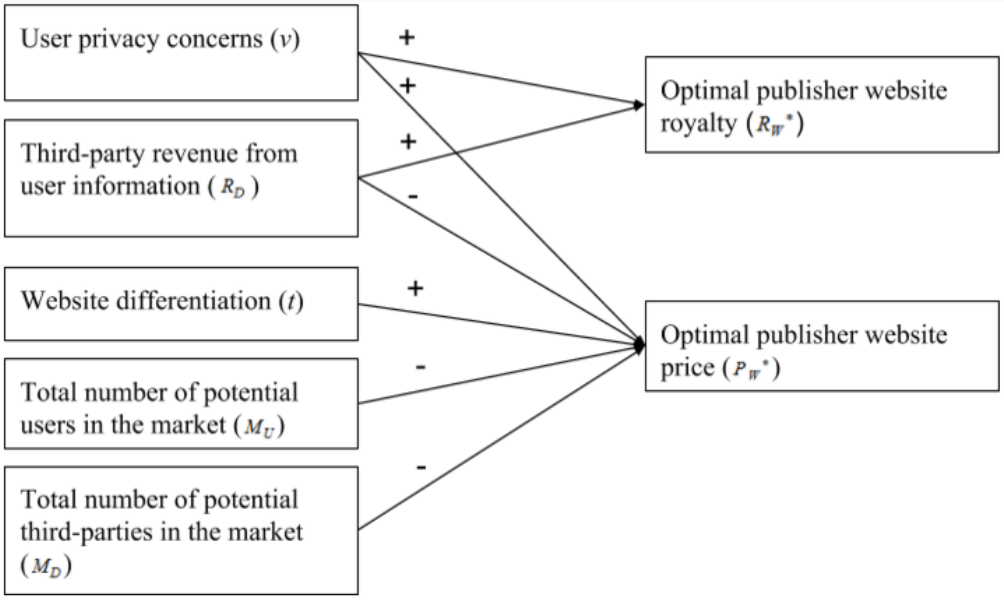


Figure 3.2. Effect of Model Parameters on Optimal Publisher Website Decision Variables

3.3.1.1 Number of Third Parties

By substituting the optimal formula for R_W^* and P_W^* from Lemma 3.1 into equations (3.7) and (3.8), optimal number of users and third parties on publisher websites can be obtained. For the number of users, the assumption is that the market is covered, and each user in the market is served by exactly one of the two publisher websites. Thus, the sum of number of users on two

publisher websites is equal to the total number of users, that is[1], ${N_{U_i}}^* + {N_{U_{-i}}}^* = M_U$ and we have:

$$ {N_{U_i}}^* = {N_{U_{-i}}}^* = {N_U}^* = \frac{M_U}{2} \tag{3.12} $$

On the other hand, third parties can participate in none, one, or both websites. The following proposition provides the effect of model parameters on the number of third parties.

Proposition 3.2 *The optimal number of third parties on publisher website i,* ${N_{D_i}}^*$ *is calculated as follows:*

$$ {N_{D_i}}^* = {N_{D_{-i}}}^* = {N_D}^* = M_D \frac{M_U(R_D - v)}{4\Phi} \geq 0. \tag{3.13} $$

The following results hold for the optimal number of third parties on the publisher website:

${N_D}^*$ *decreases with users' privacy concerns* (v) *and maximum third party fixed cost* (Φ)*, while it increases with third party revenue from users* (R_D)*, the total number of potential users in the market* (M_U)*, and the total number of potential third parties in the market* (M_D)*.*

Note that in (3.13) we assume that $R_D - v \geq 0$, and we further assume that $R_D - v > 0$, so that there are positive number of third parties present on the publisher website. Additional third parties pose risks to user privacy being violated and user information being misused. Thus, publisher websites that deal with sensitive user information would tend to do most of their operations themselves, rather than engaging third parties. We would expect publisher websites

[1] The subscript notation i and $-i$ for ${N_{U_i}}^*$ and ${N_{U_{-i}}}^*$ where $-i$ is read as "not i" can also be interpreted as $i = 1$ and $-i = 2$. We utilize the current to maintain consistency with prior literature.

dealing with sensitive content to work with substantially fewer third parties. Figure 3.3 summarizes findings in Proposition 3.2.

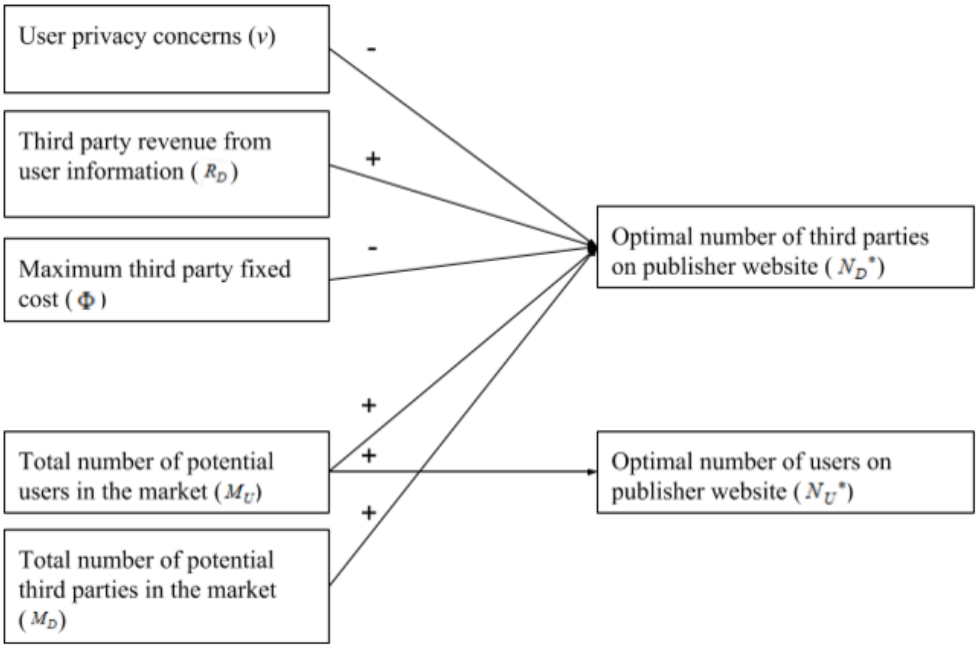


Figure 3.3. Effect of Model Parameter on Optimal Number of Users and Third Parties

3.4 Results and Analysis

3.4.1 Effect of User Privacy Concerns on Stakeholders

In this section we discuss how the model parameters affect the third party usage behavior. Proposition 3.3 examines the effect of privacy concerns on publisher websites, users, and third parties. An extension of this proposition which considers the effect of other model parameters on stakeholders is provided in the Appendix 3.2.

Proposition 3.3 *In equilibrium, the duopolistic publisher websites set the optimal royalty (R_W^*) and optimal price, (P_W^*) so as to maximize their profit, which yields optimal publisher website profit (Π_W^*), user surplus (Z_U^*), and third party surplus (Z_D^*) as follows:*

$$\Pi_{W_i}{}^* = \Pi_{W_{-i}}{}^* = \Pi_W{}^* = \frac{M_U(8\,\Phi\,t - M_U M_D(R_D - 3v)(R_D - v))}{16\,\Phi} \tag{3.14}$$

$$Z_U{}^* = \frac{\Phi(4X - 5t) + M_U M_D(R_D - 2v)(R_D - v)}{4\,\Phi} \tag{3.15}$$

$$Z_D{}^* = \frac{M_U{}^2}{16}(R_D - v)^2 \tag{3.16}$$

The following holds for optimal publisher website profit ($\Pi_W{}^*$), user surplus ($Z_U{}^*$), and third party surplus ($Z_D{}^*$)

(i) When $v < \frac{2}{3}R_D$ profit for each publisher website ($\Pi_W{}^*$) increases with user privacy concern (v) and when $\frac{2}{3}R_D < v < R_D$ it decreases with user privacy concern (v).

(ii) When $v < \frac{3}{4}R_D$ user surplus ($Z_U{}^*$) decreases with user privacy concern (v) and when $\frac{3}{4}R_D < v < R_D$ it increases with user privacy concern (v).

(iii) Third party surplus ($Z_D{}^*$) decreases with user privacy concern (v).

Figure 3.4 provides the effect of model parameters on the stakeholders.

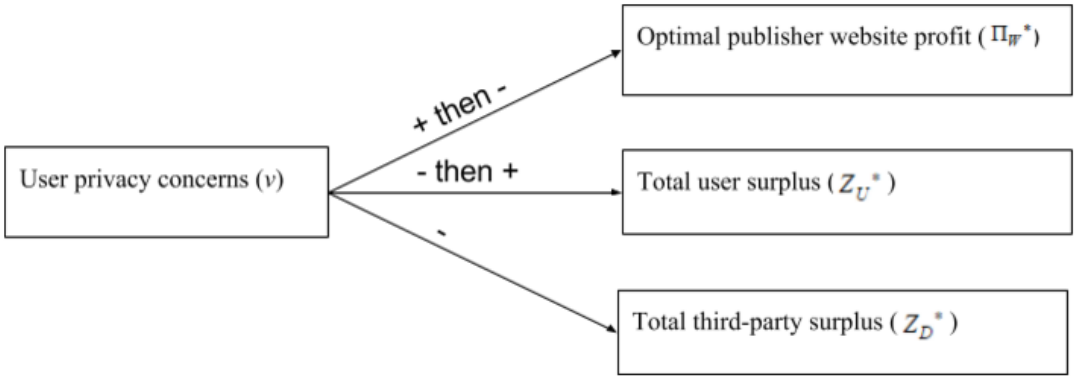


Figure 3.4. Effect of User Privacy Concerns on Stakeholders

Part *(i)* of the Proposition 3.3 provides that when the privacy concerns of users are relatively small (relative to the revenue that third parties make from user information), then publisher website profits increase as user privacy concerns increase. However, above a threshold, the profit decreases with increase in privacy concern. This implies that it is beneficial for the publisher website to have users with a moderate amount of privacy concerns. When user privacy concerns are too high, then the publisher website may not have the option of selling user information to the third parties, due to possible backlash from users. For example, Quintin (2015) reported that Healthcare.gov shared very personal information such as user "zip code, income level, smoking status, pregnancy status and more." When this was uncovered and published by a privacy watchdog, there was a backlash from users who demanded that the practice be stopped, and Healthcare.gov quickly discontinued the user information sharing. On the other hand, when user privacy concerns are on the low side, then users may not be willing to pay for publisher website subscription fees. From *(ii)* it can be seen that when user privacy concerns are relatively low, then the user surplus actually decreases with privacy concerns. However, when user privacy concerns are relatively high, then it is beneficial for users to have higher privacy concerns. In other words, the user surplus is convex with respect to user privacy concerns. The implication of this finding is that it is best for users if their privacy concerns are either very low, or very high. The results from *(iii)* is intuitive, as the third parties utilize user information, and if the users are concerned about this, the publisher website's response would be to cut the third party usage, and this would hurt the third parties.

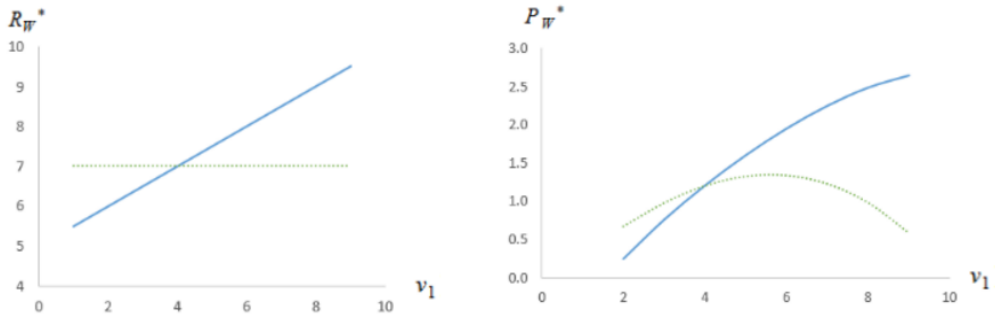
3.4.2 Asymmetry in User Privacy Concerns

Until now, we have considered symmetric publisher websites. In reality, publisher websites may face different user privacy concerns, perhaps as a result of differing brand reputations or

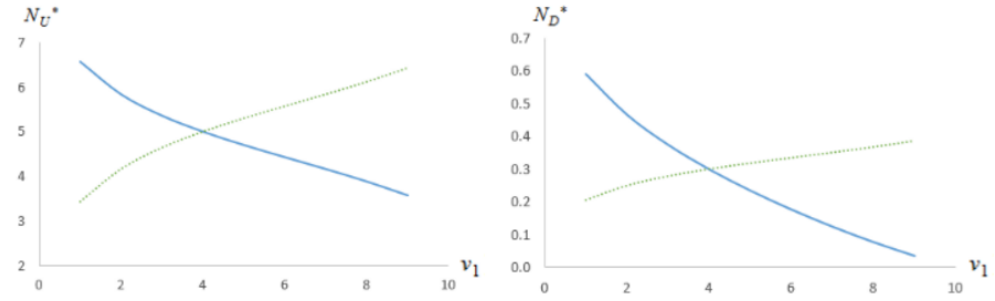
sensitivity of user information. In this section we numerically examine the asymmetric version of the base model with respect to privacy concerns. The details of the asymmetric model are provided in Appendix 3.3. We assume that the privacy concern for publisher websites 1 and 2 are v_1 and v_2, respectively. Here, we analyze the effect of changes in one publisher website's user privacy concerns on royalties and prices, the number of users, the number of third parties, and profit for each publisher website. We analyze these for a numerical example in Figure 3.5, where the privacy concern for publisher website 1 varies over the range 1 to 9, while the privacy concern for publisher website 2 is held constant at v_2= 4.

As shown in Figure 3.5.a, the publisher website 1's royalty increases as privacy concerns of their users increase, but is unchanged for publisher website 2. Figure 3.5.b provides that publisher website price is a non-linear combination of both websites' user privacy concerns. Publisher website 1 will charge higher prices as user privacy concerns increase. Changes in publisher website 1's user privacy concern also affects publisher website 2's prices, as publisher website 2's price initially increases and then decreases as website 1's user privacy concerns increase.

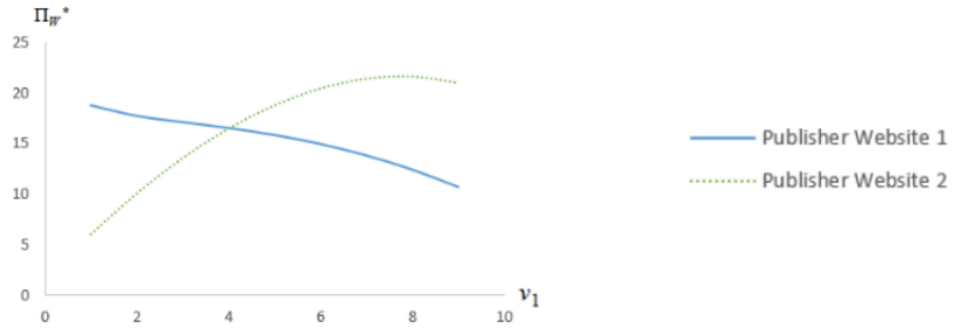
As seen in Figures 3.5.c and 3.5.d, the optimal number of users and third parties declines for publisher website 1 and increases for publisher website 2 over the entire test range. The combined effect of changes in both price and quantity is illustrated in publisher website profit, as presented in Figure 3.5.e. When website 1's user privacy concerns increase, its profit declines. However, publisher website 2's profits are also impacted by increases in website 1's user privacy concerns, as their profits initially increase, and then decrease.



a) Optimal publisher website royalties b) Optimal publisher website prices



c) Optimal number of users d) Optimal number of third parties



e) Optimal publisher website profits

Figure 3.5. Effect of Changes in v_1 when $v_2 = 4$ on the Two Publisher Websites

Sufficiently asymmetric user privacy concerns result in two different business models for the publisher websites. When website 1's user privacy concerns are high, publisher website 1 has a smaller niche market of customers willing to pay high publisher website prices in exchange

for privacy protection. Publisher website 2, which faces relatively lower user privacy concerns, has a larger mass market of customers who are willing to have their privacy violated in exchange for lower publisher website prices. The key insight is that different user privacy concerns faced by firms cause different business models to be adopted. In this sense, business model adoption can be seen, at least in part, as a reaction to the user privacy concerns each firm faces in the environment.

3.4.3 Effect of User Privacy Concerns on Third Party Market Concentration

Regulators are concerned about high industry concentration within third parties, especially with respect to concentrated user information and the possibility of re-identification. In a similar situation of concentrated user information, although in a different context, the Office of the Privacy Commissioner of Canada (OPC) ruled that Bell Canada's tracking of users' smartphone activity on an opt-out basis, rather than on an opt-in basis, was a violation of users' privacy due to inadequate consent (Dobby 2015). We can see from the Bell Canada case that high industry concentration of users' information, especially without their knowledge or adequate consent, would be of great concern to regulators and privacy watchdogs. A more comprehensive user profile is more valuable in terms of the potential to exploit users through the publisher website's improved ability to re-identify users and combine user information. Thus, higher industry concentration as a result of user information being sent to a few third parties from many publisher websites, constitutes a serious privacy concern. In this section, we analyze the effect of privacy concerns on third party market concentration.

In studying the third party market concentration, we consider two cases: 1. third parties with homogenous shares of the market, and 2. third parties with non-homogenous shares of the market. We use the Herfindahl-Hirschman Index (HHI) as a recognized measure for market

concentration. Here, we report only the main result of the analysis, and the details of the calculations are provided in Appendix 3.4. We find that for the homogeneous market share case, the market concentration increases with user privacy concerns. We also find that higher barriers to entry, such as an increase in the minimum allowable privacy and security level of the third parties, will also result in fewer third parties (by definition) and higher industry concentration.

Next, we reconsider the asymmetric model described in Section 3.4.2, where user privacy concerns for the two websites are not necessarily equal. We use the number of third parties on the two publisher websites to calculate the third party market shares. In analyzing the findings in Appendix 3.4, we note that the HHI in the non-homogeneous case is affected by two factors. First, HHI depends on the number of third parties, consistent with the homogenous case. As it can be seen from Figure 3.5.d in the previous section, the net effect of an increase in publisher website 1's user privacy concerns is that the total number of third parties decrease, and thus the HHI would increase as website 1's user privacy concerns increase. Second, HHI also depends on the market share of third parties, and this is impacted by the differences in the number of websites that each third party serves. We find that the HHI is maximized when the number of third parties on two websites are slightly different.

3.4.4 Implications for Policymakers: Taxation

In this section we consider the effect of regulatory organizations that can force taxes on users or third parties. We consider symmetric taxes, that is royalty and price taxes that are the equal for both publisher websites. We model this by performing the following transformations in the base model. The changes are made to the user utility equations (3.1) and (3.2) and third party profit equation (3.5), but not to the publisher profit equation (3.9). Taxes on royalties and on prices are shown by T_{R_W} and T_{P_W}, respectively.

$$P_{W_i} \rightarrow P_{W_i}(1 + T_{P_W}), \qquad \forall i = 1,2 \qquad -1 < T_{P_W} < 1 \qquad (3.17)$$

$$R_{W_i} \rightarrow R_{W_i}(1 + T_{R_W}), \qquad \forall i = 1,2 \qquad -1 < T_{R_W} < 1 \qquad (3.18)$$

The taxation can represent risk reserves that might be set by policymakers to be used in the case of an adverse incident, or can be seen as setting standards on information handling that might make the transactions harder and more costly. The taxations can also take negative values, meaning that the policymaker can take action to make the processes easier or less costly through subsidies. Lemma 3.2 provides the optimal publisher website royalty and price when these taxations are included in the base model.

Lemma 3.2 *When taxations are possible, the publisher websites in duopoly choose the following royalties and publisher website price in equilibrium.*

$$R_{W_i}{}^* = R_{W_{-i}}{}^* = R_W{}^* = \frac{R_D(1+T_{P_W})+v(1+T_{R_W})}{2(1+T_{P_W}+T_{R_W}+T_{P_W}T_{R_W})} \geq 0 \qquad (3.19)$$

$$P_{W_i}{}^* = P_{W_{-i}}{}^* = P_W{}^* = \frac{4\Phi t(1+T_{P_W})(1+T_{R_W})-M_D M_U(R_D(1+T_{P_W})-v(1+T_{R_W}))^2}{4\Phi(1+T_{P_W})^2(1+T_{R_W})} \geq 0 \qquad (3.20)$$

The proof for this Lemma follows the proof for Lemma 3.1, and by making transformation (3.17) in user utility equations (3.1) and (3.2), and transformation (3.18) in third party profit equation (3.5). Using Lemma 3.2, Propositions 3.4 provides the effect of taxations on the publisher website decision variables of optimal publisher website royalty and price.

Proposition 3.4

(i) The optimal royalty price ($R_W{}^*$) decreases with taxation on third party royalties (T_{R_W}) and publisher website price (T_{P_W}).

(ii) The optimal publisher website price (P_W^*) increases with taxation on third party royalties (T_{R_W}) and decreases with taxation on publisher website price (T_{P_W}).

Proposition 3.4 yields several interesting insights. The publisher website will decrease the royalty price as a response to taxation on user revenues. The intuition for this is that taxation on the subscription price will reduce the number of users on the publisher website. In order to maximize profit, the publisher website increases number of third parties on the publisher website by reducing the royalties, and this increases the publisher website's profit from third parties. Similarly, the publisher website will decrease the royalty price as a result of taxation on third party revenues. Due to reduced third party interest because of taxation, publisher website increases third party incentives to join the publisher website by decreasing the royalties. The publisher website tries to regain the lost demand due to taxation on user revenues through user price reductions. However, the publisher website will increase price as taxation on third parties increases. This is because taxation on third parties decreases the publisher website revenue from third parties, and publisher website tries to replenish this by increasing the user price. Figure 3.6 provides the effect of taxations on optimal publisher website royalty and price.

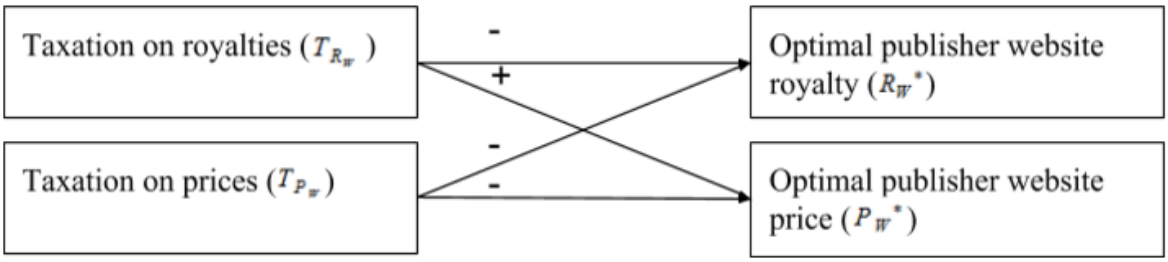


Figure 3.6. Effect of Taxations on Optimal Publisher Website Decision Variables

We next study the effect of taxations on website profit, user surplus, and third party surplus in Propositions 3.5 and 3.6.

Proposition 3.5 *When the publisher website chooses royalty ($R_W{}^*$) and the optimal publisher website price ($P_W{}^*$) so as to maximize profit, then:*

(i) When $v < \frac{R_D}{\sqrt{3}} \frac{1+T_{P_W}}{1+T_{R_W}}$ *publisher website profit ($\Pi_W{}^*$) increases with taxation on royalties (T_{R_W}) and when* $\frac{R_D}{\sqrt{3}} \frac{1+T_{P_W}}{1+T_{R_W}} < v < 1$ *it decreases with taxation on royalties (T_{R_W}).*

(ii) When $v < \frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}}$ *user surplus ($Z_U{}^*$) decreases with taxation on royalties (T_{R_W}), and when* $\frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}} < v < 1$ *it increases with taxation on royalties (T_{R_W}).*

(iii) Total third party surplus ($Z_D{}^*$) decreases with taxation on royalties (T_{R_W}).

It can be seen from *(i)* that the taxation on royalties can increase or decrease the publisher website profit and user surplus, and this depends on the privacy concerns of the users. When users are not very concerned about their privacy, taxation can improve publisher website profit. However, if users are concerned about their privacy, then taxation would decrease publisher website profit.

The effect of taxation on users is the opposite of its effect on the publisher websites. Taxation is beneficial for users with high privacy concerns, and is detrimental for users with low privacy concerns. Moreover, there is no range for privacy concerns in which the taxation on royalties can increase both publisher website profit and user surplus. Thus the regulator needs to decide which player they want to benefit, and what is the cost of doing that on the other player. Taxation on royalties always decreases the third party surplus.

Proposition 3.6 *When the publisher website chooses royalty ($R_W{}^*$) and the optimal publisher website price ($P_W{}^*$) so as to maximize profit, then:*

(i) Publisher website profit ($\Pi_W{}^*$) decreases with taxation on price (T_{P_W}).

(ii) When $v < \frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}}$ user surplus ($Z_U{}^*$) increases with taxation on price (T_{P_W}), and when $\frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}} < v < 1$ it decreases with taxation on price (T_{P_W}).

(iii) Total third party surplus ($Z_D{}^*$) increases with taxation on price (T_{P_W}).

It can be seen that the price taxation decreases publisher website profit. Interestingly, price taxation can either increase or decrease user surplus. It increases user surplus when user privacy concerns are low, and it increases user surplus when user privacy concerns are high. Thus while taxation on price does not benefit the publisher website, it does benefit users when their privacy concerns are low. Therefore, price taxation is a viable tool for benefiting the users when their privacy concerns are relatively low.

Figure 3.7 provides the summary of results from propositions 3.5 and 3.6.

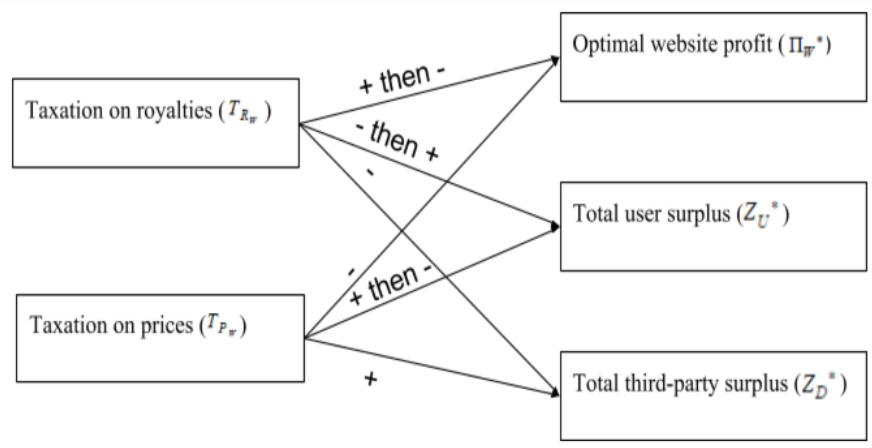


Figure 3.7. Effect of Taxations on Stakeholders

3.4.5 Implications for Policymakers: Collusion on Royalties

In this section we study the effect of collusion among publisher websites on the stakeholders. When firms collude in order to increase profits, they can set royalties and prices other than their equilibrium values. Since the duopoly market is covered in our model, analyzing collusion in prices does not provide interesting insights, and here we only study the effect of collusion in terms of royalties.

In order to analyze the effect of setting royalties other than their equilibrium values, we analyze the profits that publisher websites can obtain with and without collusion. For the case when collusion is not possible, we consider that publisher websites set royalties to its equilibrium value. For the case of collusion, we consider both publisher websites can collaboratively decide on the royalties. The details of the calculations are provided in Appendix 3.5. The profit curves for the publisher websites with respect to royalties with and without collusion for a numerical example are provided in Figure 3.8.

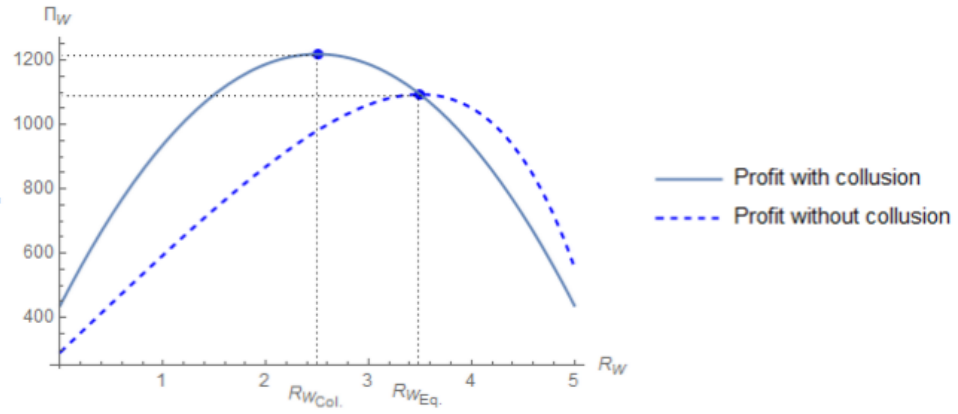


Figure 3.8. Publisher Website Profit with and without Collusion with respect to Royalties

The dashed line is the profit for publisher websites, when one of the publisher websites sets the equilibrium royalties, $R_{W_{Eq.}}$ and the other publisher website sets royalties to R_W. The

profit in this case is maximized at the equilibrium point, $R_{W_{Eq.}}$. However, if the publisher websites can collude and set identical royalty R_W, then they will decrease their royalty to $R_{W_{Col.}}$, where both firms make higher profits ($\Pi_{W_{R_{W_{Col.}}}}$) than the equilibrium profit ($\Pi_{W_{R_{W_{Eq.}}}}$). It can analytically be shown that these results hold irrespective of the parameters (refer to Appendix 3.5). In other words, the following hold:

$$R_{W_{Eq.}} > R_{W_{Col.}} \tag{3.21}$$

$$\Pi_{W_{R_{W_{Eq.}}}} < \Pi_{W_{R_{W_{Col.}}}} \tag{3.22}$$

The collusion among publisher websites results in lower royalties overall, and thus is also beneficial for the third parties. This collusion however, is not beneficial for the users, as they will be exposed to more third parties due to decrease in R_W. Intuition behind this can be explained as follows. The competition among the two publisher websites for attracting users, drives them to increase the royalties, as increasing royalties benefits user utility through reduced number of third parties (and hence lower privacy concerns for users). If both publisher websites can simultaneously reduce royalties through regulation or collusion (which simultaneously reduces user utility), then publisher website profits can be maximized and third party surplus will increase. The role of a regulatory organization who is interested in increasing user surplus would be to prevent such collusion, possibly by setting minimum required royalties. Note that this regulatory mechanism behaves differently than royalty taxation in terms of the effects on publisher website profit, third party surplus, and user surplus.

3.5 Robustness Check

In this section, we test the robustness of the model that was presented in the chapter. We compare the base duopoly model to a duopoly model with a nonlinear utility function for users, and to a monopoly version of the model.

3.5.1 Duopoly Model with Nonlinear Utility Functions

Here, we test a non-linear utility function for users, and transform the user utility functions (3.1) and (3.2) as follows:

$$U_1(y) = u_1 - ty, \qquad u_1 = X - {N_{D_1}}^2 v - P_{W_1} \tag{3.23}$$

$$U_2(y) = u_2 - t(1-y), \qquad u_2 = X - {N_{D_2}}^2 v - P_{W_2} \tag{3.24}$$

This is the case when the users get quadratic disutility from presence of third parties on the publisher website. Solving the problem with nonlinear utility function is not tractable, however, it can be solved numerically. We use numerical analysis to test if the main results that were derived from the duopoly model hold for this model as well. We report the key results from the analysis, and provide the details of the numerical results in Appendix 3.6. We find that the publisher website decision variables of royalties and prices in the model with nonlinear utility behave similarly to the base model. Specifically, the optimal publisher website royalties increase with user privacy concerns and third party revenue from user information, and the optimal publisher website prices increase with user privacy concerns and decrease with third party revenue from user information. Similar to the base model, the number of third parties decreases with user privacy concerns and increases with third party revenue from user information.

The two models however, differ in terms of the effects on publisher website profit, and user and third party surplus. Generally, the two models yield similar results for the higher range of user privacy concerns, however the nonlinear model does not account for the effects on publisher website profit and user surplus for small values of user privacy concerns.

Overall, we conclude that the base model is robust, in that the main findings are unchanged when modeling with a nonlinear utility function.

3.5.2 Monopoly Model

We have also compared the monopoly and duopoly cases. In the monopoly, there is no Hotelling's fit cost (t), instead, the users are differentiated based on their intrinsic utility for the publisher website, x, assumed to be uniformly distributed between 0 and X. The details of the monopoly and its comparison with duopoly are provided in Appendix 3.7. Comparing the monopoly model to the duopoly, we find that the decision-making behavior of the publisher websites in both models are similar, in that the effect of user privacy concerns on the royalties and prices are similar to the base model. Number of third parties on the publisher websites also behave similar to the duopoly model. For the number of users, while in the duopoly the market is covered, it may not be covered in the monopoly, and thus the results differ. The duopoly model enables us to study the effect of competition through the Hotelling's fit cost parameter.

The results on publisher website profit, user surplus, and third party surplus between duopoly and monopoly models are different, yet consistent in behavior. Similar to what we saw with the model with nonlinear utility function, the monopoly model does not account for effects on publisher website profit and user surplus for small values of user privacy concerns.

3.6 Empirical Analysis

We find partial support for the proposed model by empirically examining the number of third parties utilized by different categories of publisher websites, as well as the industry concentration of third parties. We carry out an exploratory validation study on the 100 most-visited publisher websites from each of seven different subject categories (news, arts, shopping, kids and teens, health, business, and adult) for a total of 700 publisher websites. We record the third parties utilized on each website. To better capture the structure of the third party industry, we profile the third parties and divide them based on the industry sectors. The three industry sectors are targeting/advertising (*T/A*, e.g. advertising presentation and analytics), functionality (*F*, e.g., password security, social media integration, video hosting, chat and forum services, and payment services), and performance (*P*, e.g., backup service, publisher website security, and responsiveness tools). Appendix 3.7 provides the details of empirical analysis and model validation.

Many outcomes of the model are not empirically observable in our validation study. We can, however, make predictions regarding user privacy concerns in different publisher website subject categories, and observe and compare the number of third parties and industry concentration among these categories. If the empirical study is consistent with our a priori expectations from the model, then we can conclude that the model is partially validated. Figure 3.9 illustrates our expectations regarding empirical observations based on the analytic model.

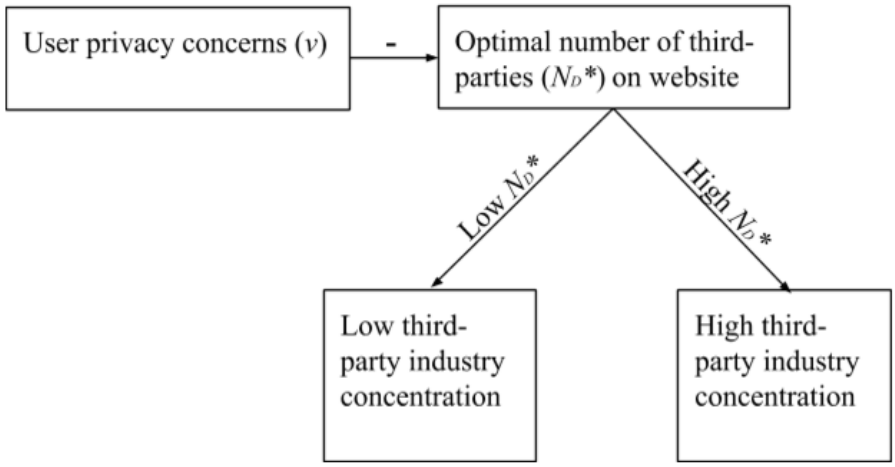


Figure 3.9. Outline of Conjectures for Empirical Validation

Noting that information sensitivity and user privacy concerns likely vary among different content subject categories, we expect the sharing behavior to differ for publisher websites with different subject categories. Comparing the three industry sectors, we find that the number of third parties used in the *T/A* industry sector to be significantly higher than both *F* and *P*. The *F* sector has significantly higher sharing than *P* in 3 out of the 7 subject categories, and overall. The *T/A*, *F*, and *P* sectors comprise 60, 20, and 15% of all third party connections. Thus, sharing across different third party sectors varies. The results suggest that user information is falling into the hands of many companies who use the information for targeting, lead generation, and advertising purposes, which provides credence to concerns raised in the literature (e.g., Krishnamurthy and Wills 2006, Mikians et al. 2012, and Valentino-DeVries et al. 2012).

In terms of publisher website subject categories, news is the category that shares with most number of third parties. Note that while the publisher website's business model is beyond the scope of this study, it may also influence the use of third parties. In the case of news, the industry has a history of revenues coming from both advertising and subscription fee business models, and in many cases, publisher websites such as the Los Angeles Times and the New York

Times employ a mixed, or freemium, model (e.g. 5 free articles per month for the Los Angeles Times and 10 free articles per month for the New York Times). On the other hand, the adult subject category has the least average sharing, followed by business and health for all three industry sectors (*T/A*, *F*, and *P*). Specifically, health websites obtain sensitive information (Quintin 2015), which is typically associated with higher privacy concern for users. These observations support the findings of the model that predicts that sharing should be lower for publisher website categories where users' privacy concerns and information sensitivity are greater.

We also examine the third party market concentration using Herfindahl-Hirschman index (HHI). We find that the *T/A* sector has the lowest HHI concentrations, followed by *P*, and then by *F* for all subject categories analyzed separately. There are relatively fewer third parties in *F* and *P* sectors, and analysis of the HHI values indicate that these industry sectors have higher number of large dominant third parties. The market concentration results are also in tandem with findings of the model, where categories with higher privacy concerns are found to have higher HHI. We do not find evidence for the popularity of publisher websites (as measured by monthly unique visitors) to have any significant effect on third party sharing.

The *T/A* sector is clearly dominant in terms of number of third parties involved. One reason behind this is that the information involved in *T/A* sector is perhaps less sensitive than the *F* and *P* sectors. Publishers tend to stick with smaller number of third parties in the *F* and *P* sectors. It is safe to assume that privacy concerns play an important role in the sharing behavior within each sector. Another reason could be that more money is potentially available for *T/A* versus *F* and *P*.

3.7 Conclusions and Directions for Future Research

In this chapter, we present a two-sided economic model to explain and analyze the decision-making of publisher websites who must balance user pricing and privacy. The publisher website needs to maintain their user base to increase profits. However, they have a secondary source of profit from third parties. A publisher website must balance this with monetization through third party information sharing and the subsequent personal privacy violations that result from this sharing, along with the associated declines in the user base due to third party monetization, on the other hand. The model describes how user privacy concerns drive publisher website decision-making and third party market structures, with higher privacy concerns driving higher industry concentration. We also explore the effect of competition and asymmetry among publisher websites, and provide insights on the different publisher website business models that arise. We also provide several policy and welfare implications, and analyze the effect of regulatory decisions such as taxation and setting minimum royalties on the stakeholders.

This chapter makes several important contributions to our understanding of the publisher website third party sharing problem. An important implication is that firms facing different user privacy concerns can be driven to one of two business models: 1. Low publisher website price and high user privacy violation for the firm facing low user privacy concerns, and 2. High publisher website price and low user privacy violation for the firm facing high user privacy concerns. Empirical results are consistent with this finding.

We also show how increased user privacy concerns decrease the number of third parties utilized by publisher websites, and how this in turn can lead to substantially higher third party industry concentration. These higher industry concentrations represent an irony with respect to the impact of user privacy concerns on publisher website third party usage. When privacy

concerns are relatively high, such as for the health and business categories, the publisher website uses fewer third parties. This is shown to cause higher industry concentration among the third parties. Thus, if a user visits multiple publisher websites in a category with high privacy concerns, then the user's data is more likely to be aggregated at these fewer number of third parties. This concentration of information sharing constitutes a privacy threat in its own right, due to concerns regarding re-identification and aggregation of data. We validate some key outcomes of the model with an empirical validation study which confirms these key model outcomes regarding user privacy concerns, third party utilization, and industry concentration.

In the empirical study, we find that user information is being shared extensively among third parties by publisher websites, and the actual third party usage behavior is consistent with the predictions of the model. Due to privacy concerns and potential for re-identification, the extent of third party sharing is of strong interest to policy makers and regulatory organizations. Also, from the publisher website administrator's point of view, sharing generates ad revenue and potentially better service. However, too much sharing will violate user privacy and reduces usage. We see these market forces being reflected in sharing levels and industry concentration measures between industry sectors and subject categories.

We examine the impact of two government taxation policies (taxation on royalty revenues and taxation on subscription revenues), and find that the impacts of a sales tax on the activity between the user and publisher website differ from a tax on the third party activity. Profit and welfare impacts of these two taxation policies are examined. We find that the impact of such actions, depend on the level of user privacy concerns, and the goal of the policy makers.

We contribute to the two-sided market literature by considering the case where one side has a negative cross-sided network effect. Traditionally, with the exception of Casadesus-Masanell and Hervas-Drane (2015), cross-sided network effects are positive for both sides (e.g., Eisenmann et al. 2006, Rochet and Tirole 2003, Parker and Van Alstyne 2005, and Anderson et al. 2013). In two-sided markets where cross-sided network effects are positive for both sides, a common strategy is to drop the price to one side in order to monetize the other. We illustrate that this strategy remains effective when one side has a negative cross-sided network effect as well. In our problem, where users have negative cross-sided network effects from third parties, the publisher website can decrease the revenue from users in order to monetize the third parties. Conversely, the publisher website's strategy could be to decrease the third party privacy violations, which decreases the revenue from third parties, and increases the revenues from users.

There are several limitations to this research. Note that different business models may dominate one publisher website category versus another. The concern is that differences in third party sharing may be due to differing business models, rather than users' privacy concerns related to the nature of the publisher website subject. While we try to avoid this issue by considering the decision-making only from the privacy point of view, focusing on both the business model and privacy concern can be a good avenue for future research. We also recognize the limitation in the exploratory validation study with respect to the business model issue. The impact of business model design on third party utilization is beyond the scope of this study, and thus our economic model is limited to the impact of privacy concerns on third party information sharing.

In the case of advertising third parties, the third parties are often the ad serving companies such as Google and Yahoo, and not the advertisers whose ad is being shown. So while the

publisher websites do not select the advertisers, they select the third parties that manage these ads. Our model does treat all third parties equally, but in reality not all third parties are identical in their level of advertising or their amount of abuse of user privacy. Treating all third parties identically is a simplification for the model, but we believe this simplification does not impair the current analysis. The treatment of third parties as non-homogeneous remains a topic for future research.

Additionally, this study only examines direct information sharing and selling. User data is sold and resold, potentially making the problem much worse than what is modeled in this chapter. Because we do not consider the interaction effect between third parties, our essay represents a lower bound on the problem, with the likelihood that results found with the model are understated. We leave the interaction among third parties for future research. Another avenue for future research would be to consider third parties that provide some service to users, and thus would actually increase user utility to some degree.

CHAPTER 4

Use and Abuse of User Privacy Preferences: Experiments on Effectiveness of Passive and Active Privacy Tools

4.1 Introduction

People spend a big portion of their time on the web. Some reports state that U.S. adults spend as much as 18% of their daily time online, and this does not include the time spent with mobile devices (eMarketer 2014). Websites provide a variety of information and services to their users (or visitors). However, not all of the information and services are provided by the website itself, but are instead outsourced to other "third party" entities. The third parties cover a wide variety of services including advertising, providing content, hosting media, and providing security and performance tools. The third parties gather user information from these websites. The scope of collected information is not limited to browsing data; for example, Quintin (2015) finds that Healthcare.gov, the U.S. governmental healthcare website, shared very sensitive user information with multiple third parties, and Krishnamurthy et al. (2011) find that most of the popular websites that have registered users leak sensitive user information to third parties.

To further aggravate the situation, there is evidence that user information sharing and aggregation is increasing over time, causing these concerns to escalate. Krishnamurthy and Wills (2009b) show in a longitudinal study that sharing and aggregation of user information has been increasing over time. Discrimination is another concern of the user information leakage, Mikians et al. (2012) and Valentino-DeVries et al. (2012) provide evidence for price and search discrimination on the web. Authorities and policy makers such as Federal Trade Commission

and the European Union are investigating these concerns and are looking for solutions (Mayer and Mitchell 2012). However, the speed of technological advancements and the disparity and span of the industry, makes it hard for such efforts to become effective. Moreover, while there is some evidence for the ineffectiveness of self-regulation by publisher websites (Culnan 2000), it is not clear if regulation can be more effective than self-regulation (Jamal et al. 2005).

An important question surrounding the privacy issues on the websites is whether or not self-regulation is effective in moderating the third party usage and making websites respect user privacy concerns, or if further regulation and intervention is required. More specifically, in this chapter, we empirically examine the following question: What is the impact of passive privacy tools such as "Do Not Track" (DNT) on third party usage? We will examine the impact on sharing of DNT in comparison to the active privacy tool AdBlock Plus (ABP, adblockplus.org), as well as the condition where the user is silent regarding privacy, which we call no restriction (NR).

4.2 Literature Review

Despite the importance and the extent of the issue of information privacy in websites, the impact of privacy tools on self-regulatory behavior has not received proper attention in the literature. Pavlou (2011), Smith et al. (2011), and Bélanger and Crossler (2011) provide comprehensive reviews of information privacy literature, as well as recommendations for future research. There is some research on information privacy in the e-commerce environment and user self-disclosure (Li 2012), but there is a gap in the literature on issues surrounding the use of third parties by websites and how it may be impacted by privacy tools.

Some literature has recently studied issues related to the use of third parties. Gopal et al. (2014) provide an experimental study on the extent of third party usage among popular websites. They examine the extent of third party usage and find that different categories of websites have varying levels of third party usage. Gopal et al. (2015a) provide an economic model for explaining and predicting the effect of user privacy concerns on the website's decision to share user information with third parties, and provide some policy-making and welfare analysis.

Within the literature on self-regulation of privacy in online settings, Jamal et al. (2005) provides an empirical analysis of the state of online privacy in the UK and US. While the UK imposes regulation on e-commerce privacy as opposed to the free market of the US, the authors do not find significant differences between UK and US firm behaviour. Bowie and Jamal (2006) study the privacy issues around e-commerce to analyze the state of self-regulation in the United States, and compare it to the regulatory environment of the European Union (UK). They focus on the privacy policies, disclosures, and effectiveness of opt-out mechanisms, and more specifically, on the self-regulatory effectiveness of webseals. They find some evidence for the effectiveness of webseals as a self-regulation mechanism. However, they conclude that some change is required in the webseals in order for them to maintain their purpose. This chapter differs from Jamal et al. (2005) and Bowie and Jamal (2006), in that we consider the privacy concerns as evidenced by third party sharing rather than by violation of stated privacy policies. This chapter is also a more general approach to online user privacy, as it considers a wide range of websites (rather than e-commerce only). We consider what happens when a user initially visits a website, rather than the self-disclosed information of registered users.

4.3 Experimental Design and Data Description

In this section we present our experimental methods. The data is collected within the first three seconds of visiting a website. The number of third parties, the number of cookies, and the proportion of secure connections utilized by the website is collected. The presumption is that an increase in the sharing intensity (e.g., sharing with more third parties and the use of more cookies) is a direct measure of privacy violation. We use 7 out of 17 website subject categories as identified by Alexa.com list of the top websites in each category, consistent with the pilot study (Gopal et al. 2015b) for this chapter.

There is currently a significant global debate with respect to how best to honor users' need for privacy (Electronic Frontier Foundation 2015). We study the effect of DNT requests on third party sharing intensity. DNT is a simple mechanism in the HTTP language that sends a request to the website to not track that user in that website or across other websites. DNT is supported by most major browsers, and was initiated in late 2010. It is a passive tool available to users, but the websites do not face regulatory pressure to obey these user-stated preferences. Websites can respond to these requests in three different ways: decrease third party usage, do nothing, or increase third party usage. We define active management as changing the third party usage (either decrease or increase) in response to a DNT request. In order to better understand the different responses that websites can have to the DNT requests, consider the categorization in Figure 4.1.

Figure 4.1 presents a conceptual model of the amount of third party sharing and attitude toward active management with respect to different DNT browser settings. On the left side of the figure, the reduction of third parties in response to DNT request may be due to the website respecting the user's request. Alternatively, it may be the fact that the number of third parties is

reduced, but more abusive third parties are used. The opposite can also occur as can be seen on the right side of the figure. That is, while the number of third parties may decrease, this does not essentially result in higher tracking of user information, as less abusive third parties may be utilized. While we acknowledge these interesting nuances, we do not have access to the actual information that is being shared with third parties, and the only available information is if a connection was made with a third party, how many cookies were used, and the proportion of secure connections.

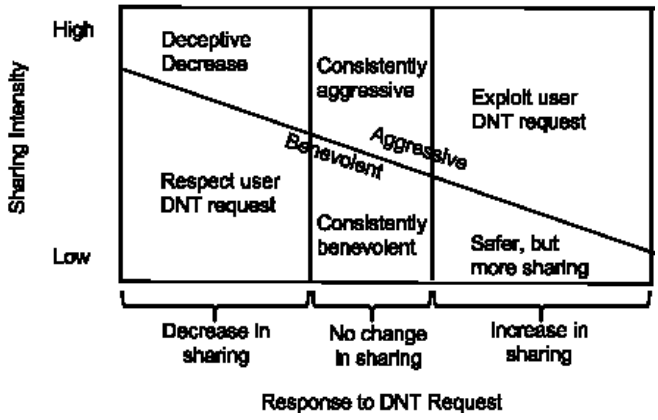


Figure 4.1. Conceptual Model of Response to DNT Request

4.3.1 Experimental Setting

There is no regulation in Canada with respect to DNT, and thus Canada provides an appropriate environment in which to conduct this study. We carried out the experiments on the 100 most-visited publisher websites from each of the 7 subject categories of websites from Alexa website rankings: Adult, Arts, Business, Health, Kids and Teens, News, and Shopping. These 7 categories were selected because they are expected to provide differences based on differing users' privacy concerns. The Alexa top 100 list was created in May 2014. The data collection was conducted on January 30 and 31, 2016 as follows. An automated script orders the browser

to go to the pages of all top 100 publisher websites in a single subject category, sequenced from most visited to least. The pages are loaded one by one, and each page is visited for the page loading time plus a 3 second wait. The connections that are made from each publisher website to the third parties are collected using Lightbeam for Firefox (Windows). The cookies and history are cleared at the start of the experiment for each subject category. ABP, an active privacy tool, is used as a baseline measure to obtain the set of relatively safe and privacy-friendly third party connections used by the website. The experiments are carried out in three scenarios:

1) without use of any privacy tools, which we call "no restriction" and label as NR;

2) having DNT enabled, labeled as DNT; and

3) having ABP enabled, labeled as ABP.

4.3.2 Measuring User Information Sharing

In this chapter, we are interested in analyzing the amount of user information being shared with third parties. However, information sharing of user data is not directly observable. We use several observable variables related to user information sharing to define factors (or components) of user information sharing. In our data, we have access to four main variables linked to the user information sharing for each website: number of third parties, number of cookies, number of connections, and number of secure connections. Number of third parties is the number of individual third parties that have been utilized by the website. Number of connections is the number of links that the website has with all the third parties, and there may be more than one connection to each third party. Number of cookies is the number of cookies that are placed on the browser computer when a website is visited. Number of secure connections is a measure of how many of the connections are secured using the https protocol. Instead of individually using

the number of connections and number of secure connections, we create a composite variable by combining these two variables as the proportion of secure connections to the total number of connections.

We use factor analysis to generate factors that describe the user information sharing. First, we normalize the data for each variable. Using the three variables, number of third parties, number of cookies, and proportion of secure connections, we are able to find two components that best describe the data. Table 4.1 provides the rotated components.

The first component (C1) is sharing intensity and focuses on the number of third parties and number of cookies, while the second component (C2) is sharing security and gives most of the weight to the proportion of secure connections. There is a small negative correlation between the two components (-0.127). These two components account for approximately 90% of the variance in the data (Table 4.2).

Table 4.1. Factor Analysis Components

Components / Variables	C1: Sharing Intensity	C2: Sharing Security
Number of Third Parties	0.917	-0.053
Number of Cookies	0.916	-0.064
Proportion of Secure Connections	-0.064	0.998
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in 3 iterations.		

Table 4.2. Variance Explained by the Components

Total Variance Explained									
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
C1	1.726	57.533	57.533	1.726	57.533	57.533	1.685	56.183	56.183
C2	.962	32.077	89.611	.962	32.077	89.611	1.003	33.428	89.611
Remainder	.312	10.389	100.000						

Extraction Method: Principal Component Analysis.

In the component loadings, we disregard the loadings with magnitude of less than 0.1 for simplicity. As a result, component 1 is a linear combination of normalized number of third parties and normalized number of cookies with loadings of 0.917 and 0.916, respectively. Component 1 describes the sharing intensity of the publisher website. For component two, we round the loading to 1 for the proportion of secure connections, thus the second component is simply the proportion of secure connections. Component 2 describes the sharing security. Throughout this chapter, we use these two components, along with the number of third parties as primary variables of interest for analysis.

4.4. Data Analysis and Results

In this section we report the results from the experiments. We use the data from the raw variables as well as the components created in Section 4.3.2.

4.4.1 Differences in Sharing Behavior Between Subject Categories

The primary focus of our work is on analyzing sharing intensity (C1). Figure 4.2 provides the average sharing intensity among different subject categories and their 95% intervals with no restriction (NR). The significance of the differences under condition NR is provided in Table 4.3. The reported p-values are for two sided, two-sample t-tests of whether the sharing intensity between category pairs are significantly different. Figure 4.2 and Table 4.3 illustrate that sharing intensity is significantly higher in websites where we would expect there to be lower privacy concerns (e.g., News websites would be expected to have lower user privacy concerns than Adult websites). This provides an indication that the privacy concerns are effective in moderating the sharing intensity, at least to a certain degree.

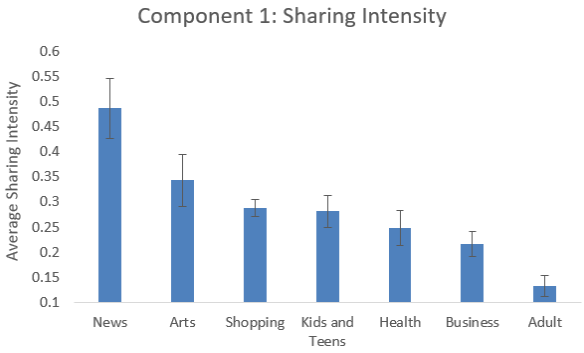


Figure 4.2. Average Sharing Intensity (C1) by Subject Category

Table 4.3. Statistical Comparison of Sharing Intensity (C1) between Subject Categories

	Arts	Shopping	Kids and Teens	Health	Business	Adult
News	0.001**	0.000**	0.000**	0.000**	0.000**	0.000**
Arts		0.077*	0.109	0.006**	0.000**	0.000**
Shopping			0.848	0.129	0.001**	0.000**
Kids and Teens				0.331	0.038**	0.000**
Health					0.203	0.000**
Business						0.000**
* Significant at p=0.1, ** Significant at p=0.05						

4.4.2 Sharing Intensity Differences Between Browsing Options

Here we analyze the differences in sharing intensity with different browsing options (NR, DNT, and ABP). We presume that sharing using ABP is benevolent. Significant increases in sharing intensity, comparing ABP to NR and DNT conditions, would indicate that users' privacy is being violated. Significant decreases in average sharing intensity would indicate that users' privacy is, on average, being respected more for a particular category. As shown in Table 4.4, the average sharing increases in 5 of 7 categories when DNT is invoked, compared to NR. Only News shows a significant increase in average sharing intensity at the 0.10 level. We also compare the publisher websites' sharing intensity before and after use of DNT and ABP through a paired t-test. These results are provided in the third portion of Table 4.4. In the paired comparison of NR

and DNT, the Kids and Teens category also shows significant difference, however, there is a decrease in sharing intensity with DNT in this case.

Average sharing intensity is significantly higher for NR compared to ABP in 5 out of 7 categories considering the average sharing intensity, and in all 7 categories considering the paired comparison. Comparing DNT to ABP, the results are the same except that Adult category also significantly decreases the average sharing intensity with ABP. Thus, we can conclude that in all categories, there is significantly higher sharing intensity for both NR and DNT compared to ABP. Additionally, we observe that there are no significant changes in sharing intensity for any category when comparing NR to DNT other than the News category, which experiences a significant increase in sharing intensity with DNT.

Table 4.4. Statistical Comparison of Sharing Intensity (C1) within Subject Categories between Website Browsing Options

		News	Arts	Shopping	Kids and Teens	Health	Business	Adult
Average Sharing Intensity	NR	0.486	0.343	0.287	0.281	0.248	0.216	0.132
	DNT	0.570	0.345	0.289	0.265	0.240	0.218	0.137
	ABP	0.246	0.203	0.247	0.160	0.172	0.186	0.107
P-Value for Average Difference	NR vs DNT	0.0709*	0.963	0.941	0.675	0.780	0.945	0.786
	NR vs ABP	0.000**	0.000**	0.055*	0.000**	0.001**	0.121	0.108
	DNT vs ABP	0.000**	0.000**	0.047**	0.000**	0.002**	0.121	0.071*
P-Value for Paired Difference	NR vs DNT	0.001**	0.862	0.698	0.017**	0.145	0.862	0.349
	NR vs ABP	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**
	DNT vs ABP	0.000**	0.000**	0.000**	0.000**	0.000**	0.002**	0.001**
* Significant at p=0.1, ** Significant at p=0.05								

Next we examine pairwise differences between categories and between browsing options. The change in average sharing intensity for the two browsing options are compared between different categories. For example, the 0.006 value for the NR-ABP comparison between Health and Business subject categories in Table 4.5 indicates that the reduction of

sharing intensity using ABP as compared to NR in the Health subject category is significantly different from reduction of sharing intensity using ABP as compared to NR in the Business subject category. This analysis shows that the way in which websites react to ABP and DNT varies across categories. In Table 4.5, we observe that 16 of 21 pairwise comparisons of NR versus ABP are statistically different at the 0.05 level, and a 17th comparison is significantly different at the 0.10 level. Comparing DNT to ABP, we observe that 15 of 21 pairwise comparisons of DNT versus ABP are statistically different at the 0.05 level, and a 16th comparison is significantly different at the 0.10 level. Comparing NR to DNT, we observe that 8 of 21 pairwise comparisons of NR versus DNT are statistically different at the 0.05 level, and a 9th comparison is significantly different at the 0.10 level. This analysis provides strong evidence that subject categories, in pairwise comparisons with other subject categories, show different reactions to ABP and DNT conditions.

Table 4.5. Statistical Comparison of Sharing Intensity (C1) Changes between Subject Categories between Website Browsing Options

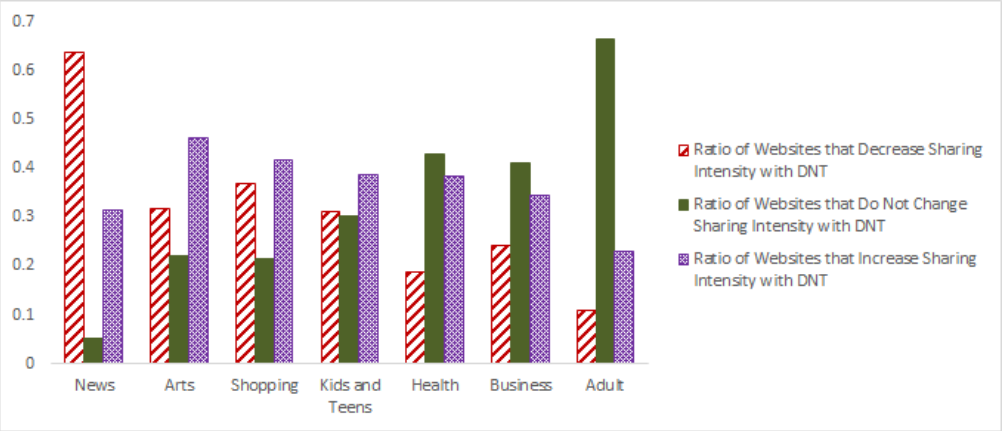
		Arts	Shopping	Kids and Teens	Health	Business	Adult
NR vs DNT	News	0.003**	0.002*	0.000**	0.000**	0.002*	0.003**
	Arts		0.990	0.158	0.418	0.985	0.785
	Shopping			0.027*	0.166	0.991	0.622
	Kids and Teens				0.342	0.117	0.015**
	Health					0.367	0.090*
	Business						0.736
NR vs ABP	News	0.005**	0.000**	0.001**	0.000**	0.000**	0.000**
	Arts		0.000**	0.534	0.013**	0.000**	0.000**
	Shopping			0.000**	0.022**	0.326	0.128
	Kids and Teens				0.084*	0.000**	0.000**
	Health					0.006**	0.002*
	Business						0.620
DNT vs ABP	News	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
	Arts		0.000**	0.212	0.003**	0.000**	0.000**
	Shopping			0.003**	0.061*	0.401	0.277
	Kids and Teens				0.129	0.001**	0.000**
	Health					0.025**	0.014**
	Business						0.889

* Significant at p=0.1, ** Significant at p=0.05

4.4.3 Effect of DNT on Sharing Behavior

Figure 4.3.a presents the proportion of websites each category that increase, decrease, or do not change their third party usage when DNT is requested. An interesting observation is that when users turn on the DNT feature, it does not always cause the websites to decrease their sharing. In fact, websites with lower privacy concerned users such as News are much more likely to actually increase the number of third parties when DNT is turned on. In fact, we find that in the News category, 45% of the websites increase their number of third parties when DNT is turned on, but 63% decreased the number cookies, resulting in 31% of News websites increasing sharing intensity, 64% decreasing it, and 5% with no change. Compare this to the Adult category where only 6% of the websites increase their number of third parties when DNT is turned on and 84% are unchanged, but 11% decreased the number cookies, 68% are unchanged and 21% increased the number of cookies, resulting in 11% of Adult websites increasing sharing intensity, 23% decreasing it, and 66% with no change.

One plausible reason for increases in sharing intensity is due to the signaling phenomenon: website advertisers may be more interested in the users who are concerned about their privacy and use DNT. Websites with more information sensitivity and higher privacy concerns generally tend to not change their third party usage when DNT is requested. Figure 4.3.b provides the changes in sharing intensity in websites that either increase or decrease their sharing intensity with DNT. All of the reported changes are significant at the 0.001 significance level.

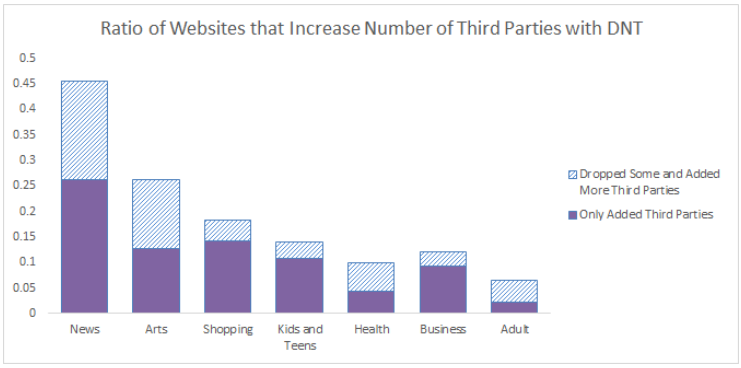


a. Effect on Ratio of Websites

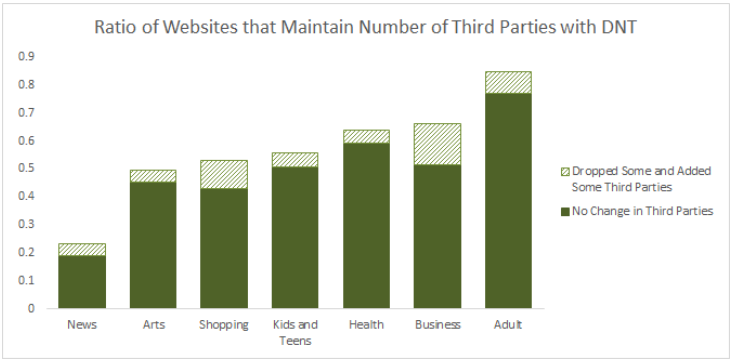
b. Effect on Sharing Intensity

Figure 4.3. Effect of DNT on Sharing Intensity by Subject Category

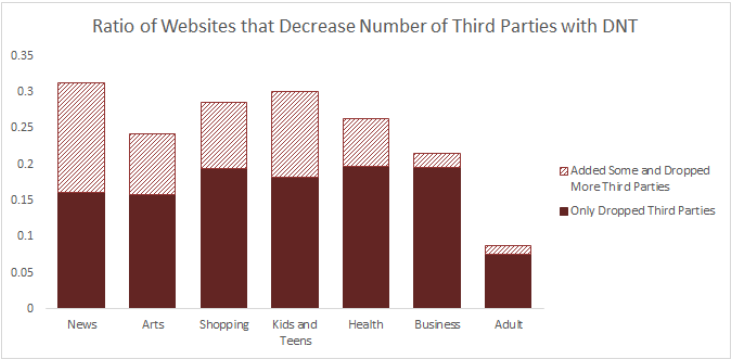
Next we examine the ratio of the websites that change the number of third parties in response to a DNT request. We first examine the websites that experienced a net increase in third party sharing. Figure 4.4.a breaks the ratio into two components: 1) the firms that both added and dropped third parties in response to the DNT request, and 2) the firms that only added third parties, but did not drop any third parties, in response to a DNT request.



a. Websites that Increase Number of Third Parties They Share with When DNT is Set



b. Websites that Maintain Number of Third Parties They Share with When DNT is Set



c. Websites that Decrease Number of Third Parties They Share with When DNT is Set

Figure 4.4. Different Behaviors among Websites with Respect to DNT

We contend that websites that only add third parties when given a DNT request are clearly abusive toward users, and would clearly be characterized in Figure 4.1 as websites that "Exploit

user DNT requests". As shown in Figure 4.4, 26.3% of all News websites in the study only added third parties when given a DNT request. The ratios for the other categories are 4.4% for Health, 9.3% for Business, 14.3% for Shopping, 12.6% for Arts, 10.8% for Kids and Teens, and 2.2% for Adult. Thus, every category showed at least some amount of user abuse for DNT requests. As shown in Figure 4.4.c, there are websites that decreased their sharing with third parties when presented with a DNT request. In Figure 4.4.c, this seemingly benevolent behavior is broken into 2 parts: websites that both add and delete third parties in response to a DNT request, versus websites that only delete third parties from their original sharing list. The latter, where the website strictly reduces the number of third parties, might be characterized as websites that "respect user DNT requests" in Figure 4.1, especially if they also decreased the number of cookies. In Table 4.6, we explore the differences seen in Figure 4.4.a. Note in Table 4.6 that 11.6% (78 out of 675 websites) were found to strictly add third parties in response to DNT. That number was 26 out of 99 for News (26.3%). Table 4.6 illustrates that a substantial segment of websites actually abuse user privacy by strictly increasing the number of third parties that information is shared with when the user presents with a DNT request. Additionally, the behavior of strictly increasing the number of third parties is clearly impacted by the subject category.

With respect to abuse of user privacy, we are uncertain regarding websites that added some third parties, yet removed more than they added. Websites that both add and remove third parties may or may not be benevolent towards the users' DNT request for privacy. In Table 4.7, we further examine the magnitude of the addition and subtraction of third parties by websites that have a net increase in the number of third parties. It can be seen in News, for example, that the average number of third parties dropped was 1.13, compared to the 10.56 average number of

third parties that were added. The 26 firms that only added third parties were clearly more abusive with respect to privacy, but a strong argument can be made that all 45 News websites that increased the number of third parties on average by a net of 9.42 were abusive toward users. The maximum increase in the number of third parties was 62 new third parties when the user presents with a passive DNT request, and the smallest maximum among all 7 categories is 15 new third parties.

Table 4.6. The Number of Websites that Added or Dropped Third Parties in Response to DNT (Compared to NR) among Websites that Increased Number of Third Parties

	Added Third Parties	Also Dropped Third Parties	Strictly Added Third Parties
News	45	19	26
Arts	25	13	12
Shopping	18	4	14
Kids and Teens	13	3	10
Health	9	5	4
Business	13	3	10
Adult	6	4	2

Table 4.7. Increased Sharing for The Websites that Increase Number of Third Parties with DNT

	Websites that Increase Number of Third Parties with DNT														
	Dropped Third Parties					Added Third Parties					Net Change				
	Count	Mean	Min	Max	STD	Count	Mean	Min	Max	STD	Count	Mean	Min	Max	STD
News	19	1.13	0	13	2.64	45	10.56	1	60	13.66	45	9.42	1	58	13.06
Arts	13	2.04	0	8	2.68	25	8.96	1	52	13.19	25	6.92	1	46	11.36
Shopping	4	0.83	0	5	1.65	18	5.00	1	20	5.94	18	4.17	1	20	5.38
Kids and Teens	3	0.92	0	8	2.29	13	3.38	1	15	3.86	13	2.46	1	7	1.76
Health	5	1.00	0	4	1.32	9	5.56	1	18	5.94	9	4.56	1	18	5.64
Business	3	0.46	0	4	1.13	13	7.31	1	62	16.67	13	6.85	1	62	16.76
Adult	4	3.50	0	11	4.76	6	6.17	1	15	6.49	6	2.67	1	7	2.25

4.4.4 Sharing Security

A priori, for sharing security, component C2, we expect to find significant differences between subject categories based on the need for secure connections with third parties. Differences

between subject categories stem from the different website business models. For example, a bank, which falls under the Business category, will have very a secure website due to the business model need to provide financial security. In contrast, News websites advertise heavily but do not process credit card data. Thus, we would a priori expect a News website's business model to need less security than a Business or Shopping website. We pose the question "Do subject categories that should have more security in fact have more security?" The answer is that we find security needs seem to vary in response to apparent business model considerations.

Figure 4.5 presents the average sharing security values by website subject category along with the 95% confidence intervals. Table 4.8 presents a statistical comparison of sharing security between subject categories. As seen in Table 4.8, there are 8 pairwise comparisons where the difference between the subject categories are statistically significant at the p=0.1 level, but the remaining 13 pairwise comparisons are not significantly different from each other. Analyzing Figure 4.5 and Table 4.8 in conjunction with each other, we can separate the subject categories into three levels of sharing security. The highest security level includes Business and Health. The middle security level includes two categories: Shopping, and Kids and Teens. The lowest security level includes News, Arts, and Adult.

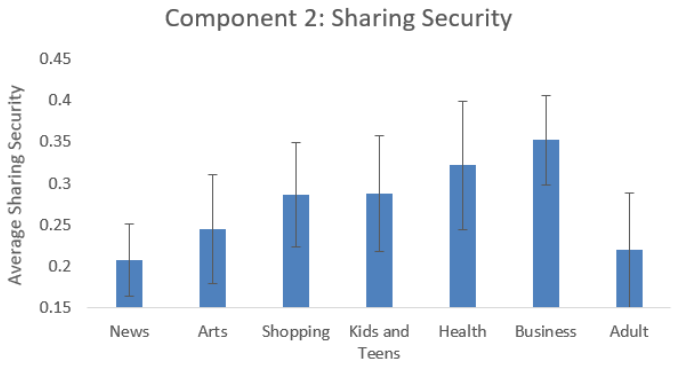


Figure 4.5. Average Sharing Security (C2) by Subject Category

Table 4.8. Statistical Comparison of Sharing Security (C2) between Subject Categories

	Arts	Shopping	Kids and Teens	Health	Business	Adult
News	0.299	0.056*	0.050*	0.007**	0.001**	0.627
Arts		0.351	0.337	0.089*	0.021**	0.679
Shopping			0.992	0.482	0.196	0.214
Kids and Teens				0.481	0.193	0.203
Health					0.544	0.051*
Business						0.012*
* Significant at p=0.1, ** Significant at p=0.05						

In Table 4.9, for the average sharing security, we observe that there are no statistically significant differences in sharing security within any subject category due to the choice of website browsing option (NR, DNT, or ABP). However, when comparing the differences in a paired t-test, we find that ABP has significantly lower sharing security than both NR and DNT in 5 out of 7 categories. It seems that the list of third parties that are blocked by ABP, on average, have a higher sharing security (use a higher proportion of secured connections) than the rest of the websites. Therefore, when these third parties are blocked by ABP, the sharing security is decreased. This effect cannot be observed through comparing the averages, due to the high variations in the data. This analysis lends support to the idea that sharing security is driven by the business model transactional needs of the website category, and is not driven by differing user privacy concerns with respect to the subject area content. It is the need for security surrounding the transactions performed on the website, such as executing credit card purchases versus reading a news article, which we believe to be driving differences observed in website security behavior between categories.

Table 4.9. Statistical Comparison of Sharing Security (C2) within Subject Categories between Website Browsing Options

		News	Arts	Shopping	Kids and Teens	Health	Business	Adult
Average Sharing Security	NR	0.208	0.245	0.287	0.287	0.321	0.352	0.220
	DNT	0.216	0.250	0.288	0.284	0.311	0.362	0.217
	ABP	0.185	0.231	0.277	0.266	0.299	0.358	0.227
P-Value for Average Difference	NR vs DNT	0.800	0.892	0.974	0.939	0.836	0.850	0.946
	NR vs ABP	0.504	0.733	0.848	0.659	0.645	0.913	0.873
	DNT vs ABP	0.363	0.637	0.822	0.715	0.797	0.937	0.821
P-Value for Paired Difference	NR vs DNT	0.052	0.152	0.629	0.359	0.259	0.201	0.057
	NR vs ABP	0.002**	0.015*	0.001**	0.004**	0.024**	0.506	0.187
	DNT vs ABP	0.000**	0.008**	0.005**	0.023**	0.021**	0.348	0.083
* Significant at p=0.1, ** Significant at p=0.05								

Combining the analysis of sharing intensity with sharing security, we can make broader statements. News and Arts both have the lowest sharing security levels and are very abusive with respect to privacy. Adult is extremely poor with respect to sharing security, but this category is very respectful with respect to privacy. Shopping and Kids and Teens are both in the middle area with respect to both privacy and security. Business and Health score well in terms of both security and privacy. Therefore, overall, we can call News and Arts the worst offenders of privacy and security, and Business and Health the best defenders of privacy and security. It is important to reiterate that no subject category comes close to the level of privacy exhibited by the Adult subject category.

4.5. Discussion and Conclusion

In this chapter, we study the impact of user privacy concerns as a self-regulatory mechanism to control the sharing intensity (number of third parties utilized and the number of cookies) by websites. Experimental analysis of sharing intensity among websites in different subject categories is performed, where user privacy concerns are believed to vary between subject

categories. We find some evidence that user privacy concerns do have a self-regulating effect on the sharing intensity. Websites in subject categories such as Health, Business, and Adult, which are expected to have users with higher privacy concerns, tend to have lower levels of sharing intensity. In contrast, websites in categories with a priori expectations of low privacy concerns, such as News and Arts, have higher sharing intensity. Our results indicate that the sharing intensity is rather low for subject categories with high privacy concerns. High sharing intensity is more common in subject categories with low privacy concerns.

We also find the effectiveness of DNT as a privacy protection tool in controlling the sharing intensity by website publishers is mixed. A significant proportion of websites in the subject categories with low privacy concerns abuse users' privacy by increasing sharing intensity when DNT is requested. For website subject categories with low user privacy concerns, the self-regulatory mechanism is not strong enough to prevent significant and abusive sharing intensity. The original intent for DNT is to provide a means for privacy-concerned users to ask websites to reduce their third party sharing. We find that sometimes third party sharing is reduced, but we also find that sometimes even more abusive sharing occurs when DNT is requested. This additional sharing suggests that the information from DNT users provides some benefit to publisher websites and third parties. The results from this study illustrate that privacy concerns are an input to a self-regulating mechanism for sharing intensity, at least to a certain degree. This indicates some optimism in that the market has some self-regulating abilities. However, the results from the effects of DNT suggest that there are challenges. The widespread existence of websites that increase their sharing intensity when DNT is requested by the user among the popular websites is particularly alarming. We believe that there is a lack of transparency between websites and users with respect to the number of third parties and cookies

used, and the lack of transparency may alter the self-regulatory mechanism of user privacy concerns. We speculate that the lack of transparency of website sharing behavior may lead to abusive sharing behavior. Our empirical analysis in a relatively regulation-free environment indicates that without regulation or transparency with respect to third party sharing, the DNT signal is often being used to abuse customer privacy - the exact opposite reason for its existence. Our analysis indicates that relatively few websites are ignoring the DNT request when NR sharing intensity is substantially higher than ABP sharing intensity. Additionally, the proportion of websites that do not react to the DNT request is dependent on the subject category, which is likely to vary with respect to users' privacy concerns.

Further study is required to explore the consequences of sharing intensity when user privacy concern is low. To the extent that regulation is warranted, regulatory policy should incorporate those aspects of self-regulation that do work to enhance overall effectiveness and protect consumers. Developing an optimal portfolio of self- and government-regulation is worthy of further investigation.

CHAPTER 5

Information Security & Cloud Suppliers: How Customer Demand Reaction Shapes Supplier Choice Strategies

5.1 Introduction

With the proliferation of cloud computing and software-as-a-service platforms, firms are changing their information and communication technology (ICT) infrastructures significantly. Cloud services provide an inexpensive and scalable strategic advantage, particularly for small and medium sized firms. However, the role of inherent risks that such ICT suppliers place on the firm's performance and how those risks impact firms' decision-making has mostly been neglected.

The recent Experian breach affecting 15 million T-Mobile customers highlights the dependence that firms have on their partners for protecting information assets (Malik 2015). While a firm may have developed excellent in-house security practices, it must still be aware of the practices of all its supply chain partners. When a wide range of firms make use of the same supplier, a single breach has the potential to affect all of these firms and their customers. Cloud computing is an emerging source of this form of shared risk of security breaches and other adverse events that occur at the supplier or business partner level. Thus, supplier selection is an important decision for many firms, and especially so in the cloud services context.

When choosing a software-as-a-service (Saas) cloud provider, there are two major models of operation; single- versus multi-tenancy. In a single-tenancy model, the applications and data

storage services are provided in individual, distinct instances for each client of the service provider. The single-tenancy design provides a layer of insulation between the client firms in the event of attacks and adverse events. In contrast, all clients share a single instance of application and data storage (e.g., shared database) in the multi-tenancy model. The appeal of the multi-tenancy model is cost savings through shared hardware, software, maintenance, and security (Ramakrishnan 2014; Bezemer and Zaidman 2010), although attacks and adverse events are more likely to affect all multi-tenant clients (Bezemer and Zaidman 2010). As an example, cloud service providers such as VMware provide both single- and multi-tenant cloud solutions. In 2014, a vulnerability affected the multi-tenant environment of VMware, where important proprietary organizational information was exposed for other tenants to view (Lennon 2014). On the other hand, an argument can be made that multi-tenancy could create opportunities to obtain better security resources compared to what would be available to the firm in a single-tenant environment. Thus, firms face the question of which type of supplier; either single- or multi-tenant, they should choose. Note that choosing either single- or multi-tenant options is equivalent to choosing a supplier that is either independent from the competitor's supplier, or sharing a single supplier. In this chapter, we study this problem based on competition among firms and different levels of risk that the suppliers may impose onto the firms.

While a shared multi-tenant cloud provider is one way in which adverse events may become synchronized across firms, there are others. Any use of common software, whether hosted locally or through an outsourced provider, provides a common means of compromise. For example, compromised automatic update servers of the Korean software company, ESTsoft, spread malware to its customers, resulting in a data breach affecting 35 million South Koreans (Hee-jin 2011; Hyung-eun 2011; The Register 2011). Further, a vulnerability in software such as

Adobe Acrobat Reader potentially impacts all users of the software (Adobe 2016; Brook 2015). Still, partnerships can provide much needed flexibility. High-tech or other highly dynamic industries may want to obtain capabilities through partnerships rather than in-house development (Barney 1999), even though sharing a supplier may expose firms and their competitor to some risk, negatively affecting an entire industry (Cleeren et al. 2008; Dyer and Singh 1998; Parmigiani et al. 2011; Wu et al. 2015). Another example of synchronized risk coming from a shared partner came in 2011 when the Epsilon breach exposed customer names and email addresses for a wide range of client organizations including Chase, Kroger, and Best Buy (Bradley 2011; Horowitz 2011; Schwartz 2011). Risk synchronization of data security events have parallels in the physical product environment. For example, when a shared supplier provided tainted peanut butter to two competing brands, research showed that not only were sales for both brands affected, but also sales across the entire product category declined (Cleeren et al. 2008). Thus, synchronization of adverse events between competing firms (and even competing brands within the same firm) is a very common and important phenomenon that affects a wide variety of strategic supply chain decisions.

The risk of failure (or success of attacks) in single- versus multi-tenant cloud computing environments is different. It will take only a single successful attack against the application to compromise all clients in a multi-tenancy model, whereas each instance would need to be independently compromised in a single-tenancy model. Thus, while moving corporate data outside the firm's boundaries requires the firm to trust the abilities of its supplier(s), the potential synchronization of attacks with the firm's competitors (e.g., through a shared, multi-tenant provider) may further complicate the choice of providers (we use the terms provider or supplier to describe the business service partner). In this work, we set out to examine when there are

advantages to choosing a shared, multi-tenant provider with a firm's competitor, and when there are not. In particular, we use game-theoretic modeling to illustrate conditions of risk and customer demand reactions to incidents that affect the strategic choice to synchronize these risks, or not, with competitors. In many cases, firms can strategically choose their suppliers, but these suppliers are likely to have different security programs and there will be different levels of risk associated with partnering with one supplier versus another. This chapter examines supplier choice under different customer demand reactions to these adverse events. We examine the equilibrium expenditures with a service provider to reduce adverse event realization and maximize firm profits. We compare the case when two firms in a duopoly setting synchronize their risk with a competitor (through a shared service provider) to the case where the firms reduce risk synchronization with a competitor (through an independent service provider).

The chapter is organized as follows. In Section 5.2, we present prior research in this area. In Section 5.3, the game theoretic duopoly model is presented, along with the modeling framework for both cases of firms investing in independent suppliers and firms investing in a shared supplier. Section 5.4 provides the comparison between the shared and independent supplier strategies along with analysis of the effects of model parameters on supplier choice. Section 5.5 illustrates how the results change when the adverse events are correlated. Section 5.6 examines the impact of coordination among firms on supplier choices. In Section 5.7, we look into asymmetric suppliers and asymmetric firms, and analyze how asymmetry impacts the firm's decision making. Section 5.8 concludes this chapter.

5.2 Literature Review

Supply chain risk literature tends to focus on risks that would disrupt the delivery or quality of manufactured goods (Tang 2006). In these settings, risk pooling has been shown to be beneficial

for inventory management under demand uncertainty while risk diversification is preferred under threat of supply chain disruption (Mak and Shen 2012). In this chapter, we are also interested in the strategic decision of supply chain design to minimize risk, although our perspective is that of risk due to unintended exposure of valuable corporate data housed at or managed by a cloud service provider.

Shared supply networks have emerged in some industries (Sturgeon and Lee 2004) where generic products or services provided to the industry allow a supplier to transfer knowledge, gained through their relationship with one lead firm, to attract additional firms in the same industry. The industry benefits from such capacity pooling; however, there remains a risk that core intellectual property may leak between client firms (Sturgeon and Lee 2004) or that a tainted ingredient makes its way into competitor products (Cleeren et al. 2008). Information sharing in the supply chain network is also a popular topic. However, the focus in this literature stream is generally on information sharing between partners to improve supply chain efficiency (e.g. Cachon and Fisher 2000; Ha and Tong 2008; Kelle and Akbulut 2005; Ojala and Hallikas 2006; Tsung 2000; Zhou and Benton 2007). Anand and Goyal (2009) discuss how competitors acquire information through either sharing or leakage.

When information assets reside outside of the direct control of the firm, any firm-developed innovation could be used to the benefit of its competitors. In the case of a managed security service provider (MSSP), the supplier learns from providing service to one firm and this naturally accrues to other clients of the MSSP (Cezar et al. 2010). Benefits have been found for sharing information about IT security across an industry as it enables better investment decisions (Gal-Or and Ghose 2005; Gordon and Loeb 2003; Hausken 2006), but it may also reduce the competitive advantage for any one firm.

Where cloud services are considered, most of the research is in the context of computing science with a focus on how to build robust infrastructure that facilitates multiple users in both single and multi-tenant environments (Bezemer and Zaidman 2010; Demirkan and Cheng 2008). Such work takes the perspective of the provider in addressing the technical design side of the problem. The impact of adverse security events on the choice of Application Service Provider (ASP) or cloud service provider has received scant attention in the literature. Garg et al. (2013) proposes a method to qualitatively rank cloud providers based on key performance indicators in an attempt to help firms select an appropriate partner. However, much of the work in this business area is focused on properly defining and enforcing service level agreements (i.e. contract negotiation and monitoring). The impact of information security needs on pricing decisions of cloud service providers is examined by August et al. (2014). Our approach considers the impact of customer demand reactions to information security leakages (breaches) on the supply chain design decisions of organizations - a decision to be made even before considering the pricing or other characteristics of various supply chain partners.

Kolfal et al. (2013) analyze the impact of customer demand changes in response to adverse IT security events, and the impact customer demand reaction to adverse IT events has on security investments. In their chapter, the security investment could be coordinated between firms, but could only be invested in the firm's own security. We build upon the work in Kolfal et al. (2013) to analyze the impact of customer demand changes in response to adverse IT security events on the supply chain design choice and the resulting effects on customer demand this choice has with respect to realized information security risks.

5.3 The Model

We consider a game-theoretic approach with two symmetric profit maximizing firms (duopoly) which have two strategic choices for their suppliers. Firms determine the supply chain configuration by choosing either independent suppliers or sharing a single supplier. Given a supply chain configuration, a firm can adjust the security or safety of their supplier by increasing their spending at the supplier. The probability of supplier ending up in either good or bad state is affected by its level of vulnerability and the amount that firms will spend on supplier security. When an adverse event occurs at a supplier, there is a probability that it will affect the firm(s) that utilize that supplier. Firms that are directly affected by an incident at the supplier suffer loss of demand. Moreover, their demand is indirectly affected, either positively or negatively, by incidents that occur at the competitor firm. These demand reactions are further explained later in this section. Figures 5.1 and 5.2 (in Sections 5.3.1 and 5.3.2 below) provide the graphical representation of the firms and suppliers in independent and shared cases, respectively. The model variables and parameters are provided in Table 5.1, and the details of the model and representative calculations are provided in Appendix 5.1.

Spending at the Supplier and Adverse Events: We assume each firm invests in the security of the one supplier it is using. We denote firm i's spending on its supplier (per unit of demand) by c_i, where $i = 1,2$. The supplier security or reliability is characterized by a vulnerability parameter v, defined by Gordon and Loeb (2003) as "the probability that a threat once realized (i.e., an attack) would be successful." Note that while the definition for vulnerability is given in the context of security breaches, the concept may be extended to apply to any adverse event or incident that may affect a supplier.

Table 5.1. Model Parameters and Variables

Notation	Definition
c_i	Firm i's spending on its supplier per unit of demand, $c_i \geq 0$
S	Set of joint supplier states
F	Set of joint firm states
f	A joint firm state, $f \in F$
ρ_i	Cascade probability for an incident at a supplier to carry through to firm i which utilizes that supplier
P_f	Probability of firms being in joint state f
$D_{i,f}$	Firm i's demand when firms are in joint state f
$Z_{i,D}$	Firm i's direct-risk elasticity of demand, $0 \leq Z_{i,D} \leq 1$
$Z_{i,C}$	Firm i's cross-risk elasticity of demand, $-1 \leq Z_{i,C} \leq 1$
$E(\Pi_i)$	Firm i's expected profit
π	Firm profit per unit of demand (excluding the spending on supplier), $\pi \geq c_i$
v	Vulnerability of the suppliers, $0 \leq v \leq 1$

Supplier State Probabilities: Each supplier can be in either good (G) or bad (B) state. Therefore, the set of joint supplier states are provided as $S = \{G, B\}$ in the case with one shared supplier, and $S = \{GG, GB, BG, BB\}$ in the case with two independent suppliers, where, for example, GB represent the case with supplier 1 (2) being in good (bad) state.

Firm State Probabilities: We assume that an effective incident at a supplier will carry over to each firm using the service from that supplier, with a fixed probability. More specifically, we assume that if an incident is realized at the supplier level, there is a cascade probability ρ_i that the firm i which is using this supplier is also affected. For now, we assume symmetry, where

$\rho_1 = \rho_2 = \rho$. We are only interested in cases where $\rho > 0$. Each firm can be in either good (g) or bad (b) state, and the set of joint firm states is provided as $F = \{gg, gb, bg, bb\}$. We denote the probability of being in each of these states as P_{gg}, P_{gb}, P_{bg}, and P_{bb}.

Demand: The adverse events that carry over to firms will affect the demand for both firms. An adverse event at a given firm has a negative effect on its demand, and can have negative or positive effect on the other firm's demand. Without loss of generality, we normalize the firm's demand for the good state to be 1 unit. The normalized demand for firm i when the firms are in joint state f is denoted by $D_{i,f}$ and is calculated as:

$$D_{i,gg} = 1, \text{ for } i = 1,2 \qquad D_{i,bb} = 1 - Z_{i,D} - Z_{i,C}, \text{ for } i = 1,2$$

$$D_{i,gb} = 1 - Z_{i,C}, \text{ for } i = 1,2 \qquad D_{i,bg} = 1 - Z_{i,D}, \text{ for } i = 1,2 \qquad (5.1)$$

where $Z_{i,D}$ is is the change in demand due to an adverse event in one's own firm, or the direct-risk elasticity of demand, and $Z_{i,C}$ is the change in demand due to an adverse event in the other firm, or the cross-risk elasticity of demand. It is important to note the difference between the direct-risk elasticity of demand versus the price elasticity of demand. While the price elasticity of demand is the change in demand due to change in price, the direct-risk elasticity of demand is the change in demand due to change in the risks or incidents associated with the product or service. A similar analogy exists between the cross-risk elasticity of demand and the cross-price elasticity of demand.

In our model, demand cannot be negative and we assume that an adverse event affecting a firm cannot increase its demand, thus $Z_{i,D} \in [0,1]$. We also assume that the cross-risk elasticity of demand effect on one firm, when an adverse event affects the other, cannot exceed the direct-

risk elasticity of demand effect for that firm in magnitude, that is $Z_{i,C} \in [-Z_{i,D}, Z_{i,D}]$. Moreover, we only consider the cases with $Z_{i,D} + Z_{i,C} \leq 1$ and $Z_{i,D} + Z_{i,C} \geq 0$ which ensure that $0 \leq D_{i,bb} \leq 1$ holds. In this section, we assume symmetry, thus $Z_{1,D} = Z_{2,D} = Z_D$ and $Z_{1,C} = Z_{2,C} = Z_C$ and we drop the firm denoting subscript, i. Following Kolfal et al. (2013), we use the terms substitutes in loss, unaffected by loss, and complements in loss for the cases of $Z_C < 0$, $Z_C = 0$, and $Z_C > 0$, respectively.

Expected Profit: It is assumed that the per unit profit excluding investment in suppliers, π, is known and fixed. Firm's marginal profit is $\pi - c_i$, and we require $\pi - c_i > 0$ as a participation constraint. The profit for firm i when firms are in joint state f (with probability P_f) is $D_{i,f}(\pi - c_i)$. In this chapter, we consider firms that maximize their expected profit. The expected profit for firm i is given as:

$$E[\Pi_i] = \sum_{f \in F} P_f \, D_{i,f} \, (\pi - c_i) = (P_{gg} D_{i,gg} + P_{bg} D_{i,bg} + P_{gb} D_{i,gb} + P_{bb} D_{i,bb}) \, (\pi - c_i) \quad (5.2)$$

In the remainder of this section, each supply chain structure alternative we consider: independent suppliers case and shared suppliers case, is discussed in detail.

5.3.1 Independent Suppliers Case

In the case of independent suppliers, each firm works with its own independent supplier. The two suppliers are independent of each other, in the sense that adverse events at one supplier do not impact the other supplier. The model setting is illustrated in Figure 5.1.

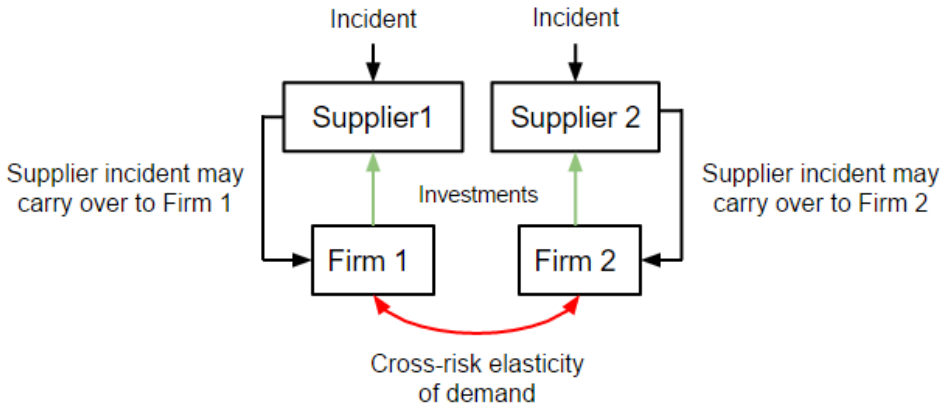


Figure 5.1. Model Setting for Two Firms with Independent Suppliers Case

Each supplier is affected by incidents according to its vulnerability v, and the amount of supplier spending by the client firm. Following the periodic model from Gordon and Loeb (2003), supplier state probabilities can be calculated as:

$$P_{GG} = (1 - \frac{v}{1+c_1})(1 - \frac{v}{1+c_2}) \qquad P_{BB} = (\frac{v}{1+c_1})(\frac{v}{1+c_2})$$

$$P_{GB} = (1 - \frac{v}{1+c_1})(\frac{v}{1+c_2}) \qquad P_{BG} = (\frac{v}{1+c_1})(1 - \frac{v}{1+c_2}) \tag{5.3}$$

When an incident occurs at a supplier, it may carry over to the client firm with a certain cascade probability ρ. Using the supplier state probabilities in (5.3) and the cascade probability, the firm joint state probabilities are calculated as:

$$P_{gg} = P_{GG} + P_{GB}(1-\rho) + P_{BG}(1-\rho) + P_{BB}(1-\rho)^2 \qquad P_{bb} = P_{BB}\,\rho^2$$

$$P_{gb} = P_{BG}\,\rho + P_{BB}\,\rho(1-\rho) \qquad P_{bg} = P_{GB}\,\rho + P_{BB}\rho(1-\rho) \tag{5.4}$$

By substituting joint firm probabilities from (5.4) and demand functions from (5.1) in the profit function (5.2) we obtain the firm's expected profit in the independent case as:

$$E[\Pi_i] = \frac{\rho\, v\,(Z_D(1+c_j)+Z_C(1+c_i))-(1+c_i)(1+c_j)}{(1+c_i)(1+c_j)}(\pi - c_i) \quad \text{for } i,j = 1,2, \text{ and } j \neq i \tag{5.5}$$

We use (5.5) to calculate the equilibrium spending for both firms. The only equilibrium supplier spending in this case is symmetric:

$$c_i{}^e = \frac{1}{2}\left(\rho\, v\, Z_C + \sqrt{\rho\, v\,\left(4\, Z_D\,(1+\pi) + \rho\, v\, Z_C{}^2\right)} - 2\right) \quad \text{for } i = 1,2 \tag{5.6}$$

which yields the expected equilibrium profit of:

$$E[\Pi_i{}^e] = \frac{2\rho\, v\, Z_D{}^2 + \rho\, v\ Z_C{}^2 - Z_C\sqrt{\rho\, v\left(4\, Z_D(1+\pi)+\rho\, v\, Z_C{}^2\right)} + 2Z_D\left(1+\pi+\rho v\, Z_C - \sqrt{4\rho\, v\, Z_D(1+\pi)+\rho\, v\, Z_C{}^2}\right)}{2Z_D} \tag{5.7}$$

We provide the details of the calculations as well as some of the properties of equilibrium supplier spending in Appendix 5.1.

5.3.2 Shared Supplier Case

The other strategic option for the firms is for each to spend on a shared supplier. The model setting for this case is provided in Figure 5.2 below.

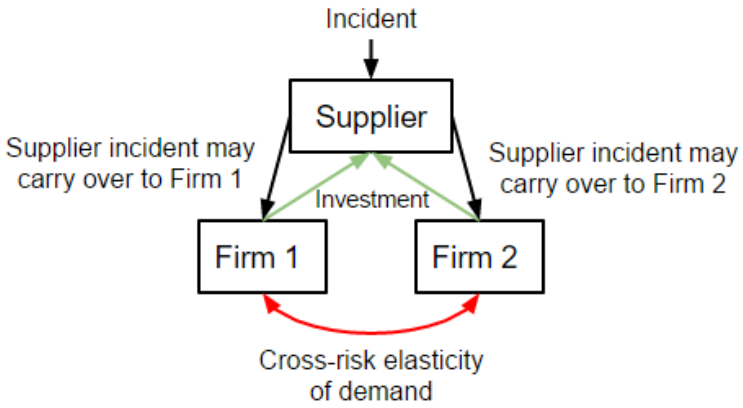


Figure 5.2. Model Setting for Two Firms with Shared Supplier Case

In this model setting, both firms spend on a single supplier, and the supplier state probabilities can be calculated following the periodic model from Gordon and Loeb (2003) as:

$$P_G = 1 - \frac{v}{1 + c_1 + c_2} \qquad P_B = \frac{v}{1 + c_1 + c_2} \tag{5.8}$$

Using supplier state probabilities in (5.8), firm state probabilities are calculated as:

$$P_{gg} = P_G + P_B(1 - P_c)^2 = 1 - \frac{(2 - \rho)\ \rho\ v}{1 + c_1 + c_2} \qquad P_{bb} = P_B \rho^2 = \frac{\rho^2\ v}{1 + c_1 + c_2}$$

$$P_{gb} = P_{bg} = P_B(1 - P_c)P_c = \frac{(1 - \rho)\ \rho\ v}{1 + c_1 + c_2} \tag{5.9}$$

By substituting the joint firm state probabilities from (5.9) and demand functions from (5.1) in the profit equation (5.2) we obtain the firms' expected profit in the shared case as:

$$E[\Pi_i] = \frac{1 + c_i + c_j - \rho\ v\ (Z_D + Z_C)}{1 + c_i + c_j}(\pi - c_i) \ \text{ for } i, j = 1,2, \ \text{ and } j \neq i \tag{5.10}$$

which can be used to calculate the equilibrium spending for both firms. The only equilibrium in this case is symmetric:

$$c_i{}^e = \frac{1}{8}\left(\rho\ v\ (Z_D + Z_C) + \sqrt{\rho\ v\ (Z_D + Z_C)\left(16\,\pi + 8 + \rho\ v(\ Z_D + Z_C)\right)}\ - 4\right) \tag{5.11}$$

which yields the expected equilibrium profit of

$$E[\Pi_i{}^e] = \pi + \frac{1}{8}\left(4 + 5\rho\ v\ (Z_D + Z_C) - 3\sqrt{\rho\ v\ (Z_D + Z_C)\left(16\,\pi + 8 + \rho\ v(\ Z_D + Z_C)\right)}\right) \tag{5.12}$$

5.4 Comparing Shared Supplier versus Independent Supplier Cases

In this section, we compare the two strategic supply chain design choices of independent suppliers versus a shared supplier to find out the conditions for which each option yields higher profit for the firms. Because the two models differ only in the supplier structure (independent or shared suppliers), we can directly compare the profit from these two choices by examining the equilibrium expected profits from sections 5.3.1 and 5.3.2. Lemma 5.1 provides the conditions for each of the strategic choices to be optimal.

Lemma 5.1 *The independent supplier choice is optimal when $L > 0$, the shared supplier choice is optimal when $L < 0$, and the two choices are equal when $L = 0$, where:*

$$L = \frac{1}{8Z_D}\Bigg(4\rho v\left({Z_D}^2 + 2\,Z_D\,Z_C + {Z_C}^2\right) - 4(2\,Z_D + Z_C)\sqrt{\rho\, v\left(4\,Z_D(1+\pi) + \rho\, v\,{Z_C}^2\right)}$$

$$+ Z_D\left(4 - 5\rho v(Z_D + Z_C) + 3\sqrt{\rho v(Z_D + Z_C)\left(8 + 16\pi + \rho v(Z_D + Z_C)\right)}\right)\Bigg)$$

While the analytical form of the regions where each of the two choices are optimal is provided, the closed-form equations for these regions are not tractable. Instead, we numerically analyze the difference between the equilibrium expected profits of two supplier choices. The findings regarding conditions of direct- and cross-risk elasticities of demand when each strategic choice is optimal are provided in the following observation.

Observation 5.1 *Comparing independent suppliers versus a shared supplier based on the direct- and cross-risk elasticities of demand, there are two possible regions for the optimal supplier choice. Firms will choose independent suppliers when both direct- and cross-risk elasticities of demand are sufficiently low and will choose the shared supplier when*

(i) direct-risk elasticity of demand is sufficiently high, or

(ii) cross-risk elasticity of demand is sufficiently large in magnitude, or

(iii) conditions (i) and (ii) both hold.

The regions in which each of the two options is optimal for a particular combination of direct- and cross-risk elasticities of demand is provided in Figure 5.3. The union of the shared supplier and independent supplier regions outlines the feasible region as presented by the dotted lines. The feasible region is characterized by the demand requirements $Z_D \leq 1$, $Z_C \leq Z_D$, $Z_C \geq -Z_D$, and $Z_C + Z_D \leq 1$. Considering both elasticities in conjunction, a clear pattern emerges as described in Observation 5.1.

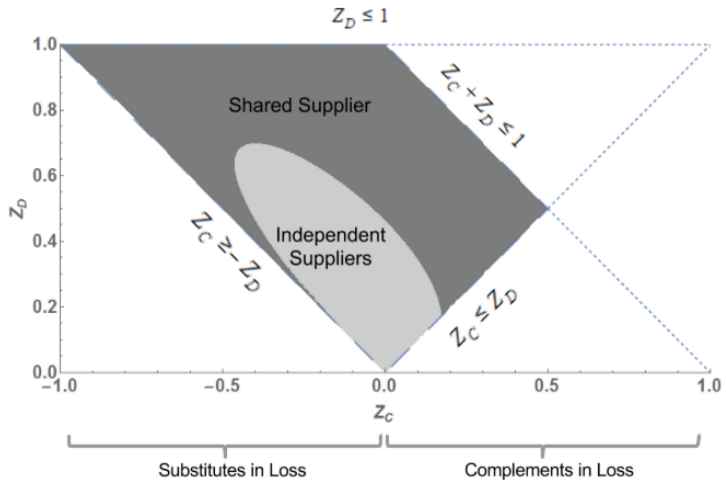


Figure 5.3. Optimality Regions for Independent Supplier versus Shared Supplier Cases

$$(\pi = 2, \rho = 0.9, v = 0.8)$$

Additionally, the decision to choose independent or shared suppliers also depends on the other factors of supplier vulnerability v, cascade probability ρ, and firm per unit profit π. We

discuss the effect of these parameters on the independent versus shared supplier choice in the remainder of this section.

Assuming that the vulnerabilities of shared and independent suppliers are equal, we find that increased (decreased) supplier vulnerability v, would expand (contract) the region where the shared supplier choice is optimal. In other words, when suppliers are more vulnerable, firms choose the shared supplier case over a wider range of direct- and cross-risk elasticities of demand. The intuition for this observation is that when suppliers are less secure, firms would rather pool their resources into one supplier in order to improve that supplier's security and reduce the impact of adverse events on themselves. This preference results in a larger region in which the shared option is preferred by the firms. We find the cascade probability to have a similar effect to that of supplier vulnerability. That is, when the cascade probability ρ is high (low), firms choose the shared supplier over a wider (smaller) range of direct- and cross-risk elasticities of demand.

Upon examination of how the firm's per unit profit π affects the decision to use a shared or independent supplier, we find that the shared supplier region expands as the firm's per unit profit increases. This implies that in industries with high profit margins, firms would share suppliers with competing firms over a wider range of direct- and cross-risk elasticities of demand. On the other hand, in industries with tight profit margins, independent suppliers would be utilized over a wider range of direct- and cross-risk elasticities of demand.

5.5 Correlated Arrivals

Until now, we considered incidents at different suppliers to be independent for the independent suppliers case. However, when the incidents have a common source, it is reasonable to argue

that incidents may be correlated across suppliers. Here, we explore the impact of adverse event correlation on the decision-making of firms. We use a correlation parameter, $\gamma \geq 0$, which is an increasing function of the correlation between the incidents. The parameter γ represents the likelihood that both firms go to the bad state simultaneously (bb). On the extreme, when γ = 0, the model is reduced to the basic model without correlated arrivals. The mathematical details of how the parameter γ impacts correlation of adverse events are provided in the Appendix 5.2, and here we report the results.

We analyze the effect of correlation on supplier decision-making. Figure 5.4 shows this effect for a representative numerical example with two different values for γ.

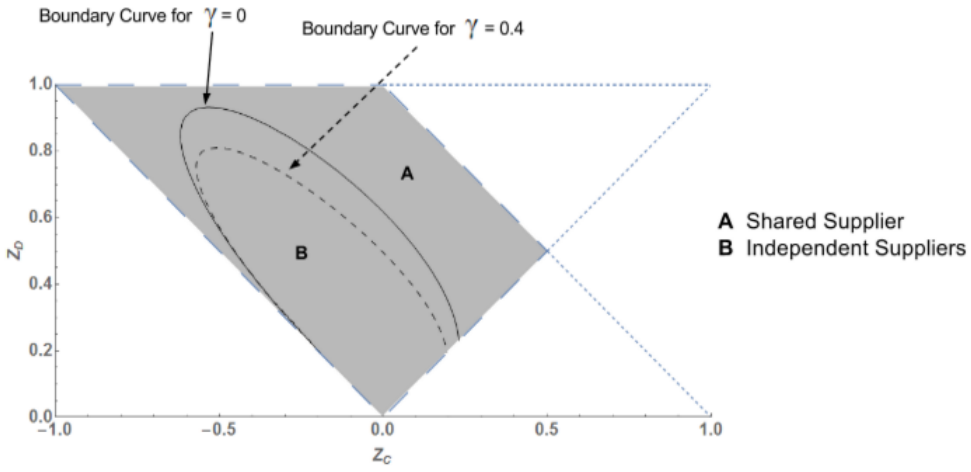


Figure 5.4. Optimality Regions for Various Adverse Event Correlations

$$(\pi = 2, \rho = 0.9, v = 0.6, \gamma = \{0, 0.4\})$$

From our numerical examples, we observe that the increase in correlation favors the shared supplier, as the region for the shared supplier increases as adverse event correlation increases. When the incidents are more likely to happen simultaneously for both firms, it makes sense for them to pool their resources as a response strategy.

5.6 Effect of Firms' Spending Coordination on Strategic Supplier Choice

In this section, we examine the effect of coordination of spending on supplier by both firms using the cases described in Section 5.3. Coordination could occur when a parent company owns two competing brands, or when regulation and industry cooperation allow two firms to coordinate spending on suppliers, regardless of whether independent or shared suppliers are selected. Note that coordination is not the same effect as pooling resources. In some cases, companies may be able to coordinate to increase or decrease the spending on suppliers (compared to the equilibrium spendings) for increased profit. In other cases, a minimum required spending, or regulated spending, may be set on each firm's security investment.

Here, we provide the analysis for the case where firms can coordinate to choose the amount of spending. The numerical analysis results are provided here, and the details of the analysis are provided in Appendix 5.3. If the firms can coordinate their spending to a given amount, then the regions of direct- and cross-risk elasticity of demand over which the shared and independent choices are optimal is as shown in Figure 5.5.

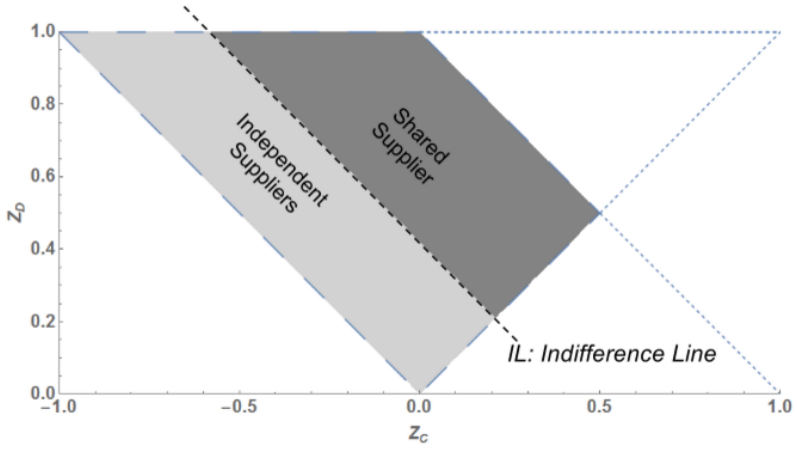


Figure 5.5. Independent versus Shared Choices under Firm Spending Coordination

$(\pi = 2, \rho = 0.9, v = 0.7)$

Similar to the results in Section 5.4, the model parameters of vulnerability, cascade probability, and per unit profit affect the optimal choices of independent and shared suppliers when firms coordinate. Observation 5.2 summarizes these findings.

Observation 5.2 *Comparing independent suppliers versus a shared supplier when firms coordinate, we find that:*

(i) as the vulnerability of the suppliers v increases, the shared supplier region expands (the indifference line IL in Figure 5.5 moves to the left).

(ii) as the cascade probability for an incident to carry over to the firms ρ increases, the shared supplier region expands (the indifference line IL in Figure 5.5 moves to the left).

(iii) as firm per unit profit (excluding the spending) π increases, the shared supplier region expands (the indifference line IL in Figure 5.5 moves to the left).

In terms of the spendings, we observe that in the shared supplier choice case, the equilibrium spendings when coordinated are always higher than the equilibrium spendings without coordination. For the independent supplier choice case, when firms are complements in loss ($Z_C > 0$), the equilibrium spendings with coordination are higher than the equilibrium spendings without coordination. However, when firms are substitutes in loss ($Z_C < 0$), then the equilibrium spendings with coordination are lower than the equilibrium spendings without coordination. These results are shown for a numerical example in Figure 5.6. The darker (lighter) shaded region in Figure 5.6 describes the area where there is an increase (decrease) in equilibrium spending with coordination.

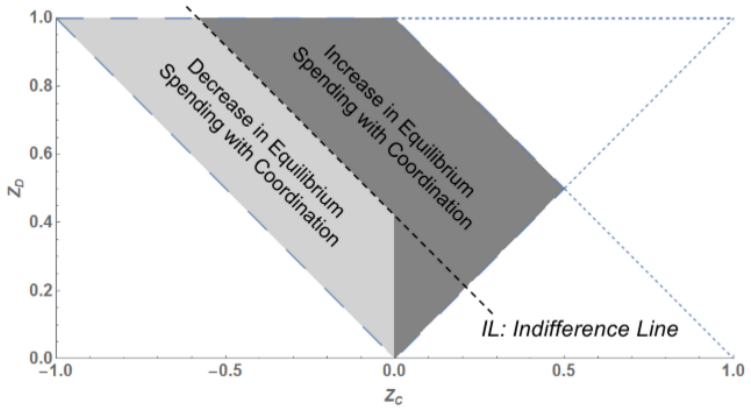


Figure 5.6. Effect of Firm Spending Coordination on Equilibrium Spendings

$(\pi = 2, \rho = 0.9, v = 0.7)$

Since consumers are negatively impacted by incidents at the firms, it can be expected that when spendings on suppliers increase, the consumer welfare would also increase. Using this argument, it can be seen from Figure 5.6 that when firms are substitutes in loss and they decide to go with independent suppliers, then the coordination negatively impacts consumers. However, in all other situations, coordination of firm spendings may actually be useful for the consumer. This result has important implications for regulators, in that regulators should only prevent coordinated spending among firms when they are substitutes in loss and there is benefit for them to use independent suppliers. In other words, this implies that when increased spending benefits the consumers, regulators could persuade firms to create alliances when there is high competition among the firms. There are many regulatory acts related to anti-competitive behavior (e.g., The Competition Act 1998, United Kingdom) which could be interpreted to prohibit these beneficial shared supplier strategies, especially when there is high competition between firms. When it comes to supplier selection, regulators should be careful to allow supply chain coordination that will benefit the consumers.

5.7 Analysis of Asymmetry

In this section we analyze the impact of asymmetry on the strategic supplier choice decision. We analyze the asymmetry with respect to two aspects: asymmetry of suppliers (see Section 5.7.1) and asymmetry of firms (see Section 5.7.2).

5.7.1 Asymmetric Shared and Independent Suppliers Cases

In this section we analyze how the asymmetry among shared and independent suppliers impacts firms' decision to go with either of these choices. We consider asymmetric shared and independent cases, where the two models may have dissimilar vulnerabilities, per unit profits, and cascade probabilities. The results in this section are based on our numerical test suites, and the results are consistent among all of the parameters tested.

Asymmetric Supplier Vulnerabilities: We consider the vulnerability of each of the independent suppliers to be v_I and the vulnerability of the shared supplier to be v_S, and then compare the decisions when these parameters are not necessarily equal. The relative vulnerability ratio is provided as $R_v = \frac{v_S}{v_I}$. We would expect that increase in R_v would favor the independent suppliers, as it means that the shared supplier is more vulnerable. Figure 5.7 illustrates in a numerical example that as the ratio R_v increases, the region in which the shared supplier case is optimal shrinks. Note that in this figure, consistent with Figure 5.3, the area outside the boundary curves (A) shows where the shared supplier is optimal, and the area inside the boundary curve (B) shows where the independent supplier is optimal. As an increase (decrease) in R_v indicates that the vulnerability in the shared case increases (decreases) compared to the independent case, and the independent suppliers become desirable over a wider (smaller) range of direct- and cross-risk elasticities of demand. In Figure 5.7, the interior of the boundary curve is optimal for the

independent supplier case, and the exterior of the boundary curve is optimal for the shared supplier case.

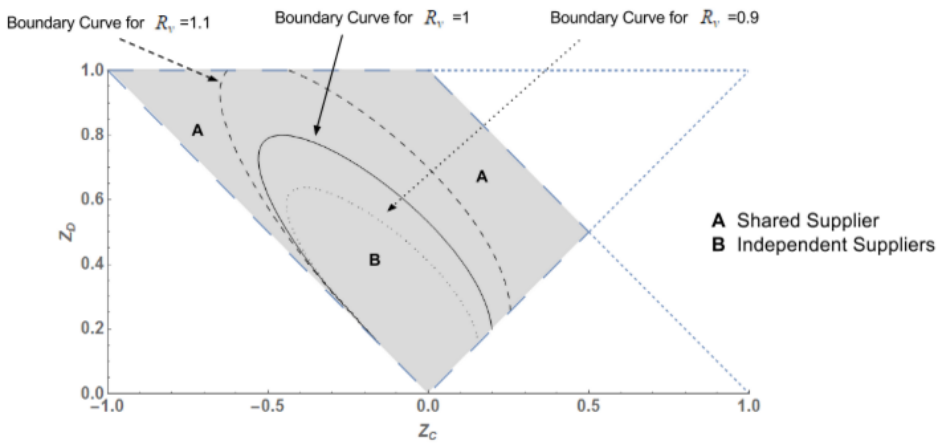


Figure 5.7. Optimality Regions for Various Relative Vulnerability Ratios

$$(\pi = 2, \rho = 0.9, v_I = 0.7, v_S = \{0.63, 0.7, 0.77\})$$

Asymmetric Per Unit Profits: Consider the case where the per unit profit of the firms depend on the type of supplier they choose. This may be the case when the cost of using shared and independent suppliers are not equal. Let the per unit profit of the firm from using independent supplier and shared supplier cases to be π_I and π_S, respectively. The relative per unit profit ratio is provided as $R_\pi = \frac{\pi_S}{\pi_I}$. Figure 5.8 illustrates in a numerical example how the independent and shared regions are affected by different ratios of R_π. It can be seen that as R_π increases, the shared region expands.

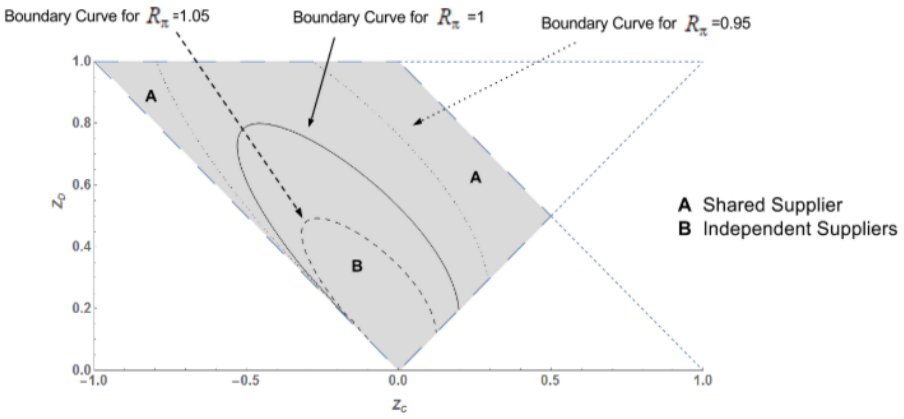


Figure 5.8. Optimality Regions for Various Relative Per Unit Profit Ratios

$$(\pi_I = 2, \pi_S = \{1.9,\ 2,\ 2.1\}, \rho = 0.9, v = 0.7)$$

Asymmetric Cascade Probabilities: We consider cascade probability for the independent suppliers to be ρ_I and the cascade probability for the shared supplier to be ρ_S, and then compare the decisions when these parameters are not necessarily equal. The relative cascade probability ratio is provided as $R_\rho = \frac{\rho_S}{\rho_I}$. Figure 5.9 illustrates in a numerical example how the independent and shared regions are impacted by the relative cascade probabilities. It can be seen that the relative cascade probability ratio has a similar effect to the relative vulnerability ratio.

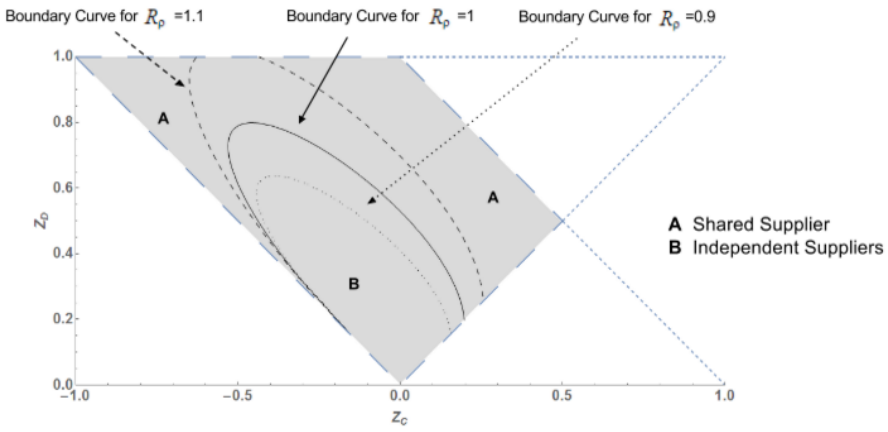


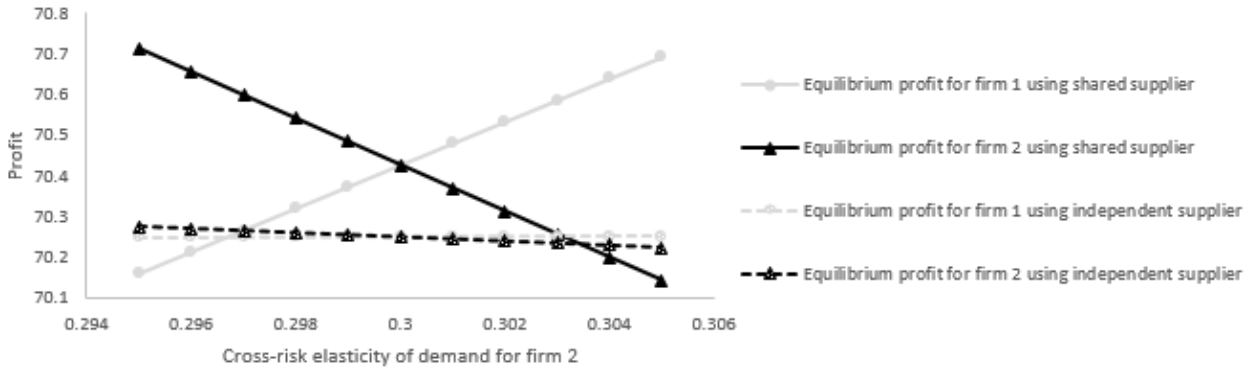
Figure 5.9. Optimality Regions for Various Relative Cascade Probability Ratios

$$(\pi = 2, \rho_I = 0.9, \rho_S = \{0.81,\ 0.9,\ 0.99\}, v = 0.7)$$

5.7.2 Asymmetric Firms

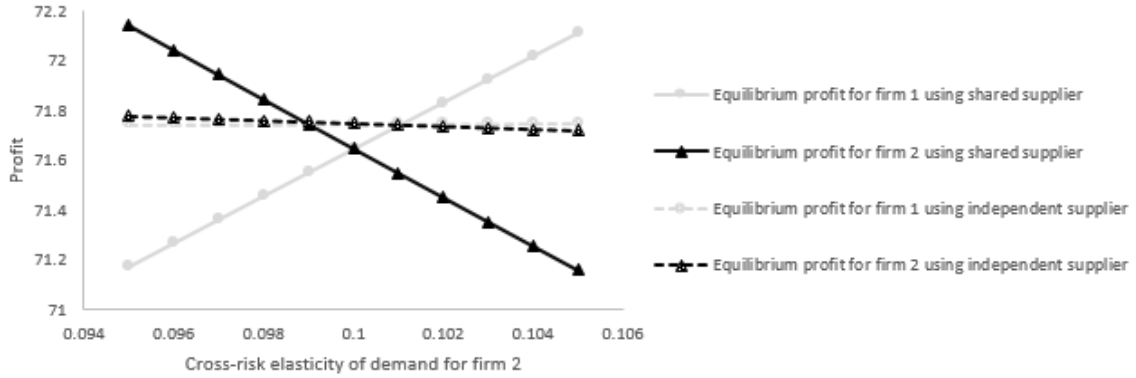
Until now, we have considered firms that are symmetric. In this section, we discuss firms with asymmetric parameters, including asymmetric customer demand reactions (direct- and cross-risk elasticities of demand), and cascade probabilities. We analyze this in a numerical test suite for a range of parameters, and present a representative numerical example here. These results are consistent among all of the parameters that we tested. We find the effect of asymmetry in direct- and cross-risk elasticities of demand, and cascade probabilities to be similar to each other. Therefore, for brevity, here we provide analysis for only one of these parameters (cross-risk elasticity of demand), and the rest of the analysis is provided in Appendix 5.4. Also note that the asymmetric cascade probabilities considered here are different from what we analyzed in Section 5.7.1. In this section, the firms, irrespective of the supplier they choose, have asymmetric cascade probabilities, whereas in Section 5.7.1, cascade probabilities are attribute of the suppliers (i.e., the asymmetric cascade probabilities belong to the suppliers).

Asymmetric Cross-Risk Elasticities of Demand: When firms are asymmetric in cross-risk elasticities of demand, we find that in both cases of shared and independent suppliers, the firm with lower cross-risk elasticity of demand spends less on IT security and gains a higher profit than the other firm. Figure 5.10 provides the two different scenarios that may arise in a representative numerical example. Figure 5.10.a illustrates the change in profits when cross-risk elasticity changes for Firm 2 and is fixed for firm 1 (at $Z_{1,C} = 0.3$). In this example, the shared option is optimal when the cross-risk elasticities are symmetric at $Z_{1,C} = Z_{2,C} = 0.3$. Figure 5.10.b shows an example in which the independent option is optimal when the cross-risk elasticities are symmetric at $Z_{1,C} = Z_{2,C} = 0.1$.



a) Case in Which the Shared Option is Optimal for Symmetric Firms

$(\pi = 75, \rho = 0.3, v = 0.5, Z_D = 0.4, Z_{1,C} = 0.4)$



b) Case in Which the Independent Option is Optimal for Symmetric Firms

$(\pi = 75, \rho = 0.3, v = 0.5, Z_D = 0.3, Z_{1,C} = 0.1)$

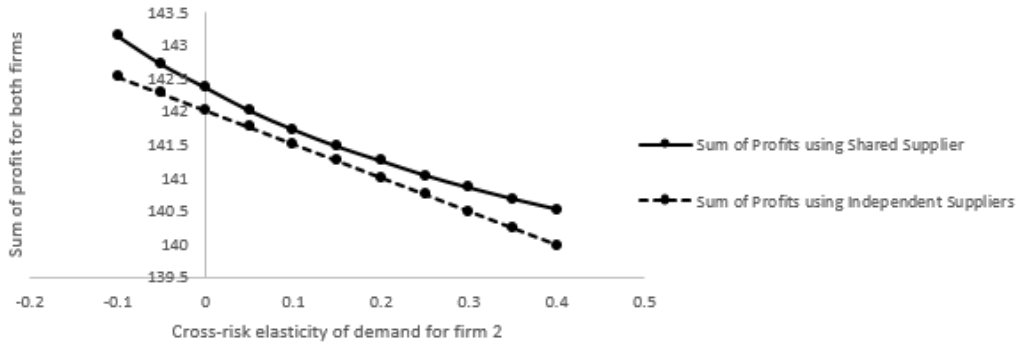
Figure 5.10. Effect of Changes in Cross-Risk Elasticity of Demand for Firm 2 on Profits

It can be seen from Figure 5.10 that asymmetry in the cross-risk elasticities of demand would decrease the profits from the shared supplier design choice for one of the firms more than it would decrease the profit from independent supplier design choice. The insight from this is that the asymmetry encourages firms to choose independent suppliers since one of the firms has less incentive to choose the shared supplier. Because it is necessary for both firms to prefer the shared supplier for that option to become viable, asymmetry in the cross-risk elasticities of

demand makes the shared option less desirable. If, in asymmetry, the shared supplier option is the optimal choice for the firms, then as firms grow more different (become more asymmetric) in cross-risk, there is a region in which they will still prefer the shared supplier choice. However, if the cross-risk of the two firms becomes too different, there is no longer incentive for them to share suppliers (Figure 5.10.a). When the optimal decision for symmetric firms is independent suppliers, then asymmetry would not change this decision (as illustrated in Figure 5.10.b).

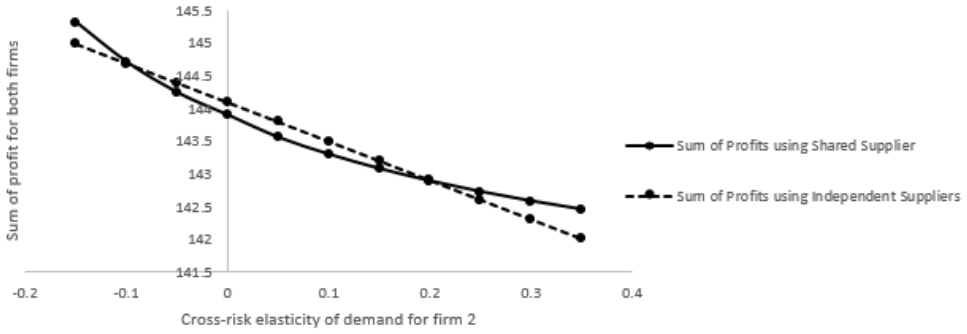
In some cases, it might be possible for a firm to make transfer payments to the competitor firm in order to make the shared option viable for them as well. For example, if both firms are run by a common parent company, the parent company may be interested in maximizing the combined profit from both firms. Figure 5.11 provides a numerical example of the sum of profits for both firms in the two scenarios presented in Figure 5.10.

In the first scenario (Figure 5.10.a), where symmetric firms would choose the shared supplier, if firms were able to make transfer payments, there can be a case where, while one firm's profit using the shared supplier is less than when using the independent supplier, the transfer payment should induce it to choose the shared supplier (Figure 5.11.a). In the second scenario (Figure 5.10.b), where firms would choose to use independent suppliers in the symmetric case, we see that a shared supplier may be preferred when firms are exceedingly different (Figure 5.11.b). In this case, the firm with the lower cross-risk elasticity could make a transfer payment to its competitor as an incentive to pool their resources through a shared supplier.



a) Sum of Profits for Case Where the Shared Option is Optimal for Symmetric Firms

$(\pi = 75, \rho = 0.3, v = 0.5, Z_D = 0.4, Z_{1,C} = 0.4)$



b) Sum of Profits for Case Where the Independent Option is Optimal for Symmetric Firms

$(\pi = 75, \rho = 0.3, v = 0.5, Z_D = 0.3, Z_{1,C} = 0.1)$

Figure 5.11. Effect of Changes in Cross-Risk Elasticity of Demand for Firm 2 on Aggregate Profits

An interesting finding here is the effect of transfer payments on the decision making of the firms. When transfer payments are not possible, we see that the asymmetry among firms is in favor of the independent supplier option (Figure 5.10). However, when transfer payments are possible, it is possible that two firms that would not share suppliers in the symmetric case, would share suppliers in the asymmetric case (Figure 5.11.b).

5.8 Discussion and Conclusions

Adverse IT security events may impact customer demand in different ways, depending on the direct- and cross-risk elasticities of demand. In a duopoly setting, this chapter presents an analysis the strategic choice of supply chain design by considering the impact of IT security breaches on customer demand. Firms must choose between selecting the same supplier as their competitor or looking for an independent supplier. We show how this choice affects the dynamics of the game, as well as how different customer demand reactions to adverse events affect the supply chain structure. To make an optimal decision, the relative differences in adverse event arrivals between supplier alternatives must be considered in conjunction with the direct- and cross-risk elasticities of demand.

Summarizing the parameters' effects on the strategic choice of independent or shared suppliers, we find the following. *Direct-risk elasticity of demand*, when low, will favor independent suppliers. As direct-risk elasticity increases, all things being equal, then the shared option becomes preferred. *Cross-risk elasticity of demand*, when large in magnitude, will favor the shared supplier design. However, if direct-risk elasticity is very large, then the shared choice becomes optimal for all values of cross-risk elasticity. *Correlation of incidents* generally favors the shared supplier option for the supply chain. *Cooperation on security spending* significantly changes the dynamics of the decision. If cooperation on spending is possible, then the shared option is optimal only at simultaneously large magnitude direct- and cross-risk elasticity of demand. *Asymmetry* in firm characteristics (such as cross-risk elasticity differences) largely favours the use of independent suppliers even for small differences between firms. However, we find that allowing for *transfer payments* between firms can, in fact, make shared suppliers more attractive for firms with greater differences.

This work provides a tool to aid managers in the strategic decision regarding choice of cloud service suppliers. Estimating the necessary parameters for different vendors is challenging. Certain parameters, such as the cascade probability, are firm-specific, depending on the firm's own security capabilities. Others, are industry-specific (eg. direct- and cross-risk elasticities) and will rely on the manager having a good feel for the level and type of competition. Finally, some parameters will be supplier- specific; a manager will need to gather information through careful reading of both supplier responses to the firm's request for proposals (RFPs) and public information about the supplier. When selecting suppliers for such important business functions, it is common to undergo an RFP process. As such, the proposals returned by vendors form an important component of information gathering for the manager. Evaluating the various responses to the call for proposals will be necessary to extract the data needed to estimate parameter values in our model, such as the cascade probability. Additional public information regarding vendor security performance can be used to estimate the likelihood of a breach or adverse event occurring.

Our results with respect to IT supply chain design both support and extend the analysis of Gal-Or and Ghose (2005), who find that cooperation and information sharing regarding IT security problems between firms is greater in competitive markets (highly negative cross-risk elasticities of demand in our model). Additionally, our results demonstrate that when the relative adverse event arrival rates for the shared supplier move downwards toward parity with the rate for independent suppliers, even firms that are complements in loss become increasingly motivated to share suppliers. This chapter illustrates the interplay between direct-risk elasticity, cross-risk elasticity, and relative adverse event arrival rates on the strategic supply chain design decision. We show that regulation and cooperation can be beneficial to the firm profits in many

cases, and generally, enhances profits in the shared supplier case. Although cooperation between competitors is often viewed negatively by regulatory agencies, results show that cooperation among competing firms with respect to spending on supplier security, if allowed, may in some cases benefit customers.

CHAPTER 6

Summary of Findings

With the growing number of people using online social networks and websites, the opportunities and challenges in this online industry are increasing. One the one hand, there are ample opportunities for the online services to monetize their platforms through advertising and selling of user information to third parties and by improving the performance of their networks, and on the other hand, users are concerned about their privacy when using such services. Chapters 2, 3 and 4 of this thesis address these respective issues. On another level, businesses and companies are vitally dependent on internet- and cloud-based services provided by suppliers, and need to strategize on maximizing the security of such services through investing in either independent or shared suppliers. We discuss this issue in Chapter 5. Here, we summarize the findings from these four essays.

In Chapter 2, we define the problem of improving propagation in malleable social networks. We propose the HMST as a framework for finding low cost network enhancements with acceptable CIT performance within a hop constraint. This framework can jointly determine nodes that should be targeted for seeding and connections that should be created in order to improve propagation performance. We propose the SLE improvement heuristic that greatly reduces the costs of the HMST solutions. The most important practical implication of this chapter is that social networks can become more effective conduits for message propagation by simultaneously considering network design changes and potential seeding points. This study also has important managerial implications, in that it shows how deliberate network

manipulation can be a potent tool for improving propagation in malleable networks, and provides an efficient method for accomplishing this.

In Chapter 3, we propose a two-sided economic model to explain and analyze the decision-making of publisher websites. We illustrate how a publisher website must balance the user information monetization through third party information sharing versus the subsequent personal privacy concerns that result from this sharing. We also study the effect of competition and asymmetry among publisher websites, and provide insights on the two different publisher website business models that arise: 1. low publisher website price and high user information sharing for the firm facing low user privacy concerns, and 2. high publisher website price and low user information sharing for the firm facing high user privacy concerns. In the empirical analysis, we find that user information is being shared extensively among third parties by publisher websites, and the actual third party usage behavior is consistent with the predictions of the model. Due to privacy concerns and potential for re-identification, the extent of third party sharing is of strong interest to policy makers and regulatory organizations. We examine the impact of two government taxation policies, and find that the impacts of these regulatory actions on profit and welfare depend on the level of user privacy concerns. This chapter also contributes to the two-sided market literature by considering the case where one side has a negative cross-sided network effect on the other, as opposed to having positive cross-sided network effects for both sides, as is the case in the literature. We find that in such markets, the subsidy strategy where one side is subsidized in order to monetize the other, can still be useful.

In Chapter 4 we study the impact of user privacy concerns as a self-regulatory mechanism to control the sharing intensity by websites. We find some evidence that user privacy concerns do have a self-regulating effect on the sharing intensity. Websites in subject categories with

higher user privacy concerns, tend to have lower levels of sharing intensity. In contrast, websites in categories with a priori expectations of low privacy concerns higher sharing intensity. We also find the effectiveness of DNT as a privacy protection tool in controlling the sharing intensity by website publishers is mixed. A significant proportion of websites in the subject categories with low privacy concerns abuse users' privacy by increasing sharing intensity when DNT is requested. Our empirical analysis in a relatively regulation-free environment indicates that without regulation or transparency with respect to third party sharing, the DNT signal is often being used to abuse customer privacy.

In Chapter 5, we study the decision making of firms when security attacks and incidents may impact firm's demand. Such incidents may impact customer demand in different ways, depending on the direct- and cross-risk elasticities of demand. In a duopoly setting, we analyze the strategic choice of firms in supply chain design by considering the impact of IT security breaches on customer demand. Firms must choose between selecting the same supplier as their competitor or an independent supplier. We show that different customer demand reactions to adverse events affect the supply chain structure by encouraging firms to invest either in independent suppliers, or to share a single supplier. To make an optimal decision, the relative differences in adverse event arrivals between supplier alternatives must be considered in conjunction with the direct- and cross-risk elasticities of demand. We find that when direct-risk elasticity of demand is low, firms will favor independent suppliers. As direct-risk elasticity increases, the shared option becomes preferred. On the other hand, when cross-risk elasticity of demand is large in magnitude (either positive and negative), firms will favor the shared supplier. However, if direct-risk elasticity is very large, then the shared choice may become optimal for all values of cross-risk elasticity of demand. We also study the impact of correlated incidents and

cooperation among firms, and find the optimal decisions for each case. This chapter provides a useful tool to aid managers in the strategic decision regarding choice of cloud service suppliers.

BIBLIOGRAPHY

Adler, M., Gibbons, P. B., and Matias, Y. 2002. "Scheduling space-sharing for internet advertising." *Journal of Scheduling* 5(2): 103-119.

Adobe. 2016. "Security Bulletins and Advisories." *Adobe*. Accessed online January 7, 2016 from: helpx.adobe.com/security.html

Akgun, I. 2011. "New formulations for the hop-constrained minimum spanning tree problem via Sherali and Driscoll's tightened Miller–Tucker–Zemlin constraints." *Computers & Operations Research* 38(1): 277–286.

Amirkhanian, Y.A., J.A. Kelly, E. Kabakchieva, T.L. McAuliffe, S. Vassileva. 2003. "Evaluation of a social network HIV prevention intervention program for young men who have sex with men in Russia and Bulgaria." *AIDS Education and Prevention* 15(3): 205-220.

Anand K. S., Goyal, M. 2009. "Strategic Information Management Under Leakage in a Supply Chain." *Management Science* 55(3): 438-52.

Anderson Jr, E. G., Parker, G. G., and Tan, B. 2013. "Platform performance investment in the presence of network externalities." *Information Systems Research* 25(1): 152-172.

August, T., Niculescu, M.F., Shin, H. 2014. "Cloud Implications on Software Network Structure and Security Risks." *Information Systems Research* 25(3): 489-510.

Barney, J. 1991. "Firm resources and sustained competitive advantage." *Journal of Management* 17(1): 99-120.

Barney, J. 1999. "How a firm's capabilities affect boundary decisions." *Sloan Management Review* 40(3): 137-145.

Barr, J. 2016. "The New York Times Begins Testing Ad Blocking Approaches." Ad Age, Accessed online April 5, 2016 from: http://adage.com/article/media/york-times-a-message-ad-blockers/302995/

Bélanger, F., & Crossler, R. E. 2011. "Privacy in the digital age: a review of information privacy research in information systems." *MIS Quarterly* 35(4): 1017-1042.

Bezemer, C. P., Zaidman, A. 2010. “Multi-tenant SaaS applications: maintenance dream or nightmare?” *Proceedings of the ACM Joint ERCIM Workshop on Software Evolution (EVOL) and International Workshop on Principles of Software Evolution (IWPSE)*: 88-92.

Bowie, N. E., & Jamal, K. 2006. “Privacy rights on the internet: self-regulation or government regulation?” *Business Ethics Quarterly* 16(03): 323-342.

Boyer, K.K., M. Swink, E.D. Rosenzweig. 2005. “Operations strategy research in the POMS journal.” *Production and Operations Management* 14(4): 442-449.

Bradley, T. 2011. “Lessons learned from the Epsilon data breach.” *PCWorld*. April 7. Accessed online June 14, 2015 from: www.pcworld.com/article/224615/lessons_learned_from_the_epsilon_data_breach.html

Brook, C. 2015. “Adobe patches 69 vulnerabilities in Reader, Acrobat, Flash.” *ThreatPost.* October 13, 2015. Accessed online January 7, 2016 at: threatpost.com/adobe-patches-69-vulnerabilities-in-reader-acrobat-flash/115005

Burt, R.S. 1987. “Social contagion and innovation: cohesion versus structural equivalence.” *American Journal of Sociology* 92(6): 1287-1335.

Cachon, G.P., Fisher, M. 2000. “Supply chain inventory management and the value of shared information.” *Management Science* 46(8):1032-1048.

Carter, B. 2012. *LinkedIn for Business: How Advertisers, Marketers and Salespeople Get Leads, Sales and Profits from LinkedIn*. Que Publishing.

Casadesus-Masanell, R., and Hervas-Drane, A. 2015. “Competing with Privacy.” *Management Science* 61(1): 229-246.

Cezar, A., Cavusoglu, H., Raghunathan, S. 2010. “Competition, speculative risks, and IT security outsourcing.” *Economics of Information Security and Privacy:* 301-320. Springer US.

Cha, M., A. Mislove, K.P. Gummadi. 2009. “A measurement-driven analysis of information propagation in the flickr social network.” *Proceedings of the 18th International Conference on World Wide Web:* 721-730.

Chellappa, R. K., and Shivendu, S. 2007. “An economic model of privacy: A property rights

approach to regulatory choices for online personalization." *Journal of Management Information Systems* 24(3): 193-225.

Chen, J., J. Stallaert. 2014. "An Economic Analysis of Online Advertising Using Behavioral Targeting." *MIS Quarterly* 38(2): 429–449.

Christakis, N. A., J. H. Fowler. 2008. "The collective dynamics of smoking in a large social network." *New England Journal of Medicine* 358(1): 2249-2258.

Cleeren, K., Dekimpe, M.G., Helsen, K. 2008. "Weathering product-harm crises." *Journal of the Academy of Marketing Science* 36: 262-270.

Clementi, A.E.F., M. Di Ianni, A. Monti, G. Rossi, R. Silvestri. 2005. "Experimental analysis of practically efficient algorithms for bounded-hop accumulation in ad-hoc wireless networks." *Proceedings of 19th IEEE international Parallel and Distributed Processing Symposium:* 247a.

Cross, R., S.P. Borgatti, A. Parker. 2002. "Making invisible work visible: using social network analysis to support strategic collaboration." *California Management Review* 44(2) 25-46.

Cross, R.L., R.D. Martin, L.M. Weiss. 2006. "Mapping the value of employee collaboration." *McKinsey Quarterly* 3: 28.

Culnan, M. J. 2000. "Protecting privacy online: Is self-regulation working?" *Journal of Public Policy & Marketing* 19(1): 20-26.

Dahl, G., L. Gouveia, C. Requejo. 2006. "On formulations and methods for the hop-constrained minimum spanning tree problem." M.G.C. Resende, P.M. Pardalos, eds. *Handbook of Optimization in Telecommunications*. Springer US. 493–515.

Deloitte. 2013. "Enterprise social networks: another tool, but not yet a panacea." *Technology, Media & Telecommunications Predictions*. Deloitte.

Demirkan, Haluk, Cheng H. K. 2008. "The Risk and Information Sharing of Application Services Supply Chain." *European Journal of Operational Research* 187(3): 765-84.

Diesner, J., T.L. Frantz, K.M. Carley. 2005. "Communication networks from the Enron email corpus: It's always about the people. Enron is no different." *Computational and Mathematical Organization Theory* 11(3): 201-228.

Dobby, C. 2015. “Bell retreats from online tracking policy.” *The Globe and Mail*, April 7, 2015. Retrieved April 9th, 2015 from: www.theglobeandmail.com/report-on-business/privacy-watchdog-urges-bell-to-change-web-tracking-policy/article23822585

Domingos, P., M. Richardson. 2001. “Mining the network value of customers.” *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining:* 57-66.

Dyer, J.H., Singh, H. 1998. “The relational view: Cooperative strategy and sources of interorganizational competitive advantage.” *The Academy of Management Review* 23(4): 660-679.

Eisenmann, T., Parker, G., and Van Alstyne, M. W. 2006. “Strategies for two-sided markets.” *Harvard Business Review* 84(10): 92.

Electronic Frontier Foundation. 2015. “Do Not Track.” Retrieved August 14, 2015 from: www.eff.org/issues/do-not-track

eMarketer, 2014. “Mobile Continues to Steal Share of US Adults' Daily Time Spent with Media.” *eMarketer*, Retrieved July 20, 2015 from: www.emarketer.com/Article/Mobile-Continues-Steal-Share-of-US-Adults-Daily-Time-Spent-with-Media/1010782

Esau, L.R., K.C. Williams. 1966. “On teleprocessing system design, part II: a method for approximating the optimal network.” *IBM Systems Journal* 5(3): 142-147.

Facebook. 2015a. “Facebook advertising targeting options.” Retrieved September 18, 2015 from: www.facebook.com/business/products/ads/ad-targeting

Facebook. 2015b. “Facebook lookalike audiences.” Retrieved September 24, 2015 from: www.facebook.com/business/a/online-sales/lookalike-audiences

Fernandes, M., L. Gouveia, S. Voß. 2007. “Determining hop-constrained spanning trees with repetitive heuristics.” *Journal of Telecommunications and Information Technology* 4: 16-22.

Financial Times 2015. Retrieved April 4, 2015, from: www.ft.com

Friedkin, N.E., E.C. Johnsen. 1990. “Social influence and opinions.” *Journal of Mathematical Sociology* 15(3-4): 193-206.

Gal-Or, E., Ghose, A. 2005. “The economic incentives for sharing security information.” *Information Systems Research* 16(2): 186-208.

Garg., S.K., Versteeg,S., Buyya, R. 2013. “A framework for ranking of cloud computing services.” *Future Generation Computer Systems* 29: 1012-1023.

Gordon, L.A., Loeb, M.P., Luchyshyn, W. 2003. “Sharing information on computer systems security: An economic analysis.” *Journal of Accounting and Public Policy* 22: 461-485.

Goldenberg, J., B. Libai, E. Muller. 2001. “Talk of the network: a complex systems look at the underlying process of word-of-mouth.” *Marketing Letters* 12(3): 211-223.

Gonçalves, B., N. Perra, A. Vespignani. 2011. “Modeling users' activity on Twitter networks: validation of Dunbar's number.” *PLoS One* 6(8): 22656.

Gopal, R., Hidaji, H., Patterson, R.A., Rolland, E., and Zhdanov, D. 2014. “Information Sharing in Web Services: An Exploratory Analysis.” *24th Workshop on Information Technology and Systems (WITS), December 17-19, 2014, Auckland, New Zealand.*

Gopal, R., Hidaji, H., Patterson, R.A., Rolland, E., and Zhdanov, D. 2015a. “How Much to Share with third parties? A Website’s Dilemma and Users’ Privacy Concerns.” *Theory in Economics of Information Systems (TEIS), March 13-15, 2015, Banff, Canada.*

Gopal, R., Hidaji, H., Patterson, R.A., Rolland, E., and Zhdanov, D. 2015b. “Self-regulation and the “Invisible Hand” of User Privacy Concerns.” *25th Workshop on Information Technology and Systems (WITS), December 12-13, 2015, Dallas, TX.*

Gouveia, L. 1995. “Using the Miller-Tucker-Zemlin constraints to formulate a minimal spanning tree problem with hop constraints.” *Computers & Operations Research* 22(9): 959-970.

Gouveia, L. 1996. “Multicommodity flow models for spanning trees with hop constraints.” *European Journal of Operational Research* 95(1): 178-190.

Guha, R., R. Kumar, P. Raghavan, A. Tomkins. 2004. “Propagation of trust and distrust.” *Proceedings of the 13th international conference on World Wide Web:* 403-412.

Ha, A. Y., Tong, S. 2008. “Contracting and information sharing under supply chain competition.” *Management Science* 54(4): 701-715.

Hausken, K., 2006. “Returns to information security investment: The effect of alternative information security breach functions on optimal investment and sensitivity to vulnerability.” *Information Systems Frontiers* 8: 338-349.

Hee-jin, K. 2011. “Nation’s worst hacking attack came from China.” *Korea Joongang Daily*. August 12. Accessed on June 14, 2015 from: koreajoongangdaily.joins.com/news/article/article.aspx?aid=2940129

Helft, M. and Vega, T. 2010. “Retargeting Ads Follow Surfers to Other Sites.“ *New York Times*, Aug 29, 2010, Accessed online June 11, 2016 from: www.nytimes.com/2010/08/30/technology/30adstalk.html?_r=0

Horowitz, M. 2011. “Spear Phishing: the real danger behind the Epsilon data breach.” *Computer World*. April 7. Accessed online June 14, 2015 from: www.computerworld.com/article/2471050/e-commerce/spear-phishing--the-real-danger-behind-the-epsilon-data-breach.html

Hyung-eun, K. 2011. “ESTsoft was ‘host’ of massive cyber caper.” *Korea Joongang Daily*. August 6. Accessed on June 14, 2015 from: koreajoongangdaily.joins.com/news/article/article.aspx?aid=2939860.

Jamal, K., Maier, M., and Sunder, S. 2005. “Enforced standards versus evolution by general acceptance: A comparative study of e-commerce privacy disclosure and practice in the United States and the United Kingdom.” *Journal of Accounting Research* 43(1): 73-96.

Jayaraman, V., R. A. Patterson, E. Rolland. 2003. “The design of reverse distribution networks: models and solution procedures.” *European Journal of Operational Research* 150(1): 128-149.

Kim, Y. S., A. Mahidadia, P. Compton, A. Krzywicki,W. Wobcke, X. Cai, M. Bain. 2012. “People-to-people recommendation using multiple compatible subgroups.” M. Tsielscher, D. Zhang, eds. *AI 2012: Advances in Artificial Intelligence*. Springer. 61-72.

Kelle, P., Akbulut, A. 2005. “The role of ERP tools in supply chain information sharing, cooperation, and cost optimization.” *International Journal of Production Economics* 93-94: 41-52.

Kempe, D., J. Kleinberg, E. Tardos. 2003. “Maximizing the spread of influence through a social network.” *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining:* 137-146.

Kolfal, B., Patterson, R. A., Yeo, M. L. 2013. “Market Impact on IT Security Spending.” *Decision Sciences* 44(3): 517-556.

Kosner, A. 2013. “Facebook is recycling your likes to promote stories you’ve never seen to all your friends.” *Forbes*. Retrieved September 17, 2015 from: www.forbes.com/sites/anthonykosner/2013/01/21/facebook-is-recycling-your-likes-to-promote-stories-youve-never-seen-to-all-your-friends/

Krishnamurthy, B., Malandrino, D., and Wills, C. E. 2007. “Measuring privacy loss and the impact of privacy protection in web browsing.” *Proceedings of the 3rd Symposium on Usable Privacy and Security*: 52-63.

Krishnamurthy, B., and Wills, C. E. 2006. “Generating a privacy footprint on the Internet.” *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*: 65-70.

Krishnamurthy, B., and Wills, C. E. 2009a. “On the leakage of personally identifiable information via online social networks.” *Proceedings of the 2nd ACM Workshop on Online Social Networks:* 7-12.

Krishnamurthy, B., and Wills, C. 2009b. “Privacy diffusion on the web: a longitudinal perspective.” *Proceedings of the 18th International Conference on World Wide Web:*541-550.

Krishnamurthy, B., Naryshkin, K., and Wills, C. 2011. “Privacy leakage vs. protection measures: the growing disconnect.” *Proceedings of the Web2.0 Security and Privacy Workshop* (2): 1-10.

Kruskal, J.B. 1956. “On the shortest spanning subtree of a graph and the traveling salesman problem.” *Proceedings of the American Mathematical Society* 7(1) 48-50.

Kumar, S., and Sethi, S. P. 2009. “Dynamic pricing and advertising for web content providers.” *European Journal of Operational Research* 197(3): 924-944.

Lennon, M. 2014. “Flaw in AirWatch by VMware Leaks Info in Multi-Tenant Environments.” *Security Week.* December 10, 2014. Accessed online January 7, 2015 from: www.securityweek.com/flaw-airwatch-vmware-leaks-info-multi-tenant-environments.

Leskovec, J., L.A. Adamic, B.A. Huberman. 2007. “The dynamics of viral marketing.” *ACM Transactions on the Web* 1(5).

Leskovec, J., L. Backstrom, R. Kumar, A. Tomkins. 2008. “Microscopic evolution of social networks.” *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 462-470.

Li, Y. 2012. “Theories in online information privacy research: A critical review and an integrated framework.” *Decision Support Systems 54*(1): 471-481.

LinkedIn. 2015. Audience Expansion. Retrieved September 24, 2015 from: help.linkedin.com/app/answers/detail/a_id/51626/ft/eng

Lv, S., L. Pan. 2014. “Influence maximization in independent cascade model with limited propagation distance.” W. Han, Z. Huang, C. Hu, H. Zhang, L. Guo, eds. *Web Technologies and Applications*. Springer. 23-34.

Mak, H-Y., Shen, Z-J. 2012. “Risk diversification and risk pooling in supply chain design.” *IIE Transactions* 44(8): 603-621.

Malandrino, D., and Scarano, V. 2013. “Privacy leakage on the Web: Diffusion and countermeasures” *Computer Networks 57*(14): 2833-2855.

Malik, O. 2015. “Why companies won’t learn from the T-Mobile/Experian hack.” *The New Yorker,* October 6, 2015, Accessed online March 16, 2016 from: www.newyorker.com/business/currency/why-companies-wont-learn-from-the-t-mobileexperian-hack.

Mayer, J. R., and Mitchell, J. C. 2012. “Third party web tracking: Policy and technology.” *2012 IEEE Symposium on Security and Privacy:* 413-427.

McAuley, J., J. Leskovec. 2012. “Learning to discover social circles in ego networks.” *Advances in Neural Information Processing Systems* 25: 548-556.

McDonald, A. M., and Cranor, L. F. 2010. “Beliefs and behaviors: Internet users’ understanding

of behavioral advertising." *Proceedings of the 2010 Research Conference on Communication, Information and Internet Policy*.

Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. 2012. "Detecting price and search discrimination on the Internet." *Proceedings of the 11th ACM Workshop on Hot Topics in Networks:* 79-84.

Moricz, M., Y. Dosbayev, M. Berlyant. 2010. "PYMK: friend recommendation at Myspace." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data:* 999-1002.

Moss, L. 2014. "Publishers try crazy idea: fewer ads, higher pricing." *Digiday.* Accessed online April 5, 2016 from: digiday.com/publishers/publishers-foregoing-online-ads/

New York Times. 2016. "NYT Subscription Options." *New York Times*, Accessed online March 8, 2016 from: international.nytimes.com/subscriptions/inyt/lp87JWF.html?currency=loonie&adxc=277709&adxa=406555&page=homepage.nytimes.com/index.html&pos=Bar1&campaignId=4L9XJ

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

Ni, Y., L. Xie, Z.Q. Liu. 2010. "Minimizing the expected complete influence time of a social network." *Information Sciences* 180(13): 2514-2527.

Ojala, M., Hallikas, J. 2006. "Investment decision-making in supplier networks: Management of risk." *International Journal of Production Economics* 104: 201-213.

Parker, G. G., and Van Alstyne, M. W. 2005. "Two-sided network effects: A theory of information product design." *Management Science 51*(10): 1494-1504.

Parmigiani, A., Klassen, R.D., Russo, M.V. 2011. "Efficiency meets accountability: Performance implications of supply chain configuration, control, and capabilities." *Journal of Operations Management* 29: 212-223.

Pavlou, P. A. 2011. "State of the information privacy literature: where are we now and where should we go?" *MIS Quarterly* 35(4): 977-988.

Pirkul, H., S. Soni. 2003. “New formulations and solution procedures for the hop constrained network design problem.” *European Journal of Operational Research* 148(1) 126-140.

Prim, R.C. 1957. “Shortest connection networks and some generalizations.” *Bell System Technical Journal* 36(6): 1389-1401.

Quintin, C. 2015. “HealthCare.gov Sends Personal Data to Dozens of Tracking Websites.” *Electronic Frontier Foundation*, Retrieved April 4, 2015 from: www.eff.org/deeplinks/2015/01/healthcare.gov-sends-personal-data

Ramakrishnan, C. 2014. “Data Leakage Detection in a Multi-Tenant Data Architecture.” *Microsoft Corporation, assignee. Patent US8850596 B2.* September 30, 2014.

Rochet, J. C., and Tirole, J. 2003. “Platform competition in two-sided markets.” *Journal of the European Economic Association* 1(4): 990-1029.

Rolland, E., R.A. Patterson, K.F. Ward. 2009. “Dynamic capabilities and e-service.” *Canadian Journal of Administrative Sciences* 26(4): 301-315.

Rolland, E., R.A. Patterson, K.F. Ward. 2010. “Boundary decision, embeddedness, and the co-creation of value: authors' response to commentary.” *Canadian Journal of Administrative Sciences-Revue* 27(1): 78.

Schoen, S. 2009. “New Cookie Technologies: Harder to See and Remove, Widely Used to Track You.” *Electronic Frontier Foundation.* Retrieved July 10, 2014 from: www.eff.org/deeplinks/2009/09/new-cookie-technologies-harder-see-and-remove-wide

Schwartz, M.J. 2011. “Epsilon fell to spear-phishing attack.” *InformationWeek DARK Reading*. April 11. Accessed online June 14, 2015 from: www.darkreading.com/attacks-and-breaches/epsilon-fell-to-spear-phishing-attack/d/d-id/1097119

Singleton, M. 2016. “The New York Times is testing pop-up ads asking users to disable ad blockers.” The Verge, March 7, 2016, Accessed online March 8, 2016 from: www.theverge.com/2016/3/7/11175250/the-new-york-times-pop-up-ad-blockers-test

Smith, H. J., Dinev, T., and Xu, H. 2011. “Information privacy research: an interdisciplinary review.” *MIS Quarterly* 35(4): 989-1016.

Somaiya, R. 2015a. “New York Times Co. Reports $16 Million Profit.” *New York Times*, Aug 6,

2015, Accessed online March 8, 2016 from: www.nytimes.com/2015/08/07/business/media/new-york-times-co-q2-earnings.html

Somaiya, R. 2015b. "Times Co. Outlines Strategy to Double Digital Revenue." *New York Times*, Oct 7, 2015, Accessed online March 8, 2016 from: www.nytimes.com/2015/10/08/business/media/times-co-outlines-strategy-to-double-digital-revenue.html?_r=0

Sturgeon, T.J., Lee, J-R. 2004. "Industry co-evolution and the rise of a shared supply-base for electronics manufacturing." *ITEC Research Paper Series.*

Tang, C. S. 2006. "Perspectives in supply chain risk management." *International Journal of Production Economics* 103(2): 451-488.

The Register. 2011. "Software maker fingered in Korean hackocalypse." *The Register*. August 12. Accessed online June 14, 2015 from: www.theregister.co.uk/2011/08/12/estsoft_korean_megahack

Tsung, F. 2000. "Impact of information sharing on statistical quality control." *IEEE Transactions on Systems, Man, & Cybernetics - Part A: Systems and Humans* 30(2): 211-216.

Turban, E., N. Bolloju, T.P. Liang. 2011. "Enterprise social networking: Opportunities, adoption, and risk mitigation." *Journal of Organizational Computing and Electronic Commerce* 21(3): 202-220.

Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., and Hennessy, M. 2009. "Americans reject tailored advertising and three activities that enable it." *Available at SSRN 1478214*.

Udemy. 2014. "How Does Facebook Suggest Friends?" Retrieved September 24, 2015 from: blog.udemy.com/how-does-facebook-suggest-friends/.

U.S. Senate, 2014. "Online Advertising and Hidden Hazards to Consumer Security and Data Privacy". *Majority and Minority Staff Report Permanent Subcommittee on Investigations, Committee on Homeland Security and Governmental Affairs.* Released in conjunction with the Permanent Subcommittee On Investigation May 12, 2014 Hearing.

Valente, T.W. 2012. "Network interventions." *Science* 337(6090): 49-53.

Valentino-DeVries, J, Singer-Vine, J, and Soltani, A. 2012. "Websites Vary Prices, Deals Based

on Users' Information." *Wall Street Journal*, December 24, 2012, Retrieved April 9, 2015 from: www.wsj.com/articles/SB10001424127887323777204578189391813881534

Ward, K.F., E. Rolland, R.A. Patterson. 2005. "Improving outpatient health care quality: understanding the quality dimensions." *Health Care Management Review* 30(4): 361-371.

Washington Times. 2015. Retrieved April 4, 2015, from: www.washingtontimes.com

Watts, D.J., J. Peretti. 2007. "Viral marketing for the real world." *Harvard Business Review* (85)5: 22-23.

Wordstream. 2011. "How Does Google Make Its Money: The 20 Most Expensive Keywords in Google AdWords." Accessed online April 5, 2016 from: www.wordstream.com/articles/most-expensive-keywords

Wu, Y., Feng, G., Wang, N., Liang, H. 2015. "Game of information security investment: Impact of attack types and network vulnerability." *Expert Systems with Applications* 42: 6132-3146.

Ye, S., S.F. Wu. 2010. "Measuring message propagation and social influence on Twitter.com." Bolc, L., M. Makowski, A. Wierzbicki, eds, *Social Informatics*. Lecture Notes in Computer Science 6430, Springer, Heidelberg, Germany. 216-231.

Zhou, H., Benton Jr., W.C. 2007. "Supply chain practice and information sharing." *Journal of Operations Management* 25: 1348-1365.

APPENDICES

Appendix 2.1: Theoretical Model of Network Structure-Cost Tradeoff

In this appendix, we provide generalized theoretical results for linear network topologies with both bidirectional and unidirectional propagation flows. First, we consider the linear network topology case with bidirectional propagation with *N* nodes. The propagation spreads in both directions. Figure A.2.1.1 shows the linear network structure for the case with $N = 25$ nodes:

Figure A.2.1.1. Linear Network of Size *N=25* with Bidirectional Propagation Flow

Assume that there is a hop constraint of *H* for reaching out to every node in the network, i.e. there is a message that needs to be sent to every node within *H* hops. In order to achieve this goal, it is possible to seed the message to nodes, or create new connections among the nodes. For any particular combination of *N* and *H*, there are a certain number of non-dominated solutions available that consist of a combination of seeds and new connections. Depending on the cost of seeds and new connections, one or many of these solutions will be optimal in terms of cost of seeding and creating new connections. The following lemma provides the solution combinations and optimal solutions in the general case.

Lemma A.2.1. Bidirectional Propagation

There needs to be at least one seed in the network for a message to propagate. The number of seeds can be between 1 and N. However, because each seed can propagate to $2(H-1)$ nodes on its sides, where $N_S \in \{1, \ldots, \left\lceil \frac{N}{1+2(H-1)} \right\rceil\}$ is the number of seeds in the solution, $N_S^{MinS} = 1$ is the

minimum number of seeds needed, and $N_S^{MaxS} = \left\lceil \frac{N}{1+2(H-1)} \right\rceil$ is the maximum number of seeds needed. For the whole network to be covered by either seeds or connections, if we have less that $\left\lceil \frac{N}{1+2(H-1))} \right\rceil$ number of seeds, there are either zero or $N - N_S(1 + 2(H - 1))$ nodes that are not reached. In order to reach all nodes, we need additional new connections.

Let N_C be the number of new connections in the non-dominated solution. Then:

$$N_C = \left\lceil \frac{Max\{0, N - N_S(1+2(H-1))\}}{1+2(H-2)} \right\rceil \quad \text{(A.2.1.1)}$$

Let C_S and C_C be the cost of seeds and new connections, respectively. The cost of a solution with N_S seeds and N_C new connections is thus calculated as:

$$Cost(N_S, N_C) = N_S C_S + N_C C_C \quad \text{(A.2.1.2)}$$

In order to find the lowest cost solution, we need to solve the following problem:

$$(N_S^*, N_C^*) = ArgMin[Cost(N_S, N_C) = N_S C_S + N_C C_C] \quad \text{(A.2.1.3)}$$

$$s.t.$$

$$N_S \in \{1, \ldots, \left\lceil \frac{N}{1+2(H-1)} \right\rceil\} \quad \text{(A.2.1.4)}$$

$$N_C = \left\lceil \frac{Max\{0, N - N_S(1+2(H-1))\}}{1+2(H-2)} \right\rceil \quad \text{(A.2.1.5)}$$

Solving this problem gives the solution for a given seeding and connection cost. The optimal solution depends on the ratio of the costs (C_S/C_C), and for different values of this ratio, we might have different optimal solutions. We first find all non-dominated solutions of the problem. The following provides the general structure of these solutions:

$$N_S^{MaxS} = \left\lceil \frac{N}{1+2(H-1))} \right\rceil \quad \text{(A.2.1.6)}$$

$$N_S^{k_1} + N_C^{k_1} = B_1 \geq \left\lceil \frac{N}{1+2(H-1))} \right\rceil \qquad \forall k_1 \in Q_1 = \{q_{k_1}, \dots, MaxS - 1\} \quad \text{(A.2.1.7)}$$

$$N_S^{k_2} + N_C^{k_2} = B_2 \geq B_1 \qquad \forall k_2 \in Q_2 = \{q_{k_2}, \dots, q_{k_1} - 1\},\ q_{k_2} < q_{k_1} \quad \text{(A.2.1.8)}$$

…

$$N_S^{k_r} + N_C^{k_r} = B_r \geq B_{r-1} \qquad \forall k_r \in Q_r = \{q_{k_r}, \dots, q_{k_{r-1}} - 1\},\ q_{k_r} < q_{k_{r-1}} \quad \text{(A.2.1.9)}$$

…

$$N_S^{k_m} + N_C^{k_m} = B_m \geq B_{m-1} \quad \forall k_m \in Q_m = \{MinS, \dots, q_{k_{m-1}} - 1\},\ MinS < q_{k_{m-1}} \quad \text{(A.2.1.10)}$$

where:

k_r is the index of number of seeds and new connections for the set of solutions with total of B_r number of seeds and new connections.

$MaxS$ is the index for number of seeds and number of new connections of the solution with maximum number of seeds.

$MinS$ is the index for number of seeds and number of new connections of the solution with minimum number of seeds.

B_r is the total of the number of seeds and new connections in each solution set Q_r.

Q_r is the set containing the indexes for all solutions with total of B_r number of seeds and new connections.

q_{k_r} is the solution with least number of seeds in the set of solutions with the total of the number of seeds and new connections of B_r, that is the set $Q_r = \{q_{k_r}, \dots, q_{k_{r-1}} - 1\}$, $q_{k_o} = MaxS$, and $q_{k_m} = MinS$.

So there are m types of solutions available in this case. The optimal solution space is as follows:

$$if \ \frac{C_S}{C_C} < \frac{\left\lceil \frac{Max\{0, N - N_S^{k_1}(1+2(H-1))\}}{1+2(H-2)} \right\rceil}{\left\lceil \frac{N}{1+2(H-1)} \right\rceil - N_S^{k_1}} \xrightarrow{Optimal\ Solution} \left(N_S^{MaxS}, N_C^{MinS} = 0\right) \quad \text{(A.2.1.11)}$$

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_1}(1+2(H-1))\}}{1+2(H-2)} \right\rceil}{\left\lceil \frac{N}{1+2(H-1)} \right\rceil - N_S^{k_1}} < \frac{C_S}{C_C}$$

$$< \frac{\left\lceil \frac{Max\{0, N - N_S^{k_2}(1+2(H-1))\}}{1+2(H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_1}(1+2(H-1))\}}{1+2(H-2)} \right\rceil}{N_S^{k_1} - N_S^{k_2}}$$

$$\xrightarrow{Optimal\ Solution} (N_S^{q_{k_1}}, N_C^{q_{k_1}}) \quad \text{(A.2.1.12)}$$

…

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_r}(1+2(H-1))\}}{1+2(H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_{r-1}}(1+2(H-1))\}}{1+2(H-2)} \right\rceil}{N_S^{k_{r-1}} - N_S^{k_r}} < \frac{C_S}{C_C}$$

$$< \frac{\left\lceil \frac{Max\{0, N - N_S^{k_{r+1}}(1+2(H-1))\}}{1+2(H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_r}(1+2(H-1))\}}{1+2(H-2)} \right\rceil}{N_S^{k_r} - N_S^{k_{r+1}}}$$

$$\xrightarrow{Optimal\ Solution} (N_S^{q_{k_r}}, N_C^{q_{k_r}}) \quad \text{(A.2.1.13)}$$

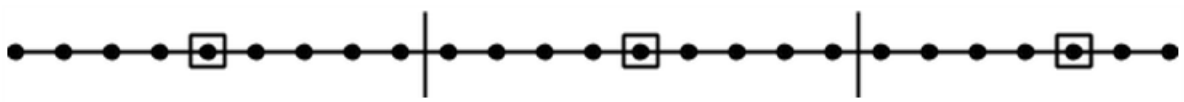
…

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_{m-1}}(1 + 2(H-1))\}}{1 + 2(H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{MinS}(1 + 2(H-1))\}}{1 + 2(H-2)} \right\rceil}{N_S^{MinS} - N_S^{k_{m-1}}} < \frac{C_S}{C_C}$$

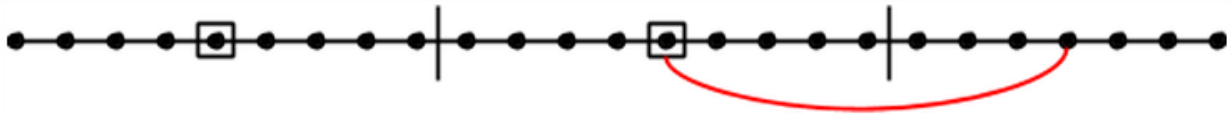
$$\xrightarrow{Optimal\ Solution} (N_S^{MinS} = 1, N_C^{MinS}) \qquad \text{(A.2.1.14)}$$

We next provide an illustrative example of the above lemma. Let's consider the case where $N = 25$ and $H = 5$. These three non-dominated solutions can be represented as follows:

A. 3 seeds and zero new connections:



B. 2 seeds and 1 new connection:



C. 1 seed and 3 new connections:

According to the lemma, we have $N_S^{MaxS} = N_S^3 = \left\lceil \frac{25}{1+2(5-1)} \right\rceil = 3$ and for the next set of problems, we have $N_S^{k_1} = 2$ and $N_C^{k_1} = 1$; and $N_S^{k_2} = 1$ and $N_C^{k_2} = 3$.

We can now analyze this problem by divide the solution space into four parts based on the C_S/C_C ratio. We have:

$$N_S^{MaxS} = N_S^3 = \left\lceil \frac{25}{1+2(5-1)} \right\rceil = 3$$

$$N_S^{k_1} + N_C^{k_1} = 3 = 3 \qquad \forall k_1 \in Q_1 = \{2\}$$

$$N_S^{k_2} + N_C^{k_2} = 4 > 3 \qquad \forall k_2 \in Q_2 = \{1\}, \qquad 1 < 2$$

So, based on the lemma, the optimal solution is as follows:

$$if \;\; \frac{C_S}{C_C} < \frac{\left\lceil \frac{Max\{0,25-2(1+2(5-1))\}}{1+2(5-2)} \right\rceil}{\left\lceil \frac{25}{1+2(5-1)} \right\rceil - 2} = 1 \xrightarrow{Optimal\ Solution} \left(N_S^{MaxS}, N_C^{MinS}\right) = (3,0)$$

$$if \;\; 1 < \frac{C_S}{C_C} < \frac{\left\lceil \frac{Max\{0,25-1(1+2(5-1))\}}{1+2(5-2)} \right\rceil - \left\lceil \frac{Max\{0,25-2(1+2(5-1))\}}{1+2(5-2)} \right\rceil}{2-1} = 2$$

$$\xrightarrow{Optimal\ Solution} \left(N_S^{k_1}, N_C^{k_1}\right) = (2,1)$$

$$if \;\; 2 < \frac{C_S}{C_C} \xrightarrow{Optimal\ Solution} \left(N_S^{MinS}, N_C^{MinS}\right) = (1,3)$$

In Figure A.2.1.2, the optimal solution is presented with respect to the cost ratio between seeds and connections. As is shown, there is a range of relative costs over which various solutions are optimal.

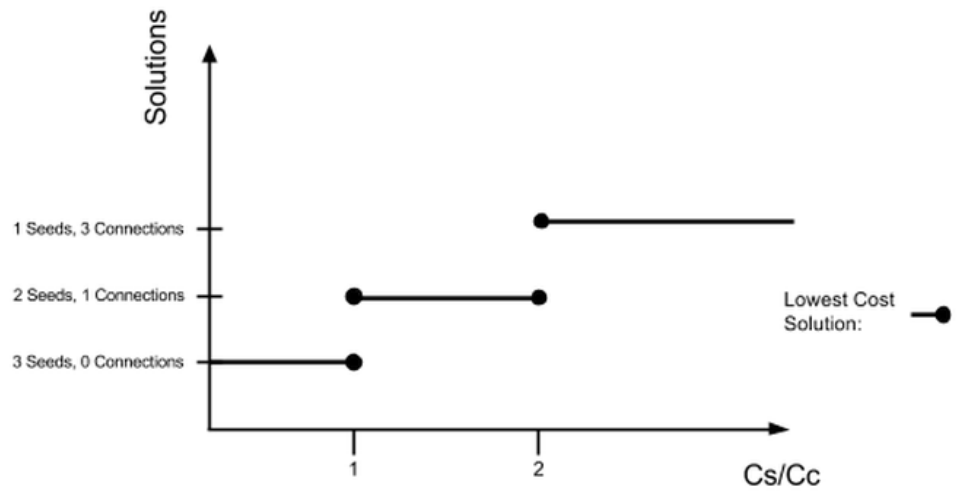


Figure A.2.1.2. Optimal Solution Structure with Respect to $\frac{C_S}{C_C}$ for a Linear Network of Size *N=25*

Next we consider the linear network topology case with unidirectional propagation with N nodes. The propagation spreads from left to right. Figure A.2.1.3 shows the network connections for the case with *N = 25* nodes:

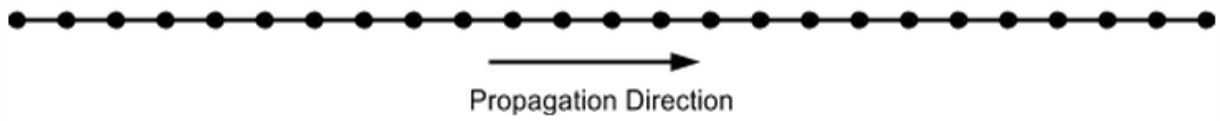


Figure A.2.1.3. Linear Network of Size *N=25* with Unidirectional Propagation Flow

Lemma A.2.1.2. Unidirectional Propagation

This case is similar to the bidirectional lemma, with the difference that each seed can propagate to $H-1$ nodes (instead of $2(H-1)$ nodes) and each new connection can propagate to $H-2$ nodes (instead of $2(H-2)$ nodes). In this case the lowest cost solution is answer to the following problem:

$$(N_S^*, N_C^*) = ArgMin[Cost(N_S, N_C) = N_S C_S + N_C C_C] \quad \text{(A.2.1.15)}$$

$s.t.$

$$N_S \in \{1, \dots, \left\lceil \frac{N}{1+2(H-1)} \right\rceil\} \quad \text{(A.2.1.16)}$$

$$N_C = \left\lceil \frac{Max\{0, N - N_S(1+(H-1))\}}{1+(H-2)} \right\rceil \quad \text{(A.2.1.17)}$$

The following provides the general structure of these solutions:

$$N_S^{MaxS} = \left\lceil \frac{N}{1+(H-1))} \right\rceil \quad \text{(A.2.1.18)}$$

$$N_S^{k_1} + N_C^{k_1} = B_1 \geq \left\lceil \frac{N}{1+(H-1))} \right\rceil \qquad \forall k_1 \in Q_1 = \{q_{k_1}, \dots, MaxS - 1\}$$

$$N_S^{k_2} + N_C^{k_2} = B_2 \geq B_1 \qquad \forall k_2 \in Q_2 = \{q_{k_2}, \dots, q_{k_1} - 1\},\ q_{k_2} < q_{k_1} \quad \text{(A.2.1.19)}$$

…

$$N_S^{k_r} + N_C^{k_r} = B_r \geq B_{r-1} \qquad \forall k_r \in Q_r = \{q_{k_r}, \dots, q_{k_{r-1}} - 1\},\ q_{k_r} < q_{k_{r-1}} \quad \text{(A.2.1.20)}$$

…

$$N_S^{k_m} + N_C^{k_m} = B_m \geq B_{m-1} \quad \forall k_m \in Q_m = \{MinS, \dots, q_{k_{m-1}} - 1\},\ MinS < q_{k_{m-1}} \quad \text{(A.2.1.21)}$$

So again in this case, there are m types of solutions available in this case. The optimal solution space is as follows:

$$if\ \ \frac{C_S}{C_C} < \frac{\left\lceil \frac{Max\left\{0, N - N_S^{k_1}(1+(H-1))\right\}}{1+(H-2)} \right\rceil}{\left\lceil \frac{N}{1+(H-1)} \right\rceil - N_S^{k_1}} \xrightarrow{Optimal\ Solution} (N_S^{MaxS}, N_C^{MinS} = 0) \quad \text{(A.2.1.22)}$$

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_1}(1 + (H-1))\}}{1 + (H-2)} \right\rceil}{\left\lceil \frac{N}{1 + (H-1)} \right\rceil - N_S^{k_1}} < \frac{C_S}{C_C}$$

$$< \frac{\left\lceil \frac{Max\{0, N - N_S^{k_2}(1 + (H-1))\}}{1 + (H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_1}(1 + (H-1))\}}{1 + (H-2)} \right\rceil}{N_S^{k_1} - N_S^{k_2}}$$

$$\xrightarrow{Optimal\ Solution} (N_S^{q_{k_1}}, N_C^{q_{k_1}}) \qquad \text{(A.2.1.23)}$$

…

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_r}(1 + (H-1))\}}{1 + (H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_{r-1}}(1 + (H-1))\}}{1 + (H-2)} \right\rceil}{N_S^{k_{r-1}} - N_S^{k_r}} < \frac{C_S}{C_C}$$

$$< \frac{\left\lceil \frac{Max\{0, N - N_S^{k_{r+1}}(1 + (H-1))\}}{1 + (H-2)} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{k_r}(1 + (H-1))\}}{1 + (H-2)} \right\rceil}{N_S^{k_r} - N_S^{k_{r+1}}}$$

$$\xrightarrow{Optimal\ Solution} (N_S^{q_{k_r}}, N_C^{q_{k_r}}) \qquad \text{(A.2.1.24)}$$

…

$$if \ \frac{\left\lceil \frac{Max\{0, N - N_S^{k_{m-1}}(1 + (H-1))\}}{1 + (H-2))} \right\rceil - \left\lceil \frac{Max\{0, N - N_S^{MinS}(1 + (H-1))\}}{1 + (H-2))} \right\rceil}{N_S^{MinS} - N_S^{k_{m-1}}} < \frac{C_S}{C_C}$$

$$\xrightarrow{Optimal\ Solution} (N_S^{MinS} = 1, N_C^{MinS}) \qquad \text{(A.2.1.25)}$$

The structure of the optimal solution in the lemmas provides that the lowest cost solutions that can propagate to all nodes within H hops can be either at the extreme solutions of all seeds

or one seed and all new connections, or at an intermediate solution with a combination of seeds and new connections. The optimal solution depends on the ratio of seeding cost to new connection cost.

We know that both propagation performance and cost of network manipulations are important to the network administrator, and the profit of the network depends on them:

$$\Pi = \gamma * Propagation\ Performance - Costs$$

Where γ is a parameter that determines the relative importance of propagation performance with respect to costs. Assume that the propagation performance measure is the probability of reaching all the nodes within H hops. Moreover, assume that the probability of the seeding, connections, and new connections are all equal to a probability $0 \leq P \leq 1$. Then the propagation performance measure is equal for all different solutions and is equal to P^N. In other words, the propagation performance of all these solutions is equal to each other, and the only difference is in the costs of these solutions.

While these lemmas are defined for very simple linear networks, there are some interesting findings in them. The lemmas state that the optimal solution can contain any number of seeds and new connections, depending on the trade-off between cost of seeding and cost of new connections, and the optimal solution is not always an extreme point.

We next discuss how the optimal solution structure changes in more complicated linear networks, where the linear communication structure is broken in one location. Figure A.2.1.4 figure provides a network with a disconnect at the last node:

Figure A.2.1.4. Linear Network of Size *N=25* with a Disconnect

Such a disconnected network has lower propagation performance than the fully connected network. The number of seeds and new connections in the non-dominated solution set, with the same propagation performance objective of reaching all nodes within H hops, is greater than or equal to the fully connected linear networks. In other words, a disconnected network never decreases the cost of creating a network, while maintaining the same propagation performance objective.

We have analyzed the effect of having a disconnect in the network at each of the 24 different locations in the 25 node linear network for both bidirectional and unidirectional propagation flow. The location of the disconnect does affect the cost structure of the optimal solution. However, there is no identifiable trend as to how the disconnect location affects the cost, and this must be analyzed case-by-case. There are cases where the disconnection results in the same solution as the fully connected linear case. We can make a series of generalized theoretical observations that are, for the most part, stating the obvious:

Observation A.2.1.1 *Having more starting connections and/or more seeds cannot increase cost of creating a network where message propagation occurs within H hops.*

Observation A.2.1.2 *Having fewer starting connections and/or fewer seeds cannot decrease the cost of creating a network where message propagation occurs within H hops.*

Observation A.2.1.3 *The lowest cost solution for propagating a message to all nodes within H hops is dependent on the exact configuration of the starting network topological design and the relative costs of creating additional seeds versus additional connections.*

Observation A.2.1.4 *For any given network structure, the lowest cost solution for propagating a message to all nodes within H hops may, depending on the relative cost of seeds versus connections, consist of a) all seeds, b) one seed and all connections, or c) some combination of multiple seeds and one or more connections. Additionally, for some network structures, only the extreme points of a) all seeds, or b) one seed and all connections, exist as non-dominated lowest cost solutions.*

Appendix 2.2: Initial Solution Algorithms

In this appendix we provide further explanation on the three initial solution techniques described in Section 2.4.

A.2.2.1 Prim

This technique is simply the Prim's algorithm for minimum spanning tree, where a constraint is added so that the number of hops from the root node to any node is not higher than a given number. The flowchart for this algorithm is as described in Figure A.2.2.1.

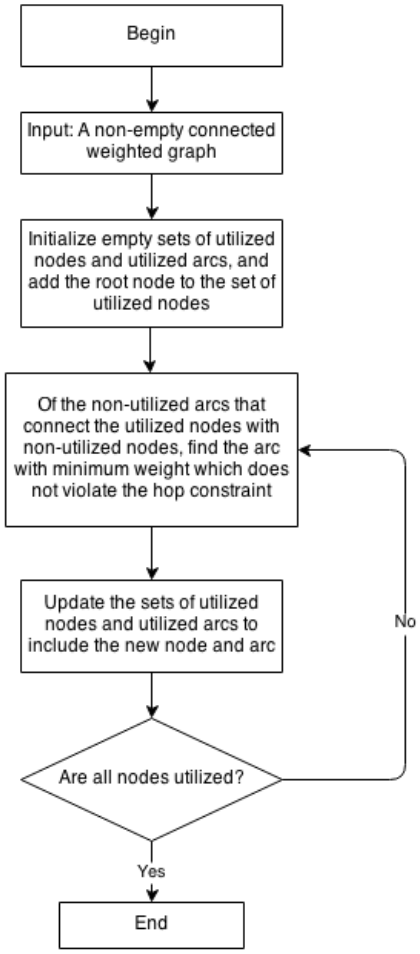


Figure A.2.2.1. Prim's Algorithm for HMST

The only difference for the case of HMST as compared to the MST is addition of the constraint on hops. We can see that this is a very crude algorithm, and does not search further in the space to find possibly better solutions, but provides a very fast solution.

A.2.2.2 FGV

FGV is the heuristic proposed by Fernandes, Gouveia, and Voß (2007) called ILA for the HMST problem, which utilizes the Esau-Williams' (EW) algorithm. The EW algorithm works as described in Figure A.2.2.2 below.

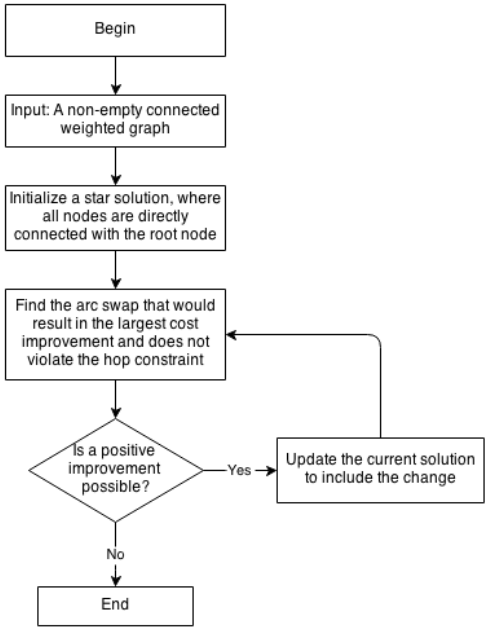


Figure A.2.2.2. EW Algorithm for HMST

A flowchart of steps of the FGV heuristic is provided in Figure A.2.2.3. The FGV technique starts with the Esau-Williams' (EW) heuristic. EW is an improvement heuristic which starting from a given solution, searches for changes that can be made to the tree so that the cost can be reduced. The FGV then proceeds by iteratively applying the EW heuristic to a subset of

the graph, which is created by inhibiting a set of arcs that can be used. If a better solution is found by using the inhibited arcs, then these arcs are permanently inhibited and the heuristic proceeds to next candidates. In order to reduce the amount of computational time, a pilot method is used for finding the candidate arcs in a directed manner.

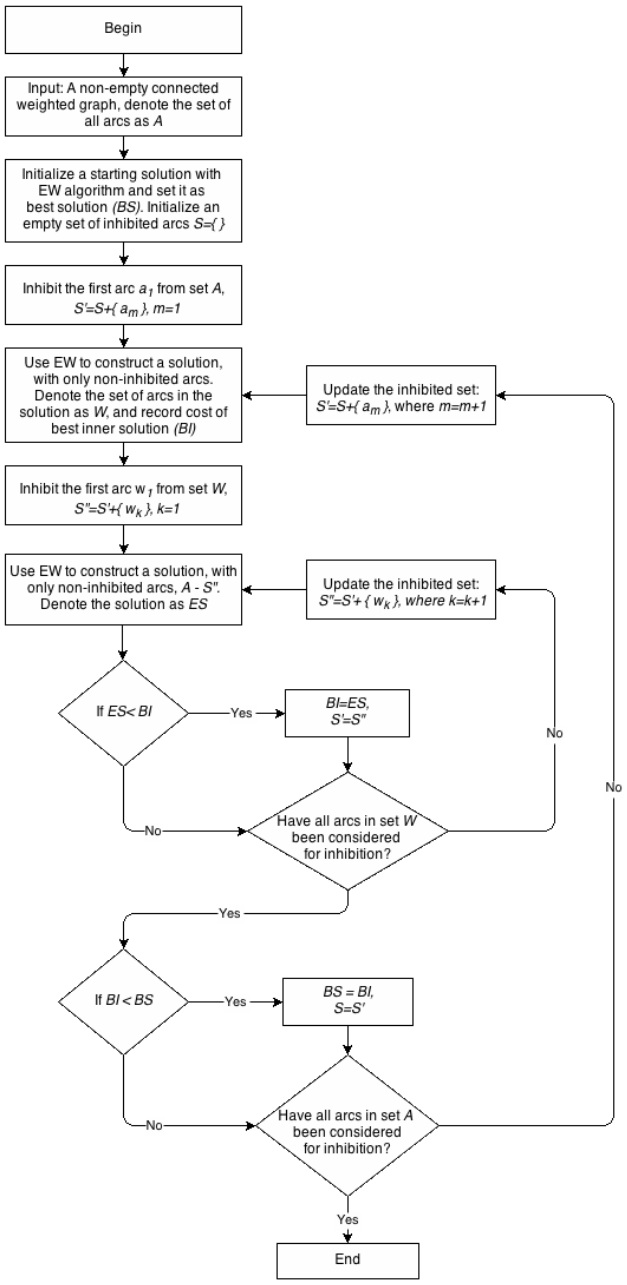


Figure A.2.2.3. FGV Algorithm for HMST

A.2.2.3 Akgun

We use Akgun's (2011) formulation to solve the HMST problem in CPLEX solver. We use the following set, decision variable, and data parameter notations.

Set definitions:

$$V = \{a\ set\ of\ nodes\ V | i, j \in V,$$
$$where\ i, j = 1, .., N, N = |V|, and\ i = 1\ is\ the\ root\ node\}$$

$$A = \{a\ set\ of\ arcs\ A | (i, j) \in A, where\ i, j \in V\}$$

Decision variable definitions:

$$X_{ij} = \begin{cases} 1 & if\ arc\ (i,j) is\ used\ in\ the\ solution \\ 0 & otherwise \end{cases}$$

$$u_i = hop\ number\ for\ node\ i, where\ u_i \in [0, N]$$

Data parameter definition:

$$C_{ij} = cost\ of\ using\ arc\ (i, j)$$

$$b_{ij} = \begin{cases} 1 & if\ arc\ (i,j)\ already\ exists \\ 0 & otherwise \end{cases}$$

$$H: Maximum\ allowable\ number\ of\ hops, where\ H \in [1, N]$$

Here, we repeat the HMST problem defined in the chapter, and then provide the formulation proposed by Akgun (2011). Let $G = (V, A)$ be a directed network with node set $V = \{1,2,\ldots,n\}$ where node is 1 defined as the root node, and A is the set of directed arcs (i, j) connecting all nodes in V. A positive arc cost C_{ij} is associated with each directed arc. Some arcs

already exist, and so there is no cost for using them in the solution. For these arcs we have $b_{ij} = 1$. Note that the data parameter b_{ij} is added for clarity, but could be omitted by adjusting C_{ij} to zero when $b_{ij} = 1$. The problem is to find the HMST in the network. The HMST is a minimum cost spanning tree with each node being at most *H* hops away from the root node.

To formulate the problem, binary variable X_{ij} is defined for each arc, denoting whether the arc (i, j) is used in the solution or not. The objective is to minimize the total cost of using the arcs in the solution, so the objective function is defined as:

$$z^* = min \sum_{(i,j)\in A, i\neq 1} C_{ij} X_{ij} (1 - b_{ij}) \quad \text{(A.2.2.1)}$$

To address the hop constraint and eliminate the possibility of a sub-tour in the solution, a hop number variable u_i is associated with each node *i*. The hop number variable determines the distance from each node to the root node in terms of number of hops. Hop number for the root node (node 1) is set to zero. The constraints from Akgun (2011) are defined as:

$$u_1 \equiv 0 \quad \text{(A.2.2.2)}$$

$$u_i \geq 1, \quad \forall\, i \in V, i \neq 1 \quad \text{(A.2.2.3)}$$

$$\sum_{j\in V, j\neq i} X_{ij} = 1, \quad \forall\, j \in V, i \neq 1 \quad \text{(A.2.2.4)}$$

$$u_i - u_j + NX_{ij} \leq N - 1, \quad \forall (i, j) \in A, j \neq 1, i \neq j \quad \text{(A.2.2.5)}$$

$$u_i \leq H, \quad \forall\, i \in V, i \neq 1 \quad \text{(A.2.2.6)}$$

$$X_{ij} \in \{0,1\}, \quad \forall\, (i, j) \in A \quad \text{(A.2.2.7)}$$

$$u_i \geq 0, \quad \forall\, i \in V \quad \text{(A.2.2.8)}$$

Constraints A.2.2.2 and A.2.2.3 determine that the hop number for the root node is set to zero and that other nodes' hop numbers are equal to or higher than 1, respectively. Constraint A.2.2.4 ensures that each node has one and only one outgoing arc, and constraint A.2.2.5 is the sub-tour elimination constraint. Finally, constraint A.2.2.6 is the hop constraint. Constraint A.2.2.7 is the binary constraint on X_{ij} and constraint A.2.2.8 is non-negativity constraint for the hop number.

It is known that Akgun's (2011) formulation is able to find good solutions in short amount of time for small to medium size problems, but finding the optimal solution even in medium size problems is not possible in a timely manner. However, this formulation is useful for finding "good feasible solutions in a short time" (Akgun 2011). A time limit of one hour is set for the solver when using Akgun's formulation. Figure A.2.2.4 shows the flowchart for this initial solution method.

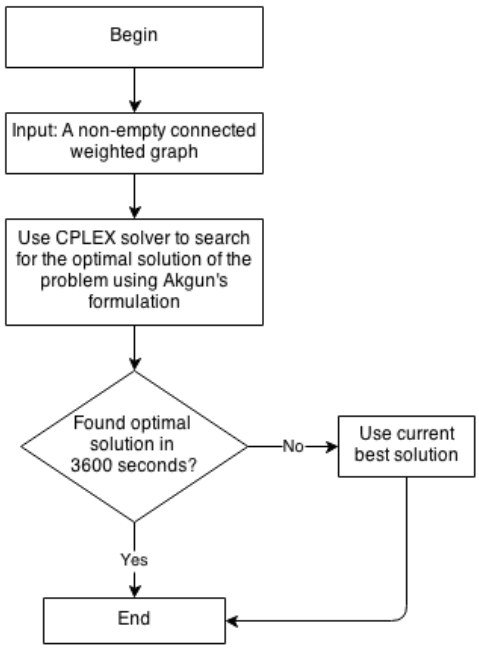


Figure A.2.2.4. Akgun Algorithm for HMST

Appendix 2.3: Propagation Model

The setting for the propagation model is as follows:

1. Seed (initially active) nodes: For the seed nodes, we use the nodes that are directly connected to the root node (node 1) in our HMST solution.
2. Propagation Probabilities: P_{ij} shows the probability of an activated node *i* activating a non-active node *j* in a single step, determined by the connection type:

   Arc type Q: $P_{ij} = P_Q + (1 - P_Q) * rand$

   where P_Q is the parameter for the existing connections, and rand is a random number between zero and one.

   Arc type L (we test the three following distributions for probabilities of type *L*):

   i. $P_{ij} = (C_{max} - C_{ij})/(C_{max} - C_{min}) * P_L$

   where C_{ij} is the connection cost between two nodes, C_{min} and C_{max} are the minimum and maximum of all connection costs respectively and P_L is a parameter of the model.

   ii. Let R_i be a set of $|R_i|$ randomly chosen nodes from a uniform distribution of the set S_t, where S_t is the set of non-active nodes at time step *t* and $R_i \subseteq S_t$, $S_t \subseteq V$, and $|R_i| \sim$ Poisson with rate λ. Node *i* will activate all nodes in set R_i. In this case P_{ij} at time step *t* is calculated as follows:

$$P_{ij} = \sum_{\beta=0}^{|S_t|} p(|R_i| = \beta) * p(activate\ node\ j \mid |R_i| = \beta)$$

where $p(activate\ node\ j\ ||R_i| = \beta)$ is the conditional probability that node *j* is activated given that there are β number of nodes to be activated in that step, and is calculated as:

$$p(activate\ node\ j\ |\ |R_i| = \beta) = \frac{\binom{|S_t|-1}{\beta-1}}{\binom{|S_t|}{\beta}} = \frac{\beta}{|S_t|}$$

where $\beta = 0,1, \ldots, |S_t|$. The first part of the argument in formula for P_{ij} provides the probability that node *i* activates β nodes. Because $|R_i| \sim$ Poisson with rate λ, this can be calculated as:

$$p(|R_i| = \beta) = \frac{\lambda^\beta e^{-\lambda}}{\beta!}$$

where $\beta = 0,1, \ldots, |S_t|$.

So for each time step *t* we have:

$$P_{ij} = \sum_{\beta=0}^{|S_t|} (\frac{\lambda^\beta e^{-\lambda}}{\beta!} * \frac{\beta}{|S_t|})$$

iii. Same as for the Poisson distribution, except that in this case $|R_i|$ is distributed as the floor (largest previous integer) of a Power-law probability distribution with lower bound $y_{min} = 1$ and rate $\alpha = 5$. In this case, the first part of the argument is defined as:

$$p(|R_i| = \beta) = \int_{y=\beta+1}^{y=\beta+2} (\frac{\alpha-1}{y_{min}})(\frac{y}{y_{min}})^{-\alpha} dy$$

where $\beta = 0,1, \ldots, |S_t|$. So for each time step *t* in this case we have:

$$P_{ij} = \sum_{\beta=0}^{|S_t|}((\int_{y=\beta+1}^{y=\beta+2} \left(\frac{\alpha-1}{y_{min}}\right)\left(\frac{y}{y_{min}}\right)^{-\alpha} dy) * \frac{\beta}{|S_t|})$$

Arc type M: $P_{ij} = P_M + (1 - P_M) * rand$

where P_M is a model parameter.

For the model parameters we use $P_Q = 0.9$ and $P_M = 0.8$. Based on a pilot study using HMST solutions for all EC and perturbed-EC problems, we find that the distribution of propagation probability for type *L* arcs does not affect the overall CIT propagation behavior of the model as long as the parameters are within an acceptable range. Thus, we present the results for this type of arcs using a Poisson distribution with rate of $\lambda = 0.01$.

Appendix 2.4: The Greedy Seeding Algorithm

We use a greedy algorithm to intelligently select initial seeds for propagation. This greedy algorithm performs as a benchmark for showing the effects of network enhancements on propagation. In the influence maximization literature, some researchers have formulated the problem of finding initial seeds that maximize propagation for different types of propagation models (Domingos and Richardson 2001; Kempe et al. 2003; Kimura et al. 2009 and Even-Dar and Shapira 2011). The influence maximization problem is applicable in case of networks that are not fully connected. In these networks, complete influence is not guaranteed when propagation starts with any set of initial nodes and each set of initial nodes will result into a certain penetration in the network. In the problems discussed in this chapter, because of the characteristics of the propagation model, the penetration will always be 100% if sufficient time is given to propagate. In other words, "given any nonempty initial target set of nodes, all nodes in the network are ultimately influenced, i.e., the complete influence is achieved" (Ni et al. 2010). Ni et al. (2010) propose another approach based on minimization of expected CIT that is more relevant to our problem. The idea behind this greedy algorithm is to find nodes that minimize the expected CIT and add them to the set of initial seeds one-by-one. So our algorithm relies on simulations of propagation to find the lowest expected CIT for the sets. The algorithm works as follows: for finding a set of initial nodes with size k we incrementally build the set by adding nodes that give the best expected CIT one at the time. This expected CIT is calculated by averaging over several propagation simulation runs. Note that the efficiency and effectiveness of this algorithm relies on the number of propagation runs and thus there is a trade-off between computational time and quality of solution. The addition of the nodes continues until we have all k number of nodes in the set.

The greedy algorithm is equivalent to the algorithm provided by Ni et al. (2010) where the target set includes all the remaining nodes. Because of this, there is no need to use a heuristic for finding the nodes to be put in the target set as their algorithm does. Another difference is in the "distance network" that should be created out of the original network that is based on a weighting method. We have used the same cost structure introduced in Section 2.6 as the weights of the distance network. This type of weighting follows the general purpose of the weighting method in Ni et al. (2010) where costs are a measure of distance between the nodes. Note that the costs in our problems also represent the social distance between the nodes in the network.

Appendix 2.5: The Problem Generation Procedure

A.2.5.1 Enron Work E-mail Networks

The Enron data includes all of the e-mail communications that exist between employees. Cumulative pairwise numbers of e-mails are counted at various points in time. The number of e-mails sent is considered the tie strength, T_{ij} being the tie strength of the connection (i, j). The costs for using the links is calculated as follows:

$$C_{ij} = 1 - \frac{Max[MaxTies, T_{ij}]}{MaxTies}$$

where $MaxTies$ is a parameter. This implies that the cost of using a connection with a tie strength of higher than $MaxTies$ is zero. The parameter is set to 100 in our calculations. The seeding cost is calculated similar to the Facebook and Twitter ego networks, that is:

$$C_{i1} = ln\ (n)$$

A.2.5.2 Twitter and Facebook Ego-networks

Twitter and Facebook data consists of directed ego-networks (EN) with data from their circles. The EN represents an existing network for which the cost of using those connections are assumed to be zero. The cost of the other arcs are calculated according to the number of mutual connections, and cost of connecting directly to the root node is set to *ln(N)*, the natural log of the size of the network.

There has been some research on estimation of connection benefits in the employee productivity literature. Hoffmann (2010) and Wu et al. (2009) have associated higher e-mail connectivity with higher financial productivity. While this line of research focuses on how new

connections improve network performance, it is silent on the cost to create a new connection. We found no previous research exploring the cost of creating new connections in an existing social network. Because the data does not include information on closeness of disconnected users, we use the concept of mutual connections as the proxy for the cost of connecting two nodes. The assumption here is that the more mutual connections two nodes have, the easier it is to create a connection between them and thus the lower the cost. The cost of creating new connections includes the effort needed for the administrator in order to create a new connection between a pair of users. The cost can also be attributed to the acceptance probability of a new connection suggestion, that is, the cost of creating a connection when there is high acceptance probability is lower than when there is low acceptance probability. In the social recommender systems literature, there is a body of literature that studies the factors that affect the success rate of such friend suggestions. For example, Moricz, Dosbayev and Berlyant (2010) study the friend recommendation system in MySpace. They find that "closeness of users' friend networks, common user interest with friends and recommended friend(s), [and] common geographical region" are the significant factors that improve the conversation between recommended users.

The set of nodes and directed arcs between nodes *i* and *j* are given in the dataset. We use the circles to select our subset of nodes and existing arcs for the HMST testing. A new node 1, the root node, is added to the set of. The existing arcs are identified by a binary variable b_{ij}, where $b_{ij} = 1$ if the arc from node *i* to *j* exists in the dataset, and $b_{ij} = 0$ otherwise. We have:

$C_{ij,i\neq 1,j\neq 1} = 0$ if $b_{ij} = 1$

Next, the directional cost of all arcs where $b_{ij} = 0$ is calculated. Let k be the number of mutual incoming or outgoing connections (excluding node 1) shared by nodes *i* and *j*, such that for all nodes $\rho \in V, where\ \rho \neq 1, \rho \neq i, \rho \neq j$:

if ($b_{i\rho} = 1$ or $b_{\rho i} = 1$) and ($b_{j\rho} = 1$ or $b_{\rho j} = 1$), then $f = f + 1$.

Otherwise, $f = 0$.

Thus, the cost of non-existing arcs is calculated as follows:

$C_{ij,i\neq1,j\neq1} = \frac{1}{f+e^{b_{ji}}} \quad \forall b_{ij} = 0, \quad$ where $e = 2.71828$.

The cost of connecting each node i directly with node 1 is:

$C_{i1} = ln\ (n)$

Because node 1 is the root node, C_{1j} does not exist.

A.2.5.3 Euclidean with Root Node in the Center (EC)

We utilize randomly generated problems to test our improvement heuristic. The problem sets are created by placing *n-1* nodes randomly in a 100 by 100 space. Node 1, the root node, is placed in the center of the space at location (50, 50). The cost for each arc between any two given nodes is calculated as the Euclidean distance between the two nodes.

A.2.5.4 Perturbed Euclidean with Root Node in the Center (Perturbed-EC)

In this case a similar setting to previous section is created. We then randomly create the arc costs in a two-step process. In the first step, the number of outgoing connections for each node $i \neq 1$, o_i, is randomly chosen integer from a uniform distribution in range $[0, n/10]$. Node *i* is

connected to the nearest o_i nodes as measured in Euclidean distance D_{ij}, where D_{ij} is the Euclidean distance between the two nodes *i* and *j*. The existing arcs are identified by a binary variable b_{ij}, where

$b_{ij} = 1$ if the arc from node *i* to *j* exists, and $b_{ij} = 0$ otherwise.

For these existing arcs, the cost is set to zero:

$C_{ij,i\neq 1,j\neq 1} = 0$ if $b_{ij} = 1$

In this case, the costs of non-existing connections (where $b_{ij} = 0$) are altered as follows. Let f be the number of mutual incoming or outgoing connections (excluding node 1) shared by nodes *i* and *j*, such that for all nodes $\rho \in V, \quad where\ \rho \neq 1, \rho \neq i, \rho \neq j$:

if ($b_{i\rho} = 1$ or $b_{\rho i} = 1$) and ($b_{j\rho} = 1$ or $b_{\rho j} = 1$), then $f = f + 1$.

Otherwise, $f = 0$.

Thus, the cost of non-existing arcs is calculated as follows:

$C_{ij,i\neq 1,j\neq 1} = \frac{1}{f+e^{b_{ji}}} * D_{ij} \quad \forall b_{ij} = 0$, where $e = 2.71828$ and $D_{ij}$is the Euclidean distance between the two nodes.

The cost of connecting each node i directly with node 1 is:

$C_{i1} = ln\ (n) * D_{ij}$

Because node 1 is the root node, C_{1j} does not exist. All the cost data are truncated to the third decimal point.

Appendix 2.6: Improved Enron Networks

We have analytically shown in Section 2.3 and Appendix 2.1 that different seeding to new connection cost will result in different optimal number of seeds and connections in a specific linear network. Here we discuss the same issue in an Enron network example. Based on Enron's e-mail activity, Figure A.2.6.1 provides a snapshot of existing Enron ESN of size 111 with the unconnected root node (node 1) in the center. This network depicts the internal e-mail activity of Enron from May 11, 1999 to December 10, 2000. We then use the proposed methodology to create an HMST in the network.

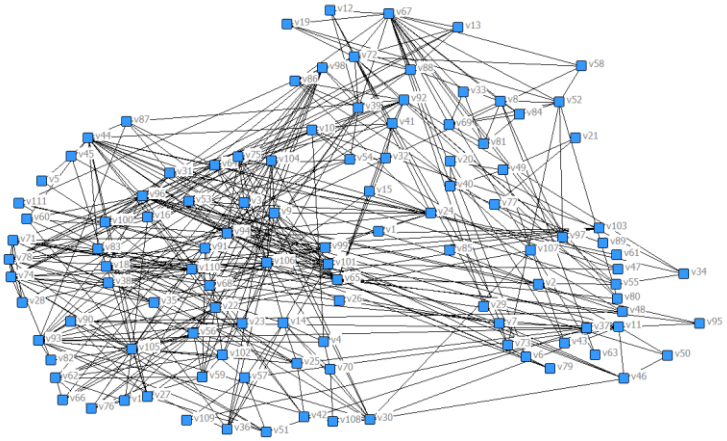


Figure A.2.6.1. An Enron Network Example of Size $N = 111$

Figure A.2.6.2 provides the same network, with added seeds and new connections in it, showing only the connections used in the HMST solution with hop of $H = 5$.

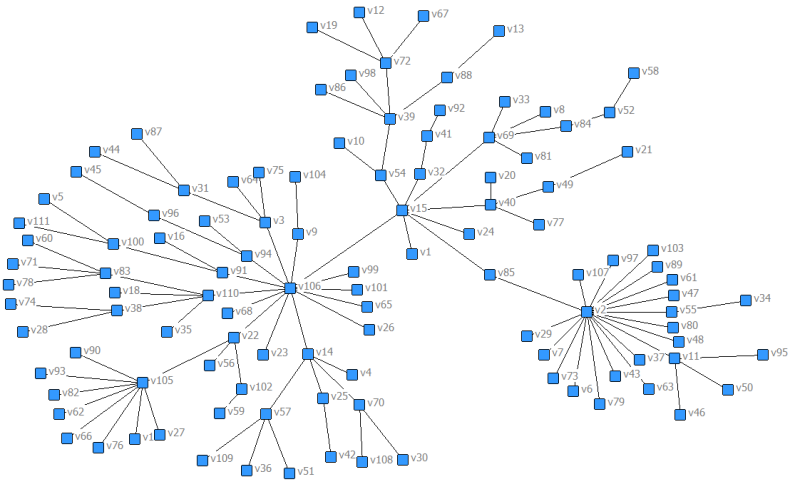


Figure A.2.6.2. Improved Enron Network of Size *N=111, H=5*

It can be seen that in the above network, all nodes can be reached from the root node within $H = 5$ hops for the HMST solution. Note that the specific network improvement is dependent on the seeding to new connection cost ratio, as is expected from our theoretical results. In this example, we have considered the seeding cost to be high. This is the case where the direct seeding might not be an effective approach for propagation, for example in the case of organizational culture, peers can have a better influence on each other rather than corporate emails. Due to high seeding costs in this example, there is only one seed, but many new connections. Next we consider different scenarios that may arise based on the different cost structures for seeding and new connection creation.

Figure A.2.6.3 provides an initial Enron network snapshot with size of $N = 59$. This network depicts the internal e-mail activity of Enron from May 11, 1999 to February 24, 2000.

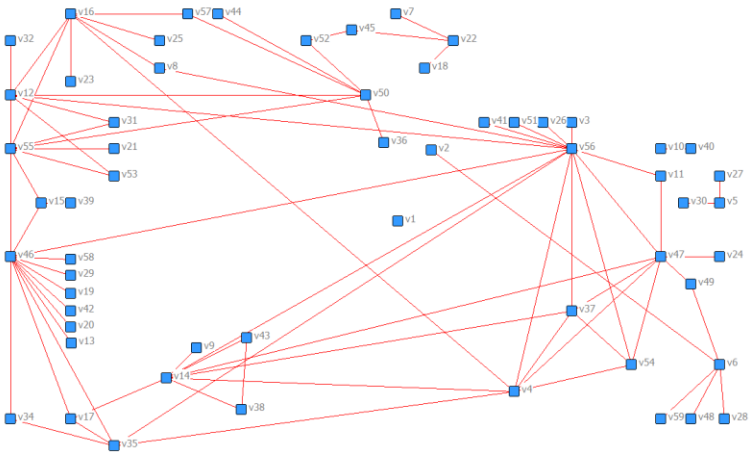


Figure A.2.6.3. An Enron Network Example of Size *N=59*

We then use the HMST approach to improve on this existing network. Figure A.2.6.4 provides the HMST-improved network.

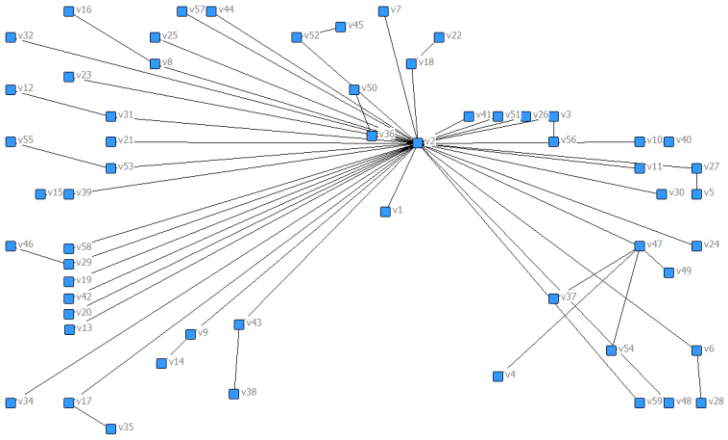


Figure A.2.6.4. Improved Enron Network of Size *N=59, H=5*

This network is for the case where the seeding cost is higher than the new connection cost. We then provide several other cases with different seeding to new connection cost ratios. Figures A.2.6.5 through A.2.6.7 provide alternative improved networks where seeding cost is equal to new connection cost for hop constraints of 3, 4 and 5.

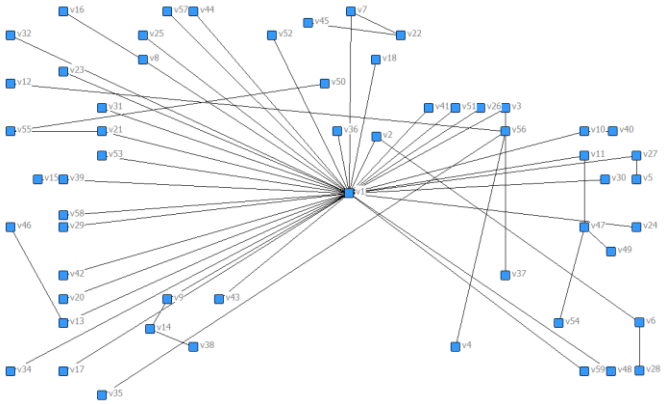


Figure A.2.6.5. Improved Enron Network of Size N=59 with $\frac{C_S}{C_C} = 1$ and H=3

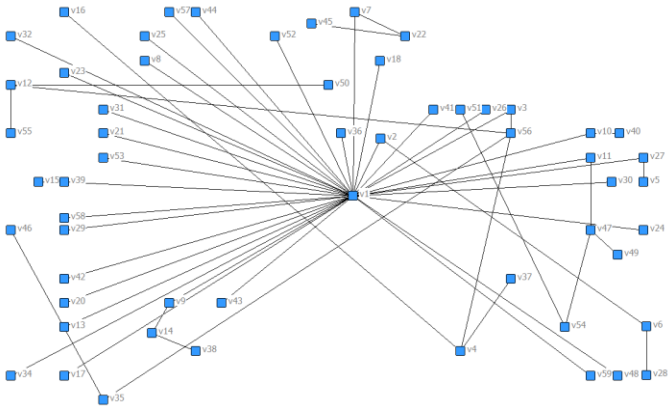


Figure A.2.6.6. Improved Enron Network of Size N=59 with $\frac{C_S}{C_C} = 1$ and H=4

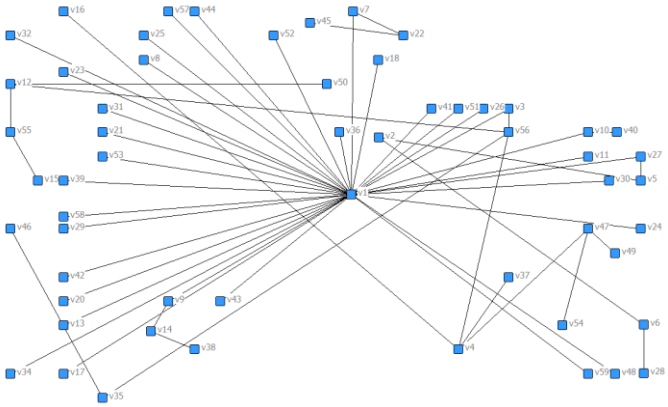


Figure A.2.6.7. Improved Enron Network of Size N=59 with $\frac{C_S}{C_C} = 1$ and H=5

It can be seen that there are more seeds than new connections in the case where the cost of seeding is the same as cost of adding new connections. In fact, in this case, there are many solutions that have equal total cost, that is, there is no difference between having a seed or a connection connecting the nodes. So the network cost of the Figure A.2.6.4. is the same as the network cost of Figure A.2.6.7, and also for Figure A.2.6.8 below.

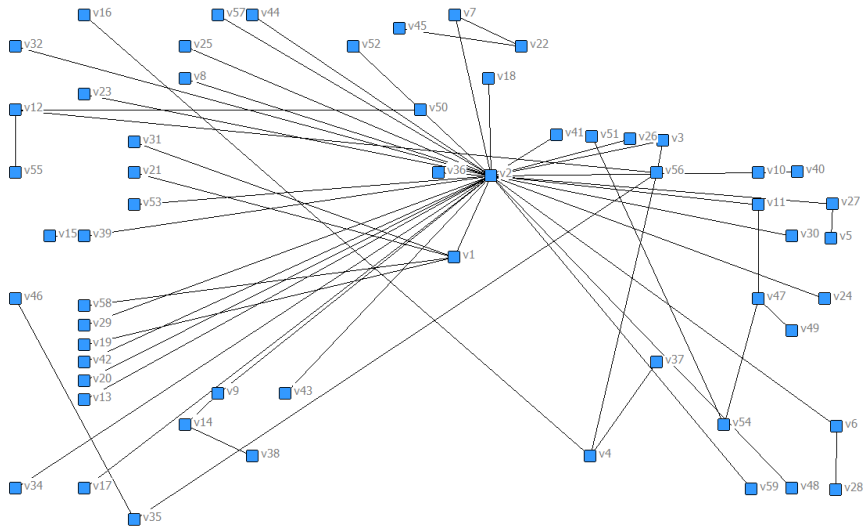


Figure A.2.6.8. An Alternative Improved Enron Network of Size N=59 with $\frac{c_S}{c_C} = 1$ and H=5

Next we provide some numerical results for the Enron problems. We use two networks from different time epochs in the lifetime of the Enron Company. Table A.2.6.1 provides the costs and solution times of the different methodologies. The 59 node network depicts the internal e-mail activity of Enron from May 11, 1999 to February 24, 2000, and the 111 node network depicts the internal e-mail activity of Enron from May 11, 1999 to December 10, 2000.

Table A.2.6.1. Average Gaps from Best Known Solutions and Times for Enron Problems

Problem Set		Averag Percentage Gap from Best Known Solution							Solution Times in Seconds						
N	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE	Prim	Akgun	FGV	Prim+ Swap	Prim+SLE	Akgun+SLE	FGV+SLE
59	3	17.0	5.1	27.5	7.3	0.0	0.0	0.0	0.2	3,600.7	959.2	0.9	74.5	3,672.9	1,031.5
	4	10.0	0.0	5.0	0.0	0.0	0.0	0.0	0.2	3,600.7	2,400.1	0.8	84.5	3,773.2	2,497.1
	5	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	3,600.8	2,937.0	0.8	4,247.2	29,838.0	22,054.2
111	3	50.4	13.6	113.9	40.3	0.0	0.0	0.0	1.2	3,601.5	17,101.8	5.5	58,737.8	59,667.3	72,489.8
	4	59.2	16.4	124.3	24.4	0.0	0.0	0.0	1.2	3,601.6	30,664.2	5.3	58,801.1	60,832.9	89,402.7
	5	48.7	11.2	98.3	26.2	0.0	0.0	0.0	1.2	3,601.6	75,111.6	5.3	57,385.6	60,848.5	133,907.3

It can be seen from the table that the proposed HMST heuristic is again able to improve the solution quality of the existing methods. This is especially observable in the larger problem of size $N = 111$.

We next use the propagation settings from Section 2.5 to estimate CIT propagation measure for these problems. The average CIT measures for the Enron network problems are provided in Table A.2.6.2.

Table A.2.6.2. Average CIT for Enron Problems

Problem Set		Average CIT					
N	H	Prim	Akgun	FGV	Prim+SLE	Akgun+SLE	FGV+SLE
59	3	2.6	3.6	2.4	3.0	3.4	3.2
	4	2.8	3.4	3.2	3.2	3.4	3.4
	5	2.4	4.8	4.2	3.4	4.4	4.6
111	3	2.6	3.0	2.8	3.2	3.2	3.0
	4	3.2	3.4	3.2	3.6	3.8	3.6
	5	4.0	4.8	4.2	4.8	4.6	4.8

It can be seen that the CIT estimation for the HMST-improved solutions are within the hop constraint, and comparable to each other.

Appendix 2.7: Computational Results on EN Problems

The complete computational results for EN problems from Table 2.2 is given in Table A.2.7.1.

Table A.2.7.1. Average Gaps from Best Known Solutions and Times for EN Problems

Problem Set				Averag Percentage Gap from Best Known Solution							Solution Times in Seconds						
N	Number of Problems	H	Source	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE	Prim	Akgun	FGV	Prim+ Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10	2	3	Twitter (1), Facebook (1)	7.1	0.0	49.3	7.1	0.0	0.0	1.9	0.0	0.4	0.1	0.0	1.9	2.0	1.8
		4		7.1	0.0	7.1	7.1	0.0	0.0	0.0	0.0	5.3	0.1	0.0	1.8	7.1	2.7
		5		7.1	0.0	8.4	7.1	1.9	0.0	1.9	0.0	25.7	0.1	0.0	2.2	28.3	2.6
11	4	3	Twitter (2), Facebook (2)	5.2	0.0	116.8	5.2	0.0	0.0	0.0	0.0	1.3	0.1	0.0	3.0	4.2	3.1
		4		2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	35.8	0.1	0.0	17.8	51.1	15.0
		5		2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	184.4	0.1	0.0	51.8	250.4	62.2
13	1	3	Twitter (1)	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	46.3	0.2	0.0	7.2	55.3	9.7
		4		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2,573.5	0.2	0.0	169.5	2,757.6	176.5
		5		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3,600.9	0.2	0.0	1,122.6	4,727.9	1,355.2
14	3	3	Twitter(2), Facebook(1)	7.5	0.0	81.7	4.8	0.0	0.0	0.0	0.0	86.2	0.3	0.0	11.2	104.1	9.5
		4		2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2,634.9	0.3	0.0	547.7	3,266.4	552.2
		5		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3,600.8	0.4	0.0	2,771.6	6,944.2	3,011.9
15	2	3	Twitter (2)	4.6	0.0	52.7	4.6	0.0	0.0	0.0	0.0	390.8	0.4	0.0	79.4	455.8	68.6
		4		2.7	0.0	2.7	2.7	0.0	0.0	0.0	0.0	3,600.7	0.4	0.0	3,457.6	7,093.5	3,752.7
		5		2.7	0.0	2.7	2.7	0.0	0.0	0.0	0.0	3,600.7	0.5	0.0	6,066.9	9,649.6	6,050.7
16	2	3	Twitter (2)	6.4	0.0	44.4	6.4	0.0	0.0	0.0	0.0	1,800.4	1.0	0.0	22.5	1,819.8	23.3
		4		5.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	1,859.5	0.9	0.0	4,150.0	4,875.8	3,132.2
		5		5.0	0.0	0.0	2.2	2.2	0.0	0.0	0.0	3,600.3	1.1	0.0	4,506.6	7,356.2	3,807.0
17	3	3	Twitter (2), Facebook (1)	7.2	0.0	24.3	7.2	0.0	0.0	0.0	0.0	2,404.5	0.9	0.0	167.4	2,448.4	133.7
		4		2.3	0.0	1.5	1.5	0.0	0.0	0.0	0.0	3,447.1	0.9	0.0	6,307.5	9,337.4	6,457.4
		5		1.5	0.0	1.5	1.5	0.0	0.0	0.0	0.0	3,600.7	1.0	0.0	7,765.4	10,214.3	7,807.6
20	2	3	Twitter (1), Facebook (1)	12.5	0.0	12.5	12.5	0.0	0.0	0.0	0.0	1,877.8	2.0	0.0	6,286.5	8,383.4	6,509.7
		4		8.3	0.0	2.8	5.6	0.0	0.0	0.0	0.0	3,600.5	2.3	0.0	6,531.4	10,168.8	6,517.8
		5		8.3	0.0	2.8	5.6	0.0	0.0	0.0	0.0	3,600.5	2.4	0.0	6,734.0	10,353.7	6,720.1
28	2	3	Twitter (1), Facebook (1)	0.7	0.0	63.5	0.7	0.0	0.0	0.0	0.0	1,808.9	9.9	0.1	5,206.8	6,822.8	5,025.8
		4		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3,600.3	18.2	0.1	8,515.3	12,115.5	8,633.7
		5		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3,600.4	19.0	0.1	8,576.6	12,155.4	8,610.3
31	1	3	Facebook (1)	14.4	1.9	4.8	5.3	0.0	0.0	0.0	0.0	3,600.3	17.7	0.2	4,501.7	7,465.8	4,535.2
		4		4.8	0.0	4.8	0.0	0.0	0.0	0.0	0.0	3,600.3	19.1	0.2	18,022.2	22,127.6	18,042.4
		5		0.0	0.0	4.8	0.0	0.0	0.0	0.0	0.0	3,600.4	25.1	0.1	18,024.5	22,152.7	18,049.7
32	2	3	Twitter (2)	2.9	0.4	83.1	2.9	0.0	0.0	0.0	0.0	3,600.4	31.1	0.1	13,715.4	15,447.3	10,271.7
		4		0.4	0.0	50.4	0.4	0.0	0.0	0.0	0.0	3,600.4	45.2	0.1	19,016.6	20,868.5	17,400.3
		5		0.4	0.0	0.4	0.0	0.0	0.0	0.0	0.0	3,600.5	45.3	0.2	18,819.0	21,671.0	18,066.5
33	1	3	Facebook (1)	36.5	0.0	113.2	29.8	0.0	0.0	0.0	0.0	3,600.3	22.7	0.2	16,124.9	13,036.6	12,634.3
		4		13.3	0.0	79.9	13.3	0.0	0.0	0.0	0.0	3,600.2	25.6	0.1	19,519.9	21,546.5	18,073.8
		5		0.0	0.0	33.3	0.0	0.0	0.0	0.0	0.0	3,600.3	32.9	0.1	19,525.6	23,138.5	19,559.1
36	1	3	Facebook (1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3,600.3	18.1	0.2	17,517.3	22,713.6	19,133.4
		4		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3,600.4	23.5	0.2	21,016.1	24,618.8	21,042.7
		5		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3,600.3	25.9	0.2	21,019.6	24,721.1	21,045.3
38	1	3	Facebook (1)	21.6	10.8	159.9	21.6	0.0	0.0	0.0	0.1	3,600.3	43.8	0.2	7,392.3	9,135.4	8,583.3
		4		21.6	0.0	100.0	0.0	0.0	0.0	0.0	0.1	3,600.5	54.3	0.3	19,153.0	22,129.1	18,684.5
		5		0.0	0.0	89.2	0.0	0.0	0.0	0.0	0.1	3,600.3	99.2	0.2	22,032.0	25,635.9	22,130.5
40	1	3	Twitter (1)	9.6	0.0	200.0	6.5	0.0	0.0	0.0	0.1	3,600.2	40.6	0.4	23,054.7	26,615.4	23,169.4
		4		0.0	0.0	200.0	0.0	0.0	0.0	0.0	0.1	3,600.2	55.7	0.2	23,073.7	26,617.4	23,072.5
		5		0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.1	3,600.2	78.5	0.3	23,120.8	26,619.1	23,099.5
43	1	3	Twitter (1)	9.2	0.4	172.6	4.3	0.0	0.4	0.4	0.1	3,600.3	207.3	0.5	8,249.0	4,002.3	13,378.0
		4		7.5	0.0	88.8	1.8	0.0	0.0	0.0	0.1	3,600.3	185.1	0.5	21,027.9	24,626.6	21,377.6
		5		15.0	0.0	98.7	8.9	7.0	0.0	7.0	0.1	3,600.3	323.1	0.5	24,036.5	28,331.3	21,453.9
45	1	3	Twitter (1)	66.0	0.0	229.7	27.3	1.3	0.0	1.3	0.1	3,600.2	199.5	0.8	9,131.5	10,649.2	7,621.4
		4		55.4	0.6	150.8	13.7	0.0	0.6	0.3	0.1	3,600.3	439.9	0.5	23,540.9	25,733.5	22,490.3
		5		81.8	3.9	12.2	10.3	0.0	3.9	6.3	0.1	3,600.3	595.8	0.5	25,047.7	25,639.5	22,630.2
82	1	3	Facebook (1)	27.4	0.0	92.5	22.7	0.0	0.0	0.0	0.7	3,600.6	4,708.7	3.4	32,389.6	33,052.6	33,215.7
		4		0.0	0.0	46.0	0.0	0.0	0.0	0.0	0.8	3,600.7	8,016.2	2.3	44,121.2	47,909.0	52,215.8
		5		0.0	0.0	46.0	0.0	0.0	0.0	0.0	0.7	3,600.8	10,182.5	2.1	44,104.7	47,811.4	54,386.9
85	1	3	Twitter (1)	75.2	18.4	367.8	35.8	35.8	0.0	4.6	0.8	3,601.2	7,312.6	5.5	233,841.2	28,644.5	39,889.5
		4		48.9	0.0	197.1	18.9	0.0	0.0	0.0	0.8	3,600.7	10,197.0	3.5	44,128.2	46,317.8	54,310.8
		5		10.2	0.6	196.5	10.2	0.0	0.0	0.0	0.8	3,600.8	15,128.1	2.3	45,714.9	49,395.9	57,443.9
86	1	3	Facebook (1)	55.8	82.2	317.3	36.2	0.0	2.7	2.7	0.9	3,600.7	7,692.4	4.0	46,213.0	49,724.1	53,797.0
		4		73.3	18.9	217.7	12.0	0.0	0.0	0.0	0.8	3,600.8	10,789.7	3.9	46,137.9	49,817.0	56,912.8
		5		235.0	34.9	219.9	81.6	12.9	7.7	0.0	0.9	3,600.7	16,691.6	4.2	46,121.7	49,922.6	62,812.8
92	1	3	Twitter (1)	140.4	61.2	265.6	70.7	0.0	1.3	1.3	1.1	3,601.8	74,793.0	7.4	17,361.2	13,426.3	85,925.0
		4		177.8	58.9	209.3	57.2	0.7	0.0	0.0	1.4	3,602.3	150,890.3	10.3	27,139.5	28,334.5	177,985.1
		5		179.4	7.5	89.8	36.1	0.0	1.3	4.5	1.1	3,601.0	228,755.8	7.7	47,279.7	47,868.8	275,003.9
96	1	3	Twitter (1)	31.2	10.1	385.4	6.6	0.0	0.2	0.6	1.3	3,601.0	22,096.2	8.3	51,202.0	54,759.5	72,802.9
		4		16.9	0.8	187.3	7.4	0.0	0.0	0.0	1.2	3,600.9	31,558.0	5.6	51,217.8	54,825.8	82,834.7
		5		8.0	2.0	135.1	3.4	1.6	0.0	4.3	1.2	3,601.1	52,747.7	5.7	51,270.1	54,812.5	141,240.8
124	1	3	Facebook (1)	209.2	68.5	177.0	143.6	3.8	16.8	0.0	2.5	3,601.6	78,388.4	13.9	65,492.5	69,022.2	143,850.9
		4		47.2	49.1	93.1	19.2	0.6	0.0	0.7	2.5	3,601.6	726,263.5	11.6	65,698.1	69,006.9	791,881.3
		5		21.2	5.1	44.0	12.1	0.0	0.0	0.0	2.7	3,601.5	596,452.9	18.1	65,774.6	68,993.6	662,022.7

Table A.2.7.2 provides the complete results for the average CIT in EN problems.

Table A.2.7.2. Average CIT for EN Problems

Problem Set				Average CIT						
N	Number of Problems	Source	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE
10	2	Twitter (1), Facebook (1)	3	2.4	2.5	2.3	2.4	2.3	2.5	2.4
			4	2.6	2.3	2.8	2.6	2.5	2.3	3.3
			5	2.7	2.6	4.0	2.8	2.4	2.6	3.8
11	4	Twitter (2), Facebook (2)	3	2.4	2.1	1.8	2.3	2.4	2.1	2.2
			4	2.5	2.3	2.5	2.7	2.8	2.3	2.4
			5	2.3	2.4	2.4	2.5	2.5	2.5	2.4
13	1	Twitter (1)	3	2.0	2.0	2.0	2.0	2.0	2.0	2.2
			4	2.2	2.0	2.0	2.4	2.4	2.0	2.0
			5	2.0	2.0	2.0	2.0	2.0	2.0	2.0
14	3	Twitter(2), Facebook(1)	3	2.3	2.1	2.0	2.3	2.0	2.1	2.0
			4	2.6	2.3	2.0	2.7	2.7	2.3	2.0
			5	2.8	2.3	2.2	2.8	2.8	2.3	2.2
15	2	Twitter (2)	3	2.2	2.0	2.2	2.2	2.0	2.0	2.0
			4	2.7	2.1	2.8	2.7	2.2	2.1	2.6
			5	2.6	2.0	2.5	2.5	2.1	2.0	2.5
16	2	Twitter (2)	3	2.5	2.7	2.7	2.5	2.7	2.5	2.6
			4	2.6	2.4	3.5	2.6	2.9	2.4	3.5
			5	3.0	2.4	3.7	2.2	2.2	2.4	3.8
17	3	Twitter (2), Facebook (1)	3	2.1	2.2	1.9	2.1	2.4	2.1	2.1
			4	2.5	2.1	2.5	2.8	2.8	2.3	2.7
			5	2.8	2.4	2.5	2.8	2.6	2.4	2.4
20	2	Twitter (1), Facebook (1)	3	2.2	2.2	2.5	2.1	2.4	2.2	2.3
			4	2.6	2.8	2.6	2.7	2.6	2.7	2.8
			5	2.6	3.3	2.8	2.7	3.3	3.1	3.1
28	2	Twitter (1), Facebook (1)	3	2.5	2.2	2.0	2.6	2.4	2.2	2.3
			4	2.9	2.8	2.8	2.8	2.8	2.8	2.8
			5	2.7	2.8	3.2	2.8	2.6	3.1	3.1
31	1	Facebook (1)	3	2.6	2.4	2.2	2.6	2.6	2.8	2.6
			4	3.0	3.0	2.8	3.4	3.4	3.0	3.2
			5	3.0	3.0	3.2	3.0	3.0	3.0	3.8
32	2	Twitter (2)	3	2.4	2.6	2.3	2.4	2.6	2.2	2.4
			4	2.6	2.5	2.8	2.6	2.0	2.5	3.0
			5	2.7	2.9	3.0	2.7	2.7	2.9	2.9
33	1	Facebook (1)	3	2.4	2.8	2.0	2.6	2.6	3.2	2.4
			4	3.4	3.0	3.4	3.4	3.4	3.0	3.4
			5	4.2	4.0	3.4	4.0	4.2	4.0	4.2
36	1	Facebook (1)	3	2.0	2.0	2.0	2.0	2.0	2.0	2.0
			4	2.0	2.0	2.0	2.0	2.0	2.2	2.0
			5	2.0	2.0	2.2	2.0	2.0	2.0	2.0
38	1	Facebook (1)	3	2.4	2.4	2.2	2.4	2.2	2.2	2.2
			4	3.0	2.2	3.0	3.0	3.0	2.2	3.4
			5	3.2	3.6	2.2	3.2	3.2	3.2	3.2
40	1	Twitter (1)	3	2.8	2.0	2.0	2.0	2.0	2.0	2.0
			4	3.0	2.0	2.0	3.0	3.0	2.0	3.0
			5	3.0	2.0	2.0	3.0	3.0	2.0	2.0
43	1	Twitter (1)	3	2.4	2.4	2.2	2.8	2.6	2.2	2.6
			4	3.4	3.0	3.4	3.0	3.0	3.0	3.0
			5	3.2	4.2	3.2	3.2	3.0	4.2	3.4
45	1	Twitter (1)	3	3.0	2.6	2.2	2.4	2.4	2.6	2.4
			4	4.0	3.2	3.4	3.8	3.6	3.4	3.6
			5	4.6	4.0	3.4	4.2	4.4	3.8	3.6
82	1	Facebook (1)	3	3.0	3.0	2.0	2.8	2.6	2.6	2.2
			4	3.4	3.2	3.0	3.4	3.4	3.2	3.2
			5	3.2	2.8	3.2	3.2	3.4	2.8	3.2
85	1	Twitter (1)	3	3.0	2.4	2.6	2.6	2.8	2.2	2.6
			4	3.2	3.6	3.0	4.0	3.8	3.6	4.0
			5	3.6	3.2	3.6	3.8	4.2	4.0	4.4
86	1	Facebook (1)	3	2.8	2.2	2.0	2.8	3.0	2.8	2.8
			4	3.8	3.6	3.0	3.6	3.2	3.4	3.4
			5	4.0	4.2	3.2	4.2	4.0	4.0	4.2
92	1	Twitter (1)	3	3.0	3.8	2.8	3.2	3.4	3.2	3.0
			4	4.2	3.6	3.4	4.2	4.4	3.8	3.8
			5	5.0	4.4	5.0	4.6	4.2	4.8	4.4
96	1	Twitter (1)	3	3.0	2.8	3.2	2.8	3.4	3.2	2.8
			4	3.6	3.4	3.0	3.6	3.0	3.2	3.6
			5	3.6	3.8	3.2	3.8	3.2	4.2	4.2
124	1	Facebook (1)	3	3.0	2.6	2.4	3.0	2.8	2.8	2.4
			4	3.4	3.0	2.6	3.8	3.0	3.4	3.0
			5	4.0	3.2	4.0	4.2	3.4	3.4	3.4

Other than cost, the decision-makers may also be interested in the number of seeds, i.e. number of direct connections to the root node. While a proportion of this is taken into account in the total cost, direct analysis of the number of seeds can also be beneficial. The CIT-number of

seeds graphs for an EN problem of size *N=96* in Figure A.2.7.1 below. For the number of seeds, random and greedy seeding methods require much larger seed sets in order to reach the same CIT as compared to enhanced networks. So these methods are not acceptable when there are limitations on the total number of direct connections available.

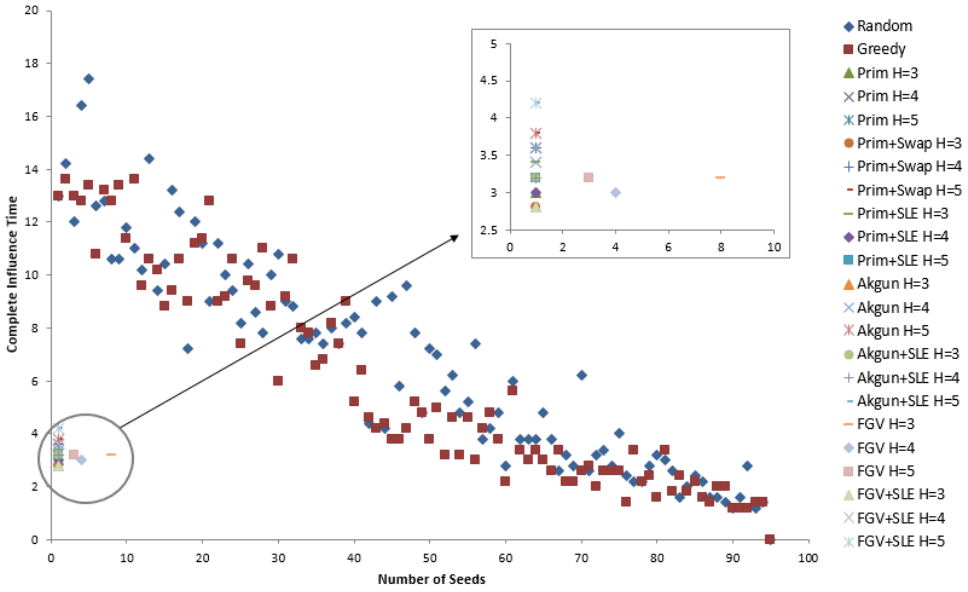


Figure A.2.7.1. Average CIT-Number of Seeds for a Twitter EN Problem of Size N=96

The analysis for CIT presented in the body of the chapter is based on the time to reach 100% penetration. A question that comes into mind is whether various methods propagate differently over time as they reach 100% penetration? To answer this, we analyzed the average penetration graph, provided in Figure A.2.7.2. This figure provides the average penetration percentage graph for all EN problems for each method. We observe that all of the methods reach 100% penetration at roughly the same time step. Moreover, the overall propagation trend is roughly the same for all methods, and that changing the propagation measurement from CIT to a less than 100% influence does not substantially change the results.

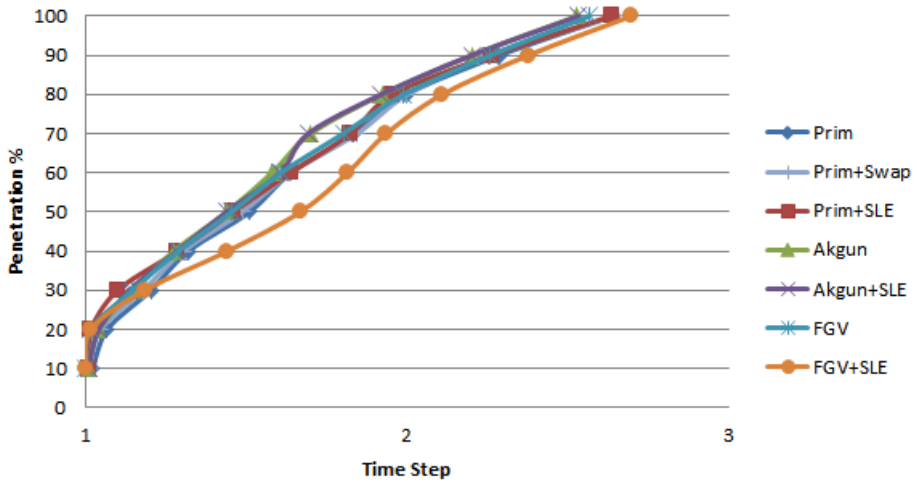


Figure A.2.7.2. Average Penetration Graph for All EN Problems

We next provide analysis on large EN problems. It is not practical to find SLE solutions for extremely large problems, but Prim and Prim+Swap solutions can be easily found. The enhanced networks of Prim and Prim+Swap still demonstrate big improvements over random and greedy seeding algorithms in large problems. Table A.2.7.3 below provides Prim and Prim+Swap HMST solution quality and times, and CIT results for these problems. HMST solution quality is provided as percentage increase in cost over the baseline. It can be seen that Prim+Swap completely dominates the Prim solutions in terms of cost quality. However, there is no clear distinction between the two methods in terms of CIT.

Table A.2.7.3. Results for Large EN Problems

			Complete Influence Time (CIT)			Improvement in CIT Over Baseline		Percentage Increase in Cost Over Baseline		HMST Computational Times in Seconds	
N	Source	H	Prim	Prim+Swap	Improvement	Prim	Prim+Swap	Prim	Prim+Swap	Prim	Prim+Swap
171	Facebook	3	3.0	3.2	(0.2)	1.2	1.0	260.5	140.3	6.5	43.7
		4	3.4	3.2	0.2	0.8	1.0	158.4	60.2	6.4	41.9
		5	4.0	4.2	(0.2)	0.2	0.0	49.8	0.0	6.4	43.9
181	Twitter	3	3.2	3.2	0.0	1.6	1.6	412.7	160.9	8.1	56.8
		4	4.0	4.2	(0.2)	0.8	0.6	228.9	50.9	8.2	56.8
		5	4.8	4.8	0.0	0.0	0.0	123.0	0.0	8.3	56.4
211	Twitter	3	3.0	3.2	(0.2)	1.6	1.4	439.9	161.8	12.7	94.8
		4	4.2	4.4	(0.2)	0.4	0.2	152.8	28.4	12.9	89.7
		5	4.4	4.6	(0.2)	0.2	0.0	59.0	0.0	12.9	91.3
221	Facebook	3	3.0	2.8	0.2	1.0	1.2	127.7	89.5	13.8	67.1
		4	3.6	3.0	0.6	0.4	1.0	86.2	55.0	13.3	66.9
		5	4.0	4.0	0.0	0.0	0.0	6.5	0.0	13.7	66.5
222	Twitter	3	3.0	3.0	0.0	1.0	1.0	157.7	28.3	15.0	108.4
		4	3.8	4.0	(0.2)	0.2	0.0	57.5	1.2	15.1	105.5
		5	4.0	3.6	0.4	0.0	0.4	42.9	0.0	15.3	103.2
225	Twitter	3	3.0	3.0	0.0	1.6	1.6	244.7	76.0	15.8	110.9
		4	4.0	3.4	0.6	0.6	1.2	111.9	12.7	15.9	112.8
		5	4.6	4.2	0.4	0.0	0.4	39.6	0.0	15.9	112.7
226	Twitter	3	3.2	3.0	0.2	1.2	1.4	197.9	76.8	15.8	115.3
		4	4.0	3.8	0.2	0.4	0.6	49.9	10.6	16.0	112.4
		5	4.2	4.4	(0.2)	0.2	0.0	47.2	0.0	16.3	114.0
227	Twitter	3	3.0	3.2	(0.2)	1.0	0.8	115.2	34.2	16.1	111.4
		4	3.4	3.6	(0.2)	0.6	0.4	29.6	0.0	15.6	110.2
		5	4.0	4.0	0.0	0.0	0.0	21.4	1.3	16.1	81.4
237	Twitter	3	3.0	3.4	(0.4)	1.8	1.4	378.8	189.4	18.5	131.8
		4	4.6	4.2	0.4	0.2	0.6	119.0	41.8	18.6	132.9
		5	4.4	4.8	(0.4)	0.4	0.0	35.3	0.0	18.1	130.0
238	Facebook	3	3.0	3.4	(0.4)	1.0	0.6	109.2	74.8	17.2	116.0
		4	3.6	3.6	0.0	0.4	0.4	47.4	10.3	17.8	85.5
		5	4.0	4.0	0.0	0.0	0.0	33.9	0.0	17.9	117.7
243	Twitter	3	3.0	3.2	(0.2)	1.4	1.2	348.2	138.4	20.1	142.7
		4	4.2	3.4	0.8	0.2	1.0	165.3	38.5	19.8	139.6
		5	4.4	4.4	0.0	0.0	0.0	135.4	0.0	19.6	140.2
248	Twitter	3	3.0	3.0	0.0	1.0	1.0	45.0	11.2	20.9	108.8
		4	3.4	3.2	0.2	0.6	0.8	3.5	0.5	21.3	147.4
		5	4.0	4.0	0.0	0.0	0.0	2.2	0.0	21.1	189.2
481	Facebook	3	3.0	3.0	0.0	1.4	1.4	1088.2	399.6	145.6	1090.5
		4	4.0	4.0	0.0	0.4	0.4	235.5	79.6	149.1	1019.4
		5	4.2	4.4	(0.2)	0.2	0.0	69.7	0.0	142.6	706.7
711	Facebook	3	3.0	3.0	0.0	1.0	1.0	846.6	133.1	470.3	3345.8
		4	3.8	3.2	0.6	0.2	0.8	103.5	8.8	472.1	3248.2
		5	4.0	4.0	0.0	0.0	0.0	47.1	0.0	460.4	3253.8
770	Facebook	3	3.0	3.0	0.0	1.4	1.4	496.7	245.6	598.1	4201.6
		4	4.0	3.8	0.2	0.4	0.6	386.2	93.5	597.2	4168.7
		5	4.2	4.4	(0.2)	0.2	0.0	46.5	0.0	588.2	2970.4

In order to compare the available HMST solutions to the seeding only methods, we graph the solutions from Prim and Prim+Swap along with random and greedy seeding solutions. Figures A.2.7.3 and A.2.7.4 provide average CIT-Cost and average CIT-number of seeds graphs for an EN problem from Facebook with size *N=481*.

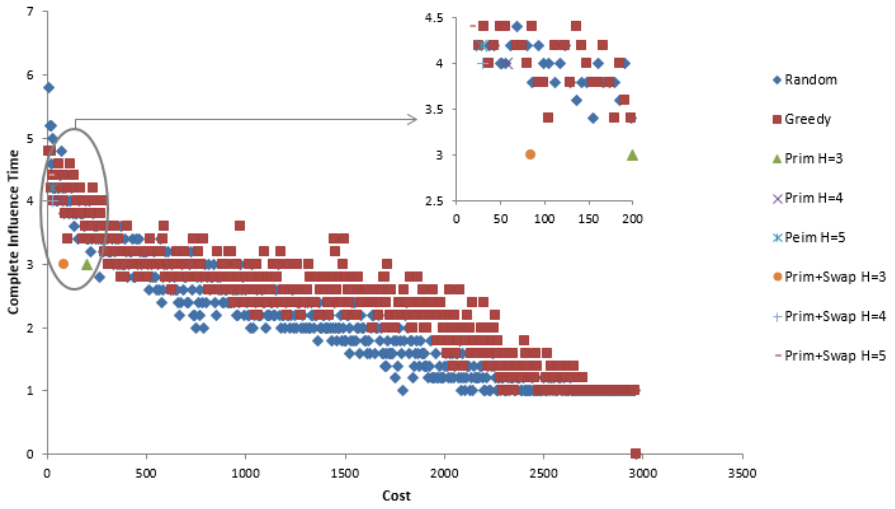


Figure A.2.7.3. Average CIT-Cost for a Facebook EN Problem of Size N=481

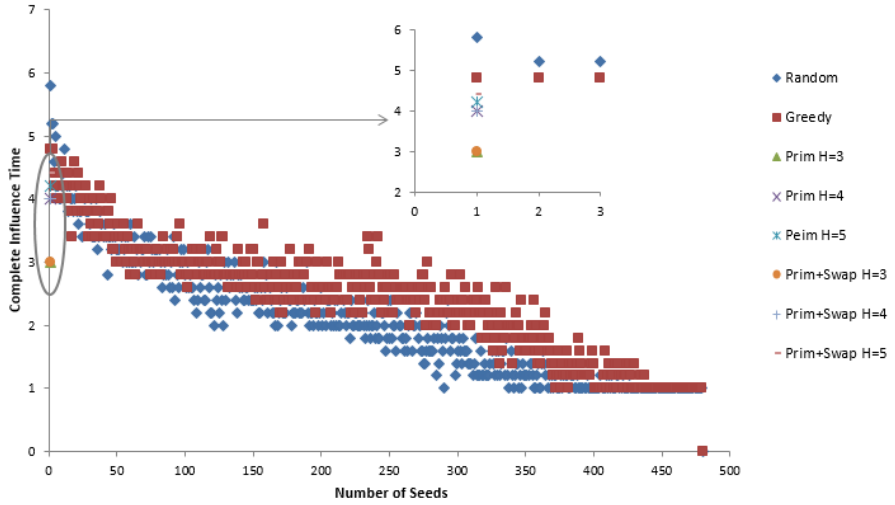


Figure A.2.7.4. Average CIT-Number of Seeds for a Facebook EN Problem of Size N=481

There is quite a difference between improved and unimproved networks, and the seeding-only methods require a lot more initial seeds to reach the propagation level of the improved networks. On the other hand, it can be seen that the greedy method is not performing better than the random seeding for EN problems of this size. It seems that the greedy algorithm cannot find

nodes that can improve the CIT without causing the total cost to be higher than the random seed selection.

Appendix 2.8: Computational Results on Random Problems

A.2.8.1 Random Euclidean Problems

For the random Euclidean problems with root node in the center (EC), the costs are simply the Euclidean distance between nodes located randomly on a 100 by 100 space, where node 1 is located in the center at point (50,50). The settings of the experiments are similar to those of EN problems in Section 2.6. For the randomly generated problems, the network sizes (number of nodes including the root node) tested are *N = 10, 20, 30, 40, 50, 70, 90*. The HMST solution costs and times in EC problems are summarized in Table A.2.8.1. Solution quality is presented as the percentage gap from the best known solution. The results in the table are averages over the 5 test problems for each *N* and *H* combination.

Table A.2.8.1. Average Gaps from Best Known Solutions and Times for EC Problems

		Percentage Gap From Best Known Solution							Solution Times in Seconds						
N	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE	Prim	Akgun	FGV	Prim+Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10	3	5.3	0.0	0.0	0.5	0.0	0.0	0.0	0.01	0.12	0.33	0.01	3.38	1.75	1.95
	4	5.0	0.0	0.0	0.4	0.4	0.0	0.0	0.00	0.12	0.32	0.00	2.30	1.70	1.81
	5	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.00	0.12	0.39	0.00	2.83	1.61	1.80
20	3	17.2	0.0	2.1	4.1	0.4	0.0	0.7	0.01	0.42	11.28	0.04	5.92	4.38	14.83
	4	17.4	0.0	0.9	6.0	1.1	0.0	0.0	0.01	0.83	15.43	0.04	5.92	4.73	18.80
	5	9.8	0.0	0.9	4.2	1.1	0.0	0.0	0.01	0.45	14.80	0.03	5.58	4.27	18.28
30	3	36.2	0.0	2.1	9.7	1.0	0.0	0.5	0.04	6.66	97.27	0.18	7.79	14.35	104.23
	4	30.6	0.0	1.7	8.6	1.2	0.0	0.5	0.04	132.30	136.95	0.14	8.14	139.99	143.82
	5	14.4	0.0	0.6	2.6	0.9	0.0	0.3	0.04	5.28	153.01	0.16	8.15	12.83	160.00
40	3	64.1	0.0	2.7	18.2	1.0	0.0	0.7	0.08	2,297.22	508.88	0.46	15.05	2,312.00	521.52
	4	45.8	0.0	3.2	8.4	1.3	0.0	0.3	0.08	1,912.09	738.35	0.43	15.39	1,925.77	751.16
	5	32.1	0.3	2.5	7.1	2.0	0.2	0.9	0.08	1,583.66	910.65	0.41	14.35	1,597.25	923.20
50	3	90.8	0.4	3.5	22.2	1.2	0.0	1.6	0.16	3,600.57	1,879.85	0.85	24.70	3,623.76	1,902.07
	4	58.4	0.1	2.7	13.9	1.6	0.1	0.9	0.16	3,600.62	3,693.58	0.82	23.57	3,623.47	3,715.61
	5	46.0	0.2	1.9	10.0	2.9	0.0	0.1	0.16	3,600.74	4,842.06	0.89	23.64	3,623.34	4,863.92
70	3	106.8	1.1	3.1	24.9	1.5	0.5	0.1	0.44	3,600.60	15,600.30	2.54	55.68	3,653.26	15,650.65
	4	105.7	0.9	3.1	24.1	1.9	0.3	1.2	0.44	3,600.67	29,702.11	2.62	55.20	3,653.55	29,752.61
	5	77.8	0.9	2.5	13.7	1.3	0.5	0.2	0.43	3,600.71	48,564.96	2.92	54.89	3,654.17	48,614.50
90	3	163.8	3.6	3.2	33.6	0.9	0.3	0.7	0.94	3,600.83	90,385.48	5.40	109.35	3,704.45	90,485.30
	4	111.9	1.8	1.9	18.0	1.4	0.0	0.6	0.94	3,600.85	182,895.09	4.91	107.51	3,704.58	182,998.09
	5	107.1	1.2	1.2	16.4	1.7	0.7	0.3	0.93	3,600.84	300,200.75	6.57	108.98	3,705.58	300,299.94

In the Euclidean problems, Akgun and FGV algorithms perform well and we can see that they are on average very close to the best known solution. However, our improvement heuristic does find improvements for larger problems and the additional computational time is small. The improvement heuristic is also able to improve the bad solution of Prim to a few percent of the best known solution.

In order to better present the improvements, in Table A.2.8.2 we compare the algorithm results to the best of all previously known techniques (i.e., we compare to the best solution from Prim, FGV and Akgun). Negative values in parenthesis in Table A.2.8.2 show solutions that are worse than the best of all previously known techniques.

Table A.2.8.2. Average Improvement Over Best of Known Solutions for EC Problems

		Percentage Improvement (Deterioration)			
N	H	Prim+Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10	3	(0.49)	0.00	0.00	0.00
	4	(0.40)	(0.40)	0.00	0.00
	5	0.00	0.00	0.00	0.00
20	3	(4.11)	(0.36)	0.00	(0.67)
	4	(6.00)	(1.08)	0.00	(0.01)
	5	(4.18)	(1.06)	0.00	(0.03)
30	3	(9.71)	(0.99)	0.00	(0.53)
	4	(8.57)	(1.22)	0.00	(0.49)
	5	(2.55)	(0.94)	0.00	(0.32)
40	3	(18.15)	(1.03)	0.00	(0.72)
	4	(8.38)	(1.32)	0.00	(0.33)
	5	(6.78)	(1.68)	0.04	(0.60)
50	3	(21.76)	(0.82)	0.35	(1.23)
	4	(13.74)	(1.45)	0.02	(0.71)
	5	(9.88)	(2.71)	0.10	0.00
70	3	(23.89)	(0.68)	0.30	0.66
	4	(23.53)	(1.37)	0.21	(0.75)
	5	(12.74)	(0.44)	0.34	0.57
90	3	(30.67)	1.40	1.95	1.60
	4	(16.37)	0.00	1.41	0.82
	5	(16.03)	(1.30)	(0.30)	0.01

Using propagation model from Section 2.5 and computational settings in Section 2.6 we run propagation simulations to find the average CIT for the networks. For the random problems, we run 5 simulations for each of the 5 problems for each combination of *N* and *H,* so each

reported number is an average of these 25 propagation runs. Table A.2.8.3 provides the average CIT values for the EC problems.

Table A.2.8.3. Average CIT for EC Problems

		Average CIT						
N	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE
10	3	2.7	2.7	2.7	2.6	2.8	2.6	2.7
	4	3.4	3.4	3.4	3.4	3.4	3.4	3.4
	5	3.8	3.7	3.5	3.7	3.7	3.7	3.7
20	3	3.2	3.2	3.2	3.2	3.2	3.1	3.1
	4	4.0	4.0	3.9	3.6	4.0	3.8	3.8
	5	4.7	5.2	5.0	4.8	4.9	5.1	5.0
30	3	3.2	3.2	3.0	3.2	3.0	3.2	2.9
	4	4.4	4.2	4.2	4.2	4.2	4.2	4.0
	5	5.7	5.4	5.2	5.5	5.4	5.5	5.4
40	3	3.6	3.3	3.3	3.2	3.3	3.4	3.3
	4	4.4	4.2	4.1	4.2	4.4	4.3	4.3
	5	5.9	5.5	5.3	5.2	5.7	5.6	5.5
50	3	3.6	3.4	3.3	3.1	3.3	3.3	3.5
	4	4.6	4.9	4.5	4.5	4.6	4.5	4.5
	5	5.8	5.5	5.8	5.6	5.6	5.6	5.6
70	3	3.8	3.4	3.7	3.3	3.5	3.5	3.6
	4	4.9	4.6	4.6	4.2	4.8	4.7	4.6
	5	5.8	5.4	5.6	5.6	5.9	5.5	5.6
90	3	3.7	3.6	3.6	3.2	3.7	3.8	3.5
	4	4.8	4.8	4.8	4.5	4.8	5.0	4.7
	5	6.0	5.6	6.1	5.6	5.9	6.0	5.8

Figures A.2.8.1, A.2.8.2 and A.2.8.3 provide the CIT-cost, CIT-number of seeds and penetration graphs for Euclidean problems. We can see from the graphs that the HMST-altered solutions have much better quality in terms of both cost and CIT. Moreover, the solutions of the heuristic, while having lower cost, propagate faster as well (have lower CIT). However, this is not always the case. Another trade-off is in that the solutions with lower hop constraint have better CIT, but also come at a higher cost.

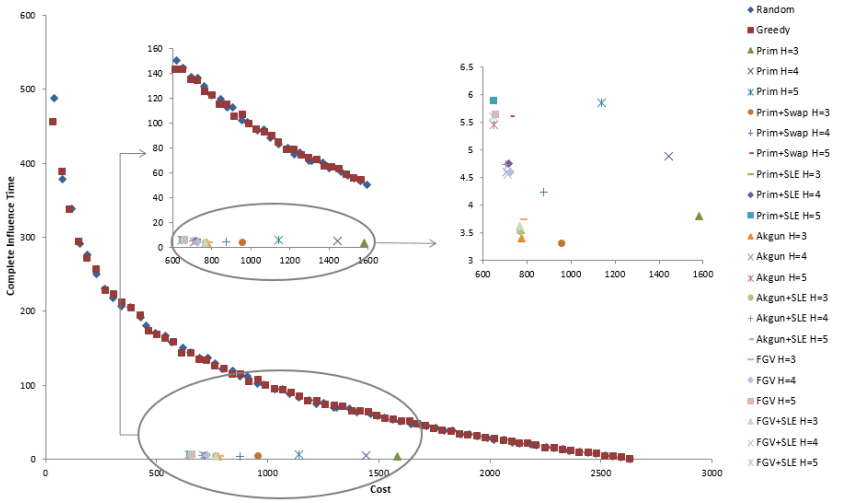


Figure A.2.8.1. Average CIT-Cost for EC Problems of Size N=70

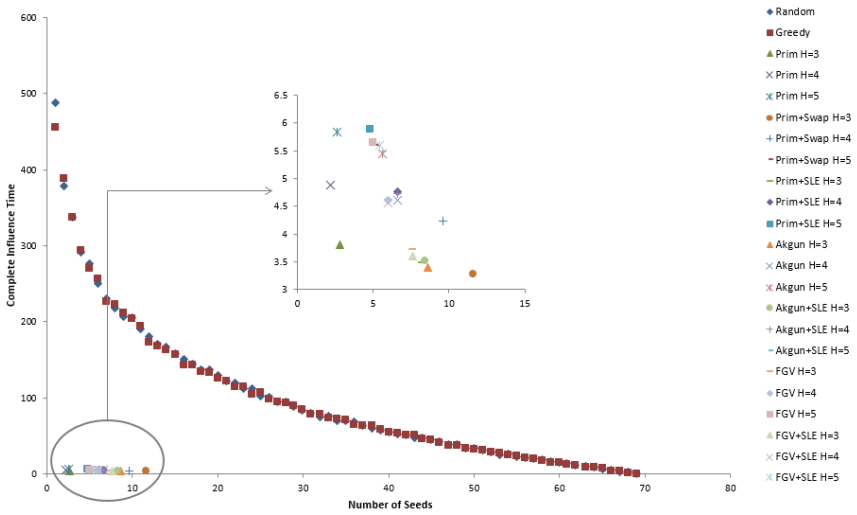


Figure A.2.8.2. Average CIT-Number of Seeds for EC Problems of Size N=70

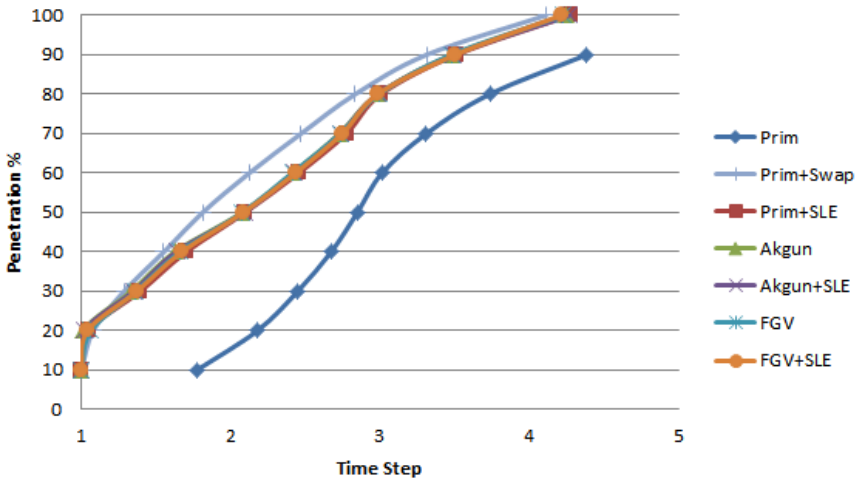


Figure A.2.8.3. Average Penetration for EC Problems (All *N* and all *H* Combined)

Table A.2.8.4 summarizes the results for experiments on larger EC problems.

Table A.2.8.4. Results for Large EC Problems

		Complete Influence Time (CIT)			Improvement in CIT Over Baseline		Percentage Increase in Cost Over Baseline		HMST Computational Times in Seconds	
N	H	Prim	Prim+Swap	Improvement	Prim	Prim+Swap	Prim	Prim+Swap	Prim	Prim+Swap
100	3	3.8	2.6	1.2	2.2	3.4	206.9	60.1	1.2	7.8
	4	5.4	4.2	1.2	0.6	1.8	132.8	15.1	1.2	7.8
	5	6.0	5.4	0.6	0.0	0.6	86.7	0.0	1.2	7.7
200	3	4.0	2.4	1.6	2.2	3.8	332.5	64.5	9.6	63.4
	4	5.2	4.0	1.2	1.0	2.2	248.6	19.5	9.6	62.9
	5	6.2	5.2	1.0	0.0	1.0	186.8	0.0	9.5	80.6
300	3	3.8	2.6	1.2	3.0	4.2	372.4	55.9	32.5	216.0
	4	5.4	4.4	1.0	1.4	2.4	320.1	17.9	32.3	216.2
	5	6.8	5.4	1.4	0.0	1.4	255.0	0.0	32.3	217.3
400	3	4.2	3.2	1.0	2.4	3.4	397.5	76.7	76.9	653.2
	4	5.0	3.8	1.2	1.6	2.8	353.2	39.8	76.8	657.0
	5	6.6	5.4	1.2	0.0	1.2	285.5	0.0	76.8	802.1
500	3	3.8	3.0	0.8	2.4	3.2	421.4	66.9	150.9	1,011.4
	4	4.8	4.0	0.8	1.4	2.2	334.3	19.4	150.3	1,287.8
	5	6.0	6.2	(0.2)	0.2	0.0	327.3	0.0	150.4	1,316.1
600	3	4.0	2.6	1.4	2.2	3.6	399.8	47.4	260.8	1,751.4
	4	4.6	3.8	0.8	1.6	2.4	366.4	26.0	260.5	1,775.5
	5	6.2	4.4	1.8	0.0	1.8	350.8	0.0	260.0	1,794.4
700	3	4.0	3.0	1.0	2.6	3.6	436.8	60.4	414.5	3,539.7
	4	5.6	3.6	2.0	1.0	3.0	415.5	33.5	412.7	2,794.9
	5	6.6	4.8	1.8	0.0	1.8	386.7	0.0	412.1	3,590.5
800	3	4.0	2.8	1.2	2.6	3.8	422.4	37.2	616.9	4,148.2
	4	5.8	4.0	1.8	0.8	2.6	393.1	31.4	616.0	5,383.9
	5	6.6	5.0	1.6	0.0	1.6	362.2	0.0	615.1	4,283.6
900	3	3.8	3.0	0.8	2.0	2.8	426.1	60.6	874.8	7,546.0
	4	4.8	4.2	0.6	1.0	1.6	486.0	41.3	880.9	6,144.5
	5	5.8	5.0	0.8	0.0	0.8	468.4	0.0	877.5	6,201.1
1000	3	3.4	3.0	0.4	3.2	3.6	618.8	82.4	1,202.5	8,252.9
	4	4.8	4.2	0.6	1.8	2.4	519.1	39.6	1,200.4	8,429.7
	5	6.6	5.6	1.0	0.0	1.0	532.6	0.0	1,202.9	10,855.6
1100	3	3.8	3.0	0.8	2.8	3.6	694.8	85.9	1,606.8	10,982.3
	4	5.2	4.4	0.8	1.4	2.2	600.3	46.3	1,600.8	11,254.3
	5	6.6	5.2	1.4	0.0	1.4	564.8	0.0	1,599.1	14,559.5
1200	3	4.0	3.0	1.0	2.0	3.0	480.5	50.2	2,080.0	14,324.3
	4	4.8	3.8	1.0	1.2	2.2	486.6	26.1	2,081.7	14,650.1
	5	6.0	5.0	1.0	0.0	1.0	443.0	0.0	2,079.4	18,853.7
1300	3	4.2	2.6	1.6	1.8	3.4	505.9	46.9	2,653.7	18,290.4
	4	5.0	3.8	1.2	1.0	2.2	486.9	18.1	2,652.1	19,197.1
	5	6.0	5.0	1.0	0.0	1.0	454.2	0.0	2,648.2	24,422.0
1400	3	4.6	3.0	1.6	1.4	3.0	490.2	47.9	3,311.7	22,774.3
	4	5.6	4.6	1.0	0.4	1.4	451.5	24.3	3,309.9	23,833.0
	5	6.0	5.2	0.8	0.0	0.8	422.9	0.0	3,303.0	29,999.2
1500	3	4.0	2.8	1.2	2.2	3.4	846.3	149.8	4,081.4	28,163.1
	4	5.0	4.0	1.0	1.2	2.2	785.4	71.7	4,072.4	28,911.7
	5	6.2	5.6	0.6	0.0	0.6	700.7	0.0	4,063.5	37,877.8

A.2.8.2 Random Perturbed-Euclidean Problems

In perturbed-EC problems, the cost structure is based on Euclidean distance between nodes that are randomly located in a 100 by 100 space, and there are some perturbations as follows. The

costs of up to 10% of randomly selected closest arcs to a node are set to zero. The connection cost between any two disconnected nodes is recalculated based on the number of mutual connections that they have. The cost of directly connecting to the root node is calculated as the Euclidean distance times the natural logarithm of size of the network (*ln(N)*), so the seeding cost is on average higher than connection costs. Details of the problem generation were previously described in Appendix 2.5. The experiment settings are similar to those of EC problems introduced in the last subsection. The HMST solution costs and times in perturbed-EC problems are summarized in Table A.2.8.5.

Similar to the results for EN problems, it can be seen that for the perturbed-EC problems, the proposed heuristic improvement results in substantial cost savings for problems with perturbed-EC data, and the savings are increased as the problem size increases. Akgun's formulation is able to find good solutions for problems of size 30 and smaller, and FGV finds good solutions for networks of size 10 only.

Table A.2.8.5. Average Gaps from Best Known Solutions and Times for Perturbed-EC Problems

		Average Percentage Gap From Best Known Solution							Average Solution Times in Seconds						
N	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE	Prim	Akgun	FGV	Prim+ Swap	Prim+SLE	Akgun+SLE	FGV+SLE
10	3	5.0	0.0	0.8	2.4	0.0	0.0	0.0	0.01	0.13	0.18	0.02	1.75	1.57	1.74
	4	2.6	0.0	0.8	1.2	0.0	0.0	0.0	0.00	0.15	0.13	0.00	1.56	1.60	1.51
	5	3.0	0.0	0.4	1.4	0.0	0.0	0.0	0.00	0.14	0.12	0.00	2.35	1.54	1.43
20	3	10.3	0.0	8.6	3.2	0.0	0.0	0.3	0.01	12.45	2.75	0.04	6.16	16.01	6.46
	4	10.2	0.0	7.1	2.7	0.3	0.0	0.3	0.01	789.77	4.40	0.05	5.80	793.67	8.17
	5	9.3	0.0	1.9	2.3	0.2	0.0	0.2	0.01	71.01	5.87	0.03	6.33	74.57	9.38
30	3	14.6	0.1	16.4	4.9	0.2	0.0	0.3	0.04	3,509.74	23.80	0.20	19.48	3,521.50	36.80
	4	16.2	0.4	10.7	4.3	0.4	0.0	0.4	0.04	3,600.33	38.38	0.18	20.07	3,616.61	55.69
	5	18.5	0.0	7.4	5.4	0.1	0.0	0.1	0.04	1,786.62	48.42	0.15	14.41	1,798.76	60.72
40	3	16.3	1.8	25.4	3.7	0.0	0.1	0.3	0.08	3,600.28	103.01	0.39	62.17	3,650.58	160.86
	4	28.5	1.5	21.4	5.3	0.3	0.2	0.8	0.08	3,600.27	195.67	0.41	157.36	3,719.28	343.40
	5	21.6	1.0	27.1	5.2	0.2	0.1	0.4	0.08	3,600.28	268.69	0.41	1,038.90	4,457.96	1,814.59
50	3	27.8	4.1	27.4	6.2	0.0	0.4	0.5	0.16	3,600.28	674.11	0.78	693.53	4,560.54	2,296.94
	4	22.0	6.1	30.2	6.0	0.4	0.5	0.5	0.16	3,600.30	873.32	0.80	13,724.03	17,475.00	14,434.47
	5	27.7	3.8	30.1	5.7	0.1	0.2	0.5	0.16	3,600.30	993.14	0.93	19,436.61	18,785.63	20,439.62
70	3	32.6	9.9	42.3	6.3	0.2	1.1	0.8	0.43	3,600.38	6,887.82	2.51	28,840.51	26,194.59	31,835.99
	4	40.5	9.3	45.0	7.3	0.0	1.3	1.6	0.44	3,600.44	10,265.13	2.41	38,105.70	40,413.95	48,044.26
	5	74.0	14.2	84.6	10.7	0.0	1.6	3.9	0.43	3,600.46	12,168.77	2.73	37,972.89	40,933.29	49,763.52
90	3	33.2	9.6	30.2	3.1	0.4	0.8	1.0	0.91	3,600.68	33,949.36	5.23	48,035.25	49,287.16	80,113.81
	4	36.0	8.9	31.8	4.0	0.3	0.6	0.8	0.91	3,600.69	57,504.75	5.27	48,147.77	51,727.58	105,638.16
	5	45.9	15.0	59.1	5.0	0.8	2.6	2.0	0.90	3,600.71	74,322.41	5.50	48,178.50	51,743.84	122,464.06

Table A.2.8.6 provides the propagation results for the perturbed-EC problems. The settings are similar to those of EC problems introduced in previous subsection. While there is some variation in the CITs for different methods, all of them are within an acceptable range of CIT performance as compared to the respective hop constraints (H). These results confirm that the HMST structure in the network has succeeded in effectively propagating the message to all users within the hop constraint. Figures A.2.8.4 and A.2.8.5 provide the average CIT-cost and CIT-number of nodes for perturbed-EC problems.

Table A.2.8.6. Average CIT for Perturbed-EC Problems

		Average CIT						
N	H	Prim	Akgun	FGV	Prim+ Swap	Prim+ SLE	Akgun+ SLE	FGV+ SLE
10	3	2.4	2.6	2.4	2.4	2.4	2.6	2.5
	4	3.7	3.4	3.1	3.4	3.1	3.4	3.4
	5	3.7	3.7	4.0	3.4	4.0	3.8	3.7
20	3	2.8	2.7	2.5	2.9	2.5	2.8	2.7
	4	3.9	3.7	3.6	3.5	3.6	3.7	3.9
	5	4.7	4.6	4.5	4.5	4.5	4.5	4.8
30	3	3.0	3.0	2.4	2.9	2.4	3.1	3.0
	4	3.8	3.9	3.4	3.7	3.4	4.0	4.0
	5	5.1	4.9	4.4	4.7	4.4	5.0	4.9
40	3	3.1	2.8	2.4	3.0	2.4	3.1	3.1
	4	4.1	3.9	3.4	4.0	3.4	4.3	4.2
	5	5.0	4.9	4.3	4.9	4.3	5.0	5.2
50	3	3.0	3.1	2.7	3.1	2.7	3.0	3.2
	4	4.4	3.9	3.2	4.0	3.2	3.8	3.9
	5	5.4	5.1	4.2	4.8	4.2	5.0	5.0
70	3	3.0	2.9	2.4	3.1	2.4	3.3	3.1
	4	4.2	4.0	3.1	3.9	3.1	4.3	3.9
	5	4.7	4.6	4.0	4.9	4.0	4.8	4.7
90	3	3.0	2.9	2.6	3.0	2.6	3.2	3.1
	4	4.2	3.8	3.2	3.7	3.2	4.0	4.1
	5	4.8	4.4	3.8	4.5	3.8	4.7	4.6

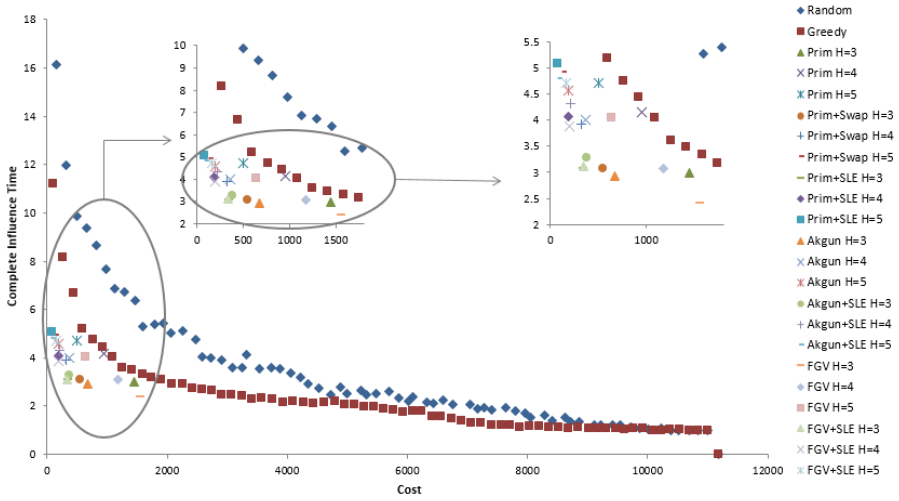


Figure A.2.8.4. Average CIT-Costs for Perturbed-EC Problems of Size N=70

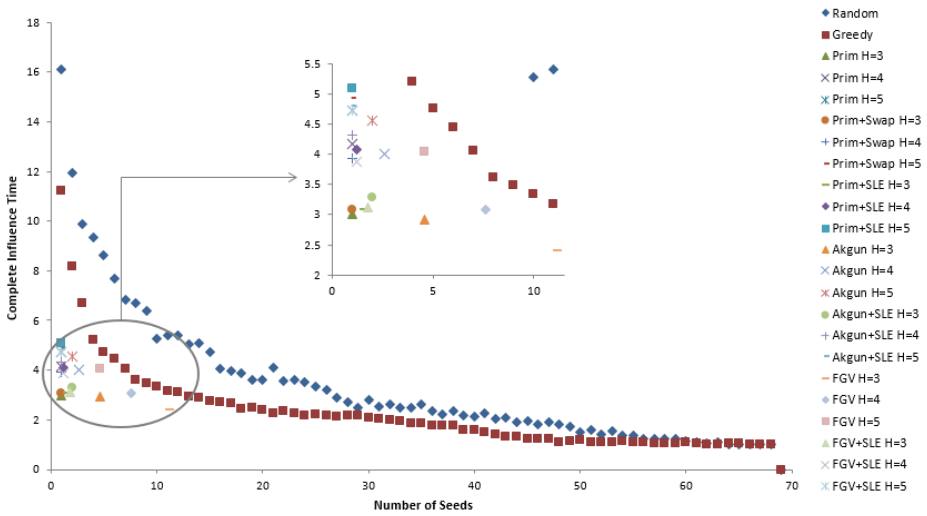


Figure A.2.8.5. Average CIT-Number of Seeds for Perturbed-EC Problems of Size N=70

Figure A.2.8.6 provides the average penetration graph for the perturbed-EC problems. It is observed that although all of the methods reach complete penetration at roughly the same time step, the FGV initially penetrates much faster and Prim penetrates much slower than the other methods. The reason for this is that FGV method generally produces solutions with higher numbers of seeds, and that Prim generally has fewer seed nodes than the other methods (see Figure A.2.8.5). However, the overall trend is again roughly the same for all methods.

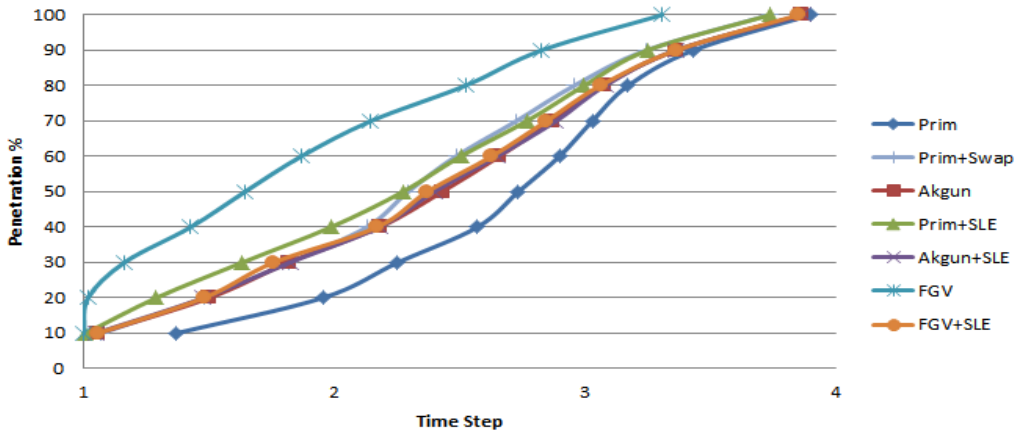


Figure A.2.8.6. Average Penetration for Perturbed EC Problems (All *N* and all *H* combined)

All of the prior analysis was performed on problem sizes from *10* to *90* nodes (note that *90* nodes is slightly higher than the very high end of all prior HMST testing in the literature). But how can we extend the analysis to extremely large problems and find how the HMST solution impacts propagation in those problems? In order to find the answer, next we provide results on 15 very large perturbed-EC networks with sizes of *100* to *1500* nodes, generating solutions for $H = 3, 4$ and *5* for each problem. Finding HMST solutions for extremely large problems is not viable using Akgun, FGV and our 3-step improvement heuristic, but solutions can be found using Prim and Prim+Swap within a reasonable amount of time. Table A.2.8.7 provides the HMST cost and computational times along with propagation results for these extremely large networks using Prim and Prim+Swap solutions.

In Table A.2.8.7, we report results for comparative improvement between the three levels of hop constraints. The baseline for CIT improvement comparisons is the worst CIT (highest) for the same problem over the three hop constraint levels, and the baseline for cost improvement comparison is the best cost (lowest) for the same problem over the three hop constraint levels. A good solution has high improvement in CIT and low percentage increase in cost over the baselines, and a bad solution would obtain little or no improvement in CIT at substantial cost. We observe that the Prim+Swap method provides substantial improvements in terms of cost in every case. Additionally, in terms of propagation, Prim+Swap yields better solutions than Prim in terms of CIT in 22 out of 45 cases, the same CIT in 17 cases, and a worse CIT in only 6 cases. We conclude that Prim+Swap is better than Prim in terms of both cost and CIT. Larger improvement in CIT is found in 100 node problem (improvement of 2 time steps) as compared to the 1500 node problem (improvement of 1 time step). Comparing across the three hop constraint levels for the Prim+Swap method, we can see that the cost increases compared to the baseline for

all cases, but this increase is more pronounced for the smaller (thus, harder to achieve) values of hop constraint. While cost of improving the complete influence time by only one step is very high, this substantial trade-off may be well-justified in certain cases.

Table A.2.8.7. Cost and Propagation Results for Large Perturbed-EC Problems

		Complete Influence Time (CIT)			Improvement in CIT Over Baseline		Percentage Increase in Cost Over Baseline		HMST Computational Times in Seconds	
N	H	Prim	Prim+Swap	Improvement	Prim	Prim+Swap	Prim	Prim+Swap	Prim	Prim+Swap
100	3	3.0	3.0	0.0	2.0	2.0	1,760.6	345.6	1.2	7.7
	4	4.0	3.8	0.2	1.0	1.2	835.2	178.2	1.2	9.7
	5	5.0	5.0	0.0	0.0	0.0	178.9	0.0	1.2	9.6
200	3	3.0	3.0	0.0	1.2	1.2	2,249.5	292.7	9.7	61.9
	4	4.0	3.6	0.4	0.2	0.6	1,024.6	82.0	9.7	78.1
	5	4.2	4.2	0.0	0.0	0.0	753.8	0.0	9.7	60.8
300	3	3.0	3.0	0.0	1.6	1.6	4,480.5	392.3	32.8	209.6
	4	4.0	3.6	0.4	0.6	1.0	1,761.6	154.0	32.6	265.2
	5	4.2	4.6	(0.4)	0.4	0.0	666.7	0.0	32.4	321.1
400	3	3.0	3.0	0.0	1.8	1.8	4,170.4	211.1	77.9	501.4
	4	4.0	3.0	1.0	0.8	1.8	2,507.3	227.0	77.6	500.7
	5	4.8	4.4	0.4	0.0	0.4	1,187.8	0.0	77.2	495.5
500	3	3.0	3.0	0.0	1.4	1.4	7,417.9	469.9	152.2	971.8
	4	4.0	4.0	0.0	0.4	0.4	2,701.2	177.9	151.6	1,243.9
	5	4.4	4.2	0.2	0.0	0.2	1,225.9	0.0	150.9	1,241.5
600	3	3.2	3.0	0.2	1.0	1.2	7,776.2	265.1	265.1	1,715.0
	4	3.6	3.2	0.4	0.6	1.0	5,411.9	158.3	264.0	1,715.2
	5	4.2	4.0	0.2	0.0	0.2	1,682.6	0.0	261.8	2,158.7
700	3	3.2	3.0	0.2	1.0	1.2	12,849.6	430.4	420.3	2,705.9
	4	4.0	3.6	0.4	0.2	0.6	5,102.3	383.6	417.4	2,686.0
	5	4.0	4.2	(0.2)	0.2	0.0	2,326.3	0.0	415.0	3,415.1
800	3	3.0	3.0	0.0	1.2	1.2	12,577.2	452.8	628.0	4,038.9
	4	4.0	3.0	1.0	0.2	1.2	8,622.9	501.2	625.2	4,024.7
	5	4.0	4.2	(0.2)	0.2	0.0	2,752.3	0.0	622.1	4,035.5
900	3	3.2	3.2	0.0	1.4	1.4	21,007.5	691.4	896.3	5,805.7
	4	4.0	3.6	0.4	0.6	1.0	7,000.5	425.3	888.7	5,805.3
	5	4.6	4.2	0.4	0.0	0.4	1,794.7	0.0	884.6	8,917.6
1000	3	3.0	3.0	0.0	1.2	1.2	11,414.6	222.9	1,227.3	8,025.2
	4	4.0	3.0	1.0	0.2	1.2	4,553.5	120.5	1,216.7	8,034.1
	5	4.0	4.2	(0.2)	0.2	0.0	3,431.2	0.0	1,212.8	10,063.3
1100	3	3.0	3.0	0.0	1.2	1.2	12,942.3	275.9	1,632.4	10,686.9
	4	4.0	3.0	1.0	0.2	1.2	4,735.3	224.3	1,620.2	10,653.2
	5	4.0	4.2	(0.2)	0.2	0.0	3,645.8	0.0	1,615.2	10,689.9
1200	3	3.0	3.0	0.0	1.2	1.2	23,667.0	523.4	2,115.3	17,518.2
	4	4.0	3.4	0.6	0.2	0.8	10,703.3	502.6	2,102.5	14,050.7
	5	4.0	4.2	(0.2)	0.2	0.0	2,374.1	0.0	2,092.8	21,196.2
1300	3	3.0	3.0	0.0	1.0	1.0	27,203.5	649.3	2,688.8	17,469.0
	4	4.0	3.8	0.2	0.0	0.2	9,399.8	401.7	2,671.2	22,330.9
	5	4.0	4.0	0.0	0.0	0.0	1,966.2	0.0	2,658.5	22,322.8
1400	3	3.4	3.0	0.4	0.6	1.0	30,061.9	593.8	3,368.8	21,883.5
	4	4.0	3.0	1.0	0.0	1.0	9,039.9	509.2	3,337.1	22,211.5
	5	4.0	4.0	0.0	0.0	0.0	3,811.6	0.0	3,336.4	28,091.4
1500	3	3.0	3.0	0.0	1.2	1.2	21,529.4	366.8	4,135.5	26,930.1
	4	4.0	3.0	1.0	0.2	1.2	14,050.5	358.1	4,120.8	27,327.7
	5	4.2	4.0	0.2	0.0	0.2	3,852.5	0.0	4,091.1	27,250.0

Figure A.2.8.7 provides the CIT-cost graphs for Prim, Prim+Swap, and greedy and random seeding only solutions for an extremely large problem of size *N=500*. Figure A.2.8.7 illustrates that the Prim+Swap method is superior to seed-only methods in terms of both cost and CIT. Note that the initial Prim solution is only superior to the "Greedy" seed-only method for the 100 node problem. Figure A.2.8.8 provides the average CIT-number of seeds graphs for this large problem. Figure A.2.8.8 shows that similar to the case of smaller problems, greedy and random seeding-only methods require much larger initial seed sets to reach the CIT level of HMST solutions, and are thus not good solutions in this aspect.

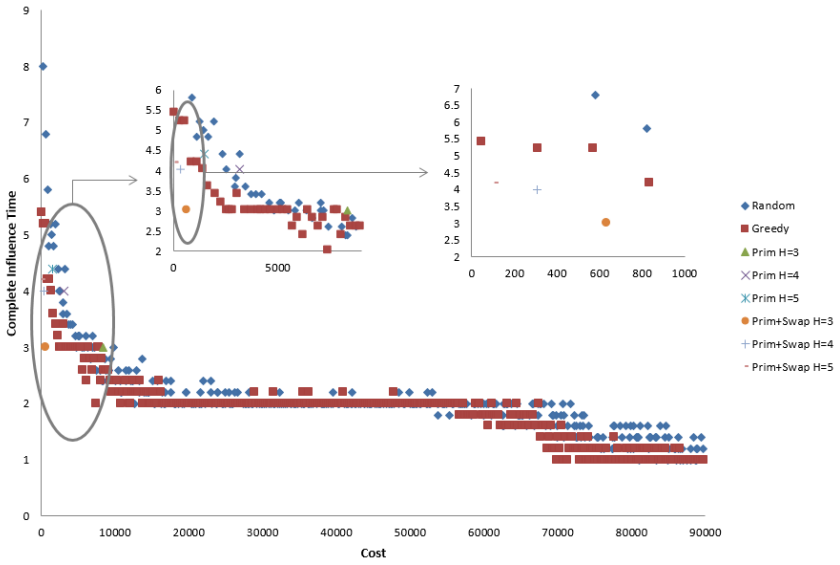


Figure A.2.8.7. Average CIT-Cost Graph for a Perturbed-EC Problem of Size N=500

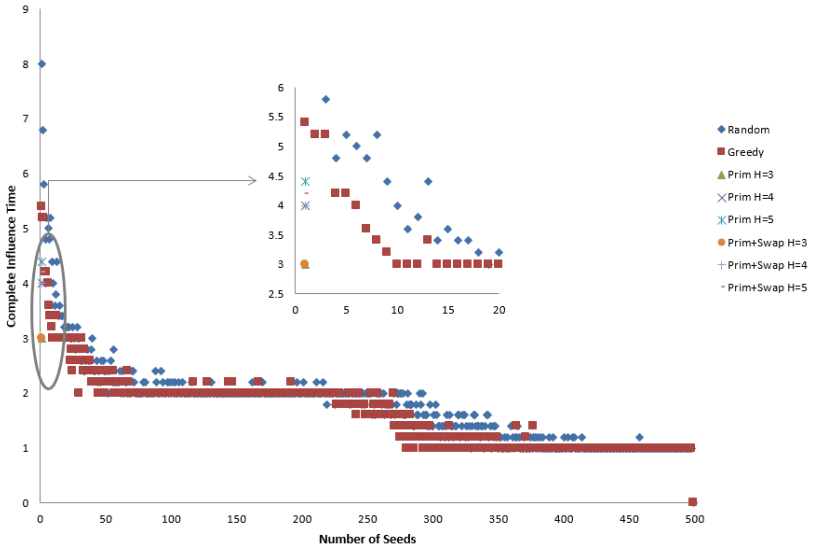


Figure A.2.8.8. Average CIT-Number of Seeds Graph for a Perturbed-EC Problem of Size N=500

Overall, our analysis on larger problems reveals the same behavior we observed for smaller problems of size *N=10* to *90* and it shows us that using the HMST solution to enhance network connections can greatly improve propagation in the network. The drawback for the large problems is that it is not easy to find HMST solutions with low cost using the algorithms that are available for smaller size problems, but slightly less cost-effective methods such as Prim and Prim+Swap can be used.

Appendix 3.1: Proofs

A.3.1.1 Proof for Lemma 3.1

The optimal publisher website royalties ${R_{W_i}}^*$, $i = 1,2$ and prices ${P_{W_i}}^*$, $i = 1,2$ satisfy the first order conditions:

$$\frac{\partial \Pi_{W_1}}{\partial R_{W_1}}\left({R_{W_1}}^*, {P_{W_1}}^*, {R_{W_2}}^*, {P_{W2}}^*\right) = \frac{\partial \Pi_{W_1}}{\partial P_{W_1}}\left({R_{W_1}}^*, {P_{W_1}}^*, {R_{W_2}}^*, {P_{W_2}}^*\right) =$$

$$\frac{\partial \Pi_{W_2}}{\partial R_{W_2}}({R_{W_1}}^*, {P_{W_1}}^*, {R_{W_2}}^*, {P_{W_2}}^*) = \frac{\partial \Pi_{W_2}}{\partial P_{W_2}}({R_{W_1}}^*, {P_{W_1}}^*, {R_{W_2}}^*, {P_{W_2}}^*) = 0 \qquad \text{(A.3.1.1)}$$

By simultaneously solving these equations, ${R_{W_1}}^* = {R_{W_2}}^* = {R_W}^*$ and ${P_{W_1}}^* = {P_{W_2}}^* = {P_W}^*$ are calculated as given in Lemma 3.1. To ensure that profit is maximized, the second order conditions must hold:

$$\frac{\partial^2 \Pi_{W_i}}{\partial {P_{W_i}}^2} = -\frac{\Phi M_U(8\Phi t - M_U M_D(R_D - 3v)(R_D - v))}{2(2\Phi t + M_U M_D v(R_D - v))^2} < 0 \qquad \text{(A.3.1.2)}$$

$$\frac{\partial^2 \Pi_{W_i}}{\partial {R_{W_i}}^2} = -\frac{M_D {M_U}^2(4\Phi t + M_U M_D v\,(R_D - v\,))(4\Phi t + M_U M_D v\,(3R_D - v\,))}{8\Phi(2\Phi t + M_U M_D v\,(R_D - v\,))^2} < 0 \qquad \text{(A.3.1.3)}$$

$$det(Hessian) = \frac{\partial^2 \Pi_{W_i}}{\partial {R_{W_i}}^2}\frac{\partial^2 \Pi_{W_i}}{\partial {P_{W_i}}^2} - \left(\frac{\partial^2 \Pi_{W_i}}{\partial P_{W_i} \partial R_{W_i}}\right)^2 = \frac{4 M_U {M_U}^3(2\Phi t + M_U M_D v(R_D - v))^2}{16(2\Phi t + M_U M_D v(R_D - v))^4} \qquad \text{(A.3.1.4)}$$

We also need the optimal number of users ${N_{U_i}}^* = {N_{U_i}}^*({R_{W_1}}^*, {P_{W_1}}^*)$ and number of third parties ${N_{D_i}}^* = {N_{D_i}}^*({R_{W_1}}^*, {P_{W_1}}^*)$ to be positive. So we need to have:

$$N_{U_i} = M_U \frac{\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D v\,(R_D - R_{W_{-i}})}{2\Phi t + M_U M_D v((R_D - R_{W_i}) + (R_D - R_{W_{-i}}))} \geq 0$$

$$\Rightarrow 2\Phi t + M_U M_D v((R_D - R_{W_i}) + (R_D - R_{W_{-i}})) \geq 0 \qquad \forall i = 1,2 \qquad \text{(A.3.1.5)}$$

$$N_{D_i} = M_D \frac{M_U(R_D - R_{W_i})(\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D v\,(R_D - R_{W_{-i}}))}{\Phi(2\Phi t + M_U M_D v((R_D - R_{W_i}) + (R_D - R_{W_{-i}})))} \geq 0$$

$$\Rightarrow R_D - R_{W_i} \geq 0 \qquad \forall i = 1,2 \qquad \text{(A.3.1.6)}$$

Throughout the chapter, we assume (A.3.1.2), (A.3.1.3), (A.3.1.5), and (A.3.1.6) to be true. (A.3.1.4) is always true. ■

A.3.1.2 Proposition 3.1

(i) By taking the derivatives of $R_W{}^*$ with respect to v and R_D we have:

$$\frac{\partial R_W{}^*}{\partial v} = \frac{1}{2} > 0 \qquad \text{(A.3.1.7)}$$

$$\frac{\partial R_W{}^*}{\partial R_D} = \frac{1}{2} > 0 \qquad \text{(A.3.1.8)}$$

It is clear from the formula for $R_W{}^*$ in Lemma 3.1 that it is independent of the other parameters. ■

(ii) We have $R_D - v > 0$ and

$$\frac{\partial P_W{}^*}{\partial v} = \frac{M_D M_U (R_D - v)}{2\Phi} > 0 \qquad \text{(A.3.1.9)}$$

$$\frac{\partial P_W{}^*}{\partial t} = 1 > 0 \qquad \text{(A.3.1.10)}$$

$$\frac{\partial P_W{}^*}{\partial R_D} = -\frac{M_D M_U (R_D - v)}{2\Phi} < 0 \qquad \text{(A.3.1.11)}$$

$$\frac{\partial P_W{}^*}{\partial M_D} = -\frac{M_U (R_D - v)^2}{4\Phi} < 0 \qquad \text{(A.3.1.12)}$$

$$\frac{\partial P_W{}^*}{\partial M_U} = -\frac{M_D (R_D - v)^2}{4\Phi} < 0 \qquad \text{(A.3.1.13)}$$

■

A.3.1.3 Proposition 3.2

Using the formula for optimal number of third parties in (3.13), we have:

$$\frac{\partial N_D^*}{\partial v} = -\frac{M_D M_U}{4\Phi} < 0 \quad \text{(A.3.1.14)}$$

$$\frac{\partial N_D^*}{\partial C_D} = -\frac{M_D M_U (R_D - v)}{4\Phi^2} < 0 \quad \text{(A.3.1.15)}$$

$$\frac{\partial N_D^*}{\partial R_D} = \frac{M_D M_U}{4\Phi} > 0 \quad \text{(A.3.1.16)}$$

$$\frac{\partial N_D^*}{\partial M_D} = \frac{M_U (R_D - v)}{4\Phi} > 0 \quad \text{(A.3.1.17)}$$

$$\frac{\partial N_D^*}{\partial M_U} = \frac{M_D (R_D - v)}{4\Phi} > 0 \quad \text{(A.3.1.18)}$$

■

A.3.1.4 Proposition 3.3

(i) Profit of each publisher website is calculated by substituting the optimal royalties and price equations (3.10) and (3.11) from Lemma 3.1 into the publisher website profit equation, and is given in equation (3.14) in Proposition 3.3. We have:

$$\frac{\partial \Pi_W^*}{\partial v} = \frac{M_D {M_U}^2 (2R_D - 3v)}{8\Phi} \quad \text{(A.3.1.19)}$$

which is positive when $v < \frac{2}{3} R_D$ and is negative when $\frac{2}{3} R_D < v$. ■

(ii) We calculate the user surplus from each publisher website as follows:

$$Z_{U_1} = \int_0^{(t+v(N_{D_2}-N_{D_1})+P_{W_2}-P_{W_1})/2t} (X - N_{D_1}\, v - P_{W_1} - ty)\, dy$$

$$Z_{U_2} = \int_{(t+v(N_{D_2}-N_{D_1})+P_{W_2}-P_{W_1})/2t}^{1} (X - N_{D_2}\, v - P_{W_2} - t(1-y))\, dy$$

$$Z_U = Z_{U_1} + Z_{U_2}$$

Solving the equation by substituting the optimal publisher website royalties and prices, we have:

$$Z_U{}^* = \frac{M_U M_D (R_D - 2v)(R_D - v) + \Phi(4X - 5t)}{4\,\Phi} \qquad \text{(A.3.1.20)}$$

Taking the derivative with respect to v we have:

$$\frac{\partial\, Z_U{}^*}{\partial\, v} = -\frac{M_U M_D (3R_D - 4v)}{4\,\Phi} \qquad \text{(A.3.1.21)}$$

which is negative when $v < \frac{3}{4}R_D$ and is positive when $\frac{3}{4}R_D < v$. ■

(iii) Here we calculate the third party surplus from each publisher website as follows:

$$Z_{D_1} = \int_0^{N_{U_1}(R_D - R_{W_1})} (N_{U_1}(R_D - R_{W_1}) - \varphi)\, d\varphi = \frac{1}{2} N_{U_1}{}^2 (R_D - R_{W_1})^2$$

$$Z_{D_2} = \int_0^{N_{U_2}(R_D - R_{W_2})} (N_{U_2}(R_D - R_{W_2}) - \varphi)\, d\varphi = \frac{1}{2} N_{U_2}{}^2 (R_D - R_{W_2})^2$$

$$Z_D = Z_{D_1} + Z_{D_2}$$

Solving the equation by substituting the optimal publisher website royalties and prices, we have:

$$Z_D{}^* = \frac{1}{16} M_U{}^2 (R_D - v)^2 \qquad \text{(A.3.1.22)}$$

Taking the derivatives, we have:

$$\frac{\partial Z_D^*}{\partial v} = -\frac{1}{8} M_U{}^2 (R_D - v) < 0 \qquad \text{(A.3.1.23)}$$

which is always negative. ■

A.3.1.5 Proposition 3.4

$$\frac{\partial R_W^*}{\partial T_{R_W}} = -\frac{R_D}{2(1+T_{R_W})^2} < 0 \qquad \text{(A.3.1.24)}$$

$$\frac{\partial R_W^*}{\partial T_{P_W}} = -\frac{v}{2(1+T_{P_W})^2} < 0 \qquad \text{(A.3.1.25)}$$

$$\frac{\partial P_W^*}{\partial T_{R_W}} = \frac{M_U M_D (R_D{}^2 (1+T_{P_W})^2 - v^2 (1+T_{R_W})^2)}{4\Phi (1+T_{P_W})^2 (1+T_{R_W})^2}$$

$$= \frac{M_U M_D (R_D (1+T_{P_W}) + v(1+T_{R_W}))(R_D (1+T_{P_W}) - v(1+T_{R_W}))}{4\Phi (1+T_{P_W})^2 (1+T_{R_W})^2} > 0 \qquad \text{(A.3.1.26)}$$

$$\frac{\partial P_W^*}{\partial T_{P_W}} = -\frac{2\Phi t (1+T_{P_W}) + M_U M_D v (R_D (1+T_{P_W}) - v(1+T_{R_W}))}{2\Phi (1+T_{P_W})^2} < 0 \qquad \text{(A.3.1.27)}$$

■

A.3.1.6 Propositions 3.5 and 3.6

(i) Using the transformations (3.17) and (3.18), the optimal profit for the website is calculated as:

$$\Pi_W{}^* = \frac{M_U (8\,\Phi\, t\, (1+T_{P_W})(1+T_{R_W}) - M_U M_D (R_D (1+T_{P_W}) - 3v(1+T_{R_W}))(R_D (1+T_{P_W}) - v(1+T_{R_W})))}{16\,\Phi (1+T_{P_W})^2 (1+T_{R_W})} \qquad \text{(A.3.1.28)}$$

Taking the derivative of profit with respect to the taxations we have:

$$\frac{\partial \Pi_{W_i}{}^*}{\partial T_{R_W}} = \frac{M_U{}^2 M_D (R_D{}^2 (1+T_{P_W})^2 - 3v^2 (1+T_{R_W})^2)}{16\,\Phi (1+T_{P_W})^2 (1+T_{R_W})^2} \qquad \text{(A.3.1.29)}$$

which is positive when $v < \frac{R_D}{\sqrt{3}} \frac{1+T_{P_W}}{1+T_{R_W}}$, and is negative when $\frac{R_D}{\sqrt{3}} \frac{1+T_{P_W}}{1+T_{R_W}} < v$.

$$\frac{\partial \Pi_{W_i}^*}{\partial T_{P_W}} = -\frac{M_U(4\Phi t(1+T_{P_W})+M_U M_D v(2R_D(1+T_{P_W})-3v(1+T_{R_W}))}{8\Phi(1+T_{P_W})^3} < 0 \quad \text{(A.3.1.30)}$$

(ii) User surplus when taxations are possible is calculated as

$$Z_U^* = \frac{\Phi(4X-5t)(1+T_{P_W})(1+T_{R_W})+M_U M_D(R_D(1+T_{P_W})-2v(1+T_{R_W}))(R_D(1+T_{P_W})-v(1+T_{R_W}))}{4\,\Phi(1+T_{P_W})(1+T_{R_W})} \quad \text{(A.3.1.31)}$$

and we have:

$$\frac{\partial Z_U^*}{\partial T_{R_W}} = -\frac{M_D M_U(R_D{}^2(1+T_{P_W})^2-2v^2(1+T_{R_W})^2)}{4\,\Phi(1+T_{P_W})\,(1+T_{R_W})^2} \quad \text{(A.3.1.32)}$$

which is negative when $v < \frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}}$ and is positive when $\frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}} < v$.

$$\frac{\partial Z_U^*}{\partial T_{P_W}} = \frac{M_D M_U(R_D{}^2(1+T_{P_W})^2-2v^2(1+T_{R_W})^2)}{4\,\Phi\,(1+T_{P_W})^2\,(1+T_{R_W})} \quad \text{(A.3.1.33)}$$

which is positive when $v < \frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}}$ and is negative when $\frac{R_D}{\sqrt{2}} \frac{1+T_{P_W}}{1+T_{R_W}} < v$.

(iii) Third party surplus when taxations are present is calculated as

$$Z_D^* = \frac{M_U{}^2}{16}(R_D(1+T_{P_W}) - v(1+T_{R_W}))^2 \quad \text{(A.3.1.34)}$$

and we have:

$$\frac{\partial Z_D^*}{\partial T_{R_W}} = -\frac{M_U{}^2\, v\,(R_D(1+T_{P_W})-v(1+T_{R_W}))}{8\,(1+T_{P_W})^2} < 0 \quad \text{(A.3.1.35)}$$

$$\frac{\partial Z_D^*}{\partial T_{P_W}} = \frac{M_U{}^2(1+T_{R_W})\, v\,(R_D(1+T_{P_W})-v(1+T_{R_W}))}{8\,(1+T_{P_W})^3} > 0 \quad \text{(A.3.1.36)}$$

■

Appendix 3.2: Extension of Proposition 3.3

In the body of the chapter, we have analyzed the effect of privacy concerns on publisher website profit, third party surplus, and user surplus in Proposition 3.3. Here, we expand the analysis to consider other model parameters. Propositions A.3.2.1, A.3.2.2, and A.3.2.3 provide these results. We do not provide the proofs for these propositions, are straightforward and can be calculated by taking the derivatives for equations (3.14), (3.15), and (3.16) for optimal website profit, user surplus, and third party surplus, respectively.

Proposition A.3.2.1: Effect of parameters on publisher website profit

(i) When $v < \frac{1}{2}R_D$ profit of each publisher website ($\Pi_W{}^*$) decreases with third party revenue from user information (R_D) and when $\frac{1}{2}R_D < v < R_D$ it increases with third party revenue from user information (R_D).

(ii) When $v < \frac{1}{3}R_D$ profit of each publisher website ($\Pi_W{}^*$) increases with third party costs (Φ) and when $\frac{1}{3}R_D < v < R_D$ it decreases with third party costs (Φ).

(iii) Profit of each publisher website ($\Pi_W{}^*$) increases with differentiation between two publisher websites (t).

(iv) Profit of each publisher website ($\Pi_W{}^*$) increases with total number of potential users in the market (M_U).

(v) When $v < \frac{1}{3}R_D$ profit of each publisher website ($\Pi_W{}^*$) decreases with total number of potential third parties in the market (M_D) and when $\frac{1}{3}R_D < v < R_D$ it increases with total number of potential third parties in the market (M_D).

Figure A.3.2.1 summarizes the Proposition A.3.2.1.

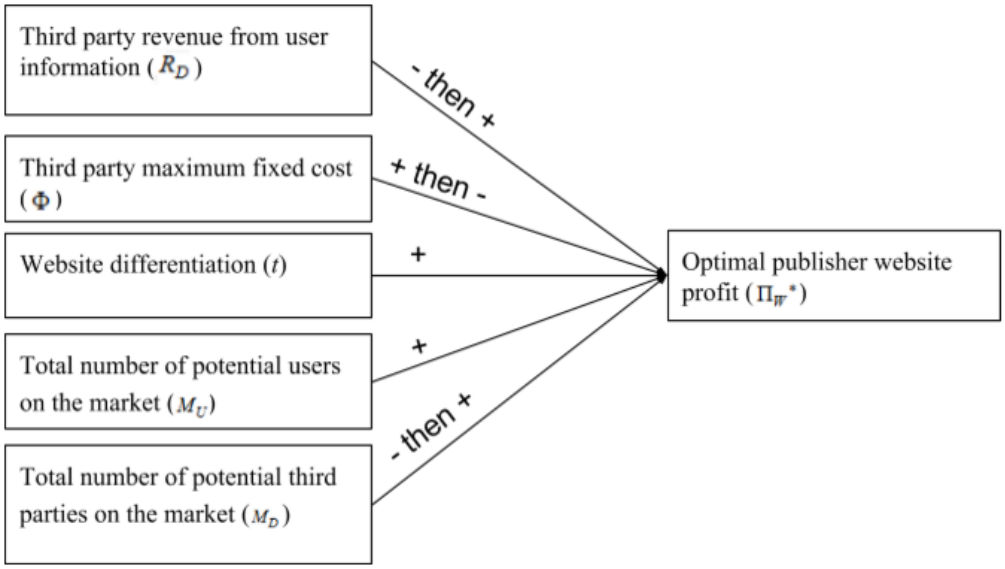


Figure A.3.2.1. Effect of Model Parameters on Publisher Website Profit

Proposition A.3.2.2: Effect of parameters on user surplus

(i) When $v < \frac{2}{3}R_D$ *user surplus* (Z_U^*) *increases with third party revenue from user information* (R_D) *and when* $\frac{2}{3}R_D < v < R_D$ *it decreases with third party revenue from user information* (R_D).

(ii) When $v < \frac{1}{2}R_D$ *user surplus* (Z_U^*) *decreases with third party fixed costs* (Φ) *and when* $\frac{1}{2}R_D < v < R_D$ *it increases with third party fixed costs* (Φ).

(iii) User surplus (Z_U^*) *decreases with publisher website differentiation* (t).

(iv) When $v < \frac{1}{2}R_D$ *user surplus* (Z_U^*) *increases with total number of users in the market* (M_U) *and total number of third parties in the market* (M_D), *and when* $\frac{1}{2}R_D < v < R_D$ *it decreases*

with total number of users in the market (M_U) and total number of third parties in the market (M_D).

Figure A.3.2.2 summarizes Proposition A.3.2.2.

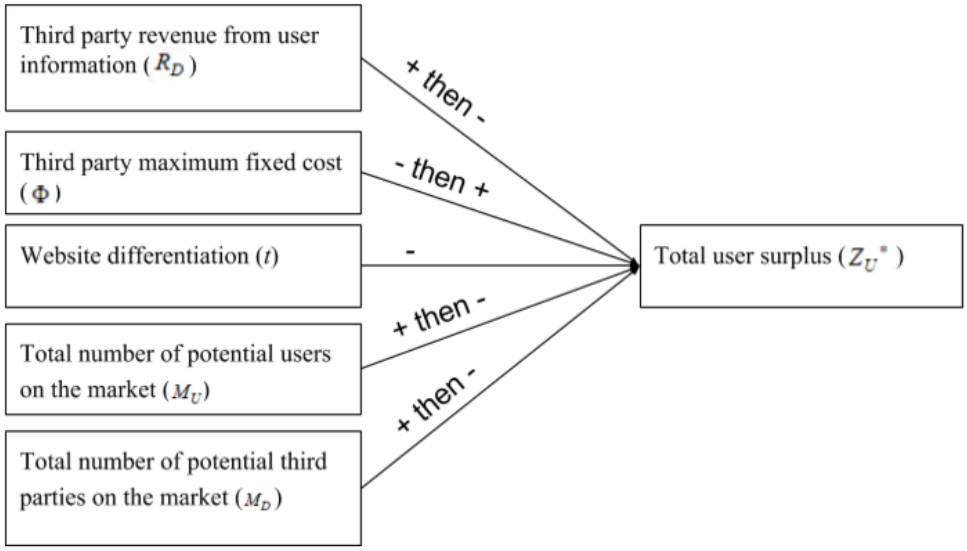


Figure A.3.2.2. Effect of Model Parameters on User Surplus

Proposition A.3.2.3: Effect of parameters on third parties

(i) Third party surplus (Z_D^*) increases with third party revenue from user information (R_D)

(ii) Third party surplus (Z_D^*) increases with total number of users in the market (M_U)

Figure A.3.2.3 summarizes Proposition A.3.2.3.

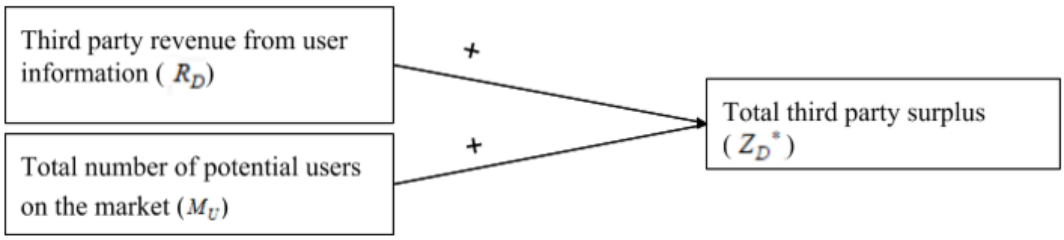


Figure A.3.2.3. Effect of Model Parameters on Third Party Surplus

Appendix 3.3: Asymmetric Model

In the asymmetric model, the two firms are asymmetric in terms of user privacy concerns. The user utility for websites in this case is as follows:

$$U_1(z) = u_1 - ty, \qquad u_1 = X - N_{D_1}v_1 - P_{W_1} \tag{A.3.3.1}$$

$$U_2(z) = u_2 - t(1-y), \qquad u_2 = X - N_{D_2}v_2 - P_{W_2} \tag{A.3.3.2}$$

The user who is indifferent between websites 1 and 2 is calculated as:

$$u_1 - t\hat{y} = u_2 - t(1-\hat{y}) \Rightarrow \hat{y} = \frac{t+(N_{D_2}v_2 - N_{D_1}v_1) + (P_{W_2} - P_{W_1})}{2t} \tag{A.3.3.3}$$

and the number of users for each publisher website i is calculated as:

$$N_{U_i} = M_U \frac{t+(N_{D_{-i}}v_{-i} - N_{D_i}v_i) + (P_{W_{-i}} - P_{W_i})}{2t} > 0 \tag{A.3.3.4}$$

The third party profit and website profit equations as well as the equation for number of third parties in this case are similar to the base model. The number of users and third parties with respect to the parameters are calculated as:

$$N_{U_i} = M_U \frac{\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D\, v_{-i}\,(R_D - R_{W_{-i}})}{2\Phi t + M_U M_D((R_D - R_{W_i})v_i + (R_D - R_{W_{-i}})v_{-i})} \tag{A.3.3.5}$$

$$N_{D_i} = M_D \frac{M_U(R_D - R_{W_i})(\Phi t + \Phi(P_{W_{-i}} - P_{W_i}) + M_U M_D\, v_{-i}\,(R_D - R_{W_{-i}}))}{\Phi(2\Phi t + M_U M_D((R_D - R_{W_i})v_i + (R_D - R_{W_{-i}})v_{-i}))} \tag{A.3.3.6}$$

Using these equations along with the website profit function, the optimal royalties and prices of the websites can be calculated as follows:

$$R_{W_i}{}^* = \frac{R_D + v_i}{2} \tag{A.3.3.7}$$

$$P_{W_i}{}^{*} = \frac{(2\Phi t+2\Phi P_{W_{-i}}+M_D M_U\, v_{-i}\,(R_D-v_{-i}))(4\Phi t+M_D M_U\,(R_D(v_i+v_{-i})-v_{-i}{}^2-R_D{}^2))}{2\Phi(8\Phi t+M_D M_U(2R_D(v_i+v_{-i})-2v_{-i}{}^2-v_i{}^2-R_D{}^2))} \geq 0 \qquad \text{(A.3.3.8)}$$

Note that the website prices are calculated based on the price from the other website, and the equilibrium price in analytically intractable when $P_{W_i}{}^{*}$ can differ from $P_{W_{-i}}{}^{*}$. It is clear from (A.3.3.8) that the publisher website i's price is nonlinear in v_i and v_{-i}. The results from the numerical analysis are provided in the body of the chapter.

Appendix 3.4: Effect of Privacy Concerns on Market Concentration

In studying the third party market concentration, we consider two cases: 1. third parties with homogenous shares of the market, and 2. third parties with non-homogenous shares of the market. We use the Herfindahl-Hirschman Index (HHI) as a recognized measure for market concentration. The HHI is generically calculated as follows:

$$HHI = \sum_{j=1}^{N_D} s_j^2 \quad \text{(A.3.4.1)}$$

where s_j is the market share of j^{th} third party.

A.3.4.1 Third parties with Homogenous Market Shares

In the symmetric duopoly model, because the publisher websites set identical royalties and prices, the third parties either participate in both publisher websites, or do not participate at all. In the homogeneous market share case, the total number of third parties on a particular publisher website i is $N_D = N_{D_1} = N_{D_2}$. When all the third parties have equal share of the market, the market share of each third party j is simply calculated as $s_j = 1/N_D$. The HHI is then calculated as:

$$HHI = \sum_{j=1}^{N_D}(1/N_D)^2 = N_D(1/N_D)^2 = 1/N_D \quad \text{(A.3.4.2)}$$

By inserting the optimal number of third parties from Proposition 3.2, we have:

$$HHI = 1/N_D = 1/[M_D \frac{M_U(R_D - v)}{4\Phi}] = \frac{4\Phi}{M_D M_U(R_D - v)} \quad \text{(A.3.4.3)}$$

It can be seen that HHI is increasing in v. In other words, the market concentration is increasing in the user privacy concerns.

To include the effect of barriers to entry, we rewrite the total number of potential third parties, M_D to be as M_D/B, where B is the level of barrier. This means that higher barriers will reduce the number of potential third parties. We can rewrite the HHI formula as:

$$HHI = 1/N_D = 1/[(M_D/B)\frac{M_U(R_D-v)}{4\Phi}] = B\frac{4\Phi}{M_D M_U(R_D-v)} \quad \text{(A.3.4.4)}$$

It can be seen that HHI is increasing in the entry barrier level, so the market concentration is increasing in the level of barrier to entry. The level of barrier to entry is higher for the third parties that operate in areas with high privacy concerns and high information sensitivity. In practice, privacy is one reason that third parties need to invest more in information technology (IT) security. These IT investments lead to higher sunk cost of entry, and are a major barrier to entry.

A.3.4.2 Third parties with Non-Homogeneous Market Shares

While previously we assumed the market shares to be homogenous for all third parties, this is not realistic in most cases. It results in a market concentration measure that is only dependent on the number of third parties utilized by the publisher websites. We now reconsider the asymmetric model described in Section 3.4.2, where v_1 varies and v_2 is held constant, using the number of third parties for the two publisher websites to calculate the third party market shares.

Let the number of third parties on publisher websites 1 and 2 be N_{D_1} and N_{D_2}, respectively. Note that since the third parties are differentiated only based on their costs, if a third party participates on the publisher website with higher privacy concern (and higher royalty), then it will also participate on the publisher website with lower privacy concerns (and lower royalty). Thus, there are a total of $Max\{N_{D_1}, N_{D_2}\}$ unique third parties active in the market. Out of these

third parties, $Min\{N_{D_1}, N_{D_2}\}$ of them participate in both publisher websites, and the rest, $Max\{N_{D_1}, N_{D_2}\} - Min\{N_{D_1}, N_{D_2}\}$ of them participate in only one publisher website (the one with lower privacy concerns). Let J_1 be the set of third parties who participate in only one publisher website, and J_2 be the set of third parties who participate in both publisher websites, where $J_1 \cap J_2 = \emptyset$. The size of J_1 is $|J_1| = Max\{N_{D_1}, N_{D_2}\} - Min\{N_{D_1}, N_{D_2}\}$, the size of J_2 is $|J_2| = Min\{N_{D_1}, N_{D_2}\}$, and $J_1 \cup J_2$ is the set of all third parties which has a size of $|J_1 \cup J_2| = Max\{N_{D_1}, N_{D_2}\}$. Third parties that are present on both publisher websites have a market size that is twice as much as those that participate in only one publisher website. For simplicity and without loss of generality, we assume the following market sizes for each third party. The market size of each third party j (q_j) depends on how many publisher websites they serve:

$$q_j = 1 \quad \forall j \in J_1 \tag{A.3.4.5}$$

$$q_j = 2 \quad \forall j \in J_2 \tag{A.3.4.6}$$

Let S be the total market size, which is calculated as the sum of relative market share for all third parties. We have:

$$S = \sum_{j \in J_1, J_2} q_j = \sum_{j \in J_1} 1 + \sum_{j \in J_2} 2 \tag{A.3.4.7}$$

The market share of each third party (s_j) is calculated as the ratio of their market size to the total market size, that is:

$$s_j = \frac{1}{S} \quad \forall j \in J_1 \tag{A.3.4.8}$$

$$s_j = \frac{2}{S} \quad \forall j \in J_2 \tag{A.3.4.9}$$

Now that the total market size and share of each third party is known, we calculate the HHI as follows:

$$HHI = \sum_{j \in J_1, J_2} {s_j}^2 = \sum_{j \in J_1} (\frac{1}{S})^2 + \sum_{j \in J_2} (\frac{2}{S})^2$$

$$= (Max\{N_{D_1}, N_{D_2}\} - Min\{N_{D_1}, N_{D_2}\})(\frac{1}{S})^2 + Min\{N_{D_1}, N_{D_2}\}(\frac{2}{S})^2 \qquad \text{(A.3.4.10)}$$

which can be calculated as

$$HHI = \frac{Max\{N_{D_1}, N_{D_2}\} + 3Min\{N_{D_1}, N_{D_2}\}}{(Max\{N_{D_1}, N_{D_2}\} + Min\{N_{D_1}, N_{D_2}\})^2} \qquad \text{(A.3.4.11)}$$

Without loss of generality, let's assume that $N_{D_2} > N_{D_1}$. It can be shown that HHI will be maximized when ${N_{D_1}}^{Max} = \frac{N_{D_2}}{3}$. Figure A.3.4.1 provides the effect of change in number of third parties on HHI values for a numerical example.

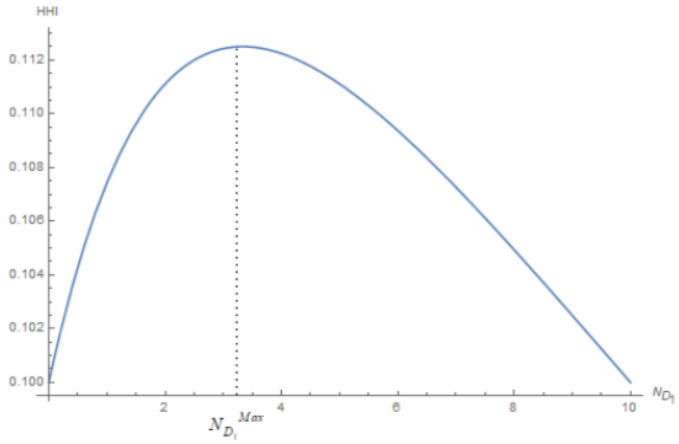


Figure A.3.4.1. HHI Values with respect to N_{D_1} when $N_{D_2} = 10$

It can be seen in the example that the HHI is not maximized where $v_1 = v_2$, where the two number of third parties are equal ($N_{D_1} = N_{D_2} = 10$), but at ${N_{D_1}}^{Max} = \frac{10}{3} \simeq 3.33$. Thus market concentration is at its highest when the number of third parties in two publisher websites are

different from each other. Next we will see how this factor directs the way market concentration is affected by user privacy concerns.

As we saw in Figure 3.5.d, $N_{D_1}{}^*$ decreases as user privacy concerns for publisher website 1 increase. Even though website 2's user privacy concerns is held steady at $v_2 = 4$, $N_{D_2}{}^*$ will be affected by changes in v_1. Here we study how these changes in number of third parties determines the market concentration. Figure A.3.4.2 provides the effect of user privacy concerns of one publisher website on HHI when the user privacy concerns of the other publisher website is fixed.

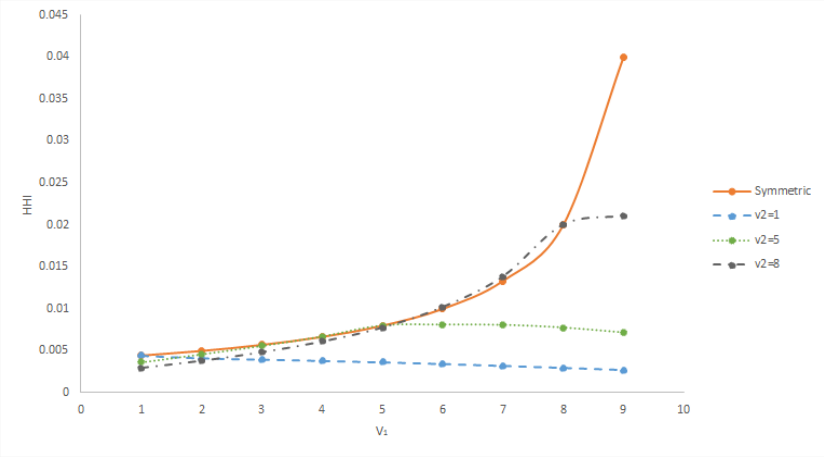


Figure A.3.4.2. HHI with respect to v_1 when v_2 is Fixed

The orange line in Figure A.3.4.2 represents the homogeneous case where both publisher websites are symmetric, and $N_{D_1} = N_{D_2}$. The other lines represent asymmetric cases where v_2 is fixed while $v_1$varies. By comparing the homogeneous case to any non-homogeneous case, it can be seen that the HHI is initially higher for the symmetric case (or equal to in when $v_2 = 1$). As v_1 increases while less than v_2, then market concentration for the asymmetric cases can become higher than the symmetric HHI. At $v_1 = v_2$ the two lines cross, for $v_1 > v_2$, again the HHI for the symmetric case is higher than the asymmetric case.

Appendix 3.5: Collusion

The calculations for collusion are provided for the case where one of the publisher websites sets the equilibrium royalties ($R_{W_{Eq.}}$) and the other sets royalties to R_W. The profit in this case is maximized when the the royalty is set to its equilibrium point, $R_{W_{Eq.}}$. However, if the publisher websites can collude and set identical royalty ($R_{W_{Col.}}$), then they can increase their profits to the collusion equilibrium ($\Pi_{W_{R_{W_{Col.}}}}$), where both firms make higher profits. From the Lemma 3.1, we know $R_{W_{Eq.}} = R_W{}^* = \frac{R_D+v}{2}$ and the equilibrium profit (Proposition 3.3) is given as:

$$\Pi_{W_{R_{W_{Eq.}}}} = \Pi_W{}^* = \frac{M_U(8\,\Phi\,t - M_U M_D(R_D-3v)(R_D-v))}{16\,\Phi} \quad \text{(A.3.5.1)}$$

For the collusion case, in the publisher website profit equation (3.9) the prices are set to their optimal values, and both publisher websites' royalties are set to R_W. The profit is calculated as:

$$\Pi_{W_{R_W}} = \frac{M_U(4\,\Phi\,t - M_U M_D({R_D}^2+2{R_W}^2+v^2-2R_D(R_W+v))}{8\,\Phi} \quad \text{(A.3.5.2)}$$

which is maximized at $R_{W_{Col.}} = \frac{R_D}{2}$ for which the profit is

$$\Pi_{W_{R_{W_{Col.}}}} = \frac{M_U(8\,\Phi\,t - M_U M_D({R_D}^2-4\,R_D v + 2\,v^2))}{16\,\Phi} \quad \text{(A.3.5.3)}$$

It can easily be shown using the formulae above that the phenomena that collusion royalties are lower than equilibrium royalties and collusion profits are higher than equilibrium profits are analytical results and hold irrespective of the parameters. In other words, the following hold:

$$R_{W_{Eq.}} > R_{W_{Col.}} \quad \text{(A.3.5.4)}$$

$$\Pi_{W_{R_{W_{Eq.}}}} < \Pi_{W_{R_{W_{Col.}}}} \quad \text{(A.3.5.5)}$$

This collusion results in setting lower royalties overall, and thus is also beneficial for the third parties, as presented in the third party surplus curve in A.3.5.1.

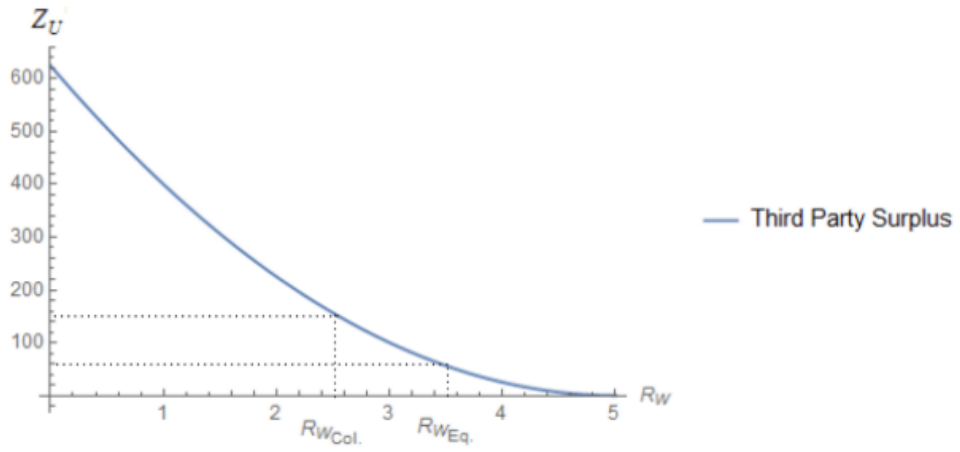


Figure A.3.5.1. Third Party Surplus with and without Collusion with respect to Royalties

When collusion is possible, the following formula provides the effect of R_W on the user surplus when prices are set to their equilibrium values (we do not consider firms colluding on price but rather only royalties), and publisher websites set equal royalties (CS_{R_W}):

$$CS_{R_W} = \frac{M_D M_U ({R_D}^2 - 4R_D\, v + v(2\, R_W + v)) + \Phi(4X - 5t)}{4\,\Phi} \quad \text{(A.3.5.6)}$$

Taking the partial derivative of the user surplus with respect to R_W we have:

$$\frac{\partial\, CS_{R_W}}{\partial\, R_W} = \frac{M_U M_D\, v}{2\,\Phi} \quad \text{(A.3.5.7)}$$

The collusion is thus not beneficial for the users, as they will be exposed to more third parties due to decrease in R_W. This can be seen for a numerical example in Figure A.3.5.2.

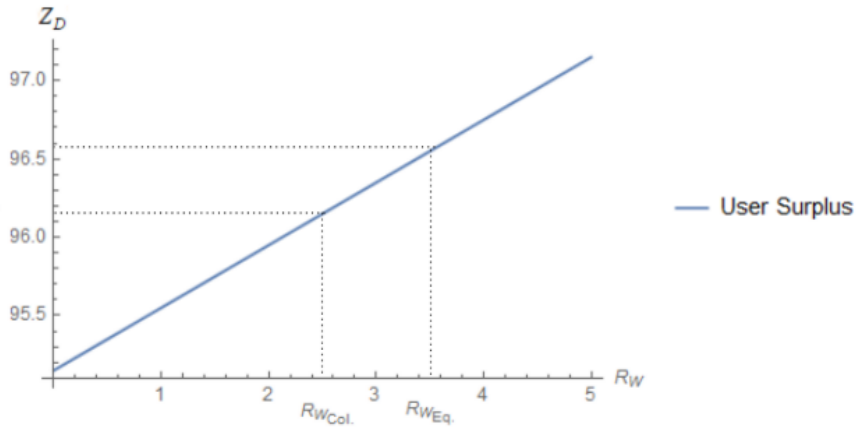


Figure A.3.5.2. User Surplus with and without Collusion with respect to Royalties

Appendix 3.6: Duopoly with Nonlinear Utility Function

For the duopoly with nonlinear utility function (NL Duopoly), the transformations (3.23) and (3.24) are made in the base model. We then analyze the behavior of the model variables with respect to the different variables. The tables A.3.6.1, A.3.6.2, and A.3.6.3 present the comparison of the behavior of parameters and variables between the base duopoly model versus the duopoly model with nonlinear utility function.

In Table A.3.6.1, it can be seen that while the behavior of some of the parameters are different in the duopoly model with nonlinear utility compared to the base model, the main results of the model in terms of user privacy concerns (v) are consistent with the base model. In Table A.3.6.2, we can see that the behavior of the number of users and third parties are entirely consistent between the two duopoly models. As described in Table A.3.6.3, the NL Duopoly model mostly picks up the effect of higher range user privacy concerns seen in the duopoly model for the publisher website profit. While we see some discrepancy among the two models, the overall conclusion is that the results for the duopoly and NL duopoly models are consistent. This is especially true for the key results with respect to user privacy concerns (v).

Table A.3.6.1. Publisher Website Decision Variables

	Changes With Respect To	Duopoly	NL Duopoly
Publisher Website Royalty R_W^*	v $(\frac{\partial R_W^*}{\partial v})$	+	+
	R_D $(\frac{\partial R_W^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial R_W^*}{\partial \Phi})$	Independent	-
	M_U $(\frac{\partial R_W^*}{\partial M_U})$	Independent	+
	M_D $(\frac{\partial R_W^*}{\partial M_D})$	Independent	+
	t $(\frac{\partial R_W^*}{\partial t})$	Independent	Independent
	T_{R_W} $(\frac{\partial R_W^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial R_W^*}{\partial T_{P_W}})$	-	-
Publisher Website Price P_W^*	v $(\frac{\partial P_W^*}{\partial v})$	+	+
	R_D $(\frac{\partial P_W^*}{\partial R_D})$	-	-
	Φ $(\frac{\partial P_W^*}{\partial \Phi})$	+	-
	M_U $(\frac{\partial P_W^*}{\partial M_U})$	-	+
	M_D $(\frac{\partial P_W^*}{\partial M_D})$	-	+
	t $(\frac{\partial P_W^*}{\partial t})$	+	+
	T_{R_W} $(\frac{\partial P_W^*}{\partial T_{R_W}})$	+	+
	T_{P_W} $(\frac{\partial P_W^*}{\partial T_{P_W}})$	-	-

Table A.3.6.2. Impact on Number of Users and Third Parties

	Changes With Respect To	Duopoly	NL Duopoly
Number of Users N_U^*	v $(\frac{\partial N_U^*}{\partial v})$	Independent	Independent
	R_D $(\frac{\partial N_U^*}{\partial R_D})$	Independent	Independent
	Φ $(\frac{\partial N_U^*}{\partial \Phi})$	Independent	Independent
	M_U $(\frac{\partial N_U^*}{\partial M_U})$	Independent	Independent
	M_D $(\frac{\partial N_U^*}{\partial M_D})$	Independent	Independent
	t $(\frac{\partial N_U^*}{\partial t})$	Independent	Independent
	T_{R_W} $(\frac{\partial N_U^*}{\partial T_{R_W}})$	Independent	Independent
	T_{P_W} $(\frac{\partial N_U^*}{\partial T_{P_W}})$	Independent	Independent
Number of Third Parties N_D^*	v $(\frac{\partial N_D^*}{\partial v})$	-	-
	R_D $(\frac{\partial N_D^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial N_D^*}{\partial \Phi})$	-	-
	M_U $(\frac{\partial N_D^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial N_D^*}{\partial M_D})$	+	+
	t $(\frac{\partial N_D^*}{\partial t})$	Independent	Independent
	T_{R_W} $(\frac{\partial N_D^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial N_D^*}{\partial T_{P_W}})$	+	+

Table A.3.6.3. Publisher Website Profit, User Surplus, and Third Party Surplus

	Changes With Respect To	**Duopoly**	**NL Duopoly**
Publisher Website Profit Π_W^*	v $(\frac{\partial \Pi_W^*}{\partial v})$	+ then -	-
	R_D $(\frac{\partial \Pi_W^*}{\partial R_D})$	- for low v + for high v	+
	Φ $(\frac{\partial \Pi_W^*}{\partial \Phi})$	+ for low v - for high v	-
	M_U $(\frac{\partial \Pi_W^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial \Pi_W^*}{\partial M_D})$	- for low v + for high v	+
	t $(\frac{\partial \Pi_W^*}{\partial t})$	+	+
	T_{R_W} $(\frac{\partial \Pi_W^*}{\partial T_{R_W}})$	+ for low v - for high v	-
	T_{P_W} $(\frac{\partial \Pi_W^*}{\partial T_{P_W}})$	-	-
User Surplus Z_U^*	v $(\frac{\partial Z_U^*}{\partial v})$	- then +	+
	R_D $(\frac{\partial Z_U^*}{\partial R_D})$	+ for low v - for high v	-
	Φ $(\frac{\partial Z_U^*}{\partial \Phi})$	- for low v + for high v	+
	M_U $(\frac{\partial Z_U^*}{\partial M_U})$	+ for low v - for high v	+
	M_D $(\frac{\partial Z_U^*}{\partial M_D})$	+ for low v - for high v	-
	t $(\frac{\partial Z_U^*}{\partial t})$	-	-
	T_{R_W} $(\frac{\partial Z_U^*}{\partial T_{R_W}})$	- for low v + for high v	+

	T_{P_W} $(\frac{\partial Z_U^*}{\partial T_{P_W}})$	+ for low v - for high v	-
Third Party SurplusZ_D^*	v $(\frac{\partial Z_D^*}{\partial v})$	-	-
	R_D $(\frac{\partial Z_D^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial Z_D^*}{\partial \Phi})$	Independent	+
	M_U $(\frac{\partial Z_D^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial Z_D^*}{\partial M_D})$	Independent	-
	t $(\frac{\partial Z_D^*}{\partial t})$	Independent	-
	T_{R_W} $(\frac{\partial Z_D^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial Z_D^*}{\partial T_{P_W}})$	+	+

Appendix 3.7: Monopoly Model

The following table compares the effect of model parameters on the key variables in the model, as well as on the publisher website profit, user and third party surplus. Tables A.3.7.1, A.3.7.2, and A.3.7.3 present the comparison of the behavior of parameters and variables between the base duopoly model versus the monopoly model. In Table A.3.7.1, it can be seen that the decision variables of royalties and prices behave similarly in the monopoly and duopoly models. Addition of the Hotelling's parameter in the duopoly model enables us to see the effect of competition on the prices. The higher the differentiation between the two publisher websites (higher t), the higher the prices. In other words, competition would decrease the prices for the publisher websites.

In Table A.3.7.2, we can see that the behavior of the number of users in the monopoly model is different from the duopoly model, because the key assumption in the duopoly model is that the market is covered. Thus, the number of users in the duopoly model is independent of the parameters. For the number of third parties, we see that the behavior of the monopoly and duopoly models are similar. In Table A.3.7.3, the publisher website profit, user surplus, and third party surplus are presented. The monopoly model picks up the effect of higher range user privacy concerns seen in the duopoly model for the publisher website profit. For user surplus, the monopoly model picks up the effect of the lower range of user privacy concerns seen in the duopoly model. While we see two different effects in the duopoly model, the pattern of results is consistent between the two models. Thus, our overall conclusion is that the results for the duopoly and monopoly models are not inconsistent. This is especially true for the key results with respect to user privacy concerns (v).

Table A.3.7.1. Publisher Website Decision Variables

	Changes With Respect To	Duopoly	Monopoly
Publisher Website Royalty R_W^*	v $(\frac{\partial R_W^*}{\partial v})$	+	+
	R_D $(\frac{\partial R_W^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial R_W^*}{\partial \Phi})$	Independent	Independent
	M_U $(\frac{\partial R_W^*}{\partial M_U})$	Independent	Independent
	M_D $(\frac{\partial R_W^*}{\partial M_D})$	Independent	Independent
	t $(\frac{\partial R_W^*}{\partial t})$	Independent	N/A
	T_{R_W} $(\frac{\partial R_W^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial R_W^*}{\partial T_{P_W}})$	-	-
Publisher Website Price P_W^*	v $(\frac{\partial P_W^*}{\partial v})$	+	+
	R_D $(\frac{\partial P_W^*}{\partial R_D})$	-	-
	Φ $(\frac{\partial P_W^*}{\partial \Phi})$	+	+
	M_U $(\frac{\partial P_W^*}{\partial M_U})$	-	-
	M_D $(\frac{\partial P_W^*}{\partial M_D})$	-	-
	t $(\frac{\partial P_W^*}{\partial t})$	+	N/A
	T_{R_W} $(\frac{\partial P_W^*}{\partial T_{R_W}})$	+	+
	T_{P_W} $(\frac{\partial P_W^*}{\partial T_{P_W}})$	-	-

Table A.3.7.2. Impact on Number of Users and Third Parties

	Changes With Respect To	Duopoly	Monopoly
Number of Users N_U^*	v $(\frac{\partial N_U^*}{\partial v})$	Independent	-
	R_D $(\frac{\partial N_U^*}{\partial R_D})$	Independent	+
	Φ $(\frac{\partial N_U^*}{\partial \Phi})$	Independent	-
	M_U $(\frac{\partial N_U^*}{\partial M_U})$	Independent	+
	M_D $(\frac{\partial N_U^*}{\partial M_D})$	Independent	+
	t $(\frac{\partial N_U^*}{\partial t})$	Independent	N/A
	T_{R_W} $(\frac{\partial N_U^*}{\partial T_{R_W}})$	Independent	-
	T_{P_W} $(\frac{\partial N_U^*}{\partial T_{P_W}})$	Independent	+
Number of Third Parties N_D^*	v $(\frac{\partial N_D^*}{\partial v})$	-	-
	R_D $(\frac{\partial N_D^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial N_D^*}{\partial \Phi})$	-	-
	M_U $(\frac{\partial N_D^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial N_D^*}{\partial M_D})$	+	+
	t $(\frac{\partial N_D^*}{\partial t})$	Independent	N/A
	T_{R_W} $(\frac{\partial N_D^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial N_D^*}{\partial T_{P_W}})$	+	+

Table A.3.7.3. Publisher Website Profit, User Surplus, and Third Party Surplus

	Changes With Respect To	Duopoly	Monopoly
Publisher Website Profit Π_W^*	v $(\frac{\partial \Pi_W^*}{\partial v})$	+ then -	-
	R_D $(\frac{\partial \Pi_W^*}{\partial R_D})$	- for low v + for high v	+
	Φ $(\frac{\partial \Pi_W^*}{\partial \Phi})$	+ for low v - for high v	-
	M_U $(\frac{\partial \Pi_W^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial \Pi_W^*}{\partial M_D})$	- for low v + for high v	+
	t $(\frac{\partial \Pi_W^*}{\partial t})$	+	N/A
	T_{R_W} $(\frac{\partial \Pi_W^*}{\partial T_{R_W}})$	+ for low v - for high v	-
	T_{P_W} $(\frac{\partial \Pi_W^*}{\partial T_{P_W}})$	-	+
User Surplus Z_U^*	v $(\frac{\partial Z_U^*}{\partial v})$	- then +	-
	R_D $(\frac{\partial Z_U^*}{\partial R_D})$	+ for low v - for high v	+
	Φ $(\frac{\partial Z_U^*}{\partial \Phi})$	- for low v + for high v	-
	M_U $(\frac{\partial Z_U^*}{\partial M_U})$	+ for low v - for high v	+
	M_D $(\frac{\partial Z_U^*}{\partial M_D})$	+ for low v - for high v	+
	t $(\frac{\partial Z_U^*}{\partial t})$	-	N/A
	T_{R_W} $(\frac{\partial Z_U^*}{\partial T_{R_W}})$	- for low v + for high v	-

	T_{P_W} $(\frac{\partial Z_U^*}{\partial T_{P_W}})$	+ for low v - for high v	+
Third Party SurplusZ_D^*	v $(\frac{\partial Z_D^*}{\partial v})$	-	-
	R_D $(\frac{\partial Z_D^*}{\partial R_D})$	+	+
	Φ $(\frac{\partial Z_D^*}{\partial \Phi})$	Independent	-
	M_U $(\frac{\partial Z_D^*}{\partial M_U})$	+	+
	M_D $(\frac{\partial Z_D^*}{\partial M_D})$	Independent	+
	t $(\frac{\partial Z_D^*}{\partial t})$	Independent	N/A
	T_{R_W} $(\frac{\partial Z_D^*}{\partial T_{R_W}})$	-	-
	T_{P_W} $(\frac{\partial Z_D^*}{\partial T_{P_W}})$	+	+

Appendix 3.8: Empirical Analysis

We find partial support for the model by empirically examining the number of third party participants utilized by publisher websites, as well as the industry concentration of third parties. Alexa Internet provides rankings for publisher websites within 17 different subject categories[2]. We carry out an exploratory validation study on the 100 most-visited publisher websites from seven (7) of these subject categories (news, arts, shopping, kids and teens, health, business, and adult) provided and ranked by Alexa website rankings. These seven categories were selected with the intention of finding subject categories for which users might reasonably be expected to have different intentions to disclose personal information and browsing behavior due to the nature of the subject content. For the study, an automated browser accessed a publisher website's home page, and the connections made from the publisher websites to third parties were recorded. We used page loading time plus a 3-second window to collect data gathered using a residential internet plan and using Lightbeam for Firefox (Windows) to record these connections.

To better capture the structure of the industry, we profile the third parties and separate them based on the industry sectors as classified by Cookiepedia.co.uk. The three industry sectors are targeting/advertising (*T/A*), functionality (*F*), and performance (*P*). For those third parties that are not profiled in Cookiepedia.co.uk, we make a judgment using available information. 1893 third party websites are identified in total, with 568 *T/A*, 487 *F*, 627 *P*, and 211 classified as unknown (*U*). Using different domain finder services[3], multiple third party websites in each sector owned by the same company are treated as a single third party for analysis, entailing 1066

[2] Alexa.com/topsites/category. The Alexa list of website categories is consistent with the Open Directory Project categories found at rdf.DMOZ.org/rdf/categories.txt.

[3] This study uses whois.domaintools.com, whois.net, and who.is.

unique owner companies comprising 442 *T/A*, 336 *F*, and 340 *P*, with some owner companies providing services in multiple categories. The number of connections made and number of cookies used follow a similar pattern to number of third parties, and so we provide the analysis based on number of third parties only. Table A.3.8.1 provides a summary of descriptive statistics of the data on number of third parties.

Table A.3.8.1. Descriptive Statistics for Number of Third Parties Used Among Websites

Subject Category	N	Targeting/Advertising				Functionality				Performace			
		Min	*Max*	*Avg.*	*StDv*	*Min*	*Max*	*Mean*	*StDv*	*Min*	*Max*	*Mean*	*StDv*
News	100	1	65	16.8	11.7	1	10	4.5	2.3	1	10	3.4	2.0
Arts	100	1	55	11.5	10.0	1	10	3.9	2.1	1	7	2.8	1.5
Shopping	100	1	51	9.4	9.2	1	8	2.9	1.6	1	8	2.7	1.6
Kids and teens	100	1	58	8.8	11.5	1	7	2.7	1.4	1	7	2.3	1.5
Health	100	1	70	8.2	11.3	1	7	2.9	1.6	1	7	2.4	1.5
Business	100	1	60	6.9	9.2	1	7	2.4	1.6	1	7	2.4	1.4
Adult	100	1	34	3.2	5.2	1	9	2.2	1.5	1	7	1.7	1.0

A.3.8.1 Observations

Noting that information sensitivity and user privacy concerns likely vary among different publisher websites, we expect the sharing behavior to differ for publisher websites with different subjects. Figure A.3.8.1 provides the sharing behavior for the top 100 publisher websites in each subject category and industry sector.

Figure A.3.8.1. Third Party Usage by Subject Categories and Industry Sectors

Table A.3.8.2 provides the statistical test results for number of third parties on different categories of websites. Since the variances are different among the categories, we use the Welch's two-tailed t-test for testing if the means are different among these websites. It can be seen from Table A.3.8.2 that the number of third parties are significantly different for most of the categories. Especially, in the *T/A* sector, news and adult categories are statistically different from other categories.

Table A.3.8.2. P-Values for Testing if Number of Third Parties Used in Different Categories of Websites are Statistically Different

		News	Arts	Shopping	Kids & Teens	Health	Business
T/A	News						
	Arts	0.001					
	Shopping	0.000	0.125				
	Kids & Teens	0.000	0.082	0.700			
	Health	0.000	0.027	0.390	0.670		
	Business	0.000	0.001	0.056	0.191	0.393	
	Adult	0.000	0.000	0.000	0.000	0.000	0.001
F	News						
	Arts	0.064					
	Shopping	0.000	0.001				
	Kids & Teens	0.000	0.000	0.299			
	Health	0.000	0.000	0.704	0.522		
	Business	0.000	0.000	0.012	0.106	0.032	
	Adult	0.000	0.000	0.002	0.024	0.006	0.551
P	News						
	Arts	0.019					
	Shopping	0.006	0.568				
	Kids & Teens	0.000	0.023	0.099			
	Health	0.000	0.061	0.210	0.690		
	Business	0.000	0.036	0.147	0.796	0.877	
	Adult	0.000	0.000	0.000	0.001	0.000	0.000

Table A.3.8.3 provides the statistical test results for number of third parties on different sectors of the industry. It can be seen from Table A.3.8.3 that the number of third parties used in the *T/A* industry sector is significantly higher than for both *F* and *P*.

Table A.3.8.3. P-values for Testing if Number of Third Parties Used in Different Industry Sectors are Statistically Different

		T/A	F
News	T/A		
	F	0.000	
	P	0.000	0.001
Arts	T/A		
	F	0.000	
	P	0.000	0.000
Shopping	T/A		
	F	0.000	
	P	0.000	0.286
Kids and Teens	T/A		
	F	0.000	
	P	0.000	0.067
Health	T/A		
	F	0.000	
	P	0.000	0.051
Business	T/A		
	F	0.000	
	P	0.000	0.925
Adult	T/A		
	F	0.080	
	P	0.007	0.008

We also examine the third party market concentration measure, using Herfindahl-Hirschman index (HHI) based on publisher websites' average monthly unique visitors in the United States for a single year period ending in March 2014 as provided by compete.com. The *T/A* sector has the lowest HHI concentrations, followed by *P*, and then by *F*. In terms of publisher website categories, we see that news and arts have the lowest industry concentration, with adult having the highest industry concentration. The HHI results are provided in Figure A.3.8.2.

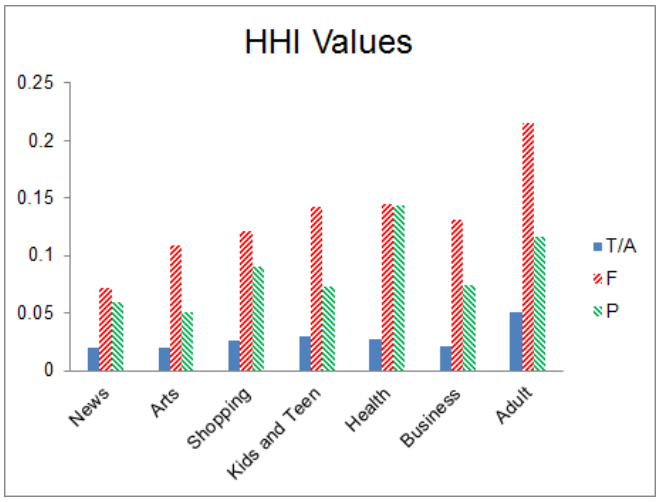


Figure A.3.8.2. HHI for Third Party Industry by Subject Categories and Industry Sectors

Appendix 5.1: Model Details

A.5.1.1 Independent Suppliers Case

We first derive the best response functions for firms' spending at equilibrium. By substituting firm state probabilities (5.4) and demand functions (5.1) in the expected profit function (5.2) we obtain the firm's expected profit as in (5.5). From the first-order conditions, we can find the best response of firm i, $c_i{}^*$, to be

$$c_i{}^* = \frac{\rho\, v\, Z_C - (1+c_j) + \sqrt{\rho\, v\, Z_D\, (1+\pi)(1+c_j)(1+c_j-\rho\, v\, Z_C)}}{1+c_j-\rho\, v\, Z_C} \quad \text{for } i, j = 1,2, \text{ and } j \neq i$$

Using best response functions, the equilibrium spendings can be calculated. The only equilibrium spending for both firms is symmetric in this case, as is provided in (5.6). Lemma A.5.1 below provides sensitivity results for equilibrium spending with respect to some model parameters.

Lemma A.5.1 *In the independent suppliers case, equilibrium spending of the firms c_i^e, is increasing in per unit profit π, direct-risk elasticity of demand Z_D, cross-risk elasticity of demand Z_C, supplier vulnerability v, and cascade probability ρ.*

Proof: These findings can easily be proven by taking the derivative of equilibrium spending (5.6) with respect to the corresponding model parameter. We omit these proofs for the sake of brevity.

While the model used in this chapter is different, the results from the independent case are similar to that of Kolfal et al. (2013). For the independent suppliers case, the best response function of the two firms for some special cases is given in Figure A.5.1.1.

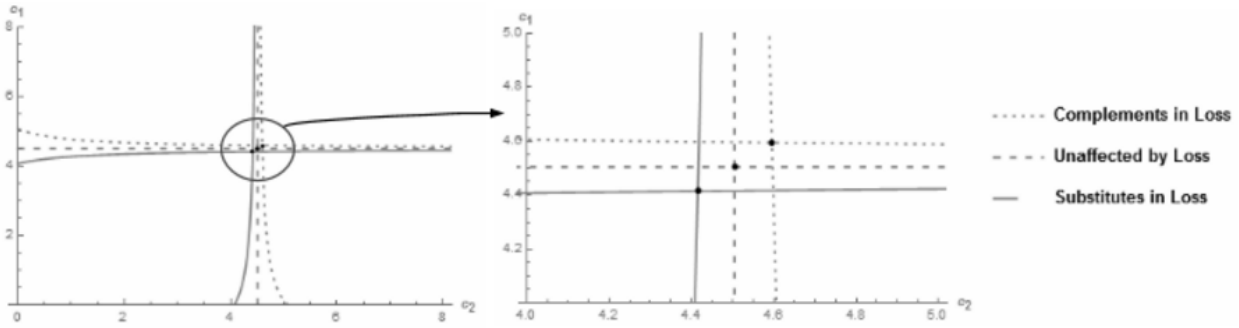


Figure A.5.1.1. Best Spending Response Function for Firms with Independent Suppliers

$(\pi = 100, \rho = 0.75, v = 0.8, Z_D = 0.5, Z_C = \{-0.3, 0, 0.3\})$

As it can be seen from Figure A.5.1.1, in the unaffected by loss case $Z_C = 0$, the two firms are independent of each other in terms of demand. In the substitutes in loss case $Z_C < 0$, firms increase their spending with the other firm's spending. In the complements in loss case $Z_C > 0$, firms decrease their spending with the other firm's spending. By substituting the equilibrium spending ${c_i}^e$ from (5.6) into the firm's expected profit (5.5), we calculate each firm's expected profit in equilibrium as provided in (5.7).

A.5.1.2. Shared Supplier Case

The expected profit function in this case is provided in (5.10). Using the first order conditions, best response functions are calculated as:

$${c_i}^* = \sqrt{\rho\, v\, (Z_D + Z_C)(1 + c_j + \pi)} - (1 + c_j) \qquad \text{for } i, j = 1,2, \text{ and } j \neq i$$

which can be used to calculate the equilibrium spending. The equilibrium in this case is again symmetric, and the spending for both firms is provided in (5.11).

The Lemma A.5.2 provides sensitivity results for equilibrium spending with respect to some model parameters.

Lemma A.5.2 *In the shared supplier case, equilibrium spending of the firms c_i^e, is increasing in per unit profit π, direct-risk elasticity of demand Z_D, cross-risk elasticity of demand Z_C, supplier vulnerability v, and cascade probability ρ. In the shared supplier case, Figure A.5.1.2 provides the best response functions for some special cases.*

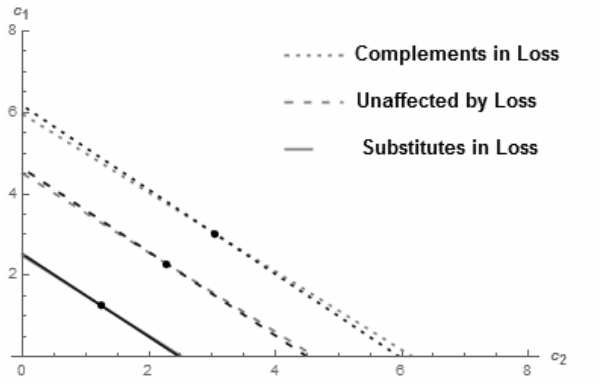


Figure A.5.1.2. Best Spending Response Functions for Firms with Shared Supplier

$(\pi = 100, \rho = 0.75, v = 0.8, Z_D = 0.5, Z_C = \{-0.3, 0, 0.3\})$

From Figure A.5.1.2, it can be seen that best response functions when there is one shared supplier are quite different from when firms have their own independent suppliers. In this case, if the other firm reduces its equilibrium spending, then the equilibrium response is to increase your own spending to make up for that difference. By substituting the equilibrium spending $c_i{}^e$ from (5.11) into the firm's expected profit (5.10), the expected profit at equilibrium can be calculated as in (5.12).

Appendix 5.2: Correlated Incidents

The parameter $\gamma \geq 0$ is added to the model by using the following transformations in firm state probabilities:

$$P_{gg}{}' = P_{gg},\ P_{bg}{}' = P_{bg} - \gamma,\ P_{gb}{}' = P_{gb} - \gamma,\ P_{bb}{}' = P_{bb} + 2\gamma \qquad \text{(A.5.2.1)}$$

For each firm, we define probabilities P_{1g} and P_{2g} firms 1 and 2 being in the good states, respectively. Note that $P_{1g} = P_{gg} + P_{gb}$ and $P_{2g} = P_{gg} + P_{bg}$. The effect of parameter γ on the correlation is presented next. Considering each firm's state as a random variable following a Bernoulli distribution that is dependent on the other firm's state, we calculate the (Pearson's) correlation coefficient. The correlation coefficient is defined as: $r = \frac{E(XY)-E(X)E(Y)}{\sqrt{\sigma(X)}\sqrt{\sigma(Y)}}$ where X and Y are Bernoulli random variables representing the state of the firms as follows:

$$X = \begin{cases} 1 & If\ firm\ 1\ is\ in\ good\ state \\ 0 & If\ firm\ 1\ is\ in\ bad\ state \end{cases} \qquad Y = \begin{cases} 1 & If\ firm\ 2\ is\ in\ good\ state \\ 0 & If\ firm\ 2\ is\ in\ bad\ state \end{cases}$$

The correlation coefficient can thus be written as: $r = \frac{P_{gg}-P_{1g}P_{2g}}{\sqrt{P_{1g}(1-P_{1g})}\sqrt{P_{2g}(1-P_{2g})}}$. To illustrate the effect of correlation parameter γ, we substitute the transformations in (A.5.2.1) into the correlation coefficient formula above. We have:

$$r = \frac{P_{gg}{}'-P_{1g}{}'P_{2g}{}'}{\sqrt{P_{1g}{}'(1-P_{1g}{}')}\sqrt{P_{2g}{}'(1-P_{2g}{}')}} = \frac{P_{gg}-(P_{1g}-\gamma)(P_{2g}-\gamma)}{\sqrt{(P_{1g}-\gamma)(1-P_{1g}+\gamma)}\sqrt{(P_{2g}-\gamma)(1-P_{2g}+\gamma)}} \qquad \text{(A.5.2.2)}$$

We are interested in cases with positive correlation among the incidents at two firms. In order to have $0 \leq r \leq 1$, we need to have:

$$P_{gb} \leq \gamma, P_{bg} \leq \gamma, \gamma \leq P_{1g} = P_{gb} + P_{gg}, \gamma \leq P_{2g} = P_{bg} + P_{gg}$$

In (A.5.2.2), it can be shown that when $P_{1g} > 2/3$ and $P_{2g} > 2/3$, then the correlation coefficient increases with γ. In our setting, the above assumptions are fairly realistic, in that firms are in the good state most of the time.

Appendix 5.3: Coordination in Spending

Here we discuss the mechanism for the cooperation in spending. We analyze how the firm profits change when the firms can set their spending to the optimal spending value R that maximizes firms' profits instead of the equilibrium spendings. We analyze several numerical test suites to provide insights on how coordination impacts the firms' strategic decision making. Based on these analyses, we find that the coordination of spending can either increase or decrease firm profits. In Figures A.5.3.1 and A.5.3.2 we provide a representative numerical example of when each of these scenarios may occur. In these figures, the black dots are equilibrium points, and the gray dots indicate points for the optimal R value. Figure A.5.3.1 illustrates that for the case of independent suppliers, firm profit can increase under coordination in both complements in loss and substitutes in loss cases. Note that in the complements in loss case, firms should spend more on security than the equilibrium spending in order to gain higher profits. This is consistent with the findings from Kolfal et al. (2013). In the substitutes in loss case, firms should spend less on security than the equilibrium spending to gain higher profits. While these observations are based on numerical examples, we find that they are consistent among the range of parameters that we are interested in.

It can be observed that in the independent case, when firms are substitutes in loss, they have an incentive to reduce spending on suppliers to increase their profits. On the other hand, for the complements in loss case, there is room to increase profits by increasing spendings on suppliers.

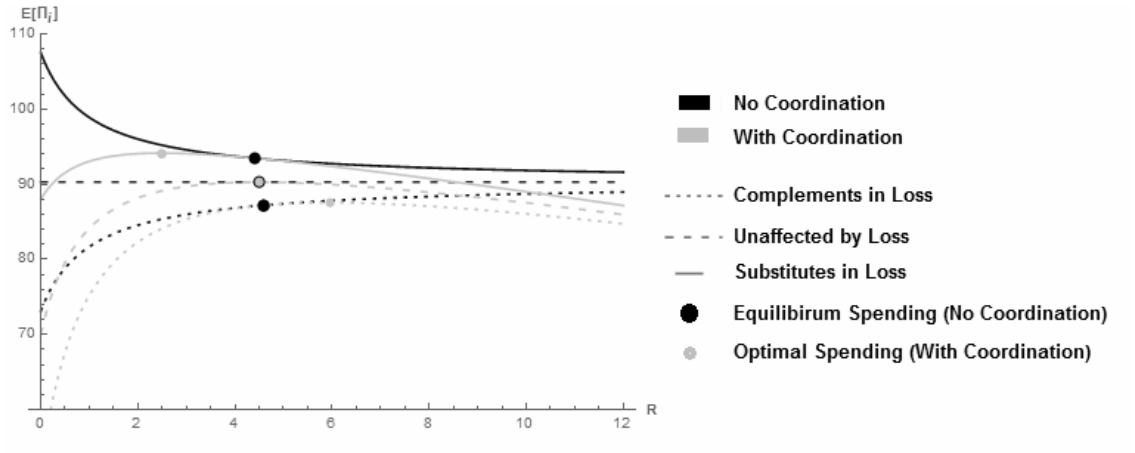


Figure A.5.3.1. Expected Profit with Respect to Spendings in the Independent Case

$(\pi = 100, \rho = 0.75, v = 0.8, Z_D = 0.5, Z_C = \{-0.3, 0, 0.3\})$

In the case of shared supplier, we find that coordination of spending between firms can increase profit over all values of cross-risk elasticity, as illustrated in Figure A.5.3.2.

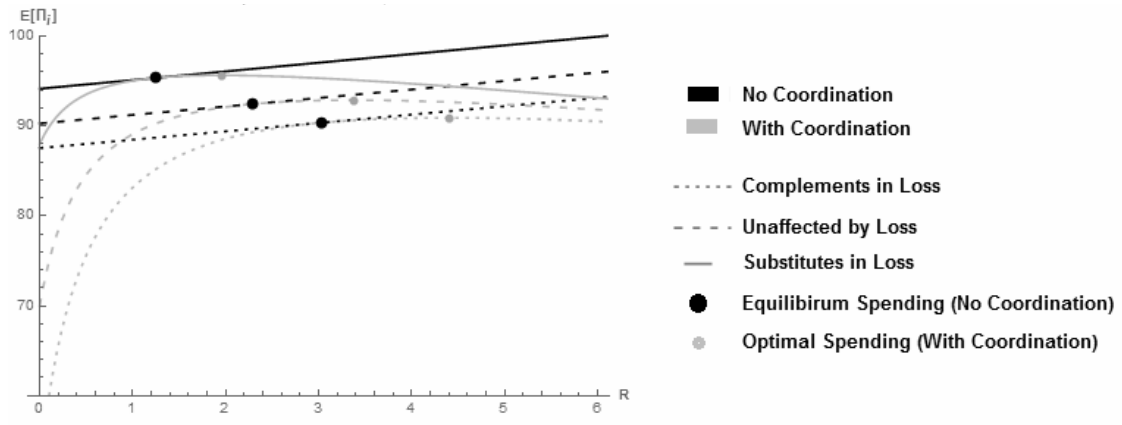


Figure A.5.3.2. Expected Profit with Respect to Spendings in the Shared Case

$(\pi = 100, \rho = 0.75, v = 0.8, Z_D = 0.5, Z_C = \{-0.3, 0, 0.3\})$

It can be seen from Figure A.5.3.2. that firms can always improve their profits over the equilibrium profit by improving their spending on security.

We now provide the proof for how the indifference line in the cooperative case (Figure 5.5) moves with the model parameters. First, we need to calculate the profit when both firms spend R in the suppliers. We can calculate the expected profit for each firm by substituting R for both c_1 and c_2. In the independent case (5.5), the substitution results in the following expected profit function:

$$E[\Pi_i]^I = E[\Pi]^I = \frac{1+R-\rho\, v\, (Z_D+Z_C)}{1+R}(\pi - R) \text{ for } i = 1,2 \quad \text{(A.5.3.1)}$$

The maximum value is calculated as:

$$Max_{E[\Pi]^I} = \pi + 1 + \rho\, v\, (Z_D + Z_C) - 2\sqrt{\rho\, v\, (Z_D + Z_C)\, (1+\pi)} \quad \text{(A.5.3.2)}$$

For the shared case (10), the substitution results in the following expected profit function:

$$E[\Pi_i]^S = E[\Pi]^S = \frac{1+2R-\rho\, v\, (Z_D+Z_C)}{1+2R}(\pi - R) \qquad \text{for } i = 1,2 \quad \text{(A.5.3.3)}$$

The maximum value is calculated as:

$$Max_{E[\Pi]^S} = \pi + \frac{1+\rho\, v\, (Z_D+Z_C)-2\sqrt{\rho\, v\, (Z_D+Z_C)\, (1+2\pi)}}{2} \quad \text{(A.5.3.4)}$$

The difference between the maximum values for the independent and shared cases is as follows:

$$Diff = Max_{E[\Pi]^I} - Max_{E[\Pi]^S}$$

$$= \frac{1+\rho\, v\, (Z_D+Z_C)-4\sqrt{\rho\, v\, (Z_D+Z_C)\, (1+\pi)}+2\sqrt{\rho\, v\, (Z_D+Z_C)\, (1+2\pi)}}{2} \quad \text{(A.5.3.5)}$$

By taking the derivative of the difference with respect to vulnerability v, cascade probability ρ, and per unit profit π, we have:

$$\frac{\partial\, Diff}{\partial\, v} = \frac{\rho\, v\, (Z_D + Z_C) - \sqrt{\rho\, v\, (Z_D + Z_C)}\, (\sqrt{4 + 4\pi} - \sqrt{1 + 2\pi})}{2v} < 0$$

$$\frac{\partial\, Diff}{\partial\, \rho} = \frac{\rho\, v\, (Z_D + Z_C) - \sqrt{\rho\, v\, (Z_D + Z_C)}\, (\sqrt{4 + 4\pi} - \sqrt{1 + 2\pi})}{2\rho} < 0$$

$$\frac{\partial\, Diff}{\partial\, v} = -\rho\, v\, (Z_D + Z_C) \left(\frac{1}{\sqrt{\rho\, v\, (Z_D + Z_C)(1 + \pi)}} - \frac{1}{\sqrt{\rho\, v\, (Z_D + Z_C)(1 + 2\pi)}} \right) < 0$$

All of the derivatives are negative, meaning that the independent region shrinks (or the shared region expands) as either of these parameters increase.

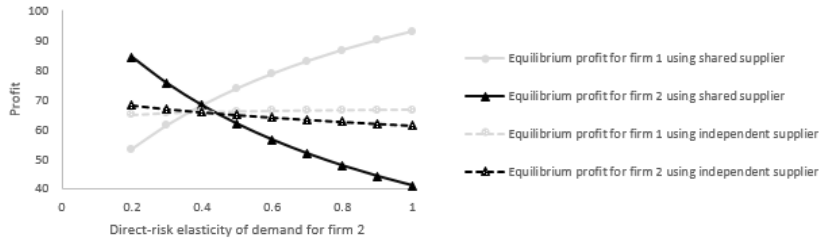
Appendix 5.4: Asymmetric Firms Analysis

A.5.4.1 Asymmetric Direct-Risk Elasticities of Demand

When firms are asymmetric in direct-risk elasticities of demand, we find that in both cases of shared and independent suppliers, all else being equal, the firm with lower direct-risk elasticity of demand spends less in IT security and gains a higher profit than the other firm. We compare the shared and independent supplier options to study how the asymmetry of direct-risk elasticities of demand affects supplier choice. We analyze a range of numerical examples, and provide one representative example here. The observations provided here are consistent among all of the examples tested. Figure A.5.4.1 provides different scenarios that may arise. Figure A.5.4.1.a illustrates the change in profits when direct-risk elasticity changes for firm 2 and is fixed for firm 1 (at $Z_{1,D} = 0.4$) for positive cross-risk elasticity of demand ($Z_{1,C} = Z_{2,C} = 0.2$). In this example, the shared option is optimal when the direct-risk elasticities are symmetric at $Z_{1,D} = Z_{2,D} = 0.4$. In Figure A.5.4.1.b, profits are shown for the case where the cross-risk elasticity of demand is negative ($Z_{1,C} = Z_{2,C} = -0.2$). In this case, when the direct risks are symmetric at $Z_{1,D} = Z_{2,D} = 0.4$ the independent option is optimal.

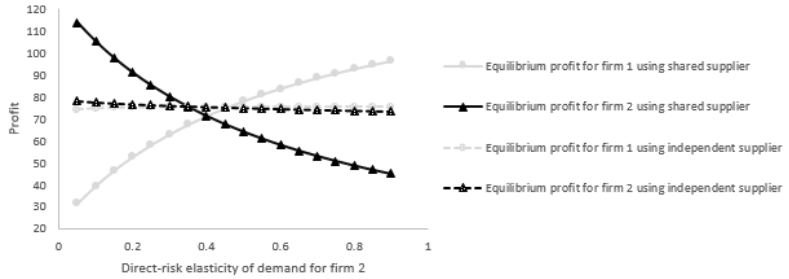
It can be seen from Figure A.5.4.1 that asymmetry in the direct-risk elasticities of demand would decrease the profits from the shared supplier for one of the firms more than it would decrease the profit from independent supplier. In other words, one of the firms has less incentive to go with the shared supplier choice. Since it is necessary for both firms to prefer the shared supplier for that option to become viable, the asymmetry is only making the shared option less desirable. If in asymmetry the shared supplier option is the optimal choice for the firms, then as firms grow more different (become more asymmetric) in direct-risk, there is a region in which

they will still prefer the shared supplier choice. However, if the direct-risk of the two firms becomes too different, there is no longer incentive for them to share suppliers (Figure A.5.4.1.a). When the optimal decision in asymmetry is to use independent suppliers, then asymmetry would not change this decision (as illustrated in Figure A.5.4.1.b).



a) Case in Which the Symmetry Shared Option is Optimal

$(\pi = 75, \rho = 0.8,\ v = 0.75, Z_{1,D} = 0.4,\ Z_C = -0.2)$



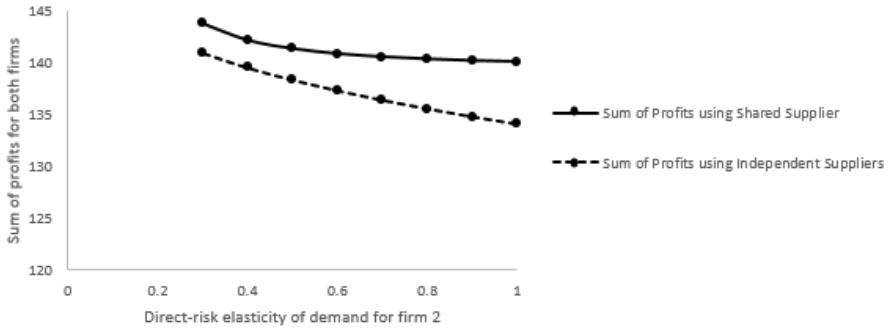
b) Case in Which the Symmetry Independent Option is Optimal

$(\pi = 75, \rho = 0.3,\ v = 0.5, Z_{1,D} = 0.4,\ Z_C = 0.1)$

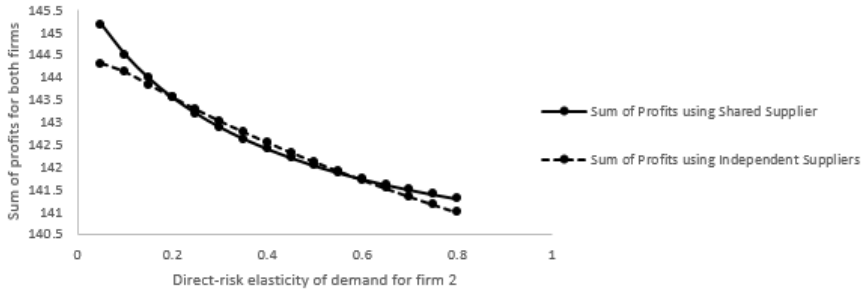
Figure A.5.4.1. Effect of Changes in Direct-Risk Elasticity of Demand for Firm 2 on Profits

In some cases, it might be possible for a firm to make transfer payments to the competitor firm, in order to make the shared option viable for them as well. For example, if both firms are run by a common parent company, the parent company may be interested in maximizing the combined profit from both firms. Figure A.5.4.2 provides a numerical example of the sum of profits for both firms in the two scenarios discussed in Figure A.5.4.1.

In the first scenario, where symmetric firms choose the shared supplier, if firms were able to make transfer payments, there can be a case where while one firm's profit using the shared supplier is less than when using the independent supplier, because of the transfer payment, it would still go with the shared supplier. It can be seen in Figure A.5.4.2.a that the sum of profits for both firms using the shared supplier is consistently higher than the sum of profits using independent suppliers.



a) Sum of Profits for Case Where the Symmetry Shared Option is Optimal
($\pi = 75, \rho = 0.8,\ v = 0.75, Z_{1,D} = 0.4,\ Z_C = 0.2$)



b) Sum of Profits for the Case Where Symmetry Independent Option is Optimal

($\pi = 75, \rho = 0.3,\ v = 0.5, Z_{1,D} = 0.4,\ Z_C = 0.1$)

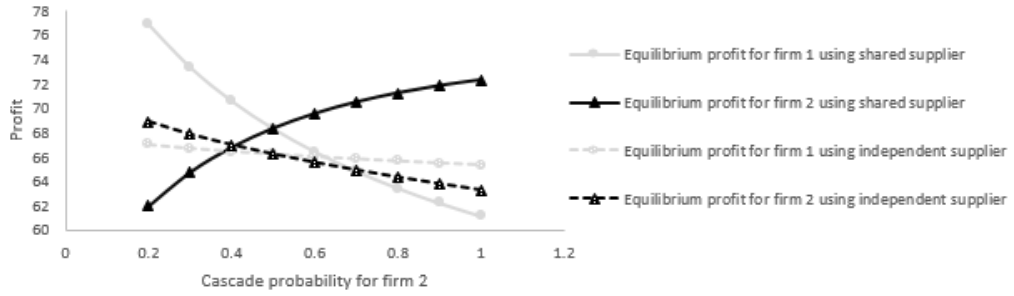
Figure A.5.4.2. Effect of Changes in Cross-Risk Elasticity of Demand for Firm 2 on Aggregate Profit

Figure A.5.4.2.b provides the sum of profits for both firms when they choose the independent supplier in the symmetric case. It can be seen that as firms move away from the symmetry (they become more heterogeneous in terms of direct-risk elasticity of demand), then there might be opportunities for the firm with lower direct-risk elasticity of demand to make transfer payments to the other firm in order to make them retain the shared supplier. This case occurs at the extreme asymmetry cases, where firms are very different from each other.

A.5.4.2. Asymmetric Cascade Probabilities

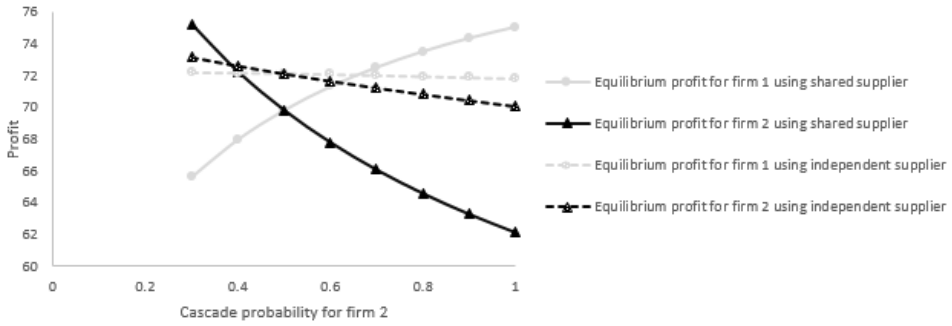
Here, we discuss firms with asymmetric cascade probabilities. We find that in both cases of shared and independent suppliers, the firm with lower cascade probability spends less in IT security and gains a higher profit than the other firm. We compare the shared and independent supplier options to study how the asymmetry affects the supplier choice. We again use a set of numerical examples to analyze the effect of asymmetry, and provide a representative example here. These findings are consistent among all of the test suites. Figure A.5.4.3 provides the different scenarios that may arise. Figure A.5.4.3.a., illustrates the change in profits when cascade probability changes for firm 2 and is fixed for firm 1 (at $\rho_1 = 0.5$). Figure A.5.4.3.b. illustrates the change in profits again when cascade probability changes for firm 2 and is fixed for firm 1 (at $\rho_1 = 0.5$), but cross-risk elasticities are changed to $Z_C = 0.3$. In this example, the independent option is optimal when the cascade probabilities are symmetric.

It can be seen from Figure A.5.4.3 that asymmetry in the cascade probabilities would decrease the profits from the shared supplier for one of the firms more than it would decrease the profit from independent supplier. The insight from this is similar to what was discussed in the previous sections.



a) Case in Which the Symmetry Shared Option is Optimal

$(\pi = 75, \rho_1 = 0.5,\ v = 0.8, Z_D = 0.5, Z_C = 0.3)$



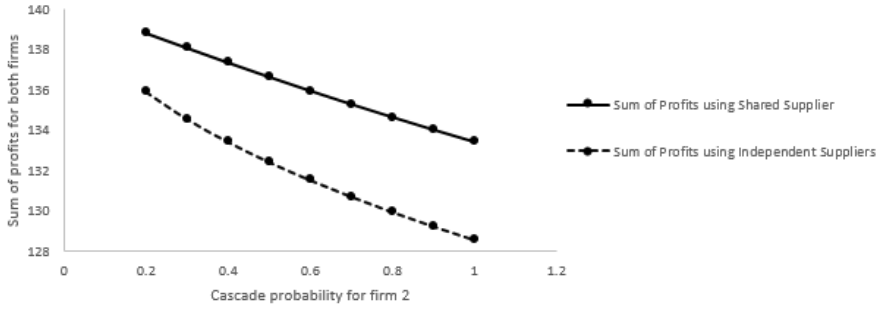
b) Case in Which the Symmetry Independent Option is Optimal

$(\pi = 75, \rho_1 = 0.5,\ v = 0.5, Z_D = 0.5, Z_C = 0.3)$

Figure A.5.4.3. Effect of Changes in Cascade Probability for Firm 2 on Profits

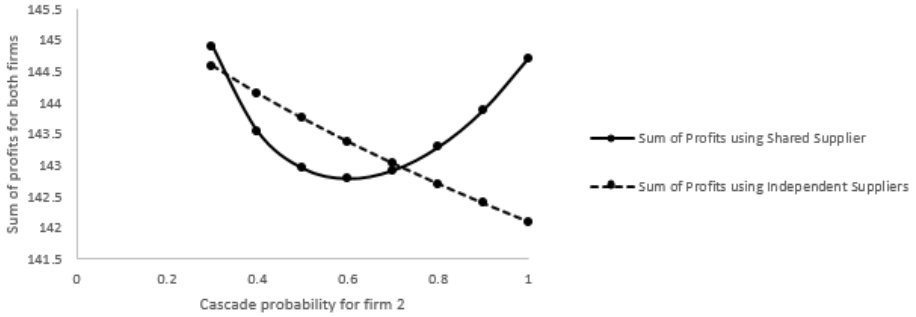
We also examine the transfer payments for this case. Figure A.5.4.4 provides a numerical example of the sum of profits for both firms in the two scenarios discussed in Figure A.5.4.3.

These results from combined firms are similar to those in Section 5.7.2. In the first scenario, where symmetric firms choose the shared supplier, if firms were able to make transfer payments, there can be a case where while one firm's profit using the shared supplier is less than when using the independent supplier, because of the transfer payment, it would still go with the shared supplier.



a) Sum of Profits for the Case in Which the Symmetry Shared Option is Optimal

$(\pi = 75, \rho_1 = 0.5,\ v = 0.8, Z_D = 0.5, Z_C = 0.3)$



b) Sum of Profits for the Case in which the Symmetry Independent Option is Optimal

$(\pi = 75, \rho_1 = 0.5,\ v = 0.5, Z_D = 0.5, Z_C = 0.3)$

Figure A.5.4.4. Effect of Changes in Cross-Risk Elasticity of Demand for Firm 2 on Aggregate Profits

Appendix Bibliography

Even-Dar, E., A. Shapira. 2007. "A Note on Maximizing the Spread of Influence in Social Networks." X. Deng, F.C. Graham, eds., *WINE 2007 Proceedings of the 3rd international conference on Internet and Network Economics.* Lecture Notes in Computer Science 4858, Springer, Heidelberg, Germany. 281–286.

Hoffmann, L. 2010. "Mine your business." *Communications of the ACM* 53(6): 18-19.

Kimura, M., K. Saito, H. Motoda. 2009. "Blocking Links to Minimize Contamination Spread in a Social Network." *ACM Transactions on Knowledge Discovery and Data Mining* 3(2) 9:1-9:23.

Wu, L., C.Y. Lin, S. Aral, E. Brynjolfsson. 2009. "Value of social network–a large-scale analysis on network structure impact to financial revenue of information technology consultants." *The 2009 Winter Conference on Business Intelligence*.