2012

Identifying high value customers in a social network: individual characteristics vs. social influence

Sang Uk Jung
University of Iowa

IDENTIFYING HIGH VALUE CUSTOMERS IN A SOCIAL NETWORK: INDIVIDUAL CHARACTERISTICS VS. SOCIAL INFLUENCE

by

Sang Uk Jung

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Business Administration
in the Graduate College of
The University of Iowa

July 2012

Thesis Supervisors: Professor Gary J. Russell
Assistant Professor Qin Zhang

ABSTRACT

Firms are interested in identifying customers who generate the highest revenues. Typically, customers are regarded as isolated individuals whose buying behavior depends solely on their own characteristics (e.g., previous purchase behavior, demographics etc.). In a social network setting, however, customer interactions can play an important role in purchase behavior.

This thesis develops a generalizable methodology to identify high-value customers in a network. Previous work on social networks has focused most attention on modeling the interaction between individuals and understanding the positions of individuals in a network (e.g., measuring the influence of an individual based on his/her degree of network centrality). Little is known about how network influence directly translates into the benefits to the firm. In this study, the importance of taking into account both an individual characteristics and network effects when measuring customer value is argued. Drawing upon the spatial statistics literature, a spatial autocorrelation model is constructed that explicitly shows how these effects interact in generating firm revenue.

This model is applied to a unique user-level dataset from a popular online gaming company in Korea. The data contain information about demographics of individual gamer, interaction between gamers, behavior within the game environment, and revenues generated by each individual. First, we propose a static model studying gamers' revenue in one period. We quantify the relative impact of an individual characteristics and network effects on revenue. The proposed static model shows better forecasts of an individual's value within a network for the firm than the benchmark models. The empirical analysis shows that individuals who are most influential in a network sense are not necessarily individuals who have the highest customer value.

Next, we incorporate the spatio-temporal aspects of social influence in a network into the static model. This model is extended to construct the spatial dynamic model to forecast revenue in a social network. Second, we account for the homophily effects by separating the contemporaneous network effects out into the contemporaneous, temporal, and spatio-temporal effects. The proposed spatial dynamic model allows us to quantify an individual value in a network in a long-term perspective. The dynamic model is shown to outperform the static, and the other benchmark models in quantifying an individual value in revenue generation to the firm. Lastly, a dynamic coevolution model to account for homophily is suggested and discussed for future research.

Abstract Approved: ________________________________
Thesis Supervisor

____
Title and Department

____
Date

____
Thesis Supervisor

____
Title and Department

____
Date

IDENTIFYING HIGH VALUE CUSTOMERS IN A SOCIAL NETWORK: INDIVIDUAL CHARACTERISTICS VS. SOCIAL INFLUENCE

by

Sang Uk Jung

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Business Administration
in the Graduate College of
The University of Iowa

July 2012

Thesis Supervisors: Professor Gary J. Russell
Assistant Professor Qin Zhang

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Sang Uk Jung

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Business Administration at the July 2012 graduation.

Thesis Committee: ___________________________________
Gary J. Russell, Thesis Supervisor

____
Qin Zhang, Thesis Supervisor

____
Thomas S. Gruca

____
Catherine A. Cole

____
Kung-Sik Chan

To Grace, Ashton, and my parents

ACKNOWLEDGMENTS

I wish to thank all my friends and colleges who helped me throughout my studies at the University of Iowa. I am extremely thankful to my advisors, Gary J. Russell and Qin Zhang, for their countless hours of support and advice that they have given to me. I would like to extend my gratitude to my thesis committee members: Thomas Gruca, Cathy Cole, and Kung-Sik Chan. Thomas Gruca encouraged my academic development from my first year and I will always appreciate the support that he has given to me. I am also very thankful to Lopo Rego in Indiana University for his mentoring.

I am very grateful to all who supported my decision to apply for a PhD program in marketing. I would like to extend special thanks to Andrew Gelman, and Liam Paninski during my Columbia University, and Chihwa Kao in Syracuse University who helped me my application to the PhD program.

I also thank my long-time friends, Junbum Park, and Byungwha Bae for providing data used in my thesis and for their collective support and friendship during the times I needed the most.

I wish to thank my parents whose support and encouragement played a big part in getting to where I am. Last, but not the least, I thank my wife Grace for her love, sacrifices and unconditional support and thank my son Ashton for giving me the abundant happiness during these challenging but exciting years in the program.

ABSTRACT

Firms are interested in identifying customers who generate the highest revenues. Typically, customers are regarded as isolated individuals whose buying behavior depends solely on their own characteristics (e.g., previous purchase behavior, demographics etc.). In a social network setting, however, customer interactions can play an important role in purchase behavior.

This thesis develops a generalizable methodology to identify high-value customers in a network. Previous work on social networks has focused most attention on modeling the interaction between individuals and understanding the positions of individuals in a network (e.g., measuring the influence of an individual based on his/her degree of network centrality). Little is known about how network influence directly translates into the benefits to the firm. In this study, the importance of taking into account both an individual characteristics and network effects when measuring customer value is argued. Drawing upon the spatial statistics literature, a spatial autocorrelation model is constructed that explicitly shows how these effects interact in generating firm revenue.

This model is applied to a unique user-level dataset from a popular online gaming company in Korea. The data contain information about demographics of individual gamer, interaction between gamers, behavior within the game environment, and revenues generated by each individual. First, we propose a static model studying gamers' revenue in one period. We quantify the relative impact of an individual characteristics and network effects on revenue. The proposed static model shows better forecasts of an individual's value within a network for the firm than the benchmark models. The empirical analysis shows that individuals who are most influential in a network sense are not necessarily individuals who have the highest customer value.

Next, we incorporate the spatio-temporal aspects of social influence in a network into the static model. This model is extended to construct the spatial dynamic model to forecast revenue in a social network. Second, we account for the homophily effects by separating the contemporaneous network effects out into the contemporaneous, temporal, and spatio-temporal effects. The proposed spatial dynamic model allows us to quantify an individual value in a network in a long-term perspective. The dynamic model is shown to outperform the static, and the other benchmark models in quantifying an individual value in revenue generation to the firm. Lastly, a dynamic coevolution model to account for homophily is suggested and discussed for future research.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

INTRODUCTION

The idea consumer behaviors, attitudes and beliefs are interdependent of others has been the topic of marketing research for over 50 years. However, there has recently been increasing popularity of social networking sites, such as Facebook and Twitter, and more data availability on social connections or interactions among people. These phenomena have led to a surge of interest among both academics and practitioners in understanding the effects of social interactions on consumer behavior in socially connected networks. While some aspects of these effects have already been addressed and somewhat understood so far, many important questions still remain unanswered.

Despite their huge popularity, many social network (SN) firms still struggle to find optimal business models for revenue generation. Following the success of Google, many SN firms have been trying to employ an advertising-driven business model. In this model, revenue is generated by cost per click (CPC) or cost per action or purchase (CPA). However, prior research shows that average revenue per user (ARPU) on SN sites is very low because people visit SN sites to socialize, not to purchase products or find information. Therefore, the advertising-driven business model has had very limited success on SN sites. Sergey Brin, Google cofounder, also noted, “I don’t think we have the killer best way to monetize social networks yet.”

Selling digital items such as music and accessories for virtual avatars is another popular business model, referred to as the digital items model. On average, the ARPU of SN firms, which use the digital items model such as DENA in Japan and Cyworld in Korea, is ten times higher than that of Facebook, which use the advertising-driven model. Consequently, many SN firms in the US have also been recently experimenting with selling digital items through social networks. While user involvement with a site directly translates into firm revenue in the advertising-driven business model (e.g., SN firms charge the cost per clicking on the ad), it is different for the digital items model (e.g. active users are not necessarily the top buyers of virtual items). Therefore, the way to identify key users who generate high values to firms should be different, depending on the business model. In this study, our discussion is focused on the digital items model.

Identifying the key users who influence others' activity and behaviors and understanding how behavioral contagion spreads through networks are vital for marketing managers due to the spillover effects of marketing interventions via the network. Marketing efforts on influential customers in a network would increase the magnitude and speed of spread in marketing effects in a way not possible in traditional marketing and would also increase the return on marketing investment (ROMI) for the future.

Most previous work on social networks has focused on modeling interdependent behaviors or outcomes amongst connected individuals and identifying influential individuals in a network inferred from their position in a network. However, little is known about how these network effects are translated into actual purchases or revenue to firms. On the other hand, a traditional marketing study typically focuses on the outcome of marketing variables, i.e., it uses customer characteristics or customer lifetime values (CLV) to measure the value of consumers. However, this measure could be over- or underestimated if network effects are not accounted for. In this study, we focus on identifying an individual's value in terms of revenue generation to a firm in a digital items business model setting. We propose a generalizable methodology to identify high-value customers in a network that takes into account both an individual's characteristics and network effects. Based upon the spatial statistics literature, a network autoregressive model is constructed to explicitly show how these two effects interact in generating firm revenue in a social network. A potential linkage between customer profiles and the rank order of a customer's influence on revenue generation is also examined.

Our study differs from the literature in several respects. First, this research takes into account both individual and network effects when measuring customer value. Because our model allows us to evaluate the interaction effects of individual characteristics and network effects, we are able to more accurately identify high-value customers in a network. Second, we examine the impact of social influence on a consumer's actual purchase. There are few empirical studies investigating the impacts of social influence on actual purchases within a social network. Finally, the current study models individual-level influence simultaneously in the overall system of a social network. Previous research has focused on social influence within a dyadic relation (Narayan and Yang 2006) or mean behaviors of a reference group (Nair, Manchanda, and Bhatia 2010).

The remainder of this thesis is organized as follows. It starts with a review of various social influence models in the sociology, economics and marketing literature. Next, the spatial autoregressive model is introduced as the main methodology in this thesis. Its uses and potential in marketing is discussed. In Chapter 2, a static spatial autoregressive model to identify customer value in a network is discussed. It is shown that individuals who are most influential in a network sense may not necessarily be individuals who have the highest customer value. By analyzing a unique user-level dataset from a popular online gaming company in Korea, it is demonstrated that the predictive performance of the model is better than benchmarks models. In Chapter 3, the extension of the static spatial model to the dynamic model is explored. We quantify individual value within a network in a long-term perspective using this dynamic approach, and discuss the model comparison of identifying customer value in a network between the static and dynamic model. Finally, Chapter 4 discusses contributions of the thesis and offers suggestions for future research.

CHAPTER I

LITERATURE REVIEW

There is a growing body of research in the area of social influence and social networks in the marketing area. While much work has been done about diffusion (Iyengar, Van den Bulte and Valente 2010; Watts and Dodds 2007; Argo, Dahl and Morales 2006, 2008; Goldenberg *et al.* 2009), word-of-mouth (Trusov, Bucklin and Pauwels 2009; Godes *et al.* 2005; Godes and Mayzlin 2004) and joint group decision making (Hartmann 2010), this study focuses on modeling interdependent decisions of individuals and identifying an individual's influence on revenue generation to firms.

This chapter provides a literature review related to the social influence in a network. Section 1.1 provides a theoretical background on why the actions of different individual are correlated and Section 1.2 discusses what interpersonal influence means in this study. Section 1.3 reviews a set of theories and methods of identifying the importance of an individual actor in a social network analysis (SNA). Section 1.4 discusses several models in the literature that focus on how interdependent behaviors or outcomes of connected individuals are modeled. In Section 1.5, we review how social networks have been constructed in prior research. Finally, Section 1.6 presents the positioning of this research in the literature. In various disciplines such as economics, sociology, psychology, physics and computer science, as well as marketing, modeling and understanding social influence in networks has been an area of interest. Since the field is vast, the focus of this study is on models and literature in the social sciences.

1.1 Why the individual customer's behavior is correlated

The notion that we are influenced by behaviors or attitudes of people around us is suggested by even the earliest theories of social influence on customer behavior. The conformity theory by Asch (1952) suggests that an individual customer evaluates his or her own opinion using significant others' opinions as a proper criterion. Thus, customers are likely to adopt behaviors similar to those of their significant others. A number of articles have shown why individuals adapt themselves to the behavior of a reference group. For example, Venkatesan (1966) and Sims (1971) find that individuals tend to conform to group norms due to group pressure. Stafford (1966) shows that individuals tend to conform to the choice of the leaders of their reference group. Kelman (1961) identifies three types of social influence on individual behavior, each characterized by a distinct set of antecedents and a distinct set of consequent conditions; compliance, identification, and internalization. Compliance involves conforming to a rule while possibly keeping one's own private beliefs. Conformity is motivated by rewards and the avoidance of punishment. Identification is a desire to be like the influencer. People adopt an opinion or a behavior because it puts them into a satisfying relationship with the influencer. Internalization refers to the belief or behavior and conforming both publicly and privately.

1.2 What we mean by "Interpersonal Influences"

Interpersonal influence can be defined simply and intuitively by the concept that an individual's behaviors , attitudes and preferences are influenced by significant others, such as friends and families. This definition could be elaborated, depending on the context. In consumer demand studies, interpersonal influence on consumers' latent preferences has been examined. These are defined as “preferences which depend on other people’s consumption”(Pollak 1976). The current study defines social interactions by outcomes such that people adopt similar behavioral outcomes as their significant others in a social structure, which takes the form of a network. A variety of terms have been used to describe the same phenomenon: “neighborhood effects” (Datcher 1982; Case 1991; Bell and Song 2007), “bandwagon effects” (Leibenstein 1950; Mason 1995), “peer influences” (Duncan, Haller and Portes 1968; Manski 1993), “conformity” (Bernheim 1991) and “contagion” (Leenders 1997; Iyengar, Van den Bulte and Valente 2010).

While the economics literature has been interested in incorporating interdependent preference among customers into a demand system, the key questions in marketing science have involved the mechanism of interdependence, why the actions of different individuals are correlated, how they may be modeled, and who is more influential on others. Each topic is covered in the following sections.

1.3 Models of Identifying Interpersonal Influence in SNA

Social network analysis (SNA) has emerged as a key technique to measure an individual’s influence in a network in the quantitative sociology area. It has been widely used to analyze social networks in various disciplines such as sociology, economics, physics, computer science and marketing. These measures are grounded on the basic assumption that people occupying some important positions in a network tend to have greater access to relevant resources and tend to have more influence on others (Freeman 1979; Keller and Barry 2003). Various measures have been suggested such as centrality, structural equivalence and structural holes, etc. (Burt 1987, 1992). For example, centrality measures such as degree, betweenness and closeness give a rough indication of the social power of an actor based on how well they are connected with others in the network. Thus, the importance of an individual actor can be inferred from his or her location in the network (Bavelas 1950; Beauchamp 1965; Freeman 1977; Opsahl, Agneessens and Skvoretz 2010).

This idea that we can measure individuals' influence on others using their location in a network has also been discussed in marketing (e.g., Iacobucci 1990, 1996, 1998; Iacobucci and Hopkins 1992; Van Den Bulte and Wuyts 2007). Social networking sites allow us to gather relational data among individuals, such as connections of friendship or the number of messages exchanged. Because these links are easily observable by the firm and researchers, it might be tempting to apply SNA directly to infer a person's importance in the network to the firm. However, measuring customer value in the revenue creation perspective is more challenging because little is known about how network influence is translated into revenues to the firm. For example, Trusov, Bodapati and Bucklin (2010) found that having many links (high degree) did not make users influential in terms of revenue creation. They illustrate the potential for large gaps in financial returns to the firm from using model-based estimates of influence versus friend count. Also, Stephen and Toubia (2010) found that the sellers who benefit the most from the network are not necessarily those who are central to the network, but rather those whose accessibility is most enhanced by the network. Iyengar, Han and Gupta (2009) attempted to quantify social influence in terms of purchase probability and revenues at the individual level using actual purchase data in a social networking site. They found significant heterogeneity of social influences, in that highly connected people tend to be negatively influenced by their friends' purchases, whereas people with moderate connections are positively influenced by such purchases. Thus, so far, it is still a big question mark for researchers to know how best to go about building models to address the identification problem of social influence in a network with respect to firm revenue.

1.4 Models of Interpersonal Influence in a Network

Interpersonal influence takes place when an individual customer conforms his or her behavior or attitude to those of significant others, which leads them to buy similar brands or have similar preferences. There have been two distinct approaches to model these interdependent phenomena in econometrics and the marketing literatures: the individual outcome model and group decision-making. While the current study focuses on how individual behaviors or outcomes are interdependent on connected others, the joint group decision model is based on a game theoretic framework to determine which combinations of actions are at possible equilibrium such that all group members are satisfied with the chosen outcomes (i.e., Hartmann 2010; Brock and Durlauf 2001; Soetevent and Kooreman 2007). The individual outcome model is again sub-categorized into linear-in-means models and the spatial autoregressive (SAR) model, depending on whether it allows for the heterogeneity of each pair of relations. In the following discussion, we focus most attention on the individual-level outcomes model.

We structure our brief overview of social interaction models in a network around whether we consider network effects as a member of subgroups within a network or as a whole across a population. Each has been referred to as the linear-in-means model and the spatial autoregressive (SAR) model. While the linear-in-means model has more tradition in social econometrics (Manski 1993), spatial analysis approaches from geography are widely applied to social network research in sociology (Leenders 2002). These two models basically adopt the same assumptions about the process of social influence that one individual's preferences or behaviors are a function of significant others' preferences or behaviors. Moreover, the social interactions are amplified through spillover effects in a network. However, there are several key differences between these two models.

The linear-in-means model basically presumes that the outcome of each individual in a group is linearly dependent upon the average outcomes and characteristics of his or her reference group (Manski 1993). Thus, social interactions are generated by group-specific averages. In this domain, individual-level variables are typically aggregated into group-level measures (Hartmann *et al*. 2008), and significant group-level variables are interpreted as the presence of neighborhood effects. One basic assumption of this model is that an individual's social interactions are grouped. This means that the populations are partitioned in N non-overlapping social groups, and individuals are only affected by people in their own groups, not by anyone outside of their groups. Since the pioneering work by Datcher (1982), much of the empirical literature on social interactions in econometrics has involved extending the general form of the linear-in-means model (Solon 1999, Durlauf and Seshadri 2003, Graham and Hahn 2005). However, these patterns of social interactions are very particular and are not likely to represent most forms of relationships between individuals, especially social networks as a whole.

Social network models using the spatial model provide a further focus on the microstructure of interactions among individuals and allow for the heterogeneity of interactions across pairs of individual actors. This stream of research is based upon Tobler's first law of geography, in that everything is related to everything else, but near entities are more related to distant entities. Since the pioneering work of Ord (1975), spatial autocorrelation models have a long tradition in geography and have been widely applied to network research in sociology and spatial econometrics. The social networks and spatial analysis approaches are mathematically very similar. This similarity is not surprising as spatial econometric approaches deal with physical space, whereas social networks address a more abstract social space. When the model refers to individuals instead of spatial units, one has a social interactions model.

In sociology, Doreian (1989) developed a network effects model, which typically specifies a network autocorrelation structure implying that individuals who are closer to each other take more similar outcomes due to the more frequent social interactions between them (Leenders 2002). Models constructed by spatial statistics formalize how observations among people relate to each other as a function of their relations. This helps capture the interdependent behaviors or outcomes in a network.

The spatial autocorrelation model has also been increasingly applied to examine the theoretical and empirical evidence of interdependent preferences in marketing and economics (Bradlow *et al*. 2005). Since the pioneering works (Pollak 1976), additional work has examined the theory and empirical evidence of interdependent preferences using spatial econometrics approaches such as interdependence in consumer expenditure allocations (Darrough, Pollak and Wales 1983; Winder and Palm 1989; Alessie and Kapteyn 1991), charitable giving (Andreoni and Scholz 1998), automobile purchases (Yang and Allenby 2003), and labor supply (Aronsson, Blomquist and Sacklen 1999).

Depending on the theory and empirical evidence, network-related interdependent or correlated outcomes in a network have been incorporated in two different ways. First, one individual's outcome may depend directly upon the outcomes of others who are linked. As such, outcomes of linked others contribute in proportion to their influence. This is formalized with the inclusion of a lagged dependent variable as an additional predictor, which is often called a simultaneously autoregressive model (SAR) with a lagged dependent variable. Alternatively, interdependence may come through error terms. This may occur when the observed dependence does not reflect a truly causal effect (homophily or unobserved heterogeneity), when there is an omitted variable, or when there exist measurement errors. This model is often called the simultaneous autoregressive model (SAR) with lagged errors. Several extended models containing both lagged dependent variables and error terms have been suggested in the network literature (e.g., Dow 1984, Doreian 1989, Friedkin 1990).

In spite of a rich development of theoretical and empirical research on interdependent outcomes within a network, few studies have been conducted identifying an individual's influence on others within a network, perhaps due to methodological challenges. The beauty of applying this approach to the social network study is two-fold. Firstly, it allows us to capture spillover effects across individuals or regions. A change in a single observation associated with any given explanatory variable will affect its own (direct impact) and potentially all other regions or individuals indirectly (indirect impact). The application of Kelejian, Travlas and Hondronyiannis (2006) examines the impact of financial contagion arising from a single country on other countries in the model using the measure of direct and indirect impacts. Anselin (1988) argues the importance of spillover effects in spatial econometrics research. In a social networking study, Katz (1953) and Bonacich (1987) apply the direct and indirect impacts approaches to measure the number of direct and indirect connections that an individual has in a network. They interpret the dependence parameter as the reflection of discount factors that creates the decay of influence for friends or peers who are located at more distant nodes. Secondly, the spatial autocorrelation model allows us to represent numerous connected customers in the overall structure of a network and model interdependent outcomes simultaneously.

For the model specifications, researchers have to decide whether social influence occurs through the autocorrelation of the dependent variable (Bell and Song 2007), through the autocorrelation of disturbances (Yang and Allenby 2003), or through a combination of the two. Anselin (2003) outlines how we can incorporate and express the direct and indirect impacts of social influence in a network in various ways with the spatial econometrics model specifications, and Leenders (2002) discusses how the various theories of social influence in sociology could be operationalized in the spatial econometrics context. Each model is discussed in Chapter 2.

1.5 Constructing Social Network

A social network is a social structure made up of individuals called “nodes” and their relationships among them are called “ties” or “edges” (Wasserman and Faust 1994). Graphs and matrices from mathematics are commonly used to represent a social network. In a graph representation, vertices and lines are used for individuals and ties respectively. A social network can also be represented by its adjacency matrix W. In a matrix form, columns and rows correspond to individuals, and each element of the matrix represents the degree of strength in the relationships between individuals. In the spatial econometrics literature, the relational structure of a network is not estimated, but is typically assumed to be known for the researcher.

The adjacency matrix *W* in a network study is not necessarily defined by geographic information, as is typically done in spatial econometrics contexts, which has given rise to a variety of discussions in the literature. For example, Anselin (1988) shows that misspecification of the neighborhood matrix by treating non-neighbors as neighbors, or vice versa leads to biased estimates (Anselin 1988). While the adjacency matrix in spatial econometrics is mostly constructed by geographical proximity, this could also be constructed by socio-demographic similarity, geo-demographic similarity, and self-reported or observed friendships. While these measures are most often binary, symmetric and non-directional (for example, whether relations between individuals exist or not), these could also be valued, asymmetric, and directional (for example, individual *i* exerts 4 times greater influence on *j* than *j* does on *i*). Despite some degree of arbitrariness in the construction of an adjacency matrix, researchers should at least provide some theoretical justification of whether a specific measure is a good proxy for capturing the structure of dependence among people.

Previous literatures have attempted to use various measures of similarity. Spatial and physical proximity have been widely used as measures that proxy the potential interdependence among consumers (e.g. Manchanda, Xie and Youn 2008; Nam, Manchanda and Chintagunta 2008; Conley and Topa 2007; Bell and Song 2007). Case (1991) shows the interdependent demand for rice consumption in Indonesia by constructing a neighbor structure using spatial adjacency. Case *et al*. (1993) also find the interdependence of public spending in neighboring regions. Smith and LeSage (2004) propose a Bayesian probit model with an individual effect using zip code information to create the neighborhood matrix and illustrate the model by applying it to the presidential election results. In marketing, Bell and Song (2007) show that customer trial behavior on an Internet retail site is affected by current customers located nearby. However, there are some drawbacks of using spatial proximity. A main limitation is that it can only capture the symmetric and non-directed social interactions among consumers. Also using adjacent zip code as a proxy for social interactions leads to the loss of information at the individual level, and the modifiable areal unit problem (*MAUP*) (Openshaw 1984).

The use of self-report data to construct social networks has much precedence in the sociology literatures (Coleman, Katz and Menzel 1966; Burt 1987; Strang and Tuma 1993; Robins, Pattison, and Elliott 2001). While sociology studies mainly focus on the social network itself, such as homophily and network structure, marketing research has more interest in the influence of social networks on customers' behaviors, such as product adoption and word-of-mouth. Using the self-reported social network among physicians, Iyengar, Van den Bulte and Valente (2010) show that there is a positive dependence in physician's adoption of a new drug in a network. Nair, Manchanda and Bhatia (2010) show that physicians' prescription behavior is significantly influenced by opinion leaders using detailed individual-level prescription data, along with self-reported social networks. Self-report measures have many advantages such as quickness and directness, but also suffer from various biases, which may affect the results, such as the social desirability bias.

The recent development of Internet technologies has allowed researchers to use rich relational data through various online sources. While most of the literature in this stream (Iyengar, Han and Gupta 2009; Goldenberg et al. 2009; Trusov, Bodapati and Bucklin 2010; Trusov, Bucklin and Pauwels 2009; Narayan and Yang 2007) explore the friendship network and measure the social influence on others in a network, Stephen and Toubia (2010) use the link formation among sellers in a large online social commerce marketplace and investigate the overall value created by adding networks.

There is little work in marketing to exploit the multiple types of links between individuals to construct social networks. Geographic proximity and demographic similarity (Yang and Allenby 2003; Choi, Hui and Bell 2010), self-reports and spatial proximity (Iyengar, Van den Bulte and Choi 2011) and spatial proximity-based social networks down-weighted by demographic distances (Sorensen 2006) have been used to explore multiple social network structures. Moreover, Yang and Allenby (2003) find that geographic proximity is somewhat more important than demographic similarity in explaining the interdependent preferences for automobile choice.

1.6 Positioning of This Research

The positioning of the current study in the context of the relevant literatures discussed in the previous sections can be summarized as follows:

1. This research measures the value of individual-level influence on others within a network, whereas spatial autocorrelation models in social network studies place more attention on modeling the overall dependence of outcomes among connected customers.
2. The current study takes into account both individual and network effects when measuring individual value in a network. Most previous studies either focus on individual positions in a network or an individual’s characteristics such as the

quantity of purchases and customer lifetime value, respectively. It is important to consider both effects in order to evaluate an individual's value in a network more accurately.

3. An individual's importance in a network is measured in terms of revenue generation to the firm using actual purchase data. There are few previous empirical studies investigating the impacts of social influence on actual purchases within a social network.
4. We propose an extension of the static network model into the dynamic model with temporal dimensions. Dynamic approaches allow us to quantify the dynamic value of individual level influence in a network from a long-run perspective.

CHAPTER II

IDENTIFYING HIGH VALUE CUSTOMERS IN A NETWORK: STATIC APPROACH

2.1 The Model

The proposed model is formulated in two steps. We first introduce the simultaneous autoregressive model (SAR), which captures the potential interdependence among observations of consumers' purchases in a network. We then describe how this model can be used to identify high value consumers in revenue generation.

2.1.1 Simultaneous Autoregressive Model (SAR)

Suppose we observe the purchases of a set of consumers ($i = 1, \ldots, N$) who are connected with one another through a network. Let us assume that individual i's purchase is influenced directly by connected others' purchases in the network. In this case, the estimates of standard ordinary least squares (OLS) will be biased and inconsistent due to the violation of independence assumptions of OLS (Whittle 1954). Because levels of dependent variable, Y depend on connected others' purchase, the correct model should account for these dependencies among people.

The simultaneous (or sometimes Spatial) Autoregressive model (SAR) provides a parsimonious form to account for these dependencies by incorporating a lagged dependent variable or autoregressive term, ρWY, on the right-hand side of standard OLS. The idea is that the level of an individual's DV may depend on a weighted sum of DV across connected others in a network. The SAR model is expressed as (1):

$$Y = \rho WY + X\beta + \varepsilon, \ \ \varepsilon \sim N(0, \sigma^2 I_n) \tag{1}$$

where ρ is the network autocorrelation parameter, which represents the overall dependence of DV among people in a network, Y is a $n \times 1$ matrix of the dependent variable, X is a $n \times k$ matrix of explanatory variables, which includes a unit vector for intercepts, and β is a $k \times 1$ vector of the regression coefficients. The matrix W is an $n\times n$ adjacency matrix, which represents the strength of potential interactions among people. The error term, ε, is assumed to follow a multivariate normal distribution with a zero mean and variance $\sigma^2 I_n$.

To construct the social network structure, we use the $n \times n$ adjacency matrix W for n number of individuals. Each element of the adjacency matrix, W_{ij} indicates the strength of interactions from individual i to j.

$$W_{ij} = \begin{cases} 0 & \text{if there is no interaction between } i \text{ and } j \\ z & \text{if the strength of interaction from } i \text{ to } j \text{ is } z \end{cases} \tag{2}$$

Since individuals are typically not considered to interact with themselves (Yang and Allenby 2003; Anselin 1988), diagonal elements of the W matrix, w_{ii} are assumed to be zero. According to the literature (Anselin 1990; LeSage and Parent 2009), the adjacency matrix W is usually row-standardized by dividing each element by row-sums to sum up to one. The normalization of the adjacency matrix allows us to interpret WY as a weighted average of neighboring values.

The parameter of the lagged dependent variable, ρ is interpreted as an overall degree of interdependence among the people in a network. Thus, $\rho = 0$ implies that there is no dependence in the observed outcomes, and it simplifies the SAR model in (1) to the standard OLS. A significant and positive ρ indicates the existence of interdependence in the observed outcomes.

We illustrate how the adjacency matrix W was contructed in the SAR model using a small sample in the following example. Assume we create a 5×5 adjacency matrix C for 5 individuals, as shown below.

$$C = \begin{pmatrix} 0\ 4\ 8\ 8\ 0 \\ 4\ 0\ 8\ 2\ 2 \\ 8\ 8\ 0\ 8\ 4 \\ 8\ 2\ 8\ 0\ 8 \\ 0\ 2\ 4\ 8\ 0 \end{pmatrix} \tag{3}$$

C_{12}=4 and C_{53}=4 represents the strength of interactions from individuals 1 to 2, and from 5 to 3, respectively. We can see here that all of the diagonal elements are assumed to be zero. Then, the adjacency matrix C is row-normalized by dividing all elements by row-sums as the following.

$$W = \begin{pmatrix} 0.00 & 0.33 & 0.66 & 0.66 & 0.00 \\ 0.42 & 0.00 & 0.85 & 0.21 & 0.21 \\ 0.55 & 0.55 & 0.00 & 0.55 & 0.27 \\ 0.57 & 0.14 & 0.57 & 0.00 & 0.57 \\ 0.00 & 0.21 & 0.43 & 0.87 & 0.00 \end{pmatrix} \tag{4}$$

We see in the following (5) that the product of matrix W and the dependent variable Y can be interpreted as a weighted value of the networked values.

$$Y = WY = \begin{pmatrix} 0.00 & 0.33 & 0.66 & 0.66 & 0.00 \\ 0.42 & 0.00 & 0.85 & 0.21 & 0.21 \\ 0.55 & 0.55 & 0.00 & 0.55 & 0.27 \\ 0.57 & 0.14 & 0.57 & 0.00 & 0.57 \\ 0.00 & 0.21 & 0.43 & 0.87 & 0.00 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} \tag{5}$$

$$= \begin{pmatrix} 0.33\mathrm{Y}_2 + 0.66Y_3 + 0.66Y_4 \\ 0.42Y_1 + 0.85Y_3 + 0.21Y_4 + 0.21Y_5 \\ 0.55Y_1 + 0.55Y_2 + 0.55Y_4 + 0.27Y_5 \\ 0.57Y_1 + 0.14Y_2 + 0.57Y_3 + 0.57Y_5 \\ 0.21Y_2 + 0.43Y_3 + 0.87Y_4 \end{pmatrix}$$

2.1.2 Natures of Social Interactions in SAR Model

To see the nature of social interactions in the SAR model, we consider the data-generating process of the model. The SAR model in (1) can be rewritten as (6).

$$(I - \rho W)Y = X\beta + \varepsilon \qquad (6)$$

Assume that $(I - \rho W)$ matrix is invertible and the absolute value of ρ is less than one. Then (6) can be expressed as the following:

$$\begin{aligned} Y &= (I - \rho W)^{-1}(X\beta + \varepsilon) \qquad (7) \\ &= \textstyle\sum_{k=0}^{\infty}(\rho W)^k\,(X\beta + \varepsilon) \\ &= (I + \rho W + \rho^2 W^2 + \rho^2 W^2 + \rho^3 W^3 + \cdots)(X\beta + \varepsilon) \end{aligned}$$

In the previous section, we construct the adjacency matrix W using the information of the first-order neighbors, those who are directly connected with each other. Mathematically, the matrix W^2 represents the second-order neighbors, those who are neighbors to the first-order neighbors. Increasingly, the W^k matrix represents the k^{th}-order neighbor meaning that people can be reached in k^{th} relational jumps. The mathematical proof of properties of the adjacency matrix W^k is provided in APPENDIX A.

Form equation (7), we can interpret the data-generating process of the SAR model as indicating that the dependent variable, Y is generated by combinations of the network effect, $(I + \rho W + \rho^2 W^2 + \rho^2 W^2 + \rho^3 W^3 + \cdots)$ and the individual characteristics effect, $X\beta$. The infinite power series of ρW in (7) allows us to capture the spillover effects into the higher-order neighbors. As it goes toward the higher-order neighbors, say k, the network effects are discounted by ρ^k. These make sense, since individuals are usually affected more by nearby individuals than distant ones. Moreover, the infinite power series of ρW allows us to capture feedback effects. Since each individual is a second-order neighbor to him or herself by definition, the impact of its own change gets back to itself after passing through its neighbors.

2.1.3 Identifying Individual Influence on Revenue Generation

The data-generating process in (7) can be rewritten as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} A_{11}\ A_{12} \ldots A_{1n} \\ A_{21}\ A_{22} \ldots A_{2n} \\ . \\ . \\ . \\ A_{n1}\ A_{n2} \ldots A_{nn} \end{pmatrix} \begin{pmatrix} X_{11}\beta_1 & X_{12}\beta_2 & \ldots & X_{1k}\beta_k \\ X_{21}\beta_1 & X_{22}\beta_2 & \ldots & X_{2k}\beta_k \\ & . & & \\ & . & & \\ & . & & \\ X_{n1}\beta_1 & X_{n2}\beta_2 & \ldots & X_{nk}\beta_k \end{pmatrix} + A\varepsilon \qquad (8)$$

where $A_{ij} = (I - \rho W)^{-1}{}_{ij}$, i, j represents each individual, and k denotes the explanatory variables.

By taking the partial derivative of Y_i with respect to X_{ik}, we can measure the impact of the i^{th}change of the k^{th} variable for individual i (LeSage and Pace, 2009).

$$\partial E(Y_i)/\partial X_{ik} = \beta_k A_{ii} \qquad (9)$$

Similarly, the derivative of Y_j with respect to X_{ik} allows us to measure the impact of the i^{th} change of the k^{th} variable on everybody else in the system.

$$\sum_{j \neq i} \partial E(Y_j)/\partial X_{ik} = \sum_{j \neq i} \beta_k A_{ji} \qquad (10)$$

That is, the diagonal elements of matrix $A(A_{ii})$ and the off-diagonal elements of matrix A ($A_{ij},\ i \neq j$) indicate the impact of i^{th}change on its own, or everybody else in the system, respectively. Thus, to compute the overall impact of i^{th}change on the system, we need to sum down the i^{th} column of matrix A. To illustrate how these spillover effects work in the SAR model, we provide a small sample in APPENDIX B as an example.

We define individual i's value to the firm in a network as the impact of the i^{th} changes of explanatory variables on the system weighted by the level of individual i's characteristics. This is formalized as follows

$$VALUE^{(i)} = \sum_{k=1}^{r} X_{ik} \times \left(\sum_j \partial E(Y_j)/\partial X_{ik}\right) = \sum_j \sum_{k=1}^{r} X_{ik}\, \beta_k A_{ji} \tag{11}$$

2.2 Model Estimation

In the current study, we estimate the SAR model in (1) with maximum likelihood. The log-likelihood for Y is in the form of (12)

$$lnL(y) = -\left(\frac{N}{2}\right) ln(\pi\sigma^2) + \ln|I - \rho W\,| - \frac{\varepsilon'\varepsilon}{2\sigma^2} \tag{12}$$

where $\rho \in (min(\omega)^{-1}, max(\omega)^{-1})$, ω is the eigenvalue of the matrix W.

When the matrix W is symmetric, Ord (1975) shows that the condition, $\rho \in (min(\omega)^{-1}, max(\omega)^{-1})$ in (13) ensures a positive definite variance-covariance matrix. More discussions about the bounds for the dependence parameter, ρ are provided in APPENDIX C. The log-determinant of $(I - \rho W)$ is the logarithm of the sum of a collection of scalar values including all eigenvalues of the adjacency matrix W, as follows:

$$|I - \rho W| = \prod_{i=1}^{n}(1 - \rho\omega_i) \rightarrow \ln|I - \rho W| = \sum_{i=1}^{n} \ln(1 - \rho\omega_i) \tag{13}$$

Pace and Barry (1997) suggest a convenient form of a concentrated log-likelihood with respect to β, and σ^2 as follows:

$$lnL(\rho) = \kappa + \ln|I - \rho W| - \frac{n}{2}\ln(\varepsilon(\rho)'\varepsilon(\rho)) \tag{14}$$

where κ is a constant that does not depend on ρ, $\varepsilon(\rho)'\varepsilon(\rho) = \varepsilon_o'\varepsilon_o - 2\rho\varepsilon_o'\varepsilon_s + \rho^2\varepsilon_s'\varepsilon_s$, $\varepsilon_o = Y - X\hat{\beta}_O$, $\varepsilon_s = WY - X\hat{\beta}_S$, $\hat{\beta}_O = (X'X)^{-1}X'Y$, $\hat{\beta}_S = (X'X)^{-1}X'WY$.

Several methods to optimize the log-likelihood in (14) with respect to ρ have been suggested, such as the vectorized approach (Pace and Barry 1997) and the closed form solution (LeSage and Pace 2009).

Parameter estimates of other coefficients can be obtained by substituting $\hat{\rho}$ into the following:

$$\hat{\beta} = (X'X)^{-1}X'Y - \hat{\rho}(X'X)^{-1}X'WY = \hat{\beta}_O - \rho\hat{\beta}_S \tag{15}$$

$$\hat{\sigma}^2 = \left(\frac{1}{N}\right)(\varepsilon_o - \hat{\rho}\varepsilon_s)'(\varepsilon_o - \hat{\rho}\varepsilon_s) \tag{16}$$

$$\hat{\Sigma} = \hat{\sigma}^2[(I - \hat{\rho}W)'(I - \hat{\rho}W)]^{-1} \tag{17}$$

2.3 Data and Measures

We used unique data on online gaming behaviors from a company in Korea. We disguise the name of the game and the company for the purpose of confidentiality. Our datasets cover the Korean online game market, which is the largest and the first that this company launched. The monthly gaming data between March 2010 and August 2010 randomly selected by the firm are used. This game was released in Korea on July, 2009 by the largest online game publisher and developer in Korea. Currently, there are 0.3 million registered users and approximately 20,000 concurrent gamers. It has been quite successful in the market and is still in the growing stage.

The model is calibrated on cross-sectional data during March 2010. In addition, a cross-sectional data for April 2010 are used for the holdout sample. Our data set comprises 1,000 gamers. We first estimate our proposed model. We then compute the individual values using equation (11). By rank ordering people based on their values to the firm, we identify people who are more valuable in revenue generation. Table H1 presents the summary statistics and the description for the primary variables of interest.

This game is a typical Massively Multiplayer Online Role-playing Game (MMORPG) in which a very large number of players interact with one another within a virtual game world. MMORPG is distinguished from single-player or small multi-player RPGs by the number of players and by the persistent virtual world that continues to exist and evolve, even while the player is away from the game. Player interaction is one of the most important aspects of MMORPG. In some cases it is vital to work together as a team due to some quests requiring a number of players together to successfully complete them. Coordinated combat or hunting is another mechanism through which people cooperate and socialize. One other social form within the game is the possibility of "guilds," organizations of groups of like-minded players who band together to achieve large goals. These guilds have a hierarchal structure of leaders who are elected to manage the guild. Complex politics often play a large part in these guilds with many decisions being made about which new members to recruit and how special items are distributed among members. Due to the importance of player interaction in MMORPG, it has been believed that friends in the network influence each other's purchase or gaming behavior.

Revenue. This game is basically free. There is no charge to download and play the game. However, players need to pay to get special or rare in-game items to enhance the gameplay (better equipment and potions) or to decorate their avatars (ornaments and accessories). The remaining 40% of revenue is from the Internet café. This Internet café pays close to 4.15 won per minute usage to the company and then charges its consumers in the café to use the computers for approximately 1500 won per hour (The won is the unit of currency in Korea). Thus, revenue can be generated from fees paid by gamers in the form of game item sales and by the owners of Internet cafés for access time. There are two revenue sources for the company: purchases by gamers and charge for spending time in online cafés. Nearly 60% of revenue is generated by gamers.

Because what we are trying to predict for the company is revenue, we construct the focal variable to the measure of revenue at the individual level, considering two revenue sources as follows:

$$Revenue = ln[\{Purchase\ of\ Game\ Items) + (4.15 * Game\ Time) * I\} + 1] \quad (18)$$

where *I* is an indicator function of whether the gamer plays the game at home (*I* = 0) or at an Internet café (*I* = 1), and the average charge of using an online café is 4.15 won per minute. Log transformation is used to get a normal distribution. Distribution of *Revenue* and log transformation of *Revenue* are provided in Figures H1 and H2, respectively. Since most in-game items to be purchased by real money are available to all gamers, there are no availability issues.

The IP addresses of gamers' log-in data allow us to identify the locations of users (Internet café or Home). This distinction is important for us because playing games only in Internet cafés could be a revenue source to the firm. Because gamers have a strong preference for game-playing locations, either in Internet cafés or at home, we use the indicator function *I* to construct the revenue variable.

Gaming Behaviors. The log-file automatically created by a company's server provides us with 13 variables that describe the user's game playing behaviors. To identify groups of variables that are related, we use the principal component analysis (PCA). The "eigenvalue-greater-than-1" criterion and scree plot in Figure H3 suggested retention of four factors.

To make factors interpretation easier, the varimax-rotated method is used. Each factor's indices are then used as new input variables. For the panel analysis in Chapter 3, we perform PCA for all observations from March to August and obtain the formula. Then using this formula, we perform PCA for each month of observation, respectively. We categorize each factor in Table H2 as follows.

- *Social Status*: Level of Character, Experience Score, Reputation, Guild Master, Time to Level Up
- *Guild Activity:* Guild Membership, Number of Guild Members, Number of Guild Wars
- *Action on People:* Death Match (DM) Rating, Number of Wanted, Number of Players Killed (PK)
- *Action on World:* Number of Monsters Hunted, Number of Building Constructed

<u>Demographic Variables.</u> *Age* and *Gender* are collected at the individual level, and *House Value*, and *Education* information at the zip level are provided by the 2005 Korean Population and Housing Census. Values of the *Gender* variable are expressed as a dummy variable: male is defined by 1, and female by 0. The *House Value* variable is used as a proxy for income, as is typically used in Korea. The *Education* variable represents the percentage of individuals with a bachelor's and/or graduate or professional degrees. Because of the collinearity with *House Value*, the *Education* variable was dropped in the study of Chapter2.

<u>Virtual Money.</u> The *Virtual Money* variable is measured by the proportion of "how much they spent" to "how much they are available" during this month to control for the service provided by the company. Both real and virtual money can be used to purchase game

tokens. However, game tokens can only be purchased by one source of money. While merchandise that could be purchased by real money mainly concern those items that could enhance the game capability, items that could be purchased by game money are mainly related to exteriors, which do not have an effect on the game itself. Game money could be earned in the process of gaming by completing quests, mining gold or hunting etc. By using real money, gamers can avoid long hours endeavoring to accumulate game money. This may trigger less motivation for players to spend more time earning more experience and improving their gaming capability. Also, this may explain the relatively low correlation between game time and payment.[1] Thus, to control for the possible influences of game money on the revenue of the firm, the *Virtual Money* variable is used as an input in our analysis.

Strength of Social Interactions. The element of the adjacency matrix, w_{ij} represents the strength of social interactions between game player *i* to *j*. In MMORPG, gamers are able to communicate with other players by sending messages. Following the sociology literature (Leenders 1997), we assume that the strength of social interaction can be captured by the strength of communication between people.

Estimation of the parameters of the SAR model using Maximum Likelihood (ML) theory is computationally expensive because of the need to compute the logarithm of the determinant $(I - \rho W)$ of a large matrix in the log-likelihood function.

From (6), we can see that the variance-covariance for the SAR model is

$$E[(I - \rho W)^{-1} \varepsilon \varepsilon' (I - \rho W)^{-1\prime}] \tag{19}$$

[1] For the balance between the increase of item sales and continuity of game participation, items that could be purchased by real money or game money are separated (but not completely).

We need to ensure that $(I - \rho W)$ is non-singular and that the product $(I - \rho W)^{-1}(I - \rho W')^{-1}$ that equals the variance-covariance matrix is positive-definite. Asymmetric relations among people are likely to be more realistic, but an asymmetric weight matrix can have complex eigenvalues. Symmetry might be too strong a requirement to constitute a realistic model of social interaction in general. However, its simplicity makes it appealing. Bavaud (2002) suggests the use of a similar matrix having the same eigenvalues as a symmetric matrix to ensure non-singularity of $(I - \rho W)$ and a positive-definite variance covariance matrix.

To ensure symmetry, we can rely on the use of the quasi-symmetry matrix, first proposed and investigated by Caussinus (1965). A quasi-symmetric matrix is similar to the symmetric matrix due to equivalent eigenvalues. Quasi-symmetric weight matrix possesses real eigenvalues (Bhatia, Kittaneh and Li 1998), which simplifies computing the logarithm of the determinant. When this quasi-symmetric weight matrix is row-normalized, the largest characteristic root is greater than zero, and the smallest characteristic root is smaller than zero (Ord 1975). These properties facilitate the maximum likelihood estimation of ρ and ensure the invertibility of the matrix $(I - \rho W)$ (Ord 1981).

The matrix $\mathrm{C_{ij}}$ is quasi-symmetric if it can be written as $\mathrm{C_{ij}} = \mathrm{A_i B_j S_{ij}}$ where $\mathrm{S_{ij}} = \mathrm{S_{ji}}$, $\mathrm{A_i} =$ row values and $\mathrm{B_j} =$ column values. In this research, we assume that A and B are identical, so that

$$\mathrm{C_{ij}} = \mathrm{B_i B_j S_{ij}} \tag{20}$$

We use the total number of message exchanges between person *i* and *j* for $\mathrm{S_{ij}}$, the total number of messages sent by individual *i* for $\mathrm{B_i}$, and the total number of messages sent by *j* for B_j. The neighbor matrix W_{ij} is obtained by dividing each element of (20) by row-sums, which allows us to interpret *WY* as a weighted average of the neighbor's value. That is, W_{ij} can be expressed as

$$W_{ij} \propto B_j S_{ij} \tag{21}$$

Equation (21) indicates that the impact of person *j* on person *i* in the network structure depends upon the activity of individual *j* (B_j) and the amount of social interaction between *i* and *j* (S_{ij}).

2.4 Empirical Analysis

The procedure of the empirical analysis involves two steps. As a first step, we compare our proposed simultaneous autoregressive model to reduced benchmark models, demonstrate better performance of our model, present parameter estimates and interpret them for the calibration sample (March, 2010). In the second step, the predictive performance of the models is verified on the holdout sample (April 2010). Finally, we discuss the managerial implications and summarize.

2.4.1 Test of Interdependence

Interdependent outcomes between individuals could be explained by different sources: direct influence from friends (lagged dependent variables) or omitted factors such as the existence of a local marketing campaign (lagged error terms). However, these are not always easy to distinguish. Because the theoretical basis for the sources of interdependent outcomes is not that clear in the current study, several statistical tests are performed for model specification which include Moran's *I* test, the Lagrange Multiplier (LM) test and the Robust LM test for error dependence, as well as the LM and the Robust LM tests for lag dependence (Anselin 2001). The null hypothesis for Moran's *I* and LM tests for error dependence is that $\lambda = 0$ in (14) when there is no spatial lag dependence, that is, $\rho = 0$ in (6) whereas the Robust LM for error dependence tests the null hypothesis that $\lambda = 0$ but does not restrict ρ to be zero. Similarly, the LM test and Robust LM test for lag dependence is the other way around. That is, the Robust LM lag test is used for checking lag dependence, correcting for the presence of error dependence.

Table H3 provides the test statistics for six tests. We can see that both simple tests of the lag and error are significant, thereby indicating the presence of dependence. The robust tests help us discriminate what types of dependence may be at work. From the test statistics of the Robust LM (error) and the Robust LM (lag), we found that the spatial lag should be responsible for the underlying-dependence process, since the test statistics in the Robust LM (lag) are significant at 1%, while the Robust LM (error) are not. This result means that when the lagged dependent variable is present, the error dependence disappears. Based on these test statistics results in Table H3, the following estimation will use the spatial lag specification as shown in (1), in which the error term is assumed to follow an *iid* error structure.

2.4.2 Model Comparisons

We evaluate the model in terms of model fits and the predictive capability of models in identifying an individual's influence on revenue generation to firms, respectively. To do so, we estimate the models for two types of data: in-sample data that occurred during March, 2010, and additional data (out-of-sample) that originate from the same source (N=1,000), but are taken from data that occurred during April, 2010. For the holdout sample, we keep using the same value for the demographic variables such as *Age*, *Gender, and House Value*, but update information for other variables, such as *Virtual Money and Gaming Behaviors*. The *W* matrix is assumed to be constant.

Evaluation of the model involves two steps. For the model fits, the proposed SAR model is compared with two other benchmark models, based on the Akaike Information Criterion (*AIC*). For evaluations of predictive capability of the model, Mean absolute deviations (*MAD*) and correlation of the rank orders between forecasts and actual observations are used, respectively. The *MAD* in the current study is simply the arithmetic mean of the absolute differences between the predicted and actual value:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{22}$$

To rank order people in terms of individual value in a network, we examine Spearman's rank correlation between forecasts by each measure and actual values.

2.4.2.1 Benchmark Models

Two benchmark models are used to evaluate the fits of the proposed SAR model. The starting point of the empirical analysis is a parsimonious model. We call the first benchmark model, the "Network" model, where we consider the network effects only. It has the following specification.

Network model: $Y = \rho WY + \varepsilon$

We refer to the second benchmark model as the "Segmentation" model, where we assume independence in observations and consider an individual characteristics effect only.

Segmentation model: $Y = X\beta + \varepsilon$

We call our proposed model the "SAR Interaction" model, which considers both network and individual characteristic effects simultaneously.

SAR Interaction: $Y = (I - \rho W)^{-1}X\beta + \varepsilon$

All models are assumed to have intercepts.

2.4.2.2 Predictive Performance

The model comparison is presented in Table H4 for the calibration and holdout samples, respectively, for 3 different models. To verify the predictive performance of the models for the measurement of individual value on revenue generation, we use two different measures: *MAD* and the correlation of rank orders between forecasts and actual observations. For the comparison of correlations, two centrality measures from the sociology literature, degree and betweenness (Freeman 1977, 1979) are used as additional benchmark measures. For details of these centrality measures, refer to APPENDIX D. Comparisons to those centrality measures are important because they only consider the network effects, whereas our proposed model considers both network and characteristics effects.

The model fit statistics in Table H4 (both in-sample and out-of-sample) indicate the following results. First, while the proposed SAR interaction model leads to an improved in-sample and out-of-sample fit, the Network Model is the worst-fitting model in terms of *MAD*. The concordance of the predictive capacity of out-of-sample data and that of in-sample data confirms good performance of the proposed model. Second, the difference of the out-of-sample fit statistics between the Segmentation Model and the SAR Interaction Model is modest in size. This may indicate that individual characteristics contribute more to individual revenue to the firms than network effects. We will return to this issue in the following section.

To see which effects drive an individual's value on revenue generation, we create a 3D scatter plot. Figure H4 displays a 3D scatter plot of an individual's value, obtained from the data-generating process of the Network Model, the Segmentation Model, and the SAR Interaction Model for in-sample data. Figure H4 shows that an individual's value in a network goes up as the network and segmentation effects go up. The diagonal trend in Figure H4 provides evidence that the segmentation effects contribute more to generating revenues.

To quantify the relationship between segmentation and network effects, we estimate the following:

$$\begin{aligned}&\textit{SAR Interaction Effects}\\&= a + b \times \textit{Network Effects} + c \times \textit{Segmentation Effects}\end{aligned} \tag{23}$$

This could be rewritten as:

$$\begin{aligned}&\textit{Network Effects}\\&= (\textit{Interaction Effects} - a) / b - c / b \times \textit{Segmentation Effects}\end{aligned} \tag{24}$$

Parameter estimates in (24) are shown in Table H5. We obtain 3.286 of c / b, which means that the segmentation effects are approximately 3 times larger than the network effects.

An analysis of the contour maps shows that the segmentation effects are dominant in revenue generation. That is, people who have more connections with others and who have a central position in a network are not necessarily those who generate revenues for the firm. These findings are consistent with previous research in marketing (Stephen and Toubia 2010; Trusov, Bodapati and Bucklin 2010).

Second, we compute Pearson correlations between the predicted and actual revenue by each model in Table H6. Results are in line with those in Table H4, indicating that our proposed model again outperforms other benchmark models. The segmentation model falls slightly behind our proposed model but much better than the other models. Spearman rank order correlations between predicted and actual revenue by different models are also provided in Table H7. Models only considering network effects do not perform as well as the complete model. This result confirms our finding that people who have more connections with others and who have a central position in a network are not necessarily those who generate the most revenue for a firm. It is noteworthy that Katz (1953) and Bonacich (1987) in a sociology study construe the network-only model as a measure of centrality of individuals in a social network. The so-called Katz-Bonacich Centrality measures the number of direct and indirect connections that an individual has in a social network.

Adopting the customer relationship management (CRM) literature, we also examine the Gini coefficient in evaluating the predictive performance of our proposed model. The Gini curve has been widely used in CRM studies to examine the predictive power of model. The idea is that the horizontal axis is based upon the predictions of the model. The vertical axis shows the actual purchase behavior of the individuals. The 45-degree line is what is expected of a method that sorts people randomly. The best forecasts show the most deviations from this random-sort line. As we start from zero on the graph, we are actually going down the list from top to bottom (best to worst). The fact that the line of our proposed model is very high in higher-order customers in Figure H5 indicates that our method in rank-order customers is particularly good at identifying people at the top of the revenue list. These are obviously the people we care most about for marketing purposes. This result is also confirmed in the revenue curve in Figure H6.

In sum, model comparisons in this section suggest that incorporating both network and segmentation effects improve the predictive power of the model for the observed correlated outcomes in revenue.

2.4.3 Parameter Estimates and Interpretation

Parameter estimates of our proposed model are given in Table H4 above. The significantly positive parameter ρ (0.368) in Table H4 indicates that there are network or spillover effects of consumer's consumptions in online games, that is, revenue to the firm in our empirical analysis. Thus, people who are nearby tend to affect each other's consumption. This finding is consistent with past work regarding neighborhood effects on automobile purchase behaviors by Yang and Allenby (2003).

A strong and positive ρ leads to spillover effects in our datasets, as shown in Table H4. Since the parameter $\rho < 1$, parameter ρ could be interpreted as a discount factor that creates the decay of influence as we move to more distant friends. That is, close friends are more influential than distant friends.

Parameter estimates in Table H4 indicate that gamers who generate more revenues to the firm are more involved with the game, have higher income, are more socialized and spend more virtual money. Even though age and gender variables are not significant, sign of parameter estimates of these variables might imply that as gamers who are younger and male, are more likely to generate revenue to the firm.

2.4.4. Profiling Customers by the Individual Level Impacts

A change in a single observation associated with any given explanatory variable will affect its own (direct impacts), potentially all other individuals indirectly (indirect impacts) and both (total impacts). Anselin (2003) provides the individual-level measure of direct, indirect and total impacts on the dependent variable as the following (Anselin 2003).

Direct Impact (i) = A_{ii} (25)

Indirect Impact (i) = $\sum_{j \neq i} A_{ji}$ (26)

Total Impact (i) = $\sum_{j} A_{ji}$ (27)

where the n by n matrix, $A = (I - \rho W)^{-1}$, and individual i.

Distributions of direct, indirect and total impacts at the individual level are shown in Figures H7, H8 and H9. Summary statistics of these three impacts are provided in Tables H8 and H9. Using individual-level measures of customer direct and indirect influences on others in terms of revenue generation to firms as a dependent variable, we profile customers by running simple regression.

Parameter estimates of OLS in Table H10 indicate that the degree of influence on oneself or others could be explained linearly by some covariates we have; thus, we can use this to describe who is more influential on other's revenue-generating decisions in a network. We use the same covariates in this analysis as those in the SAR Interaction model. These covariates accounts for a large proportion of the observed variation in the direct and indirect influences: the adjusted R^2 is 40.6% and 55.1%, respectively, and the F-statistics are highly significant.

These reveal an interesting characterization of gamers who have stronger spillover (indirect) effects, that is, influence on others, and the findings make sense intuitively. Gamers who exert a higher influence on others are the ones who are more into the game: they are more active in a guild, have more interaction with other gamers, have higher social status in the virtual world, and have more virtual money. To profile these gamers, population is divided into two groups, which have higher and lower influence on others, and the mean of each group is examined in Table H11.

2.5 Additional Issues

2.5.1 Endogeneity

Our empirical results may suffer from an endogeneity bias due to the potential simultaneous relationship between independent (various gaming behaviors) and dependent variables (purchase of in-game items and choice of gaming time), in that purchasing game items is also part of game playing. Ignoring a possible endogeneity problem leads to inconsistency of parameter estimates.

One method to account for the potential simultaneous bias is to use an instrumental variable procedure (Greene 2003). If we assume that the lagged variables are uncorrelated with the error term, simultaneous endogeneity could be corrected to some degree by using lagged variables. We use the lagged values by 1 month for *Virtual Money*, and 4 *Gaming Behaviors*, which have changed over our observation period of March through August 2010. In this specification in Table H12, all estimates of the endogeneity-corrected model in the second column are fairly robust to their counterparts of their original model in the first column. The only noticeable change is the estimated level of magnitude. We did not correct the standard errors for the instrumental variable procedure. Nevertheless, the simultaneity bias does not seem to have any significant bearing on our main findings, so standard error correction also may not be a major issue.

2.5.2 Decomposition of Revenue Sources

Revenue can be generated from two different sources for the firm: fees paid by gamers in the form of game item sales or by owners of Internet cafés for access time. We can see some differences of distribution for total revenues—game time payment, and item purchase payment—in the following Figures H10, H11, and H12. Also a higher correlation (0.486) between total revenues and item purchase payment than that (0.182) between total revenues and game time payment in Table H13 implies that item purchase payment drives more revenue to total revenue. Since the consumer purchase decision process may be different for the two decisions, we separately analyze game time payment and item purchase payment.

When we deal with game time payment and item purchase payment, there are a number of zero observations shown in Figures H11 and H12, which results in a truncated distribution for dependent variable observations. To account for some zero observations in the dependent variables, we apply the Spatial Tobit model for the analysis.

The latent regression model, when censoring occurs at zero, takes the form in (28). The observable variable y_i is defined to be equal to the latent variable y_i^* whenever value of this variable is above zero and zero otherwise.

$$y_i = \begin{cases} y_i^* & if\ y_i^* \geq 0 \\ 0 & if\ y_i^* < 0 \end{cases} \text{ where } y^* = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon \qquad (28)$$

The adjacency matrix W constructed in the previous sections is also used for the analysis. The estimation scheme of the SAR Tobit model is detailed in APPENDIX E. The parameter estimates of the SAR Tobit model using item purchase payment, and game time payment as a dependent variable are reported respectively in Table H12. Since all estimates of endogeneity correction models in Table H12 are fairly robust to their counterparts reported in their original model, we argue that endogeneity may not be a big problem in our dataset. One noticeable finding here concerns the change of magnitude of the overall dependence measures in the system with different dependent variables. Overall social influence (ρ) for the prediction of item purchase is higher than that for the prediction of game time payment. While game time payment of peers is not observable, peers' game item purchases are more visible to gamers. This empirical finding implies that people are more influenced by observable activities. Since endogeneity and different revenue sources does not seem to be a big problem in our dataset, we will focus on the proposed SAR model in section 2.1, which uses total revenue (revenue from game item purchase and from game time payment) as the dependent variable in the next chapter.

For the simplification of computing forecasts, we treat the censored observations as observed, which are replaced for the latent values of y_i^* and apply the same logic used for the SAR model to measure the individual's influence on revenue generation as in (20).

Because each model predicts different things, our interest here is how a similar rank order is predicted by each model, and why this works. Spearman rank order correlations between each model in Table H13 shows that correlations between the SAR Tobit model with Item Purchase and the SAR model with Total Revenue is higher than that between the SAR Tobit model with Game Time Payment and the SAR model with Total Revenue. This finding might be explained by the fact that in Table H14, *Item Purchase* contributes more to *Total Revenues* than *Game Time Payment*.

CHAPTER III

IDENTIFYING HIGH VALUE CUSTOMERS: DYNAMIC APPROACH

Observed correlated behaviors amongst customers in the network can occur due to different reasons such as social influence, homophily and contextual effects (Manski 1993; Moffit 2001; Hartmann *et al*. 2008). Individual behaviors tend to be affected by significant others such as friends, family or neighbors, which is referred to as social influence. Homophily occurs when individuals tend to behave similarly as their friends because they have similar background characteristics such as socioeconomic status. Alternatively, individuals tend to behave similarly when they are exposed to the common contextual effects such as marketing interventions, or economic shocks.

An example of clustered observations of smoking behaviors may help clarify these distinctions. Social influence exists if smokers acquire smoking habits from their friends. Homophily exists if people tend to select friends who are similar, in this case, smokers. Or Smokers are exposed to a common marketing mix such as cigarette advertisements.

Distinguishing these potential sources is important because each has different policy implications. Consider, for example, marketing actions targeted to some of these customers. If purchase decisions are affected by friends, then effective marketing programs aimed at the right customer not only directly increases sales by target marketing, but also indirectly has an impact on the increase of sales of his or her friends through the connections they have in a network. Homophily and contextual effects do not generate this social multiplier effect.

The proposed model in Chapter 2 shows good predictive performance. Even so, the model can be potentially improved by incorporating dynamic dimensions, which can provide us with some information (i.e., temporal spillover effects) not available from our proposed static model. This step will make the current model of identifying individuals' influence on revenue generation more general within a network.

The objective of Chapter 3 is to measure an individual's value on revenue generation from a long-term perspective. To do so, firstly it discusses how time dimensions can be incorporated into the base SAR model in (1) and suggests a general framework of the spatial dynamic panel model. Then it describes how to apply this model to quantify a cumulative individual's value in a network over time in revenue generation. In a third section, alternative models or additional issues are discussed. Finally, it concludes with a summary.

3.1. Dynamic Model

3.1.1. General Framework

Here, we utilize a general form of spatial dynamic panel model suggested by Anselin (2001) as a "time-space dynamic model" to account for homophily and peer influence, as well as temporal inertia effects.

A customer's current period behavior tends to be influenced by his or her own past period behavior due to habit formation (Ronis *et al*. 1989) or persistence (Fishbein and Ajzen 1975), and this tendency has nothing to do with social influence. Thus, the omission of this variable in our model specification may lead to the overestimation of social influence. To account for this potential serial lag effect termed as inertia in our study, ϕY_{t-1} is added as an extra variable to the SAR model.

Anselin (2001) and Yu *et al*. (2008) consider a general framework of the dynamic spatial panel model that allows for contemporaneous effects and inertia effects, as well as one-period of temporally lagged social influence in (29):

$$Y_t = \lambda W Y_t + \phi Y_{t-1} + \theta W Y_{t-1} + X_t \beta + \varepsilon_t \quad (29)$$

where Y_t is an $N \times 1$ vector of dependent variable at time period *t* for *i*=1,2,…,*N* ,and X_t is an $N \times k$ matrix with *k* covariates including intercepts at time period *t*. $k \times 1$ vector of coefficients, β is assumed to be constant over time and space. W is an $N \times N$ adjacency matrix, and ε_t is assumed to follow *iid* across *i* at *t* with zero mean and variance $\sigma_\varepsilon^2 I_N$. λ is a parameter of contemporaneous dependence, ϕ is a parameter of inertia effects, and parameter θ measures temporally lagged social influence. When the scalar parameter $\phi\ and\ \theta$ takes on a value of zero, the model in (29) is reduced to the static SAR model in Chapter 2.

Thus, the dynamic model allows us to separate the contemporaneous social influence effects, as measured by ρ in the static model into three pieces: contemporaneous, inertia and temporal social effects. The motivation of this model specification is a potential temporal dependence of social influence. Such influence may occur immediately so that we can observe that influence in the same time period, but very often social influence is assumed to take some time to operate. To account for this different manifestation, the weighted average of the consumption level of connected others during the previous time period, denoted by WY_{t-1} is added as an extra variable to equation (29). While θ in equation (29) represents a measure of pure social influence, λ captures the mixture effects of social influence and homophily, meaning that they buy the same things because they are the same type of person. Inertia effects happen when people tend to persist in their past behavior (adjustment). We also include the same variables of individual characteristics, as used in the static SAR model in Chapter 2. Due to data limitations, we do not include any contextual effects in the current model specification. Distinguishing contemporaneous social influence and the homophily effect is statistically challenging. Thus, the dynamic model allows us to capture pure social influence effects by teasing out social influence, homophily and inertia effects.

Decomposing ρ in the cross-sectional SAR model into λ, ϕ and θ here is important because it provides marketing managers with additional information about the underlying process of correlated behaviors in a network, which is not available from the static model. For example, $\phi > \lambda$ suggests that the temporal inertia effect is larger than the contemporaneous mixture effect of homophily and social influence. And $\theta > \lambda$ implies that pure social influence drives customer behaviors and is correlated more than the contemporaneous mixture effect. Thus, by comparing the magnitude of parameter estimates, we can identify what the main driver is of correlated behaviors. For the model specification, we start with a general framework of the dynamic spatial panel model in (29) and try to determine the most appropriate model for our data via the LM test.

3.1.2. Data-Generating Process

Equation (29) can be rewritten as

$$Y_t - \lambda W Y_t - \phi Y_{t-1} - \theta W Y_{t-1} = X_t \beta + \varepsilon_t \quad (30)$$

$$(I - \lambda W) Y_t - (\phi I + \theta W) Y_{t-1} = X_t \beta + \varepsilon_t \quad (31)$$

$$\{(I - \lambda W) - (\phi I + \theta W) L\} Y_t = X_t \beta + \varepsilon_t \quad (32)$$

$$(D - BL)Y_t = X_t\beta + \varepsilon_t \tag{33}$$

where the matrix L represents the time lag operator $LY_t = Y_{t-1}$, and $Q = (I - \lambda W), B = (\phi I + \theta W)$.

Let Y_c be an $N \times 1$ vector of observations during the period *c*. Also let *Y* be an $NT \times 1$ vector of cross-sections stacked by the period for *T* periods. Then *Y* for *T* periods can be expressed as a following vector form in (34).

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{pmatrix} \tag{34}$$

where $Y_c = \begin{pmatrix} Y_{1c} \\ Y_{2c} \\ \vdots \\ Y_{Nc} \end{pmatrix}$, $c = 1,2, \ldots, T,$ and Y_{ic} is individual *i*'s consumption during the period *c*, *i*=1,2,…*N*.

Let *X* be an $NT \times K$ matrix of cross-sections stacked by period. Then *X* for *T* periods can be expressed as the following matrix form:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_T \end{pmatrix} \tag{35}$$

where $X_c = \begin{pmatrix} X_{11c} & X_{12c} & \cdots & \cdots & X_{1rc} \\ X_{21c} & X_{22c} & \cdots & \cdots & X_{2rc} \\ & & \vdots & & \\ X_{N1c} & X_{N2c} & \cdots & \cdots & X_{Nrc} \end{pmatrix}$, $c = 1,2, \ldots, T. X_{ikc}$ is the k^{th} variable for individual *i* during the period *c*, *i=1,2,…,N* and *k=1,2,…,r.*

From equation (29), we can see that $(I - \lambda W)$ represents the cross-sectional dependence of an individual's behaviors or outcomes in a network, and $(\phi I + \theta W)$ represents the temporal dependence of an individual's behaviors or outcomes in a network. Thus, the spatial dynamic panel model for T periods in (26) can be rewritten using some filter matrix S as the following:

$$SY = X\beta + \varepsilon \tag{36}$$

where $S = \begin{pmatrix} Q & 0 & 0 & \dots & 0 \\ B & Q & 0 & \dots & 0 \\ 0 & B & Q & \dots & 0 \\ & \dots & \dots & \dots & \\ 0 & 0 & \dots & B & Q \end{pmatrix}$, $NT \times NT$ matrix, and $Q = (I - \lambda W)$, and $B = (\phi I + \theta W)$

For simplification, we assume that for the first period, there is only contemporaneous dependence, that is, there is no temporal dependence.

The expression for each cross-sectional observation shows more easily how contemporaneous-temporal filter matrix S works in the data-generating process. From (33), the data-generating process can be rewritten as the following.

$$SY = \begin{pmatrix} Q & 0 & 0 & \dots & 0 \\ B & Q & 0 & \dots & 0 \\ 0 & B & Q & \dots & 0 \\ & \dots & \dots & \dots & \\ 0 & 0 & \dots & B & Q \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_T \end{pmatrix} = \begin{pmatrix} X_1\beta \\ X_2\beta \\ \vdots \\ \vdots \\ X_T\beta \end{pmatrix} \tag{37}$$

Then,

$$QY_1 = X_1\beta \tag{38}$$

$$BY_1 + QY_2 = X_2\beta \tag{39}$$

$$BY_2 + QY_3 = X_3\beta \tag{40}$$

$$BY_{c-1} + QY_c = X_c\beta \tag{41}$$

That is, we can see that the S matrix represents contemporaneous and temporal dependence between dependent variables Y.

From (34), the DGP of Y can be expressed as

$$Y = S^{-1}X\beta \tag{42}$$

$$Y = \sum_{k=1}^{r} S^{-1}\, I_{NT}\beta_k X_k \tag{43}$$

where *k* represents *k^th^* explanatory variables.

S^{-1} takes the form of a lower triangular block matrix in (40), containing blocks with $NT \times N$ matrices. Details of how to solve the inverse of the block matrix S are in APPENDIX F.

$$S^{-1} = \begin{pmatrix} Q^{-1} & 0 & \cdots & 0 & 0 \\ D_1 & Q^{-1} & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ D_{T-1} & D_{T-2} & \cdots & Q^{-1} & 0 \\ D_T & D_{T-1} & & D_1 & Q^{-1} \end{pmatrix} \tag{44}$$

where $D_t = (-1)^t (Q^{-1}B)^t Q^{-1}, \quad t = 1, \ldots, T$

The comparison between (6) and (44) makes clear the difference between the cross-sectional and dynamic model we proposed. While the cross-sectional SAR model captures only contemporaneous network effects through $(I - \lambda W)^{-1}$ in (6), the dynamic SAR model captures the complicated non-linear contemporaneous network effects, inertia effects, and pure social influence effects through S^{-1} matrices in (44). We discuss more in detail about how network effects work over time through the S^{-1} matrices in the next section.

3.2 Identifying Individual Influence

Our focus here is on the identification of individual *i*'s value in a network in terms of revenue generation from a long-term perspective. As is similar to Chapter 2, the dynamic SAR model in (42) allows us to quantify the influence of an individual in a network over time by calculating the cumulative impact of a change in the i^{th} observation of the k^{th} explanatory variable at time *t* over *T* periods (*T*=1,2,…,), which leads to a change in its own future behaviors (Y_{it+T}), as well as in others future behaviors (Y_{jt+T}).

The data-generating process in (42) can be rewritten as:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{N1} \\ Y_{12} \\ \vdots \\ Y_{NT} \end{pmatrix} = \sum_{k=1}^{r} \begin{pmatrix} S_{1,1}^{-1}\ S_{1,2}^{-1} \dots S_{1,NT}^{-1} \\ S_{2,1}^{-1}\ S_{2,2}^{-1} \dots\ S_{2,NT}^{-1} \\ \vdots \\ \vdots \\ \vdots \\ S_{NT,1}^{-1}\ \ S_{NT,2}^{-1} \dots\ S_{NT,NT}^{-1} \end{pmatrix} \begin{pmatrix} X_{1k1}\beta_k \\ X_{2k1}\beta_k \\ \vdots \\ \vdots \\ \vdots \\ X_{NkT}\beta_k \end{pmatrix} + \ S^{-1}\varepsilon$$

$$= \sum_{k=1}^{r} \begin{pmatrix} Q_{1,1}^{-1} & \cdots & Q_{1,N}^{-1} & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ Q_{N,1}^{-1} & \cdots & Q_{N,N}^{-1} & & 0 & \cdots & 0 \\ & \vdots & & \ddots & & \vdots & \\ B_{11,T} & \cdots & B_{1N,T} & & B_{1,1}^{-1} & \cdots & B_{1,N}^{-1} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ B_{N1,T} & \cdots & B_{NN,T} & & B_{N,1}^{-1} & \cdots & B_{N,N}^{-1} \end{pmatrix} \begin{pmatrix} X_{1k1}\beta_k \\ X_{2k1}\beta_k \\ \vdots \\ \vdots \\ \vdots \\ X_{NkT}\beta_k \end{pmatrix} + \ P^{-1}\varepsilon \quad (45)$$

where *i*, *j* represents each individual, *k* denotes the k^{th} explanatory variable and the first column of *X* is a constant unit vector whose coefficient is the intercept.

For the i^{th} individual at time *t*, (45) would take the form of:

$$Y_{it} = \sum_{k=1}^{r}[Q_{i1}^{-1}X_{1kt}\beta_k + Q_{i2}^{-1}X_{2kt}\beta_k + \cdots + Q_{iN}^{-1}X_{iNt}\beta_k]$$

$$+ \sum_{k=1}^{r}[B_{i11}X_{1k(t-1)}\beta_k + B_{i21}X_{2k(t-1)}\beta_k + \cdots + B_{iN1(t-1)}X_{iN(t-1)}\beta_k] +$$

$$\sum_{k=1}^{r}[B_{i12}X_{1k(t-2)}\beta_k + B_{i22}X_{2k(t-2)}\beta_k + \cdots + B_{iN2}X_{iN(t-2)}\beta_k + \cdots B_{iN2}X_{iN(t-2)}\beta_k] +$$

$$\cdots + \sum_{k=1}^{r}[B_{i1T}X_{1k(t-T)}\beta_k + B_{i2T}X_{2k(t-T)}\beta_k + \cdots + B_{iNT}X_{iN(t-T)}\beta_k + \ S^{-1}\varepsilon \qquad (46)$$

Thus, this could be rewritten as a reduced form, as follows:

$$Y_{it} = \sum_j \sum_{k=1}^{r}\{Q_{ij}^{-1}X_{ikt}\beta_k\} + \sum_T \sum_j \sum_{k=1}^{r}\{B_{ijt}X_{ik(t-T)}\beta_k\} \qquad (47)$$

Similar to the cross-sectional SAR model, the impact of a change of individual *i* in the k^{th} variable at time *t* for one time period could be written as:

$$\sum_j \partial E\left(Y_{jt+1}\right)/\partial X_{ikt} = \sum_j \beta_k\{Q_{ij}^{-1} + B_{ij1}\} \qquad (48)$$

and the general form of the cumulative impact over *T* periods from a change in the explanatory variables takes the form in (49):

$$\sum_{\tau=1}^{T}\sum_j \frac{\partial E(Y_{jt+T})}{\partial X_{ikt}} = \sum_j \beta_k\{Q_{ij}^{-1} + B_{ij1} + B_{ij2} + \cdots + B_{ijT}\} \qquad (49)$$

To compute the individual customer value in a network within a dynamic setting, we start from the cross-sectional definition of individual customer value in a network used in Chapter 2. We define individual customer value on revenue generation to the firm at time *t* as the impact of the i^{th} changes of explanatory variables on the system weighted by the level of individual *i*'s characteristics, which is formalized as (50).

$$VALUE^{(i)} = \sum_{k=1}^{r} X_{ikt} \times \left(\sum_j \partial E(Y_{jt})/\partial X_{ikt}\right) \qquad (50)$$

This cross-sectional definition of individual value can be easily generalized in a panel setting with the dynamic model, which allows us to quantify a cumulative *T*-period-ahead impact of a change in *k* explanatory variables at time *t* for individual *i* in (51).

$$VALUE^{(iT)} = \sum_{k=1}^{r} X_{ikt} \times \sum_\tau\left(\sum_j \partial E(Y_{jt+\tau})/\partial X_{ikt}\right) = \sum_j \sum_{k=1}^{r} X_{ikt}\,\beta_k\{Q_{ij}^{-1} + B_{ij1} + B_{ij2} + \cdots + B_{ijT}\} \text{ where } \tau = 1,2,\ldots,T \qquad (51)$$

By analogy to the cross-sectional SAR model in Chapter 2, the diagonal element S_{ii}^{-1} in (45) represents the impact of the i^{th} change at time *t* on its own for the time window *N* that arises from both time and network dependence. Also by summing down the j^{th} column of this matrix, $\sum_{j\neq i} S_{ij}^{-1}$, it reflects the impacts of the i^{th} change at time *t* on everybody else in the system for time window *T*. That is, they sum up the contemporaneous spillover and temporal diffusion impacts that arise over time on all other people in a network

APPENDIX G illustrates how the rank order of an individual's influence on revenue generation from the dynamic spatio-temporal model may potentially shift results from the cross-sectional SAR model. Measures by the dynamic model allow us to better identify who is more influential in revenue generation with information not available from the cross-sectional SAR model.

3.3 Estimation

Estimation methods have been widely discussed in the previous literature. In this study, we use the maximum likelihood approach by Anselin (1988) and Elhorst (2001)

The conditional log-likelihood function for the dynamic spatial panel model is given as follows :

$$Log\ f_{Y_2,Y_3,\ldots,Y_T|Y_1} = -\frac{N(T-1)}{2} log(2\pi\sigma^2) + (T-1)\sum_{i=1}^{N} \log(1-\rho\bar{\omega}_i) - \frac{1}{2\sigma^2}\sum_{t=2}^{T} \varepsilon_t'\varepsilon_t \quad (52)$$

where $\varepsilon_t = Y_t - \rho WY_t - \phi IY_{t-1} - \theta WY_{t-1} - X_t\beta$, $\bar{\omega}_i, i = 1,2,\ldots,N$ represents eigenvalues of the matrix W.

The unconditional log-likelihood function is as follows:

$$Log\ f_{Y_1,Y_2,Y_3,\ldots,Y_T} = -\frac{NT}{2} log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{N} \log[(1-\lambda\bar{\omega}_i)^2 - (\phi+\theta\bar{\omega}_i)^2] + (T-1)\sum_{i=1}^{N} \log(1-\lambda\bar{\omega}_i) - \frac{1}{2\sigma^2}\sum_{t=2}^{T} \varepsilon_t'\varepsilon_t + \frac{1}{2\sigma^2}\varepsilon_1'[(Q-B)']^{-1}[Q'Q - Q'BQ^{-1}(QBQ^{-1})']^{-1}(Q-B)^{-1}\varepsilon_1 \quad (53)$$

where $\varepsilon_t = Y_t - \lambda WY_t - \phi IY_{t-1} - \theta WY_{t-1} - X_t\beta$, $\bar{\omega}_i, i = 1,2,\ldots,N$ represents eigenvalues of the matrix W, $Q = I - \lambda W, and\ B = \phi I + \theta W$.

Since T is small, the first observation contributes greatly to the overall likelihood. Thus, we use the unconditional likelihood to estimate the model instead of conditional likelihood following Elhorst (2001). Furthermore, we adopt an iterative two-step procedure suggested by Elhorst (2001). More details about estimation methods are provided by Elhorst (2001).

There are some stationary issues in spatio-temporal dynamic models. The spatio-temporal process generating the data is covariance stationary (Yu *et al*. 2008) if

$$|BA^{-1}| < 1 \tag{54}$$

3.4 Empirical Results

3.4.1 Empirical Findings

We report the parameter estimates for the models across 6 sample months, March through August, 2010 in Table H15. Our estimates for the models are stationary, since the sum of the parameters is less than one which meets the stationary requirements in (54). Specification in the first column focuses on the contemporaneous network effect, thus ignoring the temporal effect. In the second column, the lagged social influence effect is added into the model. In the third column, all three effects—contemporaneous, lagged social influence and inertia effect—are included for the model specification.

Comparing the results for the three specifications in terms of *AIC*, one sees that overall, the dynamic SAR models have a better fit than the static model. Especially the fact that the inclusion of all effects, contemporaneous, lagged social influence, and inertia effect provides the smallest *AIC* suggests evidence for a better fit of this specification. However, most of the explanatory variables are not significant in the dynamic model, which implies that the dynamic temporal effect plays a significant role in explaining the variability in the customer's revenue generation to the firm.

Parameter estimates in Table H15 imply that overall there is a contemporaneous social influence effect in our observation, even though it becomes slightly weaker when other effects are considered. Once lagged social influence and the inertia effect are included, contemporaneous social influence becomes weaker. The inclusion of the lagged social influence effect in the second column leads to a reduction in contemporaneous social influence, measured by λ from about 0.357 to 0.219, which suggests that there is less evidence for the contemporaneous effect. If the inertia effect is included, this also comes with a positive coefficient (0.273). Overall, there is evidence that both contemporaneous social influence and inertia effects strongly exist.

When both WY_{t-1} and Y_{t-1} are included jointly, lagged social influence becomes negative and significant (-0.065) in the fourth column. This empirical finding is a somewhat surprising result, indicating that peers past purchase have a negative influence on one's own current period purchases. This finding may be due to potential collinearity between WY_{t-1} and Y_{t-1}, because the coefficient on the inertia effects remains positive, in fact, it goes up. Or it may be due to model misspecification. Since the network structure in the current study is assumed to be fixed over time, the coefficient of lagged social influence in our model of the form (30) may not accurately represent the peers' buying behaviors in the past if relations among people vary over time.

There are several results that are common across different specification of models. First of all, *Age* and *Gender* do not have a significant influence on revenue generation to the firm. Second, *Guild Activity*, *Action on People* and *House Value* always have a positive influence on revenue generation. Third, the impact of the network seems to have lasting effects, since the inclusion of all effects leads to a better fit with stronger inertia effects than the social influence effect. Overall, it is clear that without WY_{t-1} and Y_{t-1} variables, the evidence for contemporaneous dependence would have been overestimated.

3.4.2 Model Comparison

We evaluate the model in terms of model fits and predictive capability of models in identifying an individual's influence on revenue generation to the firm, respectively. To do that, we estimate the models for two types of data: in-sample data that occurred from March, 2010 through July, 2010 and additional data (out-of-sample) that originate from the same source, but are taken from data that occurred during August, 2010. For the holdout sample, we keep using the same variables *Age, Gender, and House Value* but update information for the other variables such as *Virtual Money and Gaming Behaviors*. The *W* matrix is constructed using only in-sample observations assuming that the network structure is consistent over time.

Evaluation of the model involves two steps. For model fits, the proposed spatial dynamic panel model is compared based on the Akaike Information Criterion (*AIC*). For evaluations of predictive capability of the model, Mean Absolute Deviations (*MAD*) and the Gini coefficient are used, respectively. Tables H15 and H16 show that the spatial dynamic panel model utilized here performs better than the static model in terms of AIC and *MAD*. Figure H14 shows that the Gini curve of the dynamic model is on top of the Gini curve of the static model, which indicates that the dynamic model performs better in identifying high-value customers. In particular, it shows that the dynamic model does a very good job at identifying customers on the top of the revenue list, who marketing managers care most about. From *MAD* and the Gini curve, we can conclude that dynamic spatial panel models introduced here represent a good way to identify individual customer value in a network over time.

3.5 Summary

Chapter 3 aims to identify an individual's influence on revenue generation in a network from a long-term perspective by incorporating temporal dynamics into the base static SAR model. We extend the static approach proposed in Chapter 2 for measuring an individual's influence in a network to the spatial dynamic panel model, which is consistent with the treatment of regression coefficient estimates, where we view these estimates as reflecting how changes in the explanatory variables impact the dependent variable on average over the sample. An implication of our model specification is that changing the value of the explanatory variable in one person in a single time period can impact others in the same time period, as well as his or her own and other people in future time periods. Thus, the measure of influence at the individual level using the dynamic spatial panel model allows us to better identify who is more influential in a network with information not available from the cross-sectional SAR model. We empirically show that the dynamic approach predicts individual value in a network better than the static approach. Further extensions to account for distinct sources of network correlation such as homophily and contagion are necessary to avoid biased inferences from omissions.

CHAPTER IV

CONCLUSIONS

Customer behaviors are assumed to be interdependent with others in a network. By applying the spatial autoregressive model, this research develops a methodology to identify an individual’s value within a network in terms of revenue generation to the firm. The research demonstrates empirically that the proposed approach provides better forecasts compared to benchmark models. Also, dynamic approach accounting for inertia effects and temporal network effects is shown to provide better predictions than the static approach.

There are several contributions of this thesis to the marketing science literature. First, we empirically show that social influence plays an important role in customer purchase behaviors. When we take into account both an individual’s characteristics and social influence, it performs better in predicting customer purchase behaviors in a network setting. While most of previous works on social networks have focused on modeling interdependent outcomes, the proposed model shows how social influence interacts with individual characteristics and how this interaction affects actual purchase behaviors.

Second, this thesis demonstrates how spatial statistical models can be used to measure individual customer value in terms of revenue generation to the firm using both customer network information and customer characteristics information. In particular, the approach does a very good job at identifying people with higher values, who we care most about for marketing purposes. Social influence typically takes some time to have an impact. For this reason, it is important to account for the long-term effect of social influence. We show that the spatial dynamic panel model performs better than static models and other benchmark models in predicting customer value. This strong empirical evidence implies superiority of the spatial dynamic panel model as a method to rank order customers in a network in terms of revenue generation.

Third, from a managerial perspective, this thesis contributes to marketing knowledge of customer relationship management (CRM). The proposed value function of individual customer can be used as a method of targeting households. When marketers can measure the value of each individual customer in terms of revenues, they can use this information to treat households differently for target marketing.

Finally, our model can be extended in various ways. While our proposed model analyzes how social influence has an impact on customer behaviors taking the network structure given, network structure is not static but naturally changing over time. Considering the dynamics of customer behaviors and network structure simultaneously is important in understanding the phenomena of homophily and contagion in a network. To account for this issue, we may extend our proposed model by modeling an element of the weight matrix W as a function of some explanatory or dependent variable. In spite of its increasing importance, there are few empirical models due to methodological and empirical challenges, such as model identification and Galton's problem (Naroll 1961).

Second, the model can be extended to construct generalized measures of distance between individuals by using multiple types of links, such as demographic similarities, physical proximities or other relational behavior measures. These extensions are important because they can have an impact on the measurement of individual values or effects in a network.

Third, in our dynamic analysis, the elaboration of the error component has not been considered. For simplification, we assume the independence of error terms. However, there is a concern that heterogeneity across individuals might induce heteroskedasticity, such as random effects. In a future study, we can control for the presence of heterogeneity among individuals.

Real-world-like virtual worlds in online games provide huge and various research opportunities. People interact with each other as they do in the real world. Ease in obtaining various behavioral and relational data from the virtual world allows us to model interdependent behaviors and test various economic and marketing theories. To date, there has been little research on online gaming in the marketing area. To the best of my knowledge, this is the first empirical study involving online gaming in the marketing area. More research using data from the virtual world is needed.

APPENDIX A

ADJACENCY MATRIX

Let $n \times n$ matrix $A = (a_{ij})$ be the element $\mathrm{a_{ij}} = 1$ if there is an edge from vertex *i* to vertex *j* and 0 if there is no edge from vertex *i* to vertex *j*. Then, the element of $(A^k)_{ij}$ of the k^{th} power of the adjacency matrix counts the number of paths of length *k* from *i* to *j*. That is, A^k matrix represents the k^{th}-order neighbor matrix.

Let's take a simple example of adjacency matrix (*A*) which represents a first-order contiguity neighbor.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{A.1}$$

And then, $A^2 = B$ with entries b_{ij} could be obtained as follows:

$$A^2 = B = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 3 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \tag{A.2}$$

where $b_{ij} = \sum_{k=1}^{n} a_{ik}a_{kj} = a_{i1}a_{1j} + \ a_{i2}a_{2j} + \cdots + \ a_{in}a_{nj}$

Here square matrix, A^2 represents the second-order contiguity neighbor because this matrix counts the number of paths between two nodes in length two. $a_{in}a_{nj}$ contributes 1 if there is a path of length 2 from *i* to *j* passing through *n*.

Mathematical proof of this theorem is given below.

<Proof> Let *A* be the adjacency matrix, A^k. The proof is by induction on *k*. For *k* = 1, the result follows from the definition of *A*, since a path of length 1 is just an edge.

Suppose, for a particular *k*, that $(A^k)_{ij}$ is the number of paths of length *k* from *i* to *j*.

There are $(A^k)_{in}A_{nj}$ paths of length *k +1* from *i* to *j* in which *n* is the penultimate vertex.

Our result then follows because $\sum_n (A^k)_{in}A_{nj} = (A^{k+1})_{ij}$ □

APPENDIX B

ILLUSTRATION OF SPILLOVER EFFECTS OF SAR MODEL

To provide an illustration of how the SAR model generates the spillover effects, we compare predictions $\hat{Y}^{(1)}$with explanatory variables from X_1 and $\hat{Y}^{(2)}$ with explanatory variables from X_2. We use the following adjacency matrix *W*, parameter estimates, $\hat{\rho}\ and\ \hat{\beta}$ and the explanatory variables matrix *X*. Row-standardized *W* matrix is used to fit the SAR model.

$\hat{\rho} = 0.452,\ \ \hat{\beta} = [0.153\ 0.518]'$

$$W = \begin{pmatrix} 0\ 1\ 1\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 0\ 0\ 1 \\ 0\ 1\ 0\ 0\ 1\ 1\ 0 \\ 1\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 1\ 1\ 0\ 0\ 1\ 0 \\ 1\ 1\ 0\ 1\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 0\ 1\ 0 \end{pmatrix}, X_1 = \begin{pmatrix} 50\ 30 \\ 20\ 20 \\ 20\ 10 \\ 10\ 20 \\ 30\ 10 \\ 20\ 40 \\ 40\ 20 \end{pmatrix},\ X_2 = \begin{pmatrix} 50\ 30 \\ 20\ 20 \\ 20\ 10 \\ 10\ \mathbf{30} \\ 30\ 10 \\ 20\ 40 \\ 40\ 20 \end{pmatrix} \qquad \text{(B.1)}$$

Predictions from the model using the explanatory variable *X* would take the form:

$$\hat{Y} = (I - \hat{\rho} W)^{-1} X \hat{\beta}$$

In the predictions of $\hat{Y}^{(1)}$ and $\hat{Y}^{(2)}$ below, we can see that single change for observation *i* in the explanatory variable *k* produces an increase in Y_i and also in all other Y_j. Changes in X_{42} has a direct effect on $\hat{Y}_4^{(2)}$ and has an indirect or spillover effect that produces an increase in all other $\hat{Y}_j^{(2)}$.

$$\hat{Y}^{(1)} = \begin{pmatrix} 33.45 \\ 24.66 \\ 20.75 \\ 25.02 \\ 22.08 \\ 36.30 \\ 28.84 \end{pmatrix},\ \hat{Y}^{(2)} = \begin{pmatrix} 33.77 \\ 25.70 \\ 21.12 \\ 30.51 \\ 22.45 \\ 37.33 \\ 29.88 \end{pmatrix} \qquad \text{(B.2)}$$

APPENDIX C

BOUNRDS FOR THE DEPENDENCE PARAMETER (ρ)

The log-likelihood function of SAR model takes the form in (C.1), where ω is a vector of eigenvalues of the weight matrix *W*.

$$ln\, L = -\frac{n}{2} ln(\pi\sigma^2) + ln|I_n - \rho W| - \frac{e\prime e}{2\sigma^2}, \text{ and } e = y - \rho W y - X\beta \qquad \text{(C.1)}$$

The Jacobian term is simplified to the following expression, as shown in Ord (1975):

$$ln|I - \rho W| = \sum_i ln(1 - \rho\omega_i) \qquad \text{(C.2)}$$

In the current thesis, we use the symmetric matrix, so that its eigenvalues are all real. Symmetric matrix Q can be rewritten as following:

$$Q = H^{1/2} W H^{1/2} \qquad \text{(C.3)}$$

where *H* is a diagonal matrix with elements composed of the inverse of the row sums of the matrix W.

onsequently, the restriction on the parameter is of the form $\rho < 1/\omega_i$, for all *i* (Ord 1975). The resulting acceptable parameter bound for ρ is

$$\frac{1}{\omega_{\max}} \leq \rho \leq \frac{1}{\omega_{\min}} \qquad \text{(C.4)}$$

where ω_{max} and ω_{min} are the largest and smallest eigenvalues of the spatial weight matrix *W* (Anselin and Florax 1994). For row-standardized weights where the row elements sum to unity, the largest eigenvalue is 1 and $\frac{1}{\omega_{\min}} \leq -1$ which constrains the positive values of ρ to 1. Thus, we got the bounds for ρ as following:,

$$\frac{1}{\omega_{\min}} \leq -1 \leq \rho \leq 1 \qquad \text{(C.5)}$$

For row normalized matrix of *W*, -1< ρ <1 is a sufficient condition for positive-definite variance covariance matrices. The eigenvalues of non-symmetric weight matrix can be complex. However, this restriction on ρ ensures that $(\mathrm{I} - \rho \mathrm{W})$ is non-singular and thus the variance-covariance matrix, $(\mathrm{I} - \rho W)^{-1}(\mathrm{I} - \rho W')^{-1}$ is positive-definite.

APPENDIX D

CENTRALITY MEASURE

To describe an importance of individual's location in a network, centrality measure such as degree, closeness and betweenness has been widely used in sociology. These measures are typically calculated for each individual. In current study, degree and betweenness centrality measure is used for the comparison to ours. Degree centrality is defined by the number of ties between individuals (Freeman 1977, 1979), which could be expressed as following in (D.1).

$$D(i) = \sum_j x_{ij} = \sum_j x_{ji} \tag{D.1}$$

where x_{ij} represents the links from individual *i* to individual j.

In current study, indegree ($\sum_j x_{ji}$) should differ from outdegree ($\sum_j x_{ij}$) because the relations between actors are asymmetric. Both measures provide us very similar results in our analysis though.

Betweenness centrality is defined by "how between an individual is to all others in the network." (Freeman 1977, 1979). It is based on the assumption that an individual is central if it lies between others on their geodesic. To have large betweenness centrality, an individual must be between many other people on their geodesics. Let G_{jk} be the number of geodesics linking two people, *j* and *k*. For three people, *j*, *k* and *i*, $G_{jk}(i)$ is the number of geodesics between *j* and *k* and *k* that contain *i*. Betweenness centrality is then summed over all people, which take the form in (D.2):

$$B(i) = \sum_{j<k} \left(\frac{G_{jk}(i)}{G_{jk}}\right) \tag{D.2}$$

APPENDIX E

ESTIMATION SCHEME OF SAR TOBIT MODEL

$$y^{*} = (I-\rho W)^{-1}X\beta + (I-\rho W)^{-1}\varepsilon, \text{ where } y^{*} = \begin{cases} y_1^{*} & if\ y^{*} \leq 0 \\ y_2 & otherwise \end{cases}$$

The conditional posterior distribution of n_1 censored observation is assumed to follow a truncated multivariate truncated normal distribution (TMVN),

$$y_1^{*} \sim TMVN(\mu_1^{*}, \Omega_{1,1}^{*}) \text{ with mean and variance-covariance}$$

$$\mu_1^{*} = E(y_1^{*}|y_2, X, W, \beta, \rho, \sigma_\varepsilon^2) = \mu_1 - \Sigma_{1,1}^{-1}\Sigma_{1,2}(y_2 - \mu_2)$$

$$\Omega_{1,1}^{*} = var - cov(y_1^{*}|y_2, X, W, \beta, \rho, \sigma_\varepsilon^2)$$

$$= \Omega_{1,1} + (\Sigma_{1,1})^{-1}\Sigma_{1,2}\,\Omega_{2,1}$$

where

$$\Omega_{1,1} = \sigma_\varepsilon^2[(I-\rho W)'(I-\rho W)]^{-1}$$

$$\Sigma = \Omega^{-1}$$

$$\mu_1 = (I-\rho W)_{1,1}^{-1}X_1\beta$$

$$\mu_2 = (I-\rho W)_{2,2}^{-1}X_2\beta$$

We use the subscripts 1, 2 to denote an $n_1 \times n_2$ matrix, and matrices such as $\Omega_{1,1}$ would contain n_1 rows and columns, whereas $\Omega_{2,2}$ would be of dimension $n_2 \times n_2$

APPENDIX F

BLOCK LOWER TRIANGULAR MATRIX INVERSE

We will use induction method to solve P^{-1} in equation (47).

Let $\mathrm{P}_1 = (S)$. Then, $\mathrm{P}_1^{-1} = (\mathrm{S}^{-1})$.

Let $\mathrm{P}_2 = \begin{pmatrix} \mathrm{S} & 0 \\ \mathrm{V} & \mathrm{S} \end{pmatrix} = \begin{pmatrix} \mathrm{P}_1 & 0 \\ \mathrm{V} & \mathrm{S} \end{pmatrix}$ and $\mathrm{P}_2^{-1} = \begin{pmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{pmatrix}$.

Then, we get $\mathrm{SA}_{11} = \mathrm{I}$, $\mathrm{SA}_{12} = 0$, $\mathrm{VA}_{11} + \mathrm{SA}_{21} = 0$, and $\mathrm{VA}_{12} + \mathrm{SA}_{22} = \mathrm{I}$.

Therefore, we have $\mathrm{A}_{11} = \mathrm{S}^{-1}$, $\mathrm{A}_{12} = 0$, $\mathrm{A}_{22} = \mathrm{S}^{-1}$, and $\mathrm{A}_{21} = -\mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1}$.

Let $\mathrm{P}_3 = \begin{pmatrix} \mathrm{S} & 0 & 0 \\ \mathrm{V} & \mathrm{S} & 0 \\ 0 & \mathrm{V} & \mathrm{S} \end{pmatrix} = \begin{pmatrix} \multicolumn{2}{c}{\mathrm{P}_2} & \begin{matrix} 0 \\ 0 \end{matrix} \\ 0 & \mathrm{V} & \mathrm{S} \end{pmatrix}$ and

$$\mathrm{P}_3^{-1} = \begin{pmatrix} \multicolumn{2}{c}{\mathrm{P}_2^{-1}} & \begin{matrix} \mathrm{E}_{13} \\ \mathrm{E}_{23} \end{matrix} \\ \mathrm{E}_{31} & \mathrm{E}_{32} & \mathrm{E}_{33} \end{pmatrix} = \begin{pmatrix} \mathrm{S}^{-1} & 0 & \mathrm{E}_{13} \\ -\mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1} & \mathrm{S}^{-1} & \mathrm{E}_{23} \\ \mathrm{E}_{31} & \mathrm{E}_{32} & \mathrm{E}_{33} \end{pmatrix}$$

Because of the first row of P_3, $\mathrm{E}_{13} = 0$.

Since $\mathrm{E}_{13} = 0$, because of the second row of P_3, $\mathrm{E}_{23} = 0$.

Since $\mathrm{E}_{13} = 0$ and $\mathrm{E}_{23} = 0$, because of the third row of P_3, $\mathrm{E}_{33} = \mathrm{S}^{-1}$.

To get E_{31} and E_{32}, we will use the third row of P_3.

We will get $-\mathrm{V}\mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1} + \mathrm{SE}_{31} = 0$ and $\mathrm{V}\mathrm{S}^{-1} + \mathrm{SE}_{32} = 0$.

Therefore, we have $\mathrm{E}_{31} = \mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1}$ and $\mathrm{E}_{32} = -\mathrm{S}^{-1}\mathrm{V}\mathrm{S}^{-1}$.

Let $\mathrm{D}_s = (-1)^s(\mathrm{S}^{-1}\mathrm{V})^s\mathrm{S}^{-1}$.

Then, $\mathrm{P}_1^{-1} = (\mathrm{D}_0)$, $\mathrm{P}_2^{-1} = \begin{pmatrix} \mathrm{D}_0 & 0 \\ \mathrm{D}_1 & \mathrm{D}_0 \end{pmatrix}$, and $\mathrm{P}_3^{-1} = \begin{pmatrix} \mathrm{D}_0 & 0 & 0 \\ \mathrm{D}_1 & \mathrm{D}_0 & 0 \\ \mathrm{D}_2 & \mathrm{D}_1 & \mathrm{D}_0 \end{pmatrix}$.

Assume $\mathrm{P}_\mathrm{T}^{-1} = \begin{pmatrix} \mathrm{D}_0 & 0 & \cdots & 0 & 0 \\ \mathrm{D}_1 & \mathrm{D}_0 & \cdots & 0 & 0 \\ \mathrm{D}_2 & \mathrm{D}_1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & 0 \\ \mathrm{D}_{\mathrm{T}-1} & \mathrm{D}_{\mathrm{T}-2} & \cdots & \mathrm{D}_1 & \mathrm{D}_0 \end{pmatrix}$.

We know that $P_{T+1}^{-1} = \begin{pmatrix} & & P_T^{-1} & & \begin{matrix} F_{1,T+1} \\ F_{2,T+1} \\ \vdots \\ F_{T,T+1} \end{matrix} \\ F_{T+1,1} & F_{T+1,2} & \cdots & F_{T+1,T} & F_{T+1,T+1} \end{pmatrix}$.

Using the first row of P_{T+1}, we can get $F_{1,T+1} = 0$. Since $F_{1,T+1} = 0$, because of the second row of P_{T+1}, $F_{2,T+1} = 0$. Using the same argument, when $0 \leq i \leq T$, $F_{i,T+1} = 0$.

It is easy to find $F_{T+1,T+1} = S^{-1} = D_0$.

Since T-th row of P_T^{-1} is $(D_{T-1} \quad D_{T-2} \quad \cdots \quad D_1 \quad D_0)$ for $0 \leq i \leq T$, we get $VD_{T-i} + SF_{T+1,i} = 0$. Therefore, $F_{T+1,i} = -S^{-1}VD_{T-i} = D_{T-i+1}$.

By mathematical induction, we can say that

$$D_k = (-1)^k (S^{-1}V)^k S^{-1}, \; k = 0,1,\dots,T$$

$$P^{-1} = \begin{pmatrix} S^{-1} & 0 & \cdots & 0 & 0 \\ D_1 & S^{-1} & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ D_{T-1} & D_{T-2} & \cdots & S^{-1} & 0 \\ D_T & D_{T-1} & & D_1 & S^{-1} \end{pmatrix}$$

APPENDIX G

EXAMPLE FOR SPILLOVER EFFECTS OF DYNAMIC SAR MODEL

Let's take a small sample example of *n*=5 to illustrate how the dynamic SAR model shifts results of spillover effects from cross-sectional SAR model.

W, Y, and X with subscript *t*=time are given as follows:

$$W = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$Y_1 = [3.67\ \ 3.70\ \ 3.85\ \ 3.72\ \ 4.00]',\ Y_2 = [3.97\ \ 4.12\ \ 4.31\ \ 3.84\ \ 4.21]'$$

$$X_1 = \begin{bmatrix} 7.63 & 3.58 \\ 7.07 & 3.21 \\ 7.04 & 4.60 \\ 7.14 & 4.99 \\ 7.21 & 5.13 \end{bmatrix},\ X_2 = \begin{bmatrix} 7.91 & 3.70 \\ 7.27 & 3.54 \\ 7.53 & 4.88 \\ 7.47 & 5.42 \\ 7.94 & 5.89 \end{bmatrix}$$

Parameters are assumed to be $\hat{\beta}_1 = [0.15\ 0.52]'$, $\hat{\gamma} = 0.4$, $\hat{\phi} = 0.2$, $\hat{\theta} = 0.2$

For the cross-sectional SAR model at time=1 with DGP, $y = (I_n - \gamma W)^{-1} X\beta$,

$$(I_n - \gamma W)^{-1} = \begin{bmatrix} 1.02 & 0.27 & 0.22 & 0.08 & 0.08 \\ 0.05 & 1.11 & 0.05 & 0.23 & 0.23 \\ 0.04 & 0.22 & 1.04 & 0.18 & 0.18 \\ 0.21 & 0.10 & 0.25 & 1.05 & 0.05 \\ 0.02 & 0.44 & 0.02 & 0.09 & 1.09 \end{bmatrix}$$

- Impact of *i*=1,

  $3.33 = (1.02 + 0.05 + 0.04 + 0.21 + 0.02) \times (7.63 \times 0.15 + 3.58 \times 0.52)$

- Rank order of individual's impact on revenue generation is as follows:

ID	Rank Order
	5
	4

	3
	2
	1

For the dynamic SAR model with DGP, $y = P^{-1}X\beta$, $P^{-1} = \begin{bmatrix} \mathrm{S} & 0 \\ \mathrm{K} & \mathrm{S} \end{bmatrix}$

where $S = (I_n - \gamma W)^{-1}$*,* $K = -S(\phi I_n + \theta W)$S*,* P is $NT \times NT$ matrix.

$$K = \begin{pmatrix} -0.24 & -0.32 & -0.23 & -0.15 & -0.15 \\ -0.08 & -0.40 & -0.10 & -0.25 & -0.25 \\ -0.07 & -0.29 & -0.28 & -0.22 & -0.22 \\ -0.20 & -0.19 & -0.28 & -0.31 & -0.11 \\ -0.04 & -0.47 & -0.05 & -0.16 & -0.36 \end{pmatrix}$$

- Impact of *i*=1,

$$(1.02 + 0.05 + 0.04 + 0.21 + 0.02) \times (7.63 \times 0.15 + 3.58 \times 0.52) + (0.24 + 0.08$$

$$+0.07 + 0.20 + 0.04) \times (7.91 \times 0.15 + 3.70 \times 0.52)\} = 0.96 + 1.27 = 2.23$$

- Rank order of individual's impact on revenue generation is as follows:

D	Rank Order
	5
	1
	4
	2
	3

APPENDIX H

TABLES AND FIGURES

Table H1. Summary Statistics of Variables

Variable	Mean	Standard Deviation
Revenue	3.68	0.59
Age	33.65	10.49
Gender	0.76	0.42
ln(House Value)	7.42	0.41
Game Money	0.16	0.30
Experience Score	7196008	8488460
Level of Character	245.27	141.31
Time to Level Up	85.59	127.46
Reputation	97.03	109.71
Number of Guild War	2.75	4.12
Guild Membership	0.73	0.53
Master	0.04	0.16
Number of PK	0.521	5.14
DM rating	54.96	272.04
Number of Wanted	0.14	1.85
Number of Hunting	397.502	6031.09
Number of Construction	0.961	0.498
Game Time	2020.69	476.43

Note: ‘House value’ variable is measured as market value in which the subject property may transact. And this measure is widely used as a proxy for income in Korea.

Note: ‘Game Time’ is the minutes that individual gamers play the game during this month.

Note: ‘Master’ is a dummy variable where it has 1 if users are master.

Note: ‘Guild Membership’ is a continuous variable because gamers can join multiple guild.

Note: ‘PK’ is a player killing which refers to when one player’s character kills another player’s character. When PK happens in the public place such as downtown or square, they will be on the wanted list.

Table H2. Factor Loadings: Principal Component Analysis

	Component1	Component2	Compenent3	Component4
Number of Guild Member	0.920	0.057	-0.004	-0.001
Number of Guild War	0.829	-0.015	0.052	0.004
Guild	0.768	0.147	0.000	0.248
Level of Character	0.127	0.777	0.116	0.025
Experience Score	0.266	0.651	0.191	0.393
Reputation	0.044	0.609	0.486	-0.045
Master	0.012	0.517	-0.043	-0.021
Time to Level Up	-0.052	0.363	-0.107	0.046
Wanted	-0.015	-0.029	0.839	0.065
PK	-0.001	-0.148	0.714	0.165
DM Rating	0.061	0.205	0.631	-0.103
Hunting	0.055	-0.006	0.014	0.877
Construction	0.257	0.542	0.135	0.592

Note: Each factor is standardized.

Table H3. Tests for Dependence

Test	Test Statistics
Moran's I Z	2.012 (0.000)
LM (lag)	18.753 (0.000)
LM (error)	12.772 (0.000)
Robust LM (lag)	11.348 (0.000)
Robust LM (error)	5.358 (0.021)

Note: $\text{Moran's } I \text{ Z} = \frac{I - \mathrm{E}(I)}{\mathrm{SD}(I)}$ where *I* is the estimated Moran's *I* statistics, E(*I*) and SD(*I*) is the mean and standard deviation of *I* respectively, given the null of no dependence in error terms. The null hypothesis is rejected if the estimated Z is larger than 1.64.

Note: The LM-tests asymptotically follow a $\chi^2(1)$ distribution, with a critical value of 6.635 at the 1% significance level.

Note: *p*-values are reported in parenthesis.

Table H4. Summary Table of Model Fitting

		Network	Segmentation	SAR Interaction
Variable	Intercept	2.496^{***}	1.951^{***}	0.608^{***}
	Rho (ρ)	0.322^{***}		0.368^{***}
	Age (X_1)		-0.002	-0.002
	Gender (X_2)		0.046	0.042
	ln(House Value) (X_3)		0.236^{***}	0.237^{***}
	Game Money (X_4)		0.780^{***}	0.768^{***}
	Factor1 (X_5)		0.152^{***}	0.153^{***}
	Factor2 (X_6)		0.120^{***}	0.119^{***}
	Factor3 (X_7)		0.035^{***}	0.036^{***}
	Factor4 (X_8)		0.053^{**}	0.052^{**}
AIC Calibration		1114.262	8132.342	8090.469
AIC Holdout		1198.924	8622.954	8222.438
MAD Calibration		0.536	0.314	0.295
MAD Holdout		0.544	0.309	0.291

Note: *** = p<0.01, ** = p<0.05, * = p<0.10

Note: Interpretation of each factor
Factor1: *Social Status*
Factor2: *Guild Activity*
Factor3: *Action on other people*
Factor4: *Action on virtual world*

Table H5. Parameter estimates of OLS

Variable	Parameter Estimates
Intercept	-0.486***
Network	0.151***
Segmentation	0.500***

Note: *** = p<0.01, ** = p<0.05, * = p<0.10

Table H6. Pearson Correlation between Revenue and Forecast

	Revenue	R^2
SAR Interaction Model	0.61	0.37
Segmentation Model	0.55	0.30
Network Model	0.27	0.07
Degree	0.20	0.04
Betweenness	0.18	0.03

Table H7. Spearman Correlation between Revenue and Forecast

	Revenue
SAR Interaction Model	0.63
Segmentation Model	0.56
Network Model	0.29
Degree	0.26
Betweenness	0.22

Table H8. Summary Measure of Direct, Indirect and Total Effects

Variable	Direct	Indirect	Total
Age (X_1)	- 0.002	-0.002	-0.003
Gender (X_2)	0.045	0.045	0.090
ln(House Value) (X_3)	0.263	0.239	0.502
Game Money (X_5)	0.770	0.742	1.513
Factor1 (X_6)	0.152	0.146	0. 299
Factor2 (X_7)	0.122	0.117	0.240
Factor3 (X_8)	0.036	0.034	0.070
Factor4 (X_9)	0.052	0.050	0.102

Note: Total/Direct = 1.9611
Since Ratio = $\frac{\text{Total}}{\text{Direct}} = \{(N^{-1}\iota_n' S(W)\iota_n) \times \beta_r\}/\{N^{-1}tr(S(W)_{ii}) \times \beta_r\}$ = $(\iota_n' S(W)\iota_n)/\ tr(S(W)_{ii})$, β is cancelled out. So ratio of ‘Average Total Impact’ and ‘Average Direct Impact’ should be same across variables.

Note: ρ=0 implies ratio=1 which means network has no indirect impact.

Table H9. Summary Table of Distribution of Individual Level Impacts

	Min	Mean	Max	S.D.
Direct Impact	1.000	1.003	1.005	0.002
Total Impact	1.601	1.894	2.319	0.125
Indirect Impact	0.601	0.892	1.315	0.124

Table H10. Parameter Estimates of OLS

DV	Direct Impact	Indirect Impact
Variable	Parameter Estimates	Parameter Estimates
Intercept	1.000***	0.021***
Age (X_1)	-0.005	0.021
Gender (X_2)	1.383	1.042
ln (House Value) (X_3)	0.175	-1.356***
Game Money (X_5)	-0.150	-0.086
Factor1 (X_6)	0.133	0.205
Factor2 (X_7)	-1.343	6.229***
Factor3 (X_8)	-0.069	0.029*
Factor4 (X_9)	0.160	-0.750***
Adjusted R^2	0.406	0.551

Note: *** = *p*<0.01, ** = *p*<0.05, * = *p*<0.10

Table H11. Summary Statistics

	Mean for Group1	Mean for Group2
House Value	7.363	7.458
Factor1	-0.272	0.159
Factor2	-0.725	0.424

Table H12. Parameter Estimates

Variable	Total Revenue		Game Time		Item Purchase	
	Original Model	Endogeneity Corrected	Original Model	Endogeneity Corrected	Original Model	Endogeneity Corrected
Intercept	1.365***	1.409***	0.038	0.054	0.068	0.084
Rho (ρ)	0.357***	0.349***	0.283***	0.286***	0.419***	0.412***
Age	-0.001	-0.001	-0.004	-0.004	-0.001	-0.001
Gender	0.018	0.021	0.365***	0.365***	0.201**	0.202**
Home Value	0.297***	0.291***	0.076	0.075	0.251***	0.250***
Virtual Money	0.760***	0.747***	0.781***	0.786***	0.466***	0.458***
Guild Activity	0.105***	0.111***	0.111***	0.118***	0.051**	0.058**
Action on People	0.053***	0.054***	0.092*	0.094***	0.007	0.005
Social Status	0.068***	0.062***	0.146**	0.155***	0.197***	0.197***
Action on World	0.013*	0.015*	0.004	0.000	0.010	0.014
AIC	8258.563	8317.217	12588.853	12589.24	15034.209	15032.562

Note: *** = *p*<0.01, ** = *p*<0.05, * = *p*<0.10

Note: Observation is stacked from April through August, and we use the same formula to get the factor score in monthly basis

Table H13. Spearman's Rank Order Correlation between each model

	Total Revenues	Item Purchase	Game Time
Total Revenues	1.000		
Item Purchase	0.848	1.000	
Game Time	0.757	0.890	1.000

Table H14. Correlation between Total Revenue, Game Time and Item Purchase

	Total Revenue	Game Time Payment	Item Purchase
Total Revenue	1		
Game Time Payment	0.182	1	
Item Purchase	0.486	0.125	1

Table H15. Parameter Estimates of the Spatio-temporal Model (DV: Total Revenues)

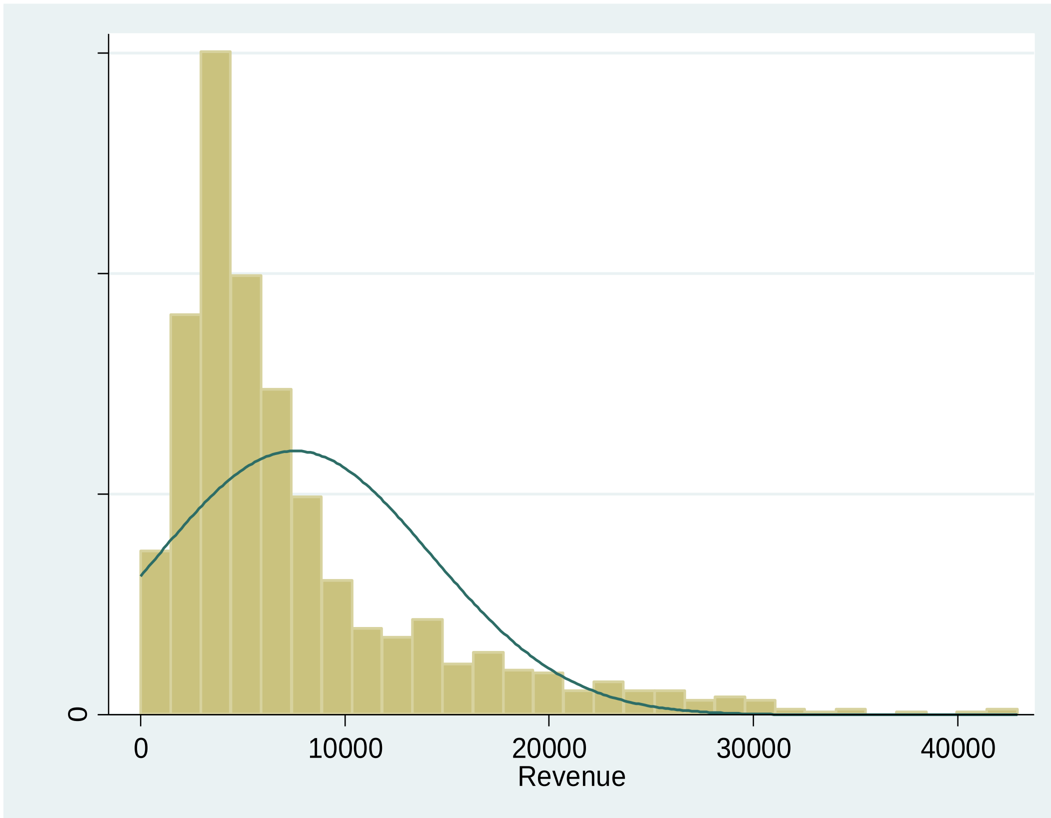
Variable	Spatial	Spatial + Spatio-temporal	Spatial + Temporal	All Effects
Intercept	1.365***	1.290***	0.095	0.091***
Inertia (ϕ)			0.273**	0.308**
Spatial Lag (λ)	0.357**	0.219**	0.135***	0.254**
Spatio-temporal Lag (θ)		0.187**		-0.065**
Age	-0.001	-0.001	-0.002	-0.005
Gender	0.018	0.022	0.010	0.019
ln (Home Value)	0.297***	0.293***	0.243*	0.253*
Virtual Money	0.760***	0.748***	0.072	0.068
Guild Activity	0.105***	0.095***	0.062***	0.065***
Action on People	0.053***	0.053***	0.030*	0.030*
Social Status	0.068***	0.069***	0.021	0.019
Action on World	0.013*	0.014*	0.025	0.025
AIC	8258.563	6617.269	4066.794	4059.518
Log-likelihood	-4119.283	-3297.635	-2023.304	-2018.759

Note: *** = p<0.01, ** = p<0.05, * = p<0.10

Table H16. Mean Absolute Deviations (MAD)

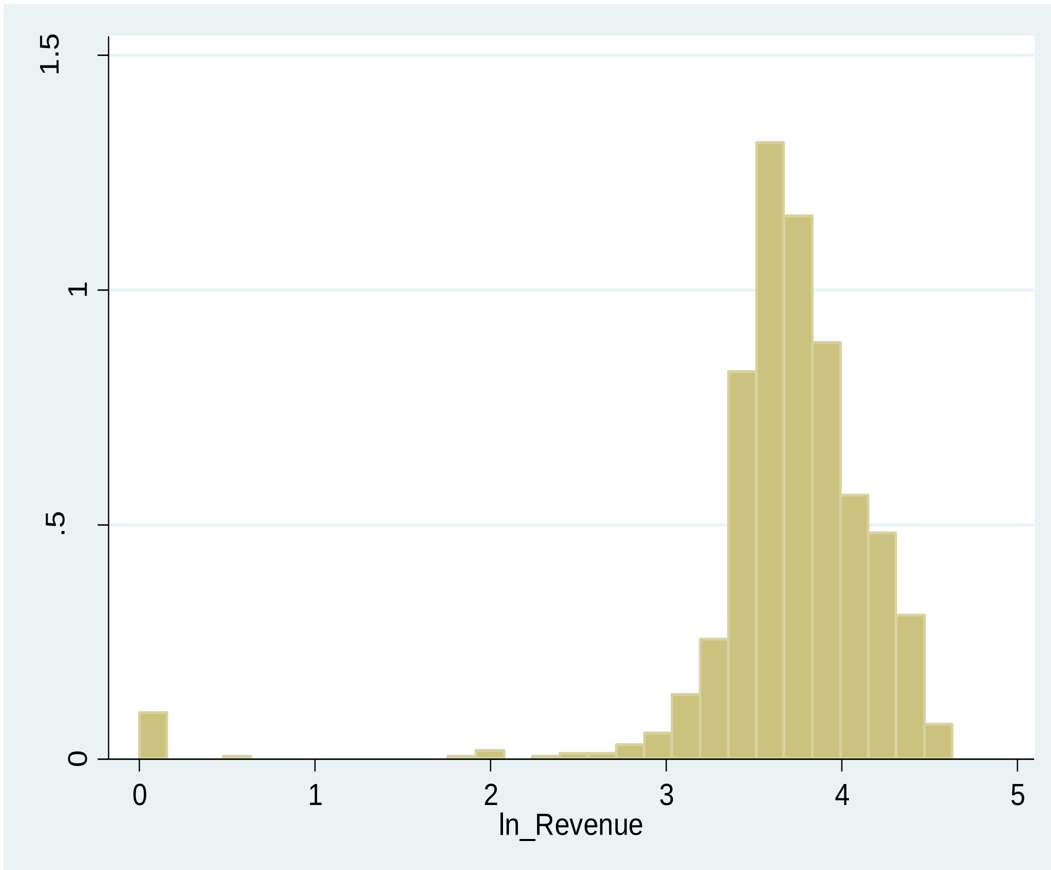
	MAD
Dynamic Model	1.518
Static Model	4.041

Figure H1. Distribution of ‘Revenue’ Variable



Note: Number of ‘zero’ observations are 16 out of 1000.

Figure H2. Distribution of Log-transformation of Revenue



Note: We use transformation as {log (Revenue)+1} for the new dependent variable

Figure H3. Scree Plot of Eigenvalues

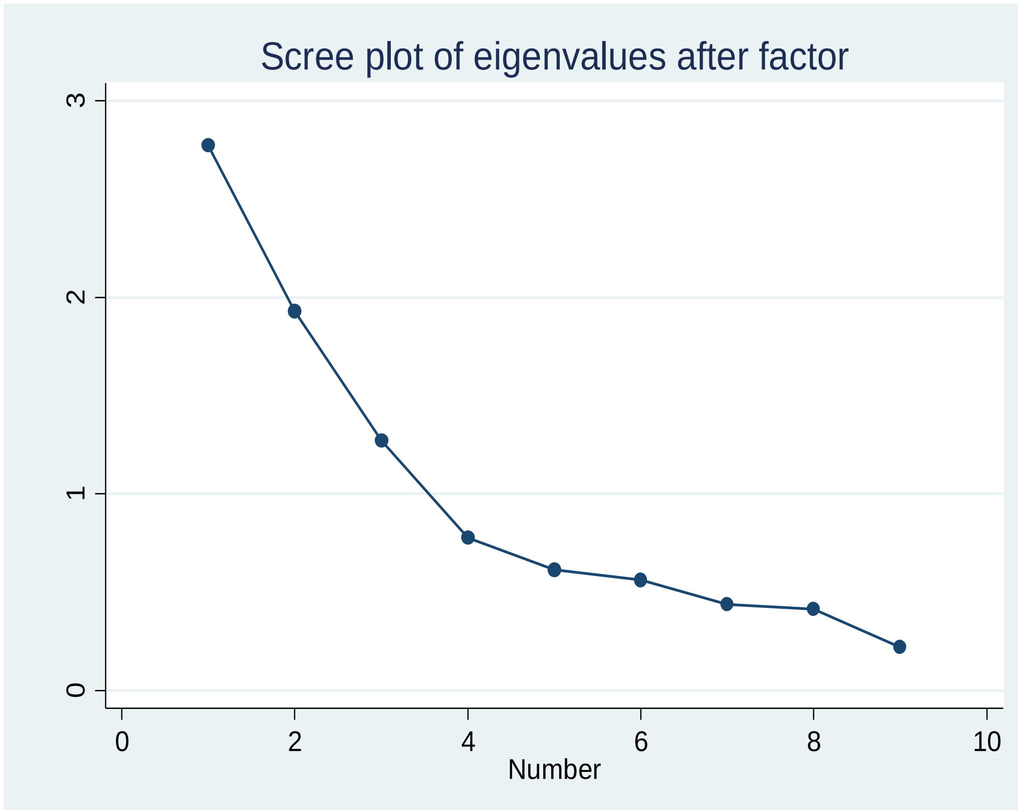


Figure H4. 3D Map of Social Influence

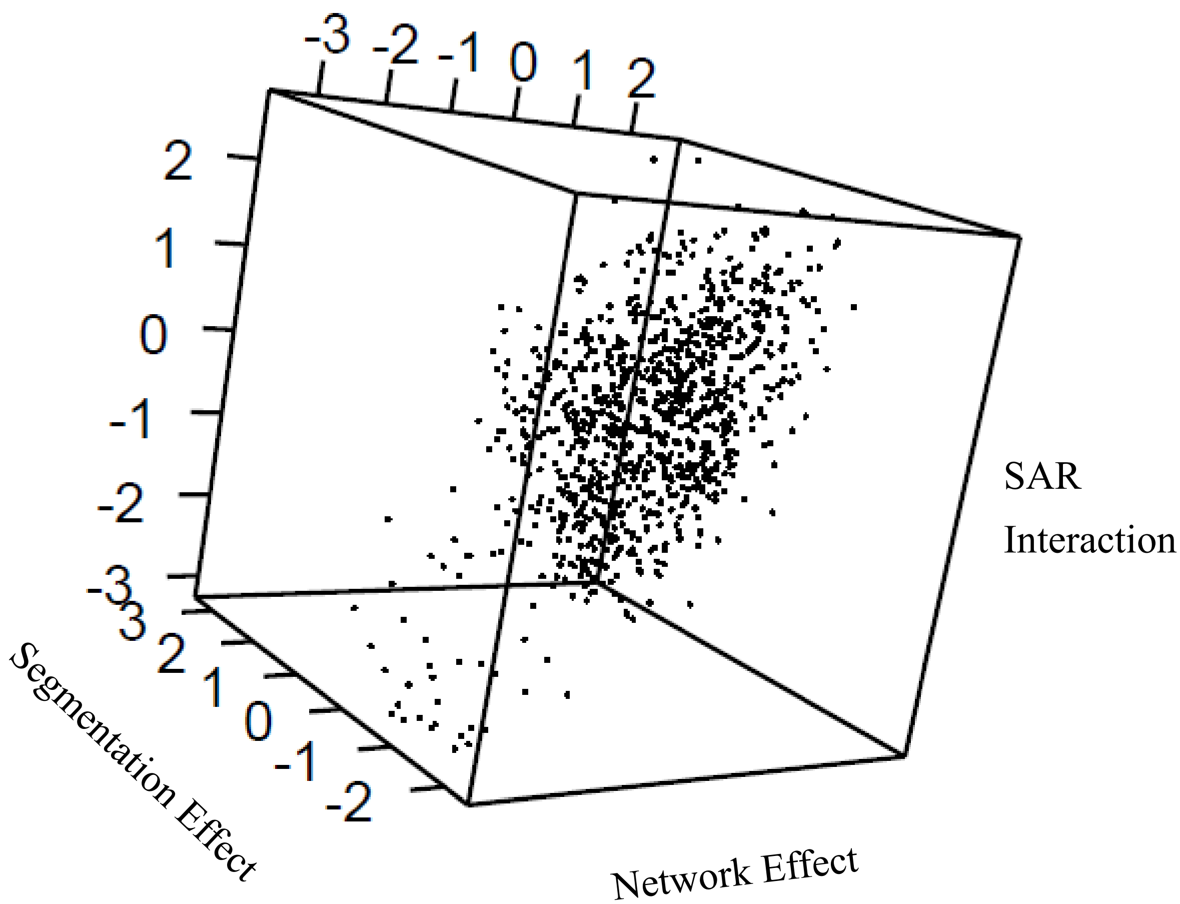


Figure H5. Gini Curve of Static Model

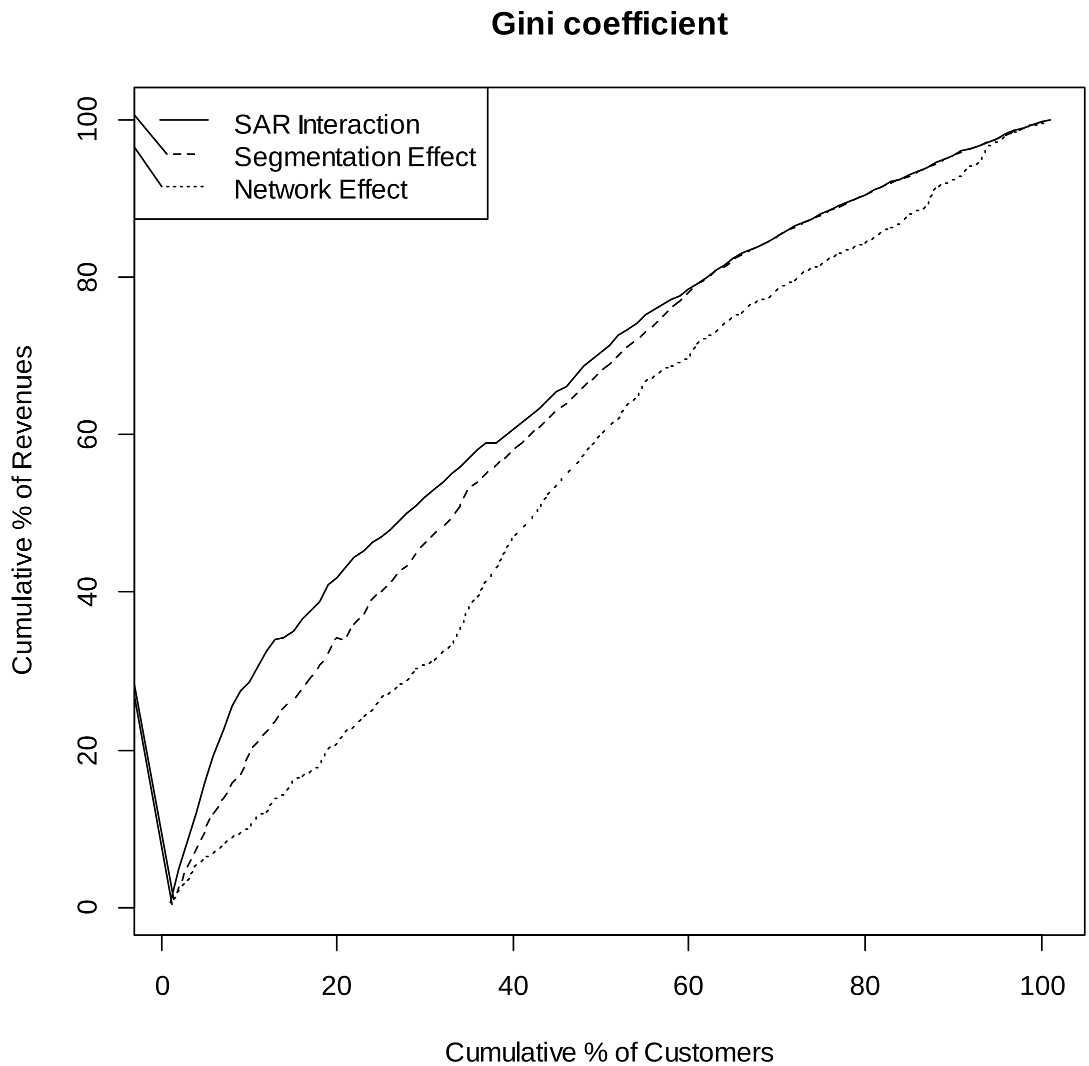


Figure H6. Revenue Curve

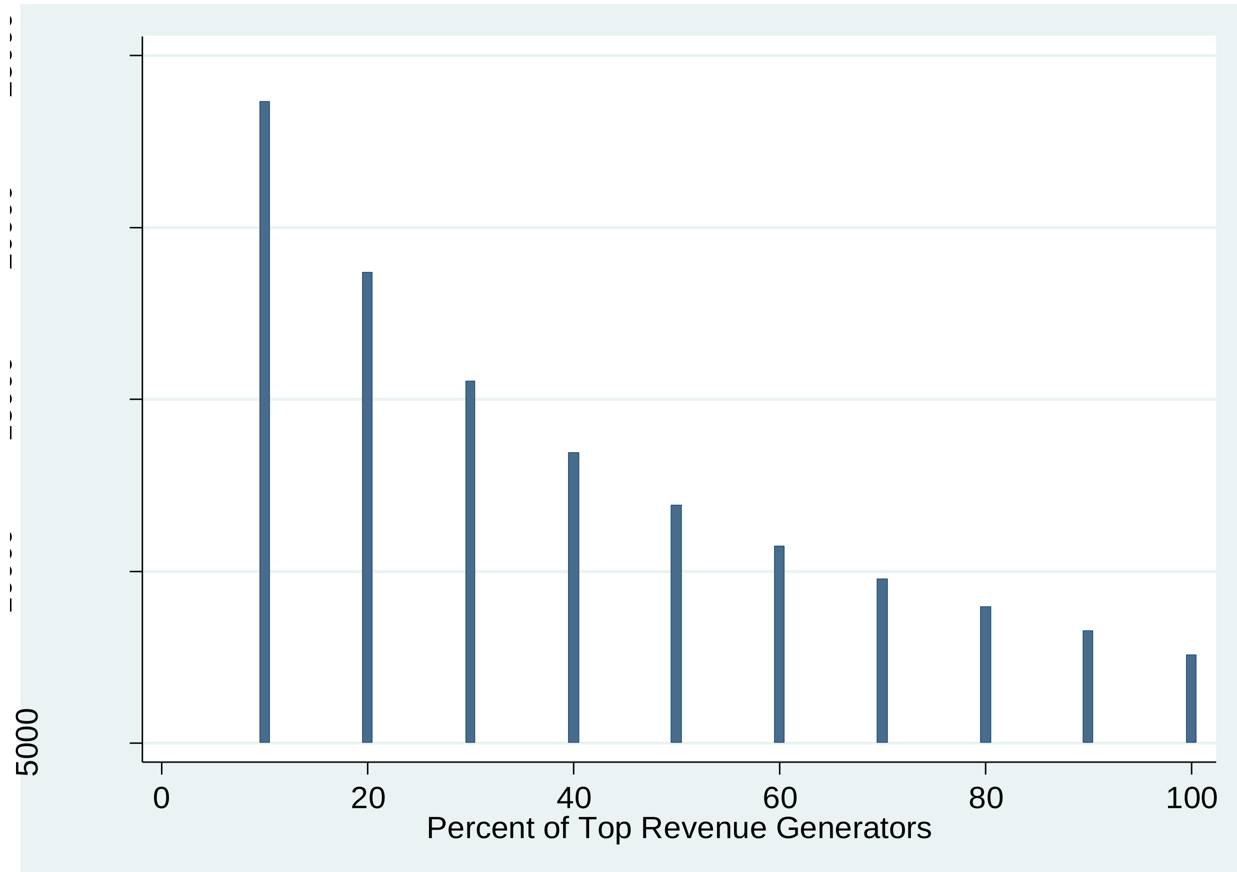


Figure H7. Distribution of Direct Impact

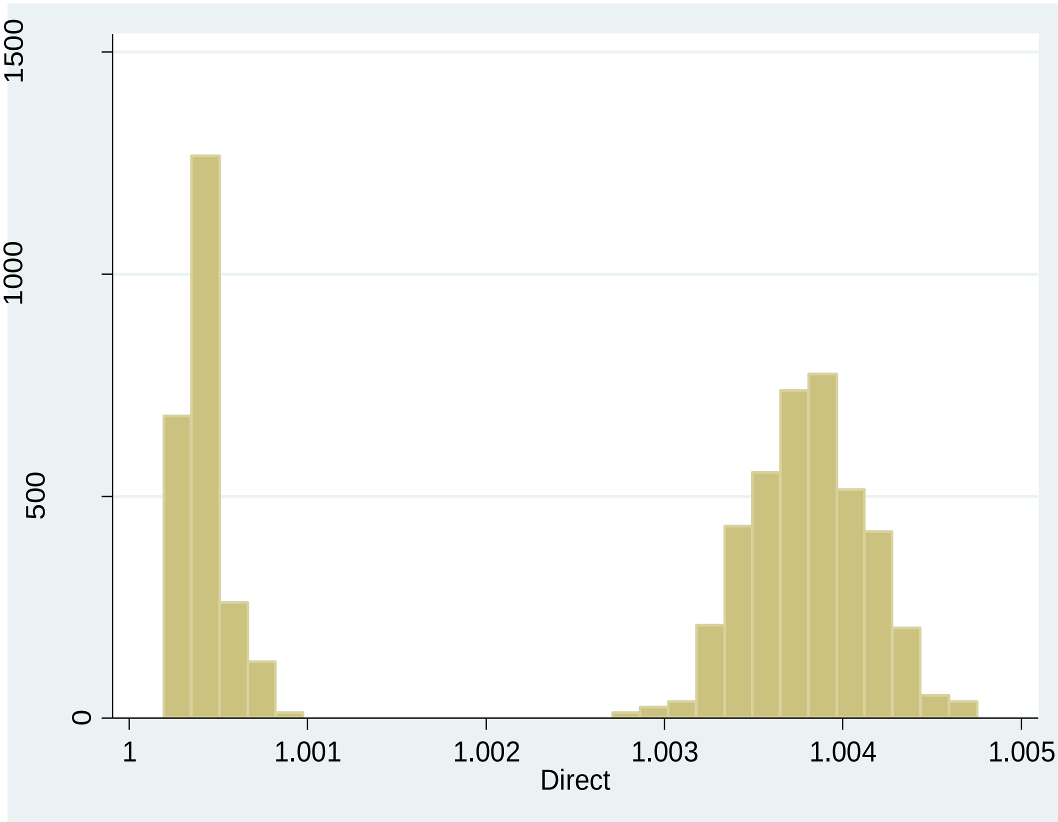


Figure H8. Distribution of Indirect Impact

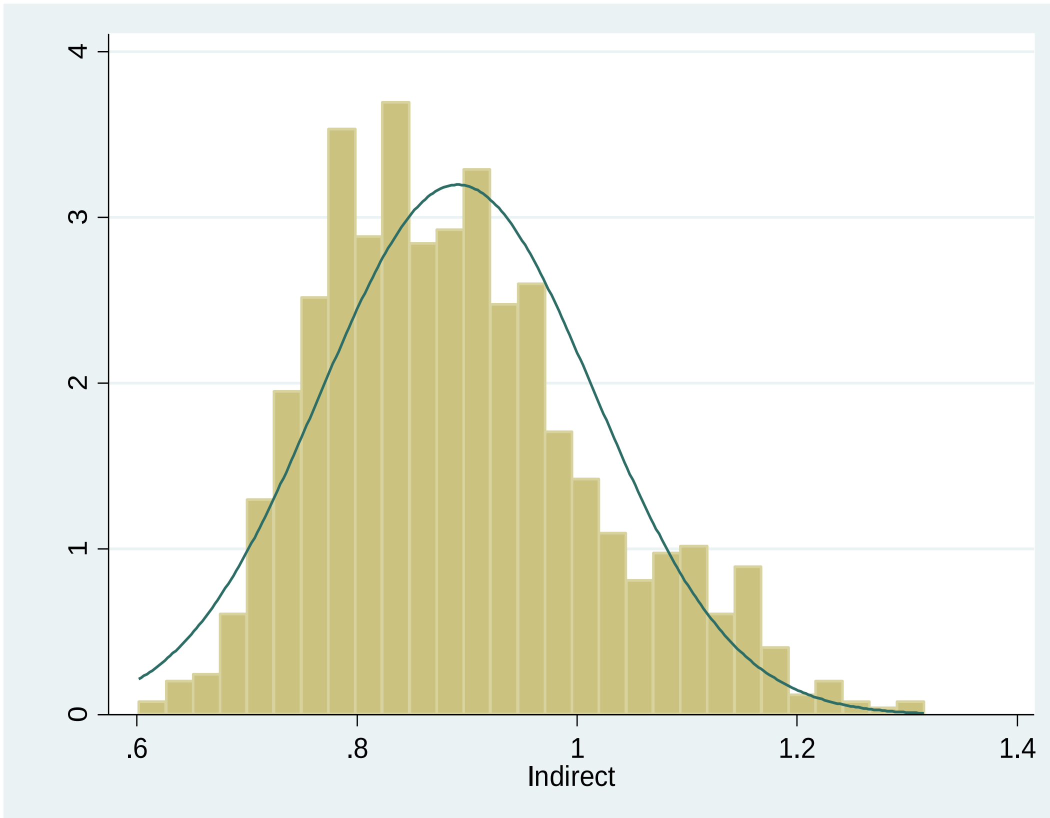


Figure H9. Distribution of Total Impact

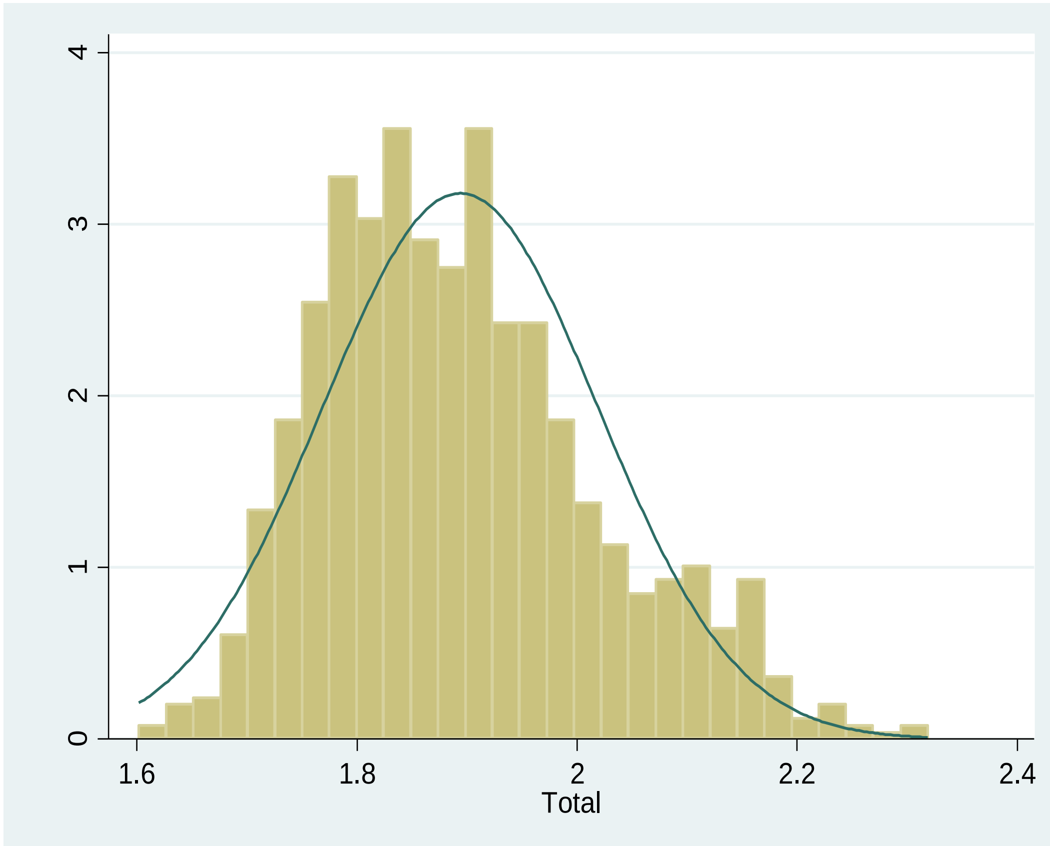


Figure H10. Distribution of Total Revenues

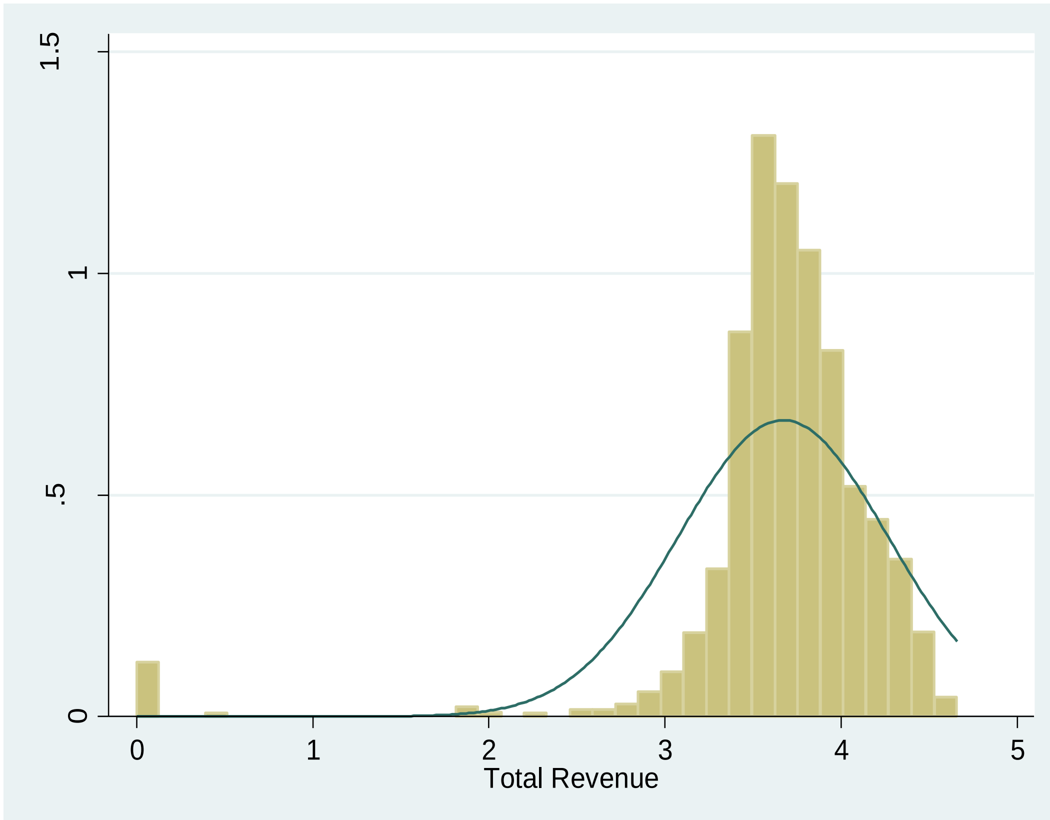


Figure H11. Distribution of Item Purchase Payment

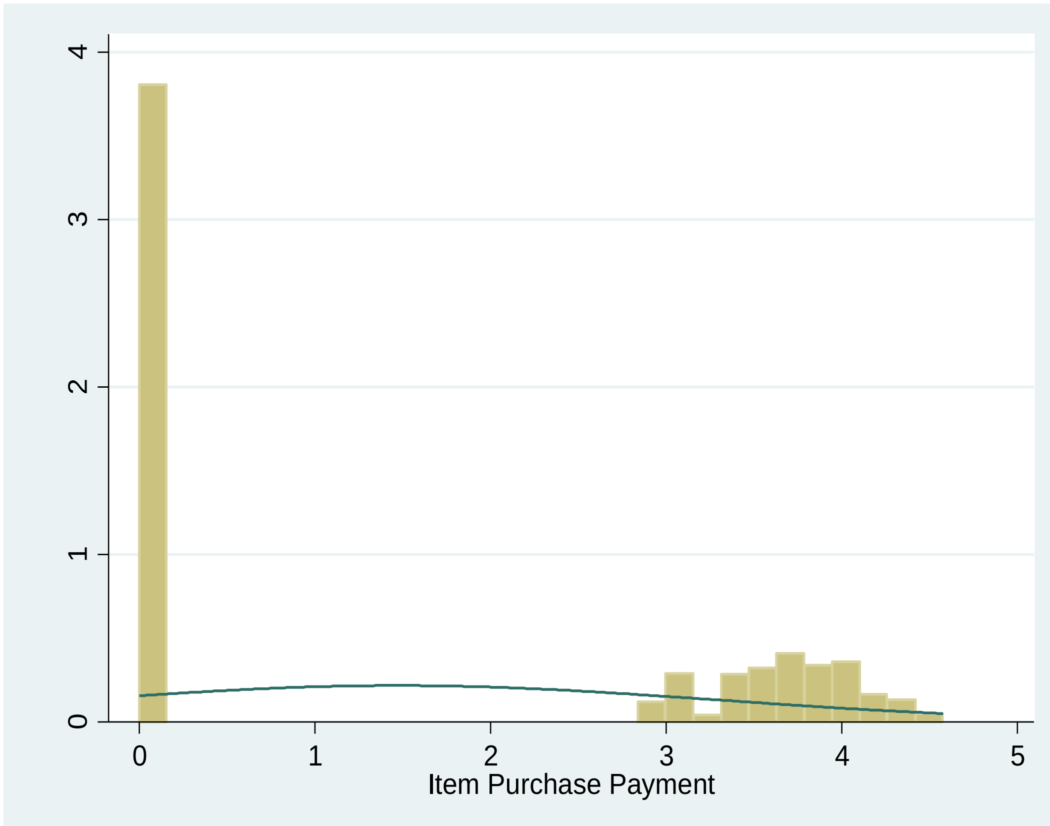


Figure H12. Distribution of Game Time Payment

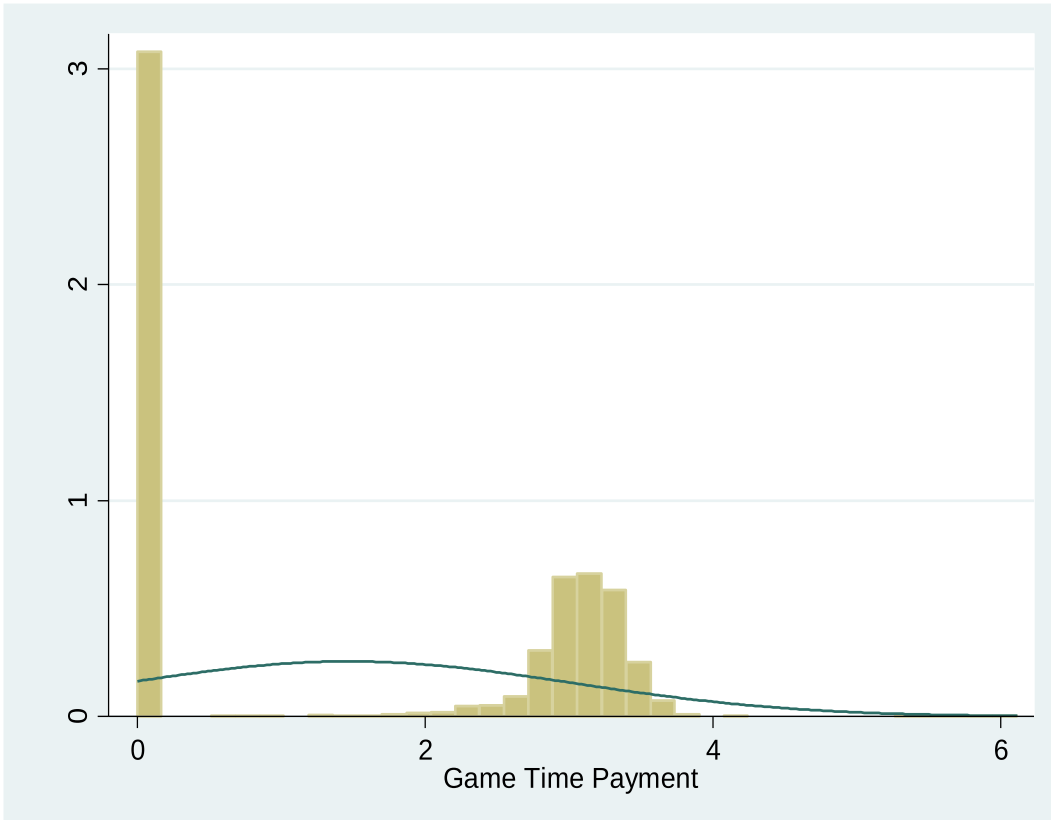
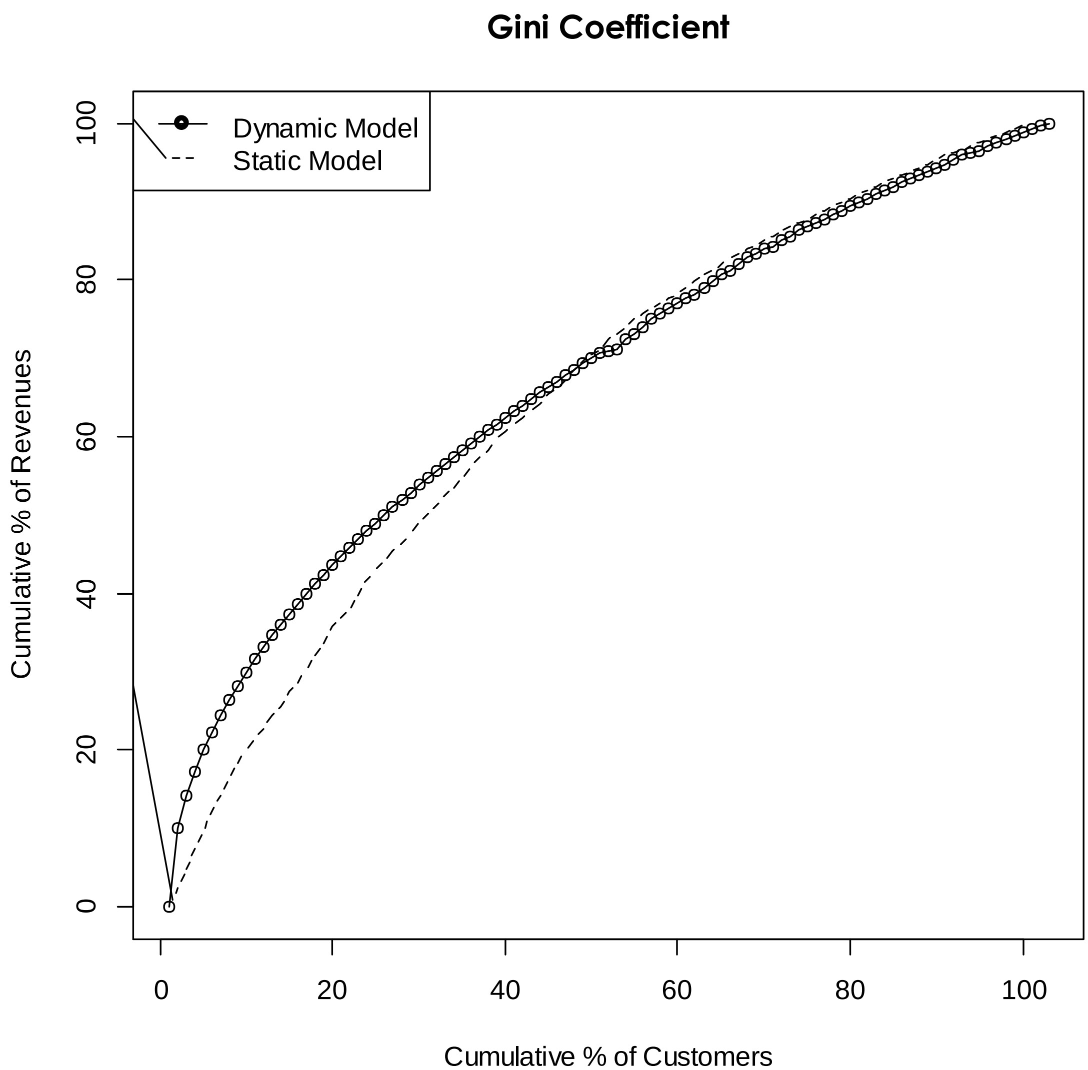


Figure H13. Gini Curve: Dynamic vs. Static Model



BIBLIOGRAPHY

Alessi, R., and Kapteyn, A. 1991. "Habit formation and interdependent preferences in the almost ideal demand system." *The Economic Journal,* 101: 404-419.

Andreoni, J., and Scholz, J.K. 1998. "An econometric analysis of charitable giving with interdependent preferences." *Economic Inquiry,* 36 (3), 410-428.

Anselin, L. 1980. "Estimation Methods for Spatial Autoregressive Structures." Regional Science Dissertation and Monograph Series 8. Field of Regional Science, Cornell University, Ithaca, N.Y.

Anselin, L. 1988. "Spatial Econometrics: Methods and Models." Dordrecht: Kluwer Academic Publishers.

Anselin, L. 1990. "Some robust approaches to testing and estimation in spatial econometrics." *Regional Science and Urban Economic,* 20, 141-163.

Anselin, L. 2001. "A Companion to Theoretical Econometrics," *Spatial econometrics,* In: Baltagi, Badi H. (Ed.), Blackwell Publishers Lte, Massachusetts, Chapter 14.

Anselin L. 2001. "Rao's score test in spatial econometrics." *Journal of Statistical Planning and Inferences,* 97, 113-139.

Anselin, L. 2003. "Spatial Externalities, Spatial Multipliers and Spatial Econometrics." *International Regional Science Review,* 26, 153-166.

Anselin, L., and A, Bera. 1998. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." In: A. Ullah and D. E. A. Giles, Eds., Handbook of Applied Economic Statistics 237–289. New York.

Anselin, L., and Griffith, D. 1988. "Do Spatial Effects Really Matter in Regression Analysis." *Regional Science Association,* 65, 11–34.

Anselin, L., J. Le Gallo, and H. Jayet. 2008. "Spatial Panel Econometrics." *The Econometrics of panel Data: Fundamentals and Recent Developments in Theory and Practice*, Springer Berlin Heidelberg.

Argo, J. J., Dahl, D. W,, and Morales, A. C. 2006. "Consumer Contamination: How Consumers React to Products Touched by Others." *Journal of Marketing,* 70 (2), 81-94.

Argo, J. J., Dahl, D. W., and Morales, A. C. 2008. "Positive Consumer Contagion: Responses to Attractive Others in a Retail Context." *Journal of Marketing Research,* 45 (6), 690-701.

Aronsson, T., Blomquist, S,. and Sacklen, H. 1999. "Identifying Interdependent Behavior in an Empirical Model of Labor Supply." *Journal of Applied Econometrics* 14(6), 607-626.

Asch, S. 1952. "Effects of Group Pressure upon the Modification and Distortion of

Judgments.” *In Readings in social psychology*, ed. Newcomb T. Swanson and E. Hartley, New York: Holt, 2-11.

Baltagi B.H., Song, S.H., Jung, B.C., and Koh, W. 2007. “Testing for Serial Correlation, Spatial Autocorrelation and Random Effects Using Panel Data." *Journal of Econometrics*, 140, 5-51.

Barry, R.P., and Pace, R.K. 1999. “A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices.” *Linear Algebra and its Applications*, 289, 41-54.

Bavelas, Alex. 1950. "Communication Patterns in Task Oriented Groups." *Journal of the Acoustical Society of America*, 57, 271-282.

Bavaud, F. 2002. “The Quasi-symmetric Side of Gravity Modelling.” *Environment and Planning A*, 34, 61-79.

Beauchamp, Murray A. 1965. "An Improved Index of Centrality." *Behavioral Science*, 10, 161-163.

Bernheim, D. 1991. "A Theory of Conformity." Department of Economics, Princeton University.

Bell, David R., and Song, Sanyoung. 2007. “Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing.” *Quantitative Marketing and Economics*, 5 (4), 361-400.

Bhatia, R., Kittaneh, F., and Li, R. 1998. “Eigenvalues of Symmetrizable Matrices.” *BIT Numerical Mathematics*, 38(1), 1-11.

Bonacich, P.B. 1987. “Power and Centrality: a Family of Measures.” *American Journal of Sociology*, 92, 1170-1182.

Bradlow, E., Bronnenberg, B., Russell, G.J., Arora, N., Bell, D., Duvvuri, S. D., Hofstede, F. Ter, Sismeiro, C., Thomadsen, R., and Yang, S. 2005. “Spatial Models in Marketing.” *Marketing Letters*, 16(3), 267-278.

Brock, William, and Durlauf, Steven. 2001. "Discrete Choice with Social Interactions." Review of Economic Studies, 68, 235-260.

Burt, Ronald S. 1987. “Social Contagion and Innovation: Cohesion versus structural equivalence.” *American Journal of Sociology*, 92, 1287-1335.

Burt, Ronald S. 1992. “Structural Holes: The Structure of Competition” Cambridge, MA: Harvard University Press.

Case, Anne C. 1991. “Spatial Patterns in Household Demand.” *Econometrica*, 59(4), 953-965.

Case, A. C., Rosen, H. S., and Hines, J. R. 1993. “Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States.” *Journal of Public Economics*, 52, 285–307.

Caussinus, H. 1966. “Contibution statistique `a l’analyse des tableaux de correlation.” *Ann.Fac.Sci.Univ.Toulouse*, 29, 177-182.

Choi, J., Hui, S. K., and Bell, S. K. 2010 “Spatiotemporal Analysis of Imitation Behavior across New Buyers at an Online Grocery Retailer.” *Journal of Marketing Research,*47, 75-89.

Coleman, James S., Elihu Katz, and Herbert Menzel. 1966. “Medical Innovation: A Diffusion Study.” Bobbs-Merrill Company, Indianapolis, IN.

Conley, T. G., and Topa, G. 2007. “Estimating Dynamic Local Interactions Models” *Journal of Econometrics,* 140, 282-303.

Cressie, N. 1993. “Statistics for Spatial Data.” New York, Wiley.

Darrough, M., Pollak., R.A., and T J. Wales. 1983. “Dynamic and Stochastic Structure: An Analysis of Three Time Series of Household Budget Studies.” *The Review of Economics and Statistics,* 65, 274-281.

Datcher, L. 1982. “Effects of Community and Family Background on Achievement.” *Review of Economics and Statistics,* 64(1), 32-41.

Debarsy, N., Ertur, C., and LeSage, J. 2011. “Interpreting Dynamic Space-time Panel Data Models.” Working Paper.

Doreian, Patrick. 1980. “Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects?” *Sociological Methods and Research,* 9, 29-60.

Duesenberry, J.S. 1949. “Income, Saving and the Theory of Consumer Behavior.” Harvard University Press, Cambridge, MA.

Duncan, O. Dudley, Archibald O. Haller, and Alejandro Portes 1968. “Peer Influences on Aspirations: Reinterpretation.” *American Journal of Sociology,* 74(2), 119-137.

Elhorst, J. P. 2001. “Dynamic Models in Space and Time.” *Geographical Analysis,* 33, 119-140.

Fishbein, M., and Ajzen, I. 1975. “Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.” Reading, MA: Addison-Wesley.

Franzese, J., and Hays, J. 2007. “Spatial-econometric Models of Cross-sectional Interdependence in Political Science Panel and Tim-series Cross-section Data.” *Political Analysi*s, 15(2), 140-164.

Freeman, L. 1977. “A Set of Measures of Centrality based upon Betweenness.” *Sociometry,* 40, 35-41.

Freeman, L. C. 1979. “Centrality in Social Networks: Conceptual Clarification.” *Social Networks,* 1, 215-239.

Godes, D., and Mayzlin, D. 2004. “Using Online Conversations to Study Word-of-Mouth Communication.” *Marketing Science,* 23(4), 545-560.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M., and Verlegh, P. 2005. "The Firm's Management of Social Interactions" *Marketing Letters*, 16(3), 415-428.

Goldenberg, Jacob, Sangman Han, Donald R. Lehmann, and Jaewon Hong 2009. "The Role of Hubs in the Adoption Process." *Journal of Marketing*, 73(March), 1-13.

Greene, W. 2003. "Econometric Analysis." Prentice-Hall, Upper Saddle River.

Griffith, D.A. (2000). "Eigenfunction Properties and Approximations of Selected Incidence Matrices Employed in Spatial Analysis." *Linear Algebra and Its Applications*, 321, 95-112.

Hartmann, W., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., and Tucker, C. 2008. "Modeling Social Interactions: Identification, Empirical Methods and Policy Implications." *Marketing Letters*, 19(3-4), 287-304.

Hartmann, W. R. 2010. "Demand Estimation with Social Interactions and the Implications for Targeted Marketing." *Marketing Science,* 29(4), 585-601.

Hays, J.C., Kachi, A., and Franzese J.R. 2010. "A Spatial Model Incorporating Dynamic, Endogenous Network Interdependence: A Political Science Application." *Statistical Methodology*, 7, 406-428.

Iacobucci, Dawn, and Nigel Hopkins 1992. "Modeling Dyadic Interactions and Networks in Marketin." *Journal of Marketing Research*, 29 (February), 5-17.

Iacobucci, Dawn 1996. Networks in Marketing. Thousand Oaks, CA: Sage Publications

Iacobucci, Dawn 1998 "Interactive Marketing and the Meganet: Network of Networks." *Journal of Interactive Marketing*, 12 (Winter), 5–16.

Iyengar, R., Han, S., and Gupta, S. 2009. "Do Friends Influence Purchases in a Social Network?" Working Paper, University of Pennsylvania.

Iyengar, R., Van den Bulte, C., and Valente, T.W. 2010. "Opinion leadership and social contagion in new product diffusion." *Marketing Science*, forthcoming.

Iyengar, R., Van den Bulte, C., and Choi, J. 2011. "Distinguishing between Drivers of Social Contagion: Insights from Combining Social Network and Co-location Data."Working Paper.

Kapteyn, A., Van De Geer, S., Van De Stadt, H., and Wasnsbeek, T. 1997. "Interdependent Preferences: An Econometric Analysis." *Journal of applied econometrics,* 12, 665-686.

Katz, L. (1953), "A New Status Index Derived from Sociometric Analysis," Psychometrika, 18, 39-43.

Kelejian, H.H., G.S. Tavlas, and G. Hondronyiannis. 2006. "A Spatial Modeling Approach to Contagion Among Emerging Economies." Open Economies Review, 17(4), 423-442.

Keller, Ed, and Barry, J. 2003. "The Influentials: One American in Ten Tells the other Nine How to Vote, Where to Eat, and What to Buy." New York: The Free Press.

Kelman, Herbert C. 1961. "Process of Opinion Change." *Public Opinion Quarterly,* 25(1), 57-78.

Lacombe, D. J. 2004. "Does Econometric Methodology Matter? An Analysis of Public Policy Using Spatial Econometric Techniques." *Geographical Analysis*, 36(2), 105-118.

Lee, Lung-fei, and Jihai Yu, 2009. "Some Recent Developments in Spatial Panel Data Models." *Regional Science and Urban Economics*, 40(5), 255-271.

Leenders, R.Th.A.J. 1997. "Longitudinal Behavior of Network Structure and Actor Attributes: Modeling Interdependence of Contagion and Selection*."* In:Doreian, P.,Stokman,F.N. (Eds), Evolution of Social Networks. Gordon and Breach, New York.

Leenders R.2002. "Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix." *Social Networks,* 24(1), 21-47.

LeSage, J. P. 1997. "Regression Analysis of Spatial Data." *Journal of Regional Analysis and Policy,* 27(2), 83-94.

LeSage, J. P., and Pace, K. 2010. "Introduction to Spatial Econometrics." Chapman and Hall/ CRC Press.

Leibenstein, H. 1950. "Bandwagon Snob, and Veblen Effects in the Theory of Consumers' Demand." *Quarterly Journal of Economics*, 64, 183-207.

Manchanda, P., Xie, Y., and Youn, N. 2008. "The Role of Targeted Communication and Contagion in Product Adoption." *Marketing Science*, 27 (6), 961–76.

Manski, C. F. 1993. "Identification of Endogenous Social Effects: the Reflection Problem." *Review of Economic Studies*, 60, 531-542.

Manski, Charles F. 2000. "Economic Analysis of Social Interactions." *Journal of Economic Perspective.* 14(3) 115-36.

Martin, R.J. 1993. "Approximations to the Determinant Term in Gaussian Maximum Likelihood Estimation of Some Spatial Models." *Communications in Statistics – Theory and Methods*, 22, 189-205.

Mason, Roger. 1995. "Interpersonal Effects on Consumer Demand in Economic Theory and Marketing Thought, 1890-1950." *Journal of Economic Issues,* 29, 871-881.

Moffitt, R. 2001. "Policy Interventions Low-level Equilibria, and Social Interactions." MIT Press.

Nair, H. S., Manchanda, P., and Bhatia, T. 2010. “Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders.” *Journal of Marketing Research*, 47, 883-895.

Nair, H., Manchanda, P., and Bhatia, T. 2010. “Asymmetric Peer Effects in Prescription Behavior: the Role of Opinion Leaders.” *Journal of Marketing Research*, 47(5).

Narayanan, S., Vishal, S., and Yang, S. 2007. “Trust Between Consumers in Online Communities: Modeling the Formation of Dyadic Relationships.” Working Paper, Stern School of Business, New York University.

Narayanan, S., and Nair, H. 2011. “Estimating Causal Installed-Base Effects: A Bias-Correction Approach.” Working Paper.

Naroll, R. 1961. “Two Solutions to Galton’s Problem.” *Philosophy of Science*, 28, 15-39.

Nam, S., P. Manchanda, and P. Chintagunta. 2008. “The Effects of Service Quality and Word of Mouth on Customer Acquisition, Retention and Usage.” Working Paper.

O’Malley, A.J., and Marsden, P.V. 2009. “The Analysis of Social Networks.” Health Services and Outcomes Research, 1, 8(4), 222-269.

Openshaw, S. 1984. “Ecological Fallacies and the Analysis of Areal Census Data.” *Environment and Planning A*, 16, 17-31.

Opsahl, T., Filip, A., and Skvoretz, J. 2010. "Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths." *Social Networks*, 32, 245.

Ord, Keith. 1975. “Estimation Methods for Models of Spatial Interaction.” *Journal of the American Statistical Association*, 70, 120-126.

Ord, Keith. 1981. “Towards a Theory of Spatial Statistics: A Comment.” *Geographical Analysis*, 13: 91-93.

Pace, R.K., and R.P. Barry. 1997. “Quick Computation of Spatial Autoregressive Estimators.” *Geographical Analysis*, 29, 232-246.

Parent, O., and J. P. LeSage. 2010. “Spatial Dynamic Panel Data Models with Random Effects Applied to Commuting Times.” *Transportation Research Part B: Methodological*, 44B, 633-645.

Parent, O., and J. P. LeSage. 2010. “A Space-time Filter for Panel Data Models Containing Random Effects.” *Computational Statistics and Data Analysis*, Forthcoming.

Pollak, Robert A. 1976. “Interdependent Preferences.” *The American Economic Revie,*. 66(3), 309-320.

Prais, S., and Winstein, C. 1954. “Trend Estimates and Serial Correlation.” Cowles Commision, Chicago, IL.

Robins, G., Pattison, P., and Elliott, P. 2001. “Network Models for Social Influence

Processes." *Psychometrika*, 66 (2), 161-90.

Ronis, D.L., J.F. Yates, and J.P. Kirscht. 1989. "Attitudes, Decisions, and Habits as Determinants of Repeated Behavior in Attitude Structure and Function." A.R. Pratkanis, S.J. Breckler, and A.G. Greenwald (eds.), Hillsdale, NJ: Lawrence Erlbaum Associates, 213-239.

Sims, J.T. 1971. "Comparison of Consumer Behavior Conformity and Independence Between Blacks and Whites: An Exploratory Study." *Proceedings:Second Annual Conference, Association for Consumer Research*, 76-81.

Smirnov, O., and Anselin, L. 2001. "Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach." *Computational Statistics and Data Analysis*, 35, 301-319.

Smith, T. E., and J. P. LeSage. 2004. "A Bayesian Probit Model with Spatial Dependencies." in James P. LeSage and R. Kelley Pace (eds.), *Advances in Econometrics: 18: Spatial and Spatiotemporal Econometrics*, Elsevier Science, 127–160.

Snijders, T. 2005. "Models for Longitudinal Network Data." In Carrington, P., Scott, J., and Wasserman, S., editors, Models and Methods in Social Network Analysis, chapter 11. Cambridge Uninversity Press, New York.

Snijders, T., Steglich, C., and Schweinberger, M. 2007. "Modeling the Co-Evolution of Networks and Behavior." In van Montfort, K., Oud, H., and Satorra, A., editors, Longitudinal Models in the Behavioral and Related Sciences, chapter 3, Lawrence Erlbaum, Mahwah, NJ.

Soetevent, A. R., and P. Kooreman. 2007. "A Discrete-choice Model with Social Interactions: With an Application to High School Teen Behavior." *Journal of Applied Econometrics*, 22(3), 599-624.

Sorensen, A. 2006. "Social Learning and Health Plan Choice." *Journal of Economics*, 37, 929-945.

Stafford, J.E. 1966. "Effects of Group Influence on Consumer Brand Preferences." *Journal of Marketing Research,* 3, 68-75.

Stephen, A. T., and Toubia, O. 2010. "Deriving Value from Social Commerce Networks." *Journal of Marketing Research*, 47 (April), 215–28.

Strang D., and Tuma, N. B. 1993. "Spatial and Temporal Heterogeneity in Diffusion." *American Journal of Sociology*, 99, 614-639.

Sun, D., R.K. Tsutakawa, P.L. Speckman. 1999. "Posterior Distribution of Hierarchical Models using CAR(1) Distributions." *Biometrika*, 86, 341-350.

Trusov, M., Bucklin, R. E., and Pauwels, K. 2009. "Effects of Word-of-Mouth versus Traditional Marketing: Findings from an Internet Social Networking Site." *Journal of Marketing*, 73, 90-102.

Trusov, M., Bodapati, A.V., and Bucklin, R.E. 2010. "Determining Influential Users in Internet Social Networks." *Journal of Marketing Research*, 47, 643-658.

Van den Bulte, C., and G. L. Lilien. 2001. "Medical Innovation Revisited: Social Contagion versus Marketing Effort." *American Journal of Sociology*, 106(5), 1409-1435.

Van den Bulte, C., and Wuyts, S. 2007. "Social Networks and Marketing." Cambridge, MA: Marketing Science Institute.

Venkatesan, M. 1966. "Experimental Study of Consumer Behavior, Conformity and Independence." *Journal of Marketing Research*, 3, 384-387.

Wasserman, S., and Faust, K. 1994. "Social Network Analysis: Methods and Applications." Cambridge University Press, New York.

Watts, D. J., and Dodds, P. S. 2007. "Influentials, Networks, and Public Opinion Formation." *Journal of Consumer Research*, 34 (December), 441-458.

Winder, C. C. A. and F. C. Palm. 1989. "Intertemporal Consumer Behavior under Structural Changes in Income." *Econometric Reviews*, 8(1), 1-88.

Whittle, P. 1954. "On Stationary Processes in the Plane." *Biometrika*, 41, 434-449.

Yang, S., and Allenby, G. 2003. "Modeling the Interdependent Consumer Preferences." *Journal of Marketing Research*, 40(3), 282-294.

Yu, J., R. de Jong, and L.F. Lee. 2008. "Quasi-Maximum Likelihood Estimators For Spatial Dynamic Panel Data with Fixed Effects When Both n and T Are Large" *Journal of Econometrics*, 146, 118-134.