ABSTRACT

Title of Document: CHOICE MODELING PERSPECTIVES ON SOCIAL NETWORKS, SOCIAL INFLUENCE, AND SOCIAL CAPITAL IN ACTIVITY AND TRAVEL BEHAVIOR

Michael Maness, Doctor of Philosophy, 2015

Directed By: Associate Professor, Cinzia Cirillo, Department of Civil and Environmental Engineering

Understanding the determinants of activities and travel is critical for transportation policymakers, planners, and engineers to design and manage transportation systems. These systems, and their externalities, are interwoven with social systems in communities, cities, regions, and societies. But discrete choice models – the predominant modeling tool for researching travel behavior and planning transportation systems – are grounded in theories of individual decision-making. This dissertation expands knowledge about the incorporation of social interactions into activity-travel choice models in the areas of social capital and social network indicators; social influence motivations and informational conformity; and misspecification errors from social network data collection.

Incorporating social capital into activity choice models involves using social capital indicators from surveys. Using a position generator question type, the role of social network occupational diversity in activity participation was explored and the performance of models using name generator and position generator data was

compared. Access to the resources embedded in diverse networks (extensity) was found to positively correlate with leisure activity participation. Compared to core network indicators from name generators, position generator indicators were typically better at predicting activity participation in a cross-validation study.

Current models of social influence in travel do not account for varying motivations for social influence such as for accuracy, affiliation, and self-concept. To test for an accuracy motivation, a latent class discrete choice model was formulated that places individuals into classes based on information exposure. Contrasting with existing work, this model showed that "more informed" households are more likely to own bicycles due to preference changes causing less sensitivity to smaller home footprints and limited incomes. A Bayesian prediction procedure was used to derive distributions of local-level equilibria and social influence elasticity.

The effect of errors in social network data collection using name and position generators is not fully understood for choice models. In a case study, the social network occupational diversity measure was robust to varying position generator lengths. Simulation experiments tested the implications of social network structure, misspecification, and small samples on social influence choice models where sample size, social influence strength, and degree of misspecification had the greatest impact.

CHOICE MODELING PERSPECTIVES ON SOCIAL NETWORKS, SOCIAL INFLUENCE, AND SOCIAL CAPITAL IN ACTIVITY AND TRAVEL BEHAVIOR

By

Michael Maness

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Associate Professor Cinzia Cirillo, Chair
Associate Professor David Lovell
Assistant Professor Christina Prell
Professor Ingmar Prucha
Professor Paul Schonfeld

Dedication

This dissertation is dedicated to my grandmothers – the late Edna Mae Thomas (1925-2013) and Ruth Ann Crudup.

Acknowledgements

I would like to acknowledge my advisor, Cinzia Cirillo, for her assistance and patience in my academic progression and research endeavors. It took a lot of trust for her to support me in the design of a dissertation topic from the bottom up. All of my colleagues in her research group have been helpful and supportive, and I have bounced a lot of ideas off of everyone from my initial research on the economics of identity to my final work here on social influence and social capital.

Additionally, I would like to thank everyone on my dissertation committee. Ingmar Prucha was a valuable resource in learning about spatial econometrics and has raised important considerations about correlations and simultaneity. David Lovell has been helpful in my simulation work and in the proper usage of parametric and non-parametric statistical tests. Christina Prell provided valuable sociology expertise and helped with my work on ERGMs and social capital. I appreciate Paul Schonfeld for being available and accommodating at such a late stage in the process. Also, I would like to thank Lei Zhang for his comments at the proposal stage which were helpful in guiding me to find a new dataset and to incorporate endogeneity in my framework.

In keeping with the theme of this dissertation, this section would not be complete without acknowledging the role of my social networks and my access to social capital. First, I would like to thank my parents for all their love and support. I would like to thank my friends and family for their support and for distracting me from my work at times. I cannot name everyone who I appreciate for their support, but I would like to particularly thank (in order of meeting them): Ankit Kavi, John Wegand, Adam Gorby, Jeb Brako, Sara Williams, Nancy Robinson, and Jennifer

Biehl. Additionally, I would like to acknowledge other researchers who have offered help and advice for this dissertation: Elenna Dugundji, Daniel Dailey, Karen White, Anna Alberini, and Erkut Ozbay. Also, Anna Petrone helped me with a proof.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Over the past two decades, travel behavior analysis has begun shifting focus from the individual (Larsen et al. 2006) to the social. Travel is an integral part of peoples' lives which shapes their residences and neighborhoods, work and economic opportunities, and activities and social networks. Travel connects and shapes people's social lives, but also, a person's social network and society at large can influence their travel behavior. Accordingly, discrete choice models, the primary modeling technique in travel behavior analysis, have begun to integrate social context into its framework (Dugundji and Walker 2005, Paez and Scott 2007). Durlauf and Ioannides (2010) define social interactions as "direct interdependences in preferences, constraints, and beliefs of individuals, which impose a social structure on individual decisions" (p.452). The constraints side of social interactions in travel has been the focus of work on intrahousehold and interhousehold activity planning and travel behavior. On the preferences and expectations side of social interactions, there has been growing interest in transportation to analyze models in which the preferences and decisions of others are incorporated into social influence models.

Being a relatively young sub-discipline in transportation, social interactions research has a variety of research foci and terminology. Social interactions research in transportation deals with three areas: (1) social cooperation, (2) social influence, and (3) social capital. Research on social interactions in travel first dealt with s*ocial cooperation* which deals with active coordination of travel and activities and generally involves intrahousehold and interhousehold planning and activity scheduling (Arentze and Timmermans 2008; Van den Berg et al. 2010, 2012; Carrasco and Miller 2006, 2009; Habib et al. 2008; Habib and Carrasco 2012). But there has been growing interest in studying s*ocial influence* where an individual's decision making process is altered by the actions, behaviors, attitudes, and beliefs of others as well as the individual's perceptions of these (Dugundji and Walker 2005, Paez and Scott 2007, Axsen et al. 2013, Bartle et al. 2013, Sherwin et al. 2014). Additionally, there has been limited research into transportation modeling and *social capital* (Carrasco and Cid-Aguayo 2012, Sadri et al. 2015), where individuals are enabled to achieve things working together that would otherwise be difficult or impossible to achieve alone (Field 2003). Thus far, research efforts have resulted in work showing that all three factors may be relevant in travel decision making.

What links these three areas is that each relies on underlying social networks. Connections with family and friends can provide social support thus promoting social capital. Car buyers are socially influenced in their decisions by advertisements from advertisers, brief connections with strangers, and the decisions and opinions of their neighbors. And the parents in a household are intricately connected with the activities of their children and their children's friends and must cooperate in the coordination of their travels. Social networks entail the social ties / connections that people make with one another in a variety of social contexts. These social ties can be strong ties such as between family members and close friends as well as weak ties such as between neighbors.

Of particular interest lately have been attempts to gain an understanding of these social interactions and use this knowledge to better understand travel and to modify behavior. The analysis of models in which social networks and social factors are incorporated into discrete choice models to understand and quantify these social effects is one avenue to accomplish this goal. But the incorporation of social interactions into discrete choice models is non-trivial since discrete choice models are grounded in theories of and methods for independent decision makers.

1.1 Who Cares about Social Interactions in Travel?

Differentiating social effects from non-social effects should be of interest to any parties attempting to understand the determinants of human behavior or to modify human behavior to accomplish certain objectives. In particular, social influence research can be used by institutions (government, non-profit, and for-profit) and behavioral researchers. When institutions are interested in changing the behavior of its members, it is important to understand the determinants of behavior. As shown in the examples above, if a government entity is interested in promoting cycling then it is important for them to know whether to invest in infrastructure or advertising campaigns. Infrastructure improvements are expensive and permanent whereas advertisements are typically cheaper and flexible in timing. Therefore it is critical to ensure that models correctly differentiate these effects, particularly for policy analysis and policy guidance.

Additionally, behavioral researchers are interested in understanding the processes that impact decision making. Qualitative and quantitative methods are used to study decision making in human populations. Qualitative methods are often useful for understanding decision making in new contexts and can be used to guide quantitative approaches. Quantitative models, which are often statistics-based, allow for researchers to examine relationship more subjectively and to account for randomness. By finding statistical methods which can differentiate the social and non-social factors, researchers can describe decision making processes more accurately and guide future research.

Choice modeling is a statistical tool which binds these two communities together. Choice modeling is often used in policy analyses to infer individuals' behavioral processes and to make predictions about the impact of different policy prescriptions. Additionally, choice models are often used to test theories of behavior in the travel behavior, economics, and marketing fields. But the incorporation of social influence into choice modeling is non-trivial and is not straightforward.

1.2 How are Social Interactions Incorporated into Choice Models?

The discrete choice model, a type of choice model, is the predominant modeling tool in travel demand modeling and travel behavior analysis. A choice model is a mathematical model of a decision process that maps aspects of the choice situation, such as individual attributes and item attributes, to a choice or choices. These models have traditionally focused on describing the choices of independent individuals through the evaluation of payoffs from choosing different alternatives. These payoffs depend on the attributes of the alternative and characteristics of the individuals. Most choice models have the following formulation[1]:

$$\begin{gathered} \mathcal{P}_{ni} = f_{ni}(x_{in}; \beta_i) + \varepsilon_{ni} \\ d(\mathcal{P}_{ni}, \forall i \in J_n) \rightarrow y_n \end{gathered} \tag{1}$$

where:

---

[1] Throughout this dissertation, payoffs will be denoted with capitalized, scripted characters, such as $\mathcal{P}$ and $\mathcal{U}$. Model parameters and error terms will be denoted by Greek letters. Variables will be denoted by Latin letters.

n ≡ an individual / decision maker

i ≡ an alternative

J_n ≡ the choice set for individual *n*

$\mathcal{P}_{ni}$ ≡ the payoff achieved by individual *n* if alternative *i* is chosen

y_n ≡ the choices made by individual *n*, denoted as the vector of size $|J|$ with all y_{ni} for all alternatives *i* in the choice set

y_{ni} ≡ a choice indicator function (typically for discrete choices, this is equal to 1 if alternative *i* is chosen by individual *n* and 0 otherwise)

f_{ni} ≡ a function which maps observable individual-level characteristics to units of payoff (observability is in reference to the modeler)

d ≡ the decision rule (or decision process) which maps the payoffs of each alternative to a choice

x_{ni} ≡ individual-level characteristics of individual *n* for alternative *i*, x_{ni} may include alternative-specific attributes (i.e. the explanatory variables)

ε_{ni} ≡ unobserved effects on individual *n* for alternative *i* (i.e. an error term)

β_i ≡ model parameters that weight the contributions of various individual-level characteristics to the payoff

In the field of transportation, random utility models of discrete choice (RUMs) are the most prevalent choice model (Ben-Akiva and Lerman 1985). In this formulation, an individual assigns a utility to each alternative then chooses the alternative with the largest utility. The modeler can observe some characteristics of the individual and each alternative but assumes that the unobserved portions of utility are randomly distributed. This assumption on the distribution of the unobservables leads to different popular formulations of RUMs such as logit, probit, nested logit, and mixed logit models. For example, the multinomial logit (i.e. conditional logit) model assumes that has ε_{ni} is type 1 extreme value distributed and thus the choice model takes the following formulation[2]:

$$\mathcal{U}_{ni} = \beta_i x_{in} + \varepsilon_{ni}, \ \forall n \in N, i \in J$$
$$y_{ni} = \begin{cases} 1 & if\ \mathcal{U}_{ni} = \max_{j \in J} \mathcal{U}_{nj} \\ 0 & otherwise \end{cases} \tag{2}$$

As stated previously, incorporating social interactions into choice models is non-trivial. It typically entails determining the type of social context and social factors at play, understanding the relevant social networks for this context, and finding measures that correlate with the decision and the social context. For a choice model, this means modifying the preferences, expectations, or constraints of the individuals based on the behavior, opinions, characteristics, and other properties of others and the individual's social network. Thus, this may change the formulation given by equation (1) as follows:

$$\mathcal{P}_{ni} = f_{ni}(x_{in}; \beta_i) + f_{ni}^{social}\left(G_n, x_{in}, m_{i(-n)}; \psi\right) + \varepsilon_{ni}$$
$$d_n\left(\mathcal{P}_{ni}, \mathcal{P}_{(-n)i}, G_n\ \forall i \in J_n(G_n, J_{-n})\right) \rightarrow y_n \tag{3}$$

where:

$-n$ ≡ all individuals / decision makers excluding individual *n*

G_n ≡ individual *n*'s social network

[2] Throughout this dissertation, coefficient vectors are assumed to be row vectors, while variable vectors are assumed to be column vectors.

$J_n(.)$ ≡ the choice set dependent on individual's n's social network and the choice sets of relevant social contacts

$\mathcal{P}_{(-n)i}$ ≡ the payoff achieved by individuals excluding individual n if alternative i is chosen

f_{ni}^{social} ≡ a function accounting for the impact on a decision of social context on preferences and expectations; this function maps characteristics of an individual, that individual's social network, and the behavior and characteristics of that individual's social contacts to units of payoff (observability is in reference to the modeler)

$d_n(.)$ ≡ the decision rule (or decision process) which maps the payoffs of each alternative as well as the payoff of relevant social contacts to the individual's choice

$m_{(-n)i}$ ≡ measures of social interaction between an individual and others, including social capital, social influence, and social cooperation, that are related to alternative i

ψ ≡ model parameters that weight the contributions of social context to the payoff

First, the preferences and expectations of an individual are modified by f_{ni}^{social} which can occur through: (1) introducing new factors that influence an individual's payoff, (2) changing the strength of influence of individual-level covariates, and (3) changing an individual's individual-level covariates that correspond to their expectations of an alternative's properties or payoffs. Second, an individual's constraints can be modified by social context through adding and removing alternatives from their choice set J_n (e.g. scheduling conflicts, social norms). Lastly, the individual's decision process may be impacted by the goals and payoffs of others thus affecting the individual's decision rule d_n.

All of these impacts are not relevant for each decision involving social interactions. Determining the functional forms for these relations depends on the social context and thus varies for modeling social capital, social influence and social cooperation. The impact on social capital will be explained first[3] in the context of activity and travel.

1.3 How is Social Capital Incorporated into Choice Modeling?

Activity diaries have begun to ask respondents to explain who they perform activities with (Habib et al. 2008, Van den Berg et al. 2010, Sener et al. 2011, Lin and Wang 2014). This has tended to be in the form of using a simple add-on question per activity: "With whom did you participate in this activity with?" The answer choices tend to be simple – e.g. none, family, friends. These efforts have spawned work to create models of social cooperation in the form of activity coordination in activity selection and duration. In activity-based models, work has been performed to explore models where cooperation and coordination of activities occur within the household and sometimes between households (Castiglione et al. 2015). But these measurement and modeling efforts have

[3] Social influence is described in sections 1.4-1.7. Social cooperation is not the focus of this dissertation so it will not be described in the context of choice modeling and it is covered in more depth in de Palma et al (2014).

suffered from a lack of understanding the social factors behind the participation in certain activities.

Incorporating social capital into studies of activity participation has the potential to more fully explain the social aspects of activity and travel. Social capital explains how our social connections bring value to our lives. As Kadushin (2012) explains “social networks have value because they allow access to resources and valued social attributes such as trust, reciprocity, and community values” (p. 164). Activity participation also allows for resource access – e.g. access to food, entertainment, and recreation. But activity participation also allows individuals to expand their social circles and provides an avenue to create trust and for reciprocity. For example, if an individual is stressed out from a long week at work, he may contact his friends to visit a sports bar over the weekend. The group gets a few drinks, watches a few games, and jokes about “old times.” Thus, the activity provides:

1. Access to food and drinks from the venue
2. Access to entertainment from the venue and conversation with friends
3. Social support for the stressed individual from his friends
4. Reciprocity – the individual may now feel obliged to hang with his other friends when they contact him in the future
5. Trust creation – his friends came to his aid and he felt comfortable sharing his feelings with them

Social networks play a vitally important role in creating and fostering social capital in this context. The individual needed to know and be able to communicate with a group of friends and acquaintances. The individual and his contacts had to be located close enough to travel to a bar to share a few hours together. Their interests had to be sufficiently close in allowing them to be interested in watching games together and to foster conversation. Examples like this are why the social network perspective and social network analysis techniques have been applied to studying social capital.

But social capital is notoriously difficult to measure (Lin 2001, Field 2003). The concept of social capital is both powerful and limiting because of its simple yet broad definition. The most popular approach in the social network perspective on social capital is to use measures of an individual’s social network as indicators of social capital (Lin 2001). Considering the previous example, the work-stressed individual may have a social network where most of his friends are of similar age (i.e. age homophily is high). Because they are at similar life stages due to their similar ages, their schedules may more easily align for a weekend bar trip. In contrast, another individual with a network that has high age heterophily may choose another activity location/type or just contact a friend by phone to gain some social support.

Measuring these indicators generally involves using two techniques: name generators / interpreters and position generators (Hennig et al. 2012). A name generator is a technique to understand core social networks or context-sensitive networks by asking individuals to describe who they are connected to in a particular context. A name interpreter may then be used to elicit characteristics about these named contacts. This technique is good for understanding the characteristics of an individual’s social contacts and the configuration characteristics of their networks. The position generator technique is used to measure the diversity and reachability of an individual’s social network and an individual’s resource access. It generally involves asking an individual if they know anyone who possesses a particular job or job type. By listing jobs of varying skill and prestige, the surveyor can gain an indication of the total diversity and reachability to (prestigious) connections. In studies of activity and travel, the name generator is the primary technique for incorporating social capital indicators into activity-travel models (Carrasco et al. 2008, Carrasco and Cid-Aguyao 2012, Kowald and Axhausen 2012, Sadri et al. 2015, Tilahun and Li 2015).

Choice models with social capital indicators expand upon the model shown in equation (1) by incorporating network characteristics as indicators of social capital effects in the following form:

$$\mathcal{P}_{ni} = f_{ni}(x_{in}; \beta_i) + r_{ni}(G_n, x_{in}, x_{i(-n)}; \psi) + \varepsilon_{ni}$$
$$d(\mathcal{P}_{ni}, \forall i \in J) \rightarrow y_n \tag{4}$$

where:

r_n ≡ a social capital function which maps indicators of social capital (e.g. homophily, network size, alter attributes) to units of payoff

G_n ≡ the social network of individual *n* – subscripted here to indicate that often this is egocentrically constructed

$x_{i(-n)}$ ≡ the characteristics of others (individuals or institutions other than individual *n*) who may impact an individual's social capital

ψ_i ≡ model parameters that weight the influences of others on an individual's payoff

In this dissertation, activity participation models with indicators of social capital are compared. Additionally, since no activity-travel studies have used the position generator previously, the role of social network occupational diversity in activity participation was explored. Access to the resources embedded in diverse networks (extensity) was found to positively correlate with leisure activity participation. Using indicators from name generator and position generator data, indicators of social network occupational diversity from the position generator were found to impact activity participation and these indicators are found to have predictive power. As compared to indicators from a name generator, for the activity types analyzed, the position generator indicators were comparable or better at predicting activity participation in a cross-validation study performed in Chapter 3.

1.4 How Does the Measurement of Social Network Indicators Impact Estimation of Social Capital and Social Influence Choice Models?

The name generator and position generator used in studies of social capital are examples of measurement techniques for acquiring social network information. This is not a problem unique to social capital studies as social influence studies also must deal with the incorporation of social networks, the types and timing of interactions, and how social networks and interactions interface in spatial dimensions. These can be difficult to model and identify from current data sources.

Social network data is collected using various questionnaire designs such as name and position generators and sampling techniques such as egocentric and snowball sampling. Additionally social network data collection can be affected by missing data and measurement error due to respondent recall, fixed-recall survey designs, and indirect sources of network data. In spite of this, applied choice models with social interactions often ignore these factors and assume that network data is accurate and complete. Gaining an understanding of the estimator properties for these models under different questionnaire design, sample sizes, and data misspecification can guide methodologist on ways to handle the misspecification, aid applied modelers in understanding potential pitfalls in their analyses, and guide the design of survey questionnaires and sampling procedures. This can aid data collection efforts when the prior intent is to estimate choice models of social interaction to test behavioral theories or make predictions.

In practice, modeling exercises in the travel behavior field tend to be applied work rather than theoretical. Specifically, the application is well ahead of the theory of how the incorporation of data from social networks impact estimation of choice models. For example, multinomial models of social influence are common but very little theoretical econometric or statistical analysis work has been done on the properties of these models[4] (Durlauf and Ioannides 2010). As such, the estimator properties of the applied models typically are not verified before their usage. Using these techniques for explanation and prediction requires that the estimator properties be more thoroughly understood.

In this dissertation, the robustness of social network data collection efforts and indicator usage are explored for both position generators and name generators / interpreters. Chapter 4 explores the sensitivity of binary probit model predictions using a measure of extensity (i.e. social network occupational diversity). In this work, the length of a position generator's occupational list varied to determine its effect on model fit and bias and variability in parameter estimates. In a case study using data from the Pew Internet Personal Networks and Community study (Hampton et al. 2009), extensity as

---

[4] Blume et al. (2011) mention theoretical work including Brock and Durlauf (2002, 2006) and Bayers and Timmins (2007).

indicated by social network occupational diversity was found to be robust to varying occupational list lengths.

Chapters 5 and 6 explore the sensitivity of binary logit model parameter estimation and model selection for models using data from a name generator/interpreter. These exercises are explored in the context of a social influence choice model[5]. This work includes:

- A simulation-based examination of the estimation properties and likelihood ratio test properties of choice models of dynamic social influence with sparse small-world social networks when the network data is misspecified due to random omission and addition of social ties
- A simulation-based examination of the estimation properties and likelihood ratio test properties of choice models of dynamic social influence with sparse small-world social network when egocentric sampling techniques are used

These simulation studies found that for small-world networks, the network shape had no impact on the quality of parameter estimation when network density and network size are held constant. Misspecification due to social tie omission and addition was found to degrade model estimation after approximately 15% to 30% of the edges have been modified. For egocentric sampling, the most important factors included network density and sample size and the strength of social influence.

[5] Although the models are dynamic models of social influence, since there are no endogenous covariates, the results are generalizable to models with contextual social influence as well as models using indicators of social capital that use alter attributes via proportions.

1.5 How Can Social Influence be Relevant in Decision Making?

Social influence is the process of altering an individual's decision making process through the actions, behaviors, attitudes, and beliefs of others as well as through an individual's perceptions of these. To clarify the concept in modeling, this section presents a hypothetical, illustrative example of various sources of influence in travel behavior.

Suppose a researcher studying cycling behavior among students and non-students makes the following observation:

College students in the US are more likely to use a bicycle than non-students.

This simple observation could have various causes. The following are several possible explanations for this observation (observability is in reference to the modeler):

1. College students tend to live on college campuses which often have amenities that are nearby. Therefore, more student trips are within the comfortable range for bike travel compared to non-student trips. Individual-level differences in travel distance and trip time (Dickinson et al. 2003) may explain differences in cycling behavior between students and non-students. These variables are typically observable to modelers. [**Observed individual-level effects**]
2. Cycling decisions depend on the choices of others because of social norms and conformity (Dill and Voros 2007). This can cause a self-perpetuating cycle of low cycling rates in neighborhoods with non-students and high cycling rates in neighborhoods with students. For example, this can lead to a situation whereby once a few people start cycling, a critical mass is reached, and cycling becomes more popular. [**Endogenous social influence effects – Conformity**]

3. Preferences for automobiles may be higher among lower income individuals compared to higher income individuals (Parkin et al. 2008). Higher income individuals have higher bicycle ownership and tend to cycle more often than lower income individuals. Since college enrollment in the US tends to increase with rising household income (Snyders and Dillow 2013), this difference in preferences may induce students to perceive cycling more favorably due to the social norms of different income groups – and perhaps more favorably than would be expected by income alone. [**Contextual social influence effects – Compliance**]
4. Infrastructure details are often not available in large scale travel datasets. If college campuses have more favorable bicycle infrastructure (e.g. bike paths, bike lanes, bikesharing programs, bicycle parking) than areas that are not near college campuses, this may lead to higher cycling rates among students. Here, an institutional environment may cause an increase in student cycling rates. [**Correlated environmental effects**]
5. Since cycling is a physical activity, a certain level of physical ability and health is needed to cycle. College students in the US tend to be less obese than non-students (Fowler-Brown et al. 2010) and since obesity correlates with health, this could explain a disparity in cycling rates. Since travel surveys tend to not measure health and ability, this may be an example of an unobservable effect which acts at the individual level. [**Correlated individual-level effects**]
6. Schools may create a stronger sense of community than an average community so the strong cohesiveness of the social networks among students may allow quicker,

stronger, and self-reinforcing dissemination of cycling behavior (Páez and Whalen 2010) as compared to the less cohesive networks in communities outside of schools. [**Social network structure**]

Each of these possible explanations requires a different policy intervention. For example, explanation #1 suggests that increasing the amenities in less dense areas would increase cycling rates, whereas explanation #2 suggests that investments in encouraging a few people to cycle (e.g. advertising campaign, bicycle loan program) would be more effective. Explanation #6 suggests that less resources likely need to be spent to encourage cycling in close-knit communities as compared to less cohesive communities.

1.6 How is Social Influence Incorporated into Choice Modeling?

In the field of travel behavior analysis, there has been interest lately in travel decision making involving social interactions, particularly social influence. Social influence has been identified as a possible factor in various travel decisions including mode choice (e.g. Dugundji and Walker 2005), cycling behavior (e.g. Sherwin et al. 2014), telecommuting (e.g. Wilton et al. 2011), vehicle ownership (e.g. Grinblatt et al. 2008), electric vehicle adoption (e.g. Axsen and Kurani 2012), pedestrian safety (e.g. Gaker et al. 2010), drunk driving (e.g. Kim and Kim 2012), and tourism (Wu et al. 2013). Choice models of social influence expand upon the model shown in equation (1) by incorporating the actions, behaviors, attitudes, and beliefs of other individuals or institutions into an individual's payoff for making a choice:

$$\mathcal{P}_{ni} = f_{ni}(x_{in};\beta_i) + s_{ni}\left(G_n, m_{-n}, x_{i(-n)};\psi\right) + e_{ni}(E_n;\mu_i) + \varepsilon_{ni} \qquad (5)$$

$$d(\mathcal{P}_{ni}, \forall i \in J) \rightarrow y_n$$

where:

s_{ni} ≡ a function which maps factors of social influence, including endogenous and contextual social influence, to units of payoff

m_{-n} ≡ the social influence indicators; actions, behaviors, attitudes, and beliefs of others (individuals or institutions other than individual *n*) that influence an individual's decision process

$x_{i(-n)}$ ≡ additional social influence indicators from the characteristics of others (individuals or institutions other than individual *n*) who influence an individual's decision process

E_n ≡ environmental factors on individual *n* (may include correlated environmental factors)

ψ_i, μ_i ≡ model parameters that weight the influences of others and environmental factors, respectively, on an individual's payoff

Travel may involve different types of social influence processes with varying motivations and generated from different types of social networks including peers, family, neighbors, colleagues, and even society at large. Incorporating these social effects into discrete choice models – which are grounded in the individual choice of independent decision makers – is non-trivial, thus making the choice of the functional form of s_{ni} important. The choice of appropriate social influence mechanisms and the associated social networks and influence sources should not be taken lightly.

This has spawned a variety of different model specifications using different social network structures and social influence processes and motivations. But there is not much guidance in the choice of model structure for social influence choice models. Additionally, the field has a variety of terminologies to describe similar models which makes comparing and contrasting research difficult. Also, understanding how social factors are incorporating into choice modeling can be a confusing and daunting task for new users. This dissertation proposes a theoretical behavioral framework to classify the various formulations of social influence choice models

Social influence choice models incorporate theories and terminology from different social science fields. Additionally, various model specifications using differing social network specifications, influence sources, and social influence types and processes have been developed. For example, network structures vary from cliques to sparse networks and the connections made can be due to similarity in social standing and interests as well as geographic proximity.

This dissertation proposes a system to describe previously built models of social influence in choice modeling as well as provide a flexible enough framework to allow new models to be described sufficiently as well. The behavioral framework developed in Chapter 7 consolidated the current state-of-the-art in a clear format. This will aid in clarifying areas for improvement and future research topics.

The framework (Figure 1) provides a behavioral basis for social influence choice models. Previous work defined endogenous and contextual effects only “in terms of [the] types of variables rather than via particular mechanisms” (Blume et al. 2011, p. 941). The framework contrasts with previous works which classified on structural terms (Manski

1993; Brock and Durlauf 2001, 2003) by emphasizing behavioral microfoundations. The framework separates the social influence mechanism from the source of its influence and explicitly acknowledges the role of social networks in the model structure.

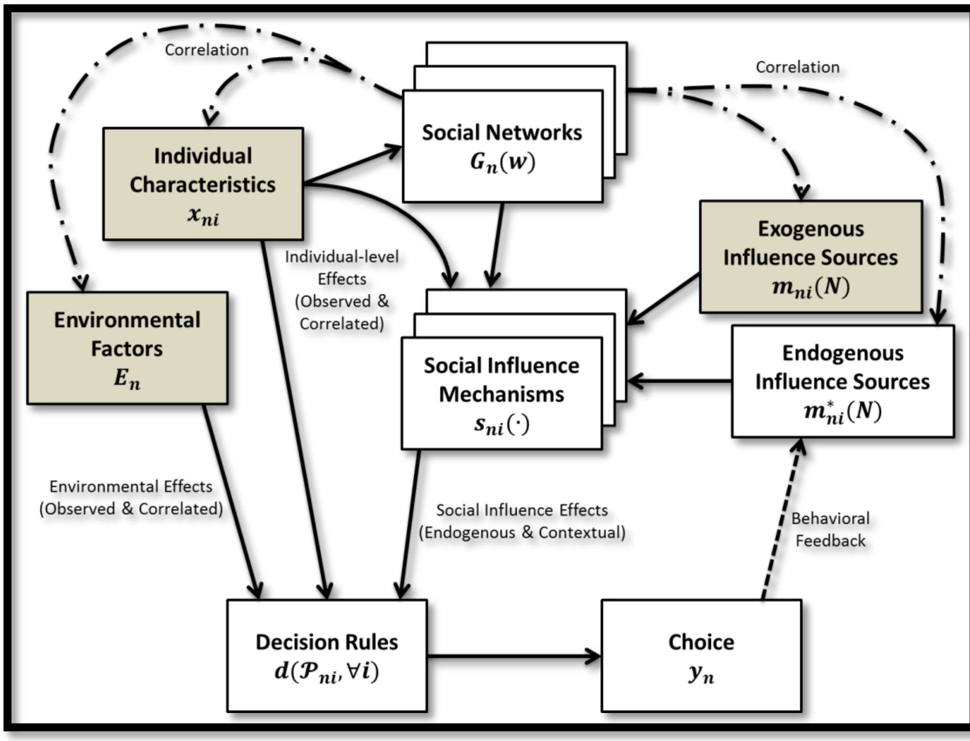

Figure 1. Generalized Framework for Choice Models of Social Influence

Using the behavioral framework as a guide, a review of social influence models of travel behavior was performed. This review showed that most models are conformity models with utility maximizing agents. This form assumes a direct-benefit effect is generated from conforming to the behavior of others (i.e. utility itself is directly increased by conforming). These models are not formulated to understand the motivations of social influence. The direct-benefit conformity formulation can be generated by different motivations; the question of why are people conforming often is not being answered. Are individuals transferring information? Are people envious of others and aspiring to obtain a similar status? Is this just a fad and people are just following the crowd? These motivations are important for understanding long-run behavior and for guiding organizations on appropriate intervention strategies to encourage behavioral change. For example, Cialdini and Goldstein (2004) describe recent research in social influence through the motivations for *accuracy*, *affiliation*, and *maintenance of a positive self-concept*. Current models either do not acknowledge the motivations for social influence or use formulations that emphasize affiliation and maintaining a positive self-concept.

Direct-benefit conformity model specifications are often relevant for behavior where imitating others provides direct benefits such as in popularity and status seeking. In contrast, if the conformity is indirect and informational, then perhaps the individual's choice set should change to include this new option or the attributes of the new alternatives should increase in attractiveness. Informational conformity occurs when individuals feel uncertain about a decision and conform to the behavior of more knowledgeable others. This is an example of a motivation for accuracy.

1.7 How Can the Motivation for Accuracy Be Incorporated into Social Influence Choice Models?

In Chapter 8, a new discrete choice model formulation is shown that works on the social influence motivation of accuracy. This informational conformity model uses a latent class discrete choice model framework where individuals are placed into different classes in accordance to their latent information on the relevant decision making topic area. Individuals in different classes have varying preferences in accordance with their informational affinities. Social influence occurs indirectly as individuals' information levels are impacted by the choices of others, but their own choice level utilities are not directly impacted by others.

The informational conformity model contrasts with the direct-benefit social influence model in equation (5) by incorporating the actions, behaviors, attitudes, and beliefs of other individuals or institutions into an individual's information acquisition function which determines their corresponding preferences in their class' choice model:

$$
\begin{gathered}
\mathcal{F}_{nc} = h_{ni}(x_{ic}; \alpha_{ic}) + s_{ni}^{\mathcal{F}}(m_{-n}, x_{i(-n)}; \psi) + \varepsilon_{nc}^{\mathcal{F}} \\
d^{\mathcal{F}}(\mathcal{F}_{nc}, \forall cJ) \rightarrow c_n \\
\mathcal{P}_{ni}^{[c]} = f_{ni}\left(x_{in}; \beta_i^{[c]}\right) + \varepsilon_{ni}^{[c]}, \qquad \forall c \\
d\left(\mathcal{P}_{ni}^{[c]}, \forall i \in J\right) \rightarrow y_n
\end{gathered}
\tag{6}
$$

where:

$\mathcal{F}_{nc}$ $\equiv$ the latent information level achieved by individual n if he is in class c

h_{nc} $\equiv$ a function which maps observable individual-level characteristics to latent units of information

$s_{ni}^{\mathcal{F}}$ $\equiv$ a function which maps factors of social influence (e.g. the actions of others) to latent units of information

c_n $\equiv$ individual n's latent information class

$d^{\mathcal{F}}$ $\equiv$ the decision rule (or decision process) which maps the latent information of each class to a latent choice of class

$.^{[c]}$ $\equiv$ denotes that this term is specific to information class c

$\varepsilon_{nc}^{\mathcal{F}}$ $\equiv$ unobserved effects on individual n for information class c

Chapter 9 describes a case study where social influence in bicycle ownership was studied using this informational conformity model. Information was signaled by city-level bicycle usage where greater usage may induce households to reevaluate their preferences towards bicycle ownership. Estimation results showed that these “more informed” households have a higher probability of owning a bike due to being less sensitive to smaller home footprints and limited incomes.

1.8 Dissertation Organization

This dissertation is broken up into three parts: social capital and social network indicators, social network data representation and estimation behavior, and behavioral motivations for social influence model formulation. Part I, which consists of Chapter 2 and Chapter 3, summarizes data collection techniques and presents a case study on the use of position generators versus name generators for activity participation modeling. The chapters of part I are briefly summarized below:

- **Chapter 2** describes theories of social network formation and structural properties of social networks. Then, this chapter summarizes the data collection techniques used in social interactions studies of travel: choice-data, qualitative data, surveys and experiments, and stochastic network models.
- **Chapter 3** compares the explanatory and predictive performance of social capital indicators in activity participation modeling. Binary probit models were analyzed and it was found that extensity / social network occupational diversity as indicated from position generator questions had significant predictive and explanatory power.

Part II presents simulation studies of the estimation properties of a common social influence choice model formulation and the effect of data collection via a name generator / interpreters. The chapters of part II are briefly summarized below:

- **Chapter 4** analyzes the sensitivity of the social network occupational diversity indicator used in Chapter 3. A sensitivity analysis of model fit and parameter estimates was performed by varying the size of the position generator used.
- **Chapter 5** details a simulation study design to analyze the effect of misspecified social networks on social influence choice model estimation. A simulation study of random additions and omissions of social ties was performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks.
- **Chapter 6** details a simulation study design to analyze some finite-sample properties of social influence choice model estimation. A simulation study of egocentric sampling was performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks.

Part III, which consists of chapters 7 through 9, explores and expands upon the behavioral underpinnings of social influence choice models. The chapters of part III are briefly summarized below:

- **Chapter 7** showcases a generalized behavioral framework for choice models of social influence. This framework stresses the important interconnection between the social influence mechanism and associated social networks and influence sources. The current state-of-practice in travel behavior studies of social influence is described as well as the limitations of current approaches. Then theories of social influence via conformity and compliance are described and these are briefly linked with the importance of social networks.

- **Chapter 8** describes a formulation of a choice model of informational conformity which uses a latent class structure. The class membership model depends on the proportion of group members exhibiting a particular behavior. Membership into the "more informed" class will cause a change in the preferences of those individuals, thus making the behavior more attractive. These "more informed" individuals are motivated to conform due to the goal of accuracy. Equilibrium properties are also derived and a Bayesian inference and policy analysis technique is described. Additionally, a technique to handle endogeneity is described by using a two-stage control function approach.
- **Chapter 9** provides a case study using the informational conformity model and the Bayesian estimation and equilibrium analysis technique described in Chapter 7. Bicycle ownership in the United States is analyzed. In contrast to existing work, this model showed that "more informed" households have a higher probability of owning a bike due to changes in preferences rather than direct benefits from others' behaviors – they were less sensitive to smaller home footprints and home ownership and more sensitive to household membership size.

This dissertation concludes with a summary of contributions and areas for future research in **Chapter 10**. Additionally, six appendices are provided:

- **Appendix A** complements work in chapters 2, 5, and 6. It describes how exponential-family random graph models (ERGMs) are formulated.

- **Appendix B** complements work in chapters 7 through 9. This appendix summarizes the specifications of different models combining choice and social influence in the areas of social influence network theory, social network analysis, statistical mechanics and social econometrics, spatial econometrics, experimental economics and game theory, the economics of identity, and travel behavior modeling.
- **Appendix C** describes how endogeneity is handled in social influence discrete choice models. This appendix supplements work in Chapter 8 and Chapter 9.
- **Appendix D** presents a generalized version of the informational conformity model described in Chapter 8. It generalizes the model by including expectation and constraint changes in addition to preference differences between different classes of informed individuals.
- **Appendix E** presents an ordered logit formulation of the bicycle ownership case study in Chapter 9. This formulation shows that the informational conformity model can be generalized to ordered choice as well as to provide a social influence explanation for the quantity of bicycles owned in a household.
- **Appendix F** presents an example formulation to complement the future research directions described in Chapter 7. A formulation of a social influence choice model with heterogeneous social influence processes is described. The choice model uses a latent class structure to allow for heterogeneity in social influence processes among different individuals in the population.
- **Appendix G** supports the model selection measures shown in Table 31 of Chapter 9.

1.9 Contributions

This dissertation makes practical and theoretical contributions to the study of social interactions in activity and travel behavior. The major contributions of this dissertation include:

1. The use of data from a position generator in a social capital model of activity selection in Chapter 3
2. The expansion of empirical evidence in favor of the robustness of position generators as an indicator of social capital via extensity of resource accessibility (i.e. social network occupational diversity) in Chapter 4
3. The testing of the implications of social network structure, social network misspecification, and finite-samples on choice models of social influence in Chapters 5 and 6
4. The creation of a generalized behavioral framework of social influence choice models in Chapter 7
5. The classification of existing research on social influence in travel according to social network, social influence mechanisms, and influence measures in Chapter 7
6. The development of a discrete choice model of informational conformity and inference and hypothesis testing procedures in Chapter 8
7. The use of an informational conformity model in an exploratory analysis of social influence in bicycle ownership in Chapter 9

Other minor contributions are mentioned in the relevant chapters.

Chapter 2: Social Networks and Data Collection in Social Interaction Studies of Travel

Table 1. Chapter 2 Summary

Background & Brief Summary	This chapter provides an overview of social networks. It begins with a look at the link generation processes and network structures. Then the data collection techniques used in social interactions studies of travel are summarized, including choice-data, qualitative data, surveys and experiments, and stochastic network models.
Motivation	Social networks link the three major areas of social interactions in activity and travel research: social capital, social influence, and social cooperation. Data collection has important implications in the robustness and design of models to represent and study the impacts of social interactions on activity and travel behavior.
Results	1. Social network theories were described in the context of link formation and network structures 2. The existing data collection technique for obtaining social interaction and social network data from the travel behavior literature is summarized 3. Strengths and weakness to the various data collection techniques are suggested
Limitations	1. More extensive reviews of social networks and data collection techniques are available in resources such as Kadushin (2012), Hennig et al. (2012), Prell (2012), and Borgatti et al. (2013)

This chapter presents an overview of the social network formation processes and structures as well as the data collections used to measure social network and thus perform analyses where social network are relevant.

2.1 Social Networks: Processes and Structures

When a modeler is thinking about appropriate social networks for their analysis, it is critical to understand: (1) why connections are made and (2) what kind of network structures are appropriate/likely?

2.1.1 Link Generation Process

The question of why connections are made is critical for understanding the importance of social networks and their effect on social interactions. Kadushin (2012) summarizes research showing that the three major motivational foundations of social networks are social safety, effectance, and status.

Social safety is important in nourishing a sense of community, affiliation, and trust (Kadushin 2012). Social networks typically provide this safety by linking individuals according to spatial proximity[6] (i.e. propinquity) and homophily. Spatial proximity describes the increased likelihood of interacting with individuals who are located close to you spatially, while homophily describes how individuals tend to associate with others who are like themselves.

Spatial proximity is a common basis for generating social networks in social interaction travel models due to the ease of measuring spatial attributes (Dugundji and Walker 2005). The open question remains of how to determine what level of spatial aggregation is appropriate for different types of social interactions. While these spatial units may be appropriate for simple transportation planning purposes, physical distance has varying implications between different topologies and build environments, near-distances and very far-distances, and even between different individuals (Robins and Daraganova 2013). Kowald et al. (2013) noted a tendency for individuals to have a majority of contacts within about an hour drive, while Matous et al. (2013) similarly

[6] Kadushin (2012) refers to spatial proximity as *propinquity*. The more familiar term of spatial proximity is used in this dissertation because only proximity in spatial categories is considered. The work in this dissertation and on social networks is generalizable to other forms of propinquity such as virtual propinquity (e.g. instant messaging, video conferencing), occupational propinquity, and acquaintance propinquity.

found that, in an infrastructure-poor region, individuals had 95 percent of their contacts within a 90-minute walk. Although spatial proximity is common in social networks, it is not guaranteed that individuals connect with their physical neighbors. Indeed, some research finds that some neighborhoods are more cohesive than others and that even the transportation network can influence this cohesion (Grannis 1998, Whalen et al. 2012).

Kadushin (2012) summarizes Verbrugge (1977) by defining homophily as "if two people have characteristics that match in a proportion greater than expected in the population from which they are drawn or network of which they are a part, then they are more likely to be connected" (p.18). The importance of this has been briefly mentioned in the transportation literature. For example, in social influence studies, the general pattern is to choose socioeconomic indicators and place individuals into groups based on these categories. In a social capital studies by Sadri et al. (2015), homophily is defined using the EI index (Krackhardt and Stern 1988) and applied to the characteristics such as marital status, age, and vehicle ownership. A contrasting approach involves quantitatively combining different aspects of an individual's socioeconomics into a measure of social similarity such as Blau space (Blau 1977, Hennig et al. 2012, Kadushin 2012) or social distance (Akerlof 1997).

Effectance contrasts with social safety by motivating individuals "to reach out and make connections where there were none" (Kadushin 2012, p.56), thus promoting the brokerage of different social groups and circles. Effectance facilitates the human desire to explore the unknown. It aids in transferring knowledge, influence, and social capital between different parts of society and can give the individuals who link these parts power and status. Axsen and Kurani (2012) and Axsen et al. (2013) mention the effect of brokers and the connections between social groups in transferring influence in the adoption of electric vehicle technology. For models emphasizing the diffusion of behavior through sparse networks, these connections are critically important to understand due to drastic changes in diffusion patterns and in the design of effective behavioral interventions. Social capital studies in sociology explore this through position generators and resource generator questions (Lin 2001).

Lastly, status entails a ranking of the power and prestige of individuals and comparisons thereof. Status can be created by organizational structures (e.g. job roles at work) and the allocation of resources (e.g. money, authority, social connections). This can encourage social interactions where individuals attempt to status seek – whether consciously or subconsciously – in order to maintain their status or seek higher status. For example, Wilton et al. (2011) mentions that, in semi-structured interviews, some employees expressed reservations about teleworking due to negative perceptions among their supervisors.

2.1.2 Network Structure

The network structure is critically impacted by the link generation process and the form of social influence. From this structure, long-run impacts of social influence are affected. Social safety, effectance, and status seeking – the primary motivations for network formation – lead to the network structural properties of dense networks, structural holes and weak ties, and pyramid/hierarchical structures, respectively (Kadushin 2012). These properties parallel some common network types that are used in research including cliques[7], small-world networks, and hierarchical networks. This section will briefly describe these network structures and concludes with a look at future development in spatial-social network overlays and two-mode networks.

A clique is a maximally dense section of a network where all individuals in the clique are connected to each other. When social networks are assumed to be reflexive large cliques, conformity models are commonly called field-effect or mean-effect models. Cliques are a good representation of small groups where it is easier to communicate with and observe the behavior of all group members. But this assumption becomes less behaviorally plausible as social group size increases since the individual is unlikely to know each person in his reference group and coordinating actions would be more difficult[8]. On the other hand, larger group sizes allow for estimates of choice percentages that are more robust to the influence of any one particular individual. Therefore, care must be undertaken when using clique structures, and modelers need to be clear about their motivations for and the limitations of using this structure.

The existence of small-worlds in human social networks is attributed to the small-world experiment (Milgram 1967) which led to the "six degrees of separation" concept. Small-world networks are sparse networks that exhibit high clustering and short average path lengths. Thus, individuals tend to form relationships such that (1) an individual's friends tend to be friends with each other but (2) "social network [also] tend to have very

[7] The nearest neighbor networks used in Goetzke (2008), Grinblatt et al (2008), and Adjeman et al (2010) are a similar conception but non-reflexive. This technique is a non-parametric technique that also creates dense networks with spatial proximity-driven link generation.

[8] An anonymous referee mentions the difficulty of coordinating and signaling average mode shares in large groups (see work by Brewer and Hensher (2000) and Murdoch et al. (2003)).

short paths between essentially arbitrary pairs of people" (Easley and Kleinberg 2010, p.32). Small-world networks are commonly viewed as due to assortative mixing (Newman and Park 2003) or preferential attachment. In assortative mixing, individuals with many social connections are attracted to other highly-connected individuals. In preferential attachment, these highly-connected individuals are not more attracted to one another, but tend to connect to low-degree nodes in the network. This is a difficulty with using small-world networks; they are sufficiently broad that researchers do not always understand which process formed them.

Hierarchical structures are generally directed social networks where influence flows from those with higher status or power to individuals with less. These commonly come in the form of status, role, or authority networks such as the example of a workplace network in Figure 2. This directed nature of the influence contrasts with the clique and small-world structures mentioned before and has implications in studies of families, workplaces, small communities, and other organizations. With richer data sources and more research on social interactions of small groups, this network structure will be used more often in travel studies.

Spatial-social network overlays refer to combining spatial features and social networks to realize the impact of geospatial factors on the structure of social networks. For example, in the spatial-social network shown in Figure 2, there are few connections across the river due to the bridge's impact on travel and physical contact. Although some individuals are directly across the river from one another, they make contact with other individuals who are farther by Euclidean distance but located on the same river bank. The implications of the formation of structural holes and weak ties in social networks –

possibly leading to small-world networks (Wong et al. 2006) –needs to be more thoroughly understood in the context of travel studies with social network data.

“A two-mode network [or bipartite network] consists of two sets of distinct units (e.g. people and events), and the relations that are measured between the two sets, e.g. participation of people in events” (Hennig et al. 2012, p.50). This could be relevant for situations where social interactions are not coming directly through direct contact between individuals but by shared events, perceptions, or influence sources. Sun et al. (2013) provides an example in which transit smart card data is used to create networks of individuals linked by the sharing of transit spaces during trips. Another example includes works deriving from the social identity perspective (Tajfel 1978, Tajfel and Turner 1979, Turner et al. 1987) where individuals in the same social category may share some ideal. This ideal type connects the individuals’ behavior by serving as a prototype of expected behavior for group members (Akerlof and Kranton 2000).

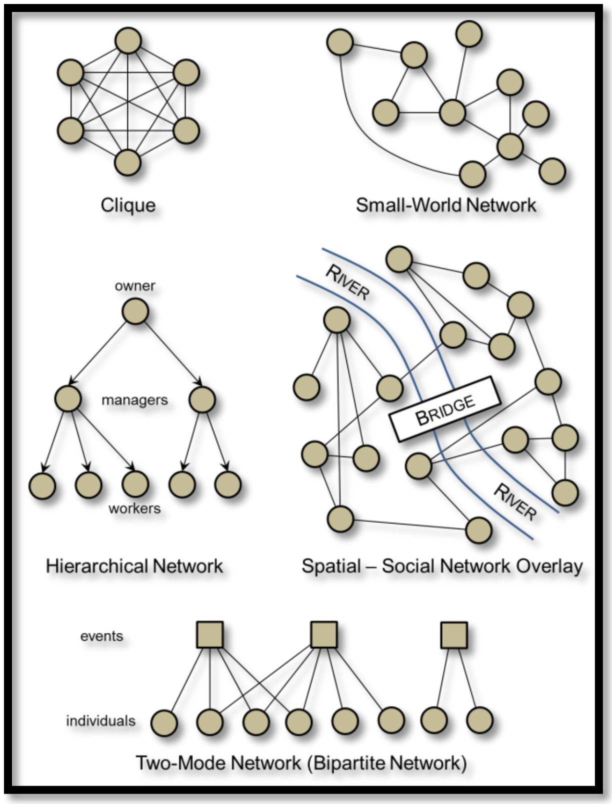

Figure 2. Examples of Network Structures

2.2 Data Collection in Social Interaction Studies of Travel

With a clearer idea of social network structures likely, modelers are faced with the task of determining the methods of social inteactions and the social network connections for their specific application. Kadushin (2012) notes the lack of "large-scale true social-interaction-network data" as a common problem across many fields. Travel behavior research is not immune from this issue as there have been limited studies collecting social network data linked with travel data (Kowald et al. 2013). Section 2.2.1 begins by noting the limitations of choice-data approaches. Section 2.2.2 explains how qualitative data can be used to guide model formation, and then section 2.2.3 looks at direct survey and experimental approaches for collection social data. Section 2.2.4 discusses stochastic network models which can be used for exploratory research or when survey methods are costly, difficult, or prohibited.

2.2.1 Choice-data Approaches

Modelers face the major problem that identifying group membership from choice-data only is difficult. For models with large cliques, most modelers take group membership and network structure as given and support their decisions often based on convenience rather than evidence. Walker et al. (2011) note concerns with their spatial group definitions due to data limitations and issues with the modifiable area unit problem and sharp spatial boundaries (Páez and Scott 2004).

Inferring group membership from data is a possible answer but modelers must be cautious with their conclusions. For example, Goetzke's (2008) study of transit mode choice limited social networks to the closest 40 neighbors, stating that increasing the network size would not significantly impact average mode share. Manski (1993, 2007) shows for linear-in-means models that using individual-level characteristics to determine group memberships – i.e. $g_n(w)$ is functionally dependent on x_{ni} – will always be "consistent with observed behaviour." This likely extends to discrete choice models but has not been clearly analyzed (Brock and Durlauf 2001).

Nonetheless, group membership has been inferred in applied work such as Walker and Li (2007) and Chen (2012) who applied latent class models to identify lifestyle groups in discrete choice decisions. Dugundji and Walker (2005) and Sidhartan et al. (2011) used goodness-of-fit measures such as log-likelihood ratio tests and non-nested tests to test various hypothesized network structures. More research is needed on goodness-of-fit measures and new data is needed with qualitative social interactions data and explicit networks and group memberships.

2.2.2 Qualitative Approaches

Qualitative study can provide guidance for modeling efforts but has seen limited use in transportation. Clifton and Handy (2003) suggest the use of interviews and focus groups in travel behavior research. Additionally, Akerlof and Kranton (2002, 2010) recommend the use of ethnographies for economic models involving group definitions and expectations of group behavior. Abdelal et al. (2009) classifies the most common techniques for measuring identity in social science studies as: surveys and interviews, content analysis, discourse analysis and ethnography, cognitive mapping, and experiments. Specific examples in transportation include:

- Axsen and Kurani (2012) identify contagion, conformity, and dissemination as possible sources of influence in electric vehicle purchasing decisions.
- Wilton et al. (2011) use semi-structured interviews to observe social influence effects from co-workers in telework decisions.
- Lovejoy and Handy (2011) study how social context affects carpooling among immigrants while Mote and Whitestone (2011) study informal commuting (slugging).
- Bartle et al. (2013) study social influence in cycling commuting through the interaction of cyclists on an online social networking site, questionnaires, and semi-structured interviews.

Qualitative methods can be used to increase the credibility of model assumptions on appropriate group memberships, group salience, and expectations of others' behavior. For example, Sherwin et al. (2014) used semi-structured interviews and thematic analysis to analyze cycling behavior in the UK. Their research found that individuals experienced direct social influence from family, friends, co-workers, and government programs. Additionally, individuals also experienced indirect social influence from seeing strangers cycle, varying cycling culture between towns, and gender norms. From this qualitative work, a modeler would have a clearer idea of the relevant influence mechanisms and social network structure for model development. Then, the modeler could use their quantitative results to determine the strength and significance of the social influence.

2.2.3 Surveys and Experiments

It is still rare for travel surveys to cover issues related to social context. Group memberships are not measured in travel surveys but may be pertinent in the implementation of social influence studies[9]. Models with sparse networks require information on individuals' social contacts in order to create valid weighting matrices. Axhausen (2008) suggests name generators for obtaining lists of contacts, including family members, friends, co-workers, and others, but cautions that name generator questions can increase respondent burden and may suffer from low response rates.

Sampling techniques for social network research in transportation falls into three broad groups:

---

[9] Woittiez and Kapteyn (1998) provide an example from labor economics.

- *Egocentric*. Egocentric sampling consists of obtaining a random sample of individuals (“egos”) then obtaining information on their direct contacts (“alters”). This has been the primary data collection technique in transportation (Carrasco et al. 2008, Carrasco and Cid-Aguayo 2012, Frei and Axhausen 2007, Larsen et al. 2008, Van den Berg et al. 2008, Scott et al. 2012).
- *Snowball Sampling*. Snowball sampling builds on egocentric sampling by proceeding to collect data directly from the alters of the initial random sample of egos. This allows for analysis of indirect contacts and the structure of networks, such as analysis of personal leisure networks (Kowald and Axhausen 2012, 2014).
- *Census*. In a census, the connections of all individuals in a network are observed. This is a rather difficult task for large populations and when population boundaries are difficult to determine. This technique has strengths in small groups and institutions or when the collection of social contacts is easily logged (e.g. social networking sites, smartphone applications).

Since it is often difficult even with explicit questions about networks to pinpoint social interaction effects, experiments which control for non-social factors are an avenue to determine whether social interactions are prevalent in travel behavior (Sunitiyoso et al. 2011). Gaker et al. (2010) explores social influence effects in automobile ownership via an information cascade experiment in which they hypothesize that “social influence in the form of an information cascade will affect whether a person buys a conventional car, buys a hybrid car, or forgoes having a car.” Subjects who were shown the prior choices of other respondents were more likely to pick the most chosen option.

2.2.4 Stochastic Network Models

Due to the extensive collection efforts required to collect large-scale social network data and privacy and ethics concerns, modelers may use random network models in simulation and agent-based models in order to create realistic imitations of real-world social networks. In Arentze et al. (2012), the authors used common concepts from social network analysis – homophily, spatial proximity, and transitivity – to create a static, stochastic, actor-based model of network formation. Dugundji and Gulyas (2003) looked at Erdős–Rényi and small-world network models to analyze the equilibrium behavior of utility maximizing agents. Observing the patterns of emergent behavior can guide future research study design to optimize resource allocations for new social interactions studies. If the structural properties of the network are only needed, modelers may use random graph models, such as Erdős–Rényi (Erdős and Rényi 1960), Barabási–Albert (Barabási and Albert 1999), and Watts–Strogatz (Watts and Strogatz 1998) models, to generate expected graph structures. Otherwise, if information on user attributes exists, modelers may use game-theoretic network models (Jackson 2010) and exponential-family random graph models (ERGMs) (Lusher et al. 2012).

2.3 Areas for Future Research

In particular, new focus can be applied to:

- Panel data collection of behavior and social networks over time will allow researchers to more accurately identify the existence of social influence effects by controlling for correlations between social networks and influence sources

- Applying random network models for policy analysis to deal with issues of privacy and ethics in social network data collection
- Exploratory work to find methods of collecting and representing social network data, such as network indicators and new influence sources that affect social capital and social influence aside from the choices of others such as attitudes, perceptions, past experiences, ideal types, and the salience of social identities

Social interactions in travel is a thriving research area in the travel behavior community, but careful consideration of the limitations of current models and data are warranted. These concerns may limit the application of these methods by institutions and policy makers, so the field must mature in the strength and accuracy of its claims with appropriate data and models with predictive capabilities. Chapter 3 presents a case study using egocentric network data to study activity participation behavior to compare the performance of different network question types. Chapters 4 through 6 will cover how choice models of social capital and social influence are affected by the design of the data collection process, including using misspecified network indicators from position generators and name generators as well as sampling technique.

Chapter 3: Comparison of Position Generators and Name Generators as Social Capital Indicators in Modeling Activity Selection

Table 2. Chapter 3 Summary

Background & Brief Summary	Existing research on social capital and activity behavior has generally used name generators to analyze core networks to understand more intimate connections. In sociology, the position generator has been used to analyze structural properties of networks and resource access. Since no activity-travel studies have used the position generator previously, a case study using the Pew Internet Personal Networks and Community survey was performed to explore the role of social network occupational diversity in activity participation. Also, the name generator and position generator indicators are compared for explaining and predicting activity participation.
Motivation	Existing research incorporating social networks and social capital into activity and travel behavior models has tended to use name generator data. But using name generators and interpreters in surveys increase respondent burden and survey length. The position generator is a technique used in the social science to study social capital and allows for measuring access to networked resources via occupational diversity of network contacts. The influence of these measures is not known in the context of activity and travel. Additionally, the dataset provides an opportunity to compare the performance of a name generator and position generator in an activity context.
Results	1. Social capital was correlated with activity participation for all of the activity types analyzed 2. Network diversity as measured from the position generator was found to be a reasonable explanatory covariate and predictive variable for activity participation 3. The network diversity indicator was found to hold more explanatory and predictive power than core network indicators from a name generator.
Limitations	1. The dataset was not designed specifically for understanding the activity process as thoroughly as focused activity diaries 2. The results cannot generate theoretical insight on its own, but over time it could become part of a portfolio of work showing practical proof of benefits from position generators in activity behavior research

In the activity-travel perspective, travel is a derived demand due to activities (Ortuźar and Willumsen 2011). Individuals connect the activities in their lives by travel because activities bring value to people's lives. This is because activities "satisfy a particular need or requirement" (Ortuźar and Willumsen 2011, p. 473). The varying requirements and needs of activities may serve to bring individuals together. Work activities bring colleagues together to collectively accomplish tasks. Leisure travel often connects individuals who are friends, family, and acquaintances to share experiences and connect socially. But some activities are often done alone, such as routine shopping and fast-food eating out.

This mixture of social networks and activity generation motivates a social capital perspective. For example, Carrasco and Cid-Aguyao (2012) explicitly mention the importance of social capital and connect social network analysis with this by emphasize the importance of "network capital" in travel behavior. Social capital is based on the premise that social networks bring value and investments in establishing and maintaining social contacts can lead to individual returns such as comfort, support, and resources (Kadushin 2012). As Lin (2001) describes:

> *Therefore, social capital can be defined as resources embedded in a social structure which are accessed and/or mobilized in purposive actions. By this definition, the notion of social capital contains three ingredients: resources embedded in a social structure; accessibility to such social resources by individuals; and use or mobilization of such social resources by individuals in purposive actions. Thus conceived, social capital contains three elements intersecting structure and action: the structural (embeddedness), opportunity (accessibility) and action-oriented (use) aspects.* (p. 12)

The strength of the social capital perspective is its generality, but this is also its curse. The three most prominent views of social capital come from Bourdieu, Coleman, and

Putnam (Field 2003). In this chapter, Lin's (2001) focus on social embeddness and measurement techniques will be emphasized. In the social sciences, the two most common approaches to measuring individual-level social capital are name generators and position generators[10] (Kadushin 2012, Lin et al. 2001).

The name generator confronts a respondent "with a specific relation and [asks] 'with whom' he or she is related in this particular way" (Hennig et al. 2012, p. 85). Lin et al (2001) explains that the name generator is often used as an indicator of social capital in one of three ways:

1. Network configuration characteristics that may indicate density, structural holes
2. Alter characteristics to indicate access to resources
3. Alter characteristics to indicate the best resources available to an individual

Name generators can be interpersonal or global. Interpersonal name generators use a context or stimulus to probe respondents about specific social contacts – these contacts are explicitly identified uniquely. An example of an interpersonal name generator is the question: "Who are the people with whom you discussed important personal matters?" (Burt 1984). In contrast, global name generators "avoid the identification of [alters]" and questions are asked which generalize an individual's social space (Hennig et al. 2012, p. 86). The following question is an example of a global name generator: "How many different friends have you had lunch or dinner with in the last six months?"

The position generator originates from a study by Lin and Dumin (1986). They studied the diversity of social networks in Buffalo by asking respondents if they

[10] Resources generators are also a popular approach (Lin 2001) but were not covered in this study.

personally knew people (at the acquaintance level or higher) who had specific occupations of varying prestige. The position generator allows for indicators of:

1. "range of accessibility to different hierarchical positions in the society"
2. "extensity or heterogeneity of accessibility to different positions"
3. "upper reachability of accessed social capital" (Lin et al 2001, p. 63)

The position generator technique is useful in measure the diversity and reachability of an individual's social network and resource access.

In empirical work from the travel-activity literature, the predominant approaches have been activity generators and interpersonal name generators. Activity generators are essentially global name generators within the context of a particular activity. These are typically used in activity diaries as it entails just adding a question about the types of contacts the respondent shared an activity with (e.g. alone, with friends, with family). Interpersonal name generators are used to gain an understanding of the characteristics of respondent's social contacts and then link this to activity content. An example of this would be to use gender homophily and age homophily among an individual's close contacts as covariates in a regression of recreational activity generation.

The position generator has seen limited usage in activity-travel studies. Thus, there is a gap in knowledge on the useful of social capital indicators from position generators. The purposes of this study are to:

1. determine if network diversity as measured by a position generator correlates with activity generation
2. determine if network diversity aids in the predicative accuracy of models of activity generation, and

3. analyze differences in models of activity generation that use name generator and interpreter data as compared to position generator data.

The Pew Internet *Personal Networks and Community* survey is used to accomplish these purposes. This survey includes both a name generator and position generator, plus it asks respondents about their frequency of visits to eight different location types. Results show that, in relation to models that use name generator measures of core network size, homophily, and alter characteristics, models using the position generator to measure social network occupational diversity or extensity have explanatory and predictive power in activity participation modeling.

3.1 Eliciting Contacts in Travel and Activity Surveys

The two primary techniques for eliciting contacts in travel and activity surveys are (1) activity generators and (2) name generators and interpreters. Activity and contact generators prompt respondents to recall their social networks in connection with particular events. For example, a survey may ask a respondent to describe their most recent shopping trip by providing location and duration information and by describing with whom the trip was undertaken with. The name generator is a survey instrument that asks respondents to list contacts in connection with a particular inquiry about contacts who share a type of relationship. With this list of contacts, a name interpreter is used to describe each contact and their connection with the respondent and other contacts.

3.1.1 Activity Generators

Activity diaries are the primary technique in travel behavior research for measuring activity-travel patterns. Over the last two decades, activity elicitation exercises in activity diaries have begun to ask questions about the individuals who share activities with respondents. This technique of acquiring social network data via correspondingly shared activities will be called *activity generators* for consistency with other approaches. This technique could also be termed as a global name generator (Hennig et al 2012), as the generators listed here do not try to uniquely identify social contacts. For example, Sener et al. (2011) used the 2007 *American Time Use Survey* which provides data on the one-day activity patterns of individuals. For each activity, the individuals with whom someone shared an activity were stated (i.e. alone, family, friends, family and friends).

In a study by Lin and Wang (2014), activity-travel diary data was collected from residents of Hong Kong. Their study used both a one-day activity generator and a global name generator[11]. The global name generator consisted of asking respondents "to report the number of social contacts (in terms of the number of people) through all means including face-to-face, phone calls, email, etc., in the past week" (p. 23). They used this question to obtain information on the sources of emotional and instrumental support and social companionship, the quantity of contacts with family and non-family, and the kinship-share of total contacts. Additional studies and analyses that use activity generators include the CHASE survey in Toronto (Habib et al. 2008, Habib and Carrasco 2011) and an activity survey in Eindhoven (Van den Berg et al. 2010, 2012; Artenze et al. 2012).

[11] Lin and Wang considered an interpersonal name generator instead but were concerned about not capturing acquintances.

3.1.2 Name Generators

Name generators are used in a number of studies in travel behavior research. This section will summarize the techniques and results from these efforts in the social capital and social cooperation research applied to activity behavior and mode choice

Carrasco et al (2008) describes the *Connected Lives Study*, an egocentric social network and activity-travel survey. This survey used a name generator and network map for non-household contacts that were:

1. "people whom you discuss important matters with, or regularly keep in touch with, or are there for you if you need help" (p. 966) [*very close contacts*]
2. people who are "more than just casual acquaintances, but not very close" (p. 966) [*somewhat close contacts*]

These categories were intended to represent the network constructs of strong and weak ties, respectively. Data collection began with a free-recall name generator using the two categories mentioned above. Then, a network map, or sociogram, with four concentric circles was used to elicit the closeness of the contacts (innermost ring was strongest and outermost was weakest) and the ties between the ego's alters. The survey also obtained information about each alter including the ego's ordered ranking of the alter's closeness, alter's home location, age, alter's relationship to the ego, alter's job, and alter's heritage. Additionally, "information about the ego's communication and interaction patterns with each alter" (p. 971) was obtained.

The *Communities in Concepción* study (Carrasco and Cid-Aguyao 2012) used a similar format to Connected Lives Study with a name generator of very close and somewhat close contacts. This survey expanded on the Connected Lives Study by asking questions about the directions of social support between the ego and alters. In the Carrasco and Cid-Aguyao (2012) paper, social support was analyzed along the dimensions of emotional, monetary, mobility, and employment support. In their analysis of the relationship of car ownership and network capital, they found “that car ownership does not directly influence the frequencies of ego–alter social interaction, and that personal network spatiality is influenced mainly by income” (p. 1081). Additionally, the analysis showed that car ownership impacted levels of social support differently depending on the context. For example, car ownership was more important for emotional support than income. In a follow-up to the Communities in Concepción study, a second wave of responses was collected four-years afterwards (Chávez et al. 2015) to understand if there were changes in ego-network composition. Some selected findings from their analysis of social tie maintenance includes that car ownership and income important for “maintaining core, non-acquaintance networks” (p. 11) and that age homophily was important in maintaining and gaining social ties.

An activity and social networks study by Sadri et al. (2015) among undergraduate students at Purdue University used a name generator for use in a study of egocentric network measures and activity generation. The name generator used was the following:

> *From time to time, most people discuss important matters with other people. Looking back over the last one month -- who are the people with whom you discussed matters important to you? Please list only those people who reside within Indiana. Write down their first name or initials.* (p. 7)

Using a name interpreter, the alter characteristics obtained included: “race, gender, age, religion, marital status, income and vehicle ownership” (p. 8). Ego-alter tie attributes included: (1) relationship length, (2) interaction frequency, and (3) spatial proximity.

Additionally, information on the existence and strength of ties between alter was also obtained from the respondent. Using zero-inflated poisson models of shared trips by activity type[12], homophily and heterogeneity of ego-networks impacted activity type in varying ways. For example, having an ego-network with more similar gender ties between ego and alters resulted in more "eating out" shared trips and more extra-curricular shared trips. As another example, greater the racial diversity of an ego's alters was correlated with greater quantities of "eating out," study, and extra-curricular shared trips.

Tilahun and Li (2015) used a name generator to study in-person meeting behavior for Minneapolis-St. Paul residents. Their survey asked for information on respondents' closest non-household contacts (up to three). Ego-alter characteristics that were obtained included: gender, age, relationship length, spatial proximity, and communication frequency by contact type. The survey also asked respondents to provide general information about their social networks including the following:

- "number of close contacts the respondent has that they communicate with at least twice on a monthly basis"
- "how many of these live within 3 blocks, 3 miles, 10 miles of their home as well as how many live in the state of Minnesota, in the U.S. or outside of the United States"
- "for three of their closest contacts that don't live with them, how many times per week they communicate by phone or other means as well as face-to-face"

[12] The activity types considered consisted of work, eating out, shopping, recreation, study, and extra-curricular.

- "the home location of these contacts, their gender, and age"
- "basic demographic information for the respondent including residence location, work related variables, and socio-demographic variables" (p. 6)

The authors analyzed frequency of contact between each ego and alter pairing (closest contact alters) with a negative-binomial model. Their analysis found that spatial proximity had little effect on face-to-face meeting except when distances were large (about 50 miles or greater). They found that communication that was not in-person may substitute for in-person meetings when individuals are sufficiently close. Additionally, their findings found that fewer face-to-face meetings occurred among between ego and alter when the alter was male or older but more face-to-face meetings occurred when there was gender and age homophily.

Kowald and Axhausen (2012) describes a snowball-sample survey instrument used to elicit information on interactions between respondents and their social contacts. A name generator of up to forty contacts[13] is used with respondents asked the following:

1. "Please list the people with whom you make plans to spend free time (examples: errands, sports, club or organized activities, cultural events, cooking together or going out to eat, taking holidays or excursions together)" and
2. "If there are other people with whom you discuss important problems, please list them here" (p. 1089)

In the name interpreter, respondents were asked to describe alter characteristics including: (1) gender, (2) age, (3) education, and (4) civil status. Respondents also

[13] Respondents are actually allowed to list additional names on an extra sheet of paper.

provided the relationships between alters as well as ego-alter attributes including spatial proximity and annual contact frequency. After the name interpreter, a network map / sociogram was used to describe future planned meetings between the ego and their alters. For a small subset of the sample, an activity diary was given which has an activity generator that is linked to the name generator. Thus, respondents could give the name of the actual respondents that they shared activities with rather than just a category (e.g. family members, friends).

Frei and Axhausen (2007) and Frei (2012) detail a similar survey with an egocentric sampling design. Two name generators were used with the following questions:

1. "Please indicate persons with whom you discuss important problems, with whom you stay in regular contact or whom you can ask for help" (p. 194) and
2. "Please indicate additional people with whom you undertake leisure activities" (p.195)

The survey then asked individuals to list any vacations / holidays they have taken with these individuals and the name of that individual. Additionally, a name interpreter was used to obtain ego-alter information on initial meeting conditions, relationship length, frequency of contact by communication method, last meeting location, and alter location.

Although not explicitly about activity and travel, Pike's (2014, 2015) studies of social influence in mode choice used a name generator to obtain up to five contacts. In her studies, each respondent was randomly assigned one of three different name generators requesting the names of:

1. “any five people who have been in your social circle over the past six months” (p.76)
2. “the five contacts you have had the most frequent regular interaction with over the past six months” (p. 77)
3. “five people in your social circle, with whom you spoke about transportation in the past six months” (p. 77)

For each alter identified, Pike asked respondents about each alter’s relationship type and length, contact frequency, commute mode, and home distance from the ego’s home. Respondents were also asked about the social connection between their egos. Pike found that the variation in name generators caused differences in the number of alters reported, the locational distribution of the alters, the relationship length distribution of alters, and the frequency of contact.

3.1.3 Position Generator

The Pew Internet *Personal Networks and Community* study (Hampton et al. 2009) was undertaken “to explore the relationship between internet and mobile phone use and the size and composition of core discussion networks” with particular emphasis on social isolation and internet/mobile phone usage. This study included eight questions for respondents about activity participation over the last month. This study has the unique feature of including both a name generator and position generator. In Hampton et al. (2009), logistic regression is used to relate the size of a person’s core network of significant contacts to their participation in eight different activity types. The position generator questions were not used in their analysis nor were structural properties of the network taken into account such as homophily.

Because both a name and position generator are used in this study, the Personal Network and Community study will be analyzed to determine the explanatory and predictive power of each generator for leisure activity generation.

3.2 Case Study Description

The dataset for this case study comes from the Pew Internet Personal Networks and Community study (Hampton et al. 2009). This survey was conducted in July and August 2008 by the Pew Internet and American Life Project.

3.2.1 Survey Design and Sampling Method

Table 3. Summary of Personal Networks and Community Survey Methodology

Time Frame	July and August 2008
Target Population	Noninstitutionalized adults living in the United States, aged 18 and older
Sampling Frame	Households with landline phones and individuals with cellular phones
Sample Design	Random digit dialing of landline and cellular phones
Sample Size	2,512 adults
Response Rate	21% (landline), 22% (cellular phone)
Use of Interviewer	Interviewer administered
Mode of Administration	Phone interview
Computer Assistance	None by respondents
Reporting Unit	One person aged 18 or older per household reports for him/herself and the entire household (landline), one person aged 18 or older reports for him/herself (cellular phone)
Time Dimension	Cross-sectional survey
Frequency	One two-month phase of collecting responses
Levels of Observation	Person, Household
Note: Statistics come from Hampton et al. (2009) and Hampton (2011)	

The survey was designed as an interviewer administered telephone survey including both landline and mobile users. Table 3 summarizes the design and administration of the survey[14]. The survey was broken up into seven modules: (1) internet usage, (2) name generator, (3) neighborhoods, (4) neighborhood group involvement, (5) position generator, (6) public spaces, and (7) household and respondent characteristics.

3.2.2 Name Generator

Two name generators were used in the survey. The first name generator probed for individuals with whom the respondent discussed important matters. Interviewers recorded up to five names, and if individuals submitted fewer than 5 names, the interviewer would attempt to probe for more names. This name generator used the following question:

> *From time to time, most people discuss important matters with other people. Looking back over the last six months — who are the people with whom you discussed matters that are important to you? If you could, just tell me their first name or even the initials of their first AND last names.*

The second name generator probed for individuals with whom the respondent felt were "especially significant." Interviewers recorded up to 5 new names and also recorded if names given from the previous name generator were also repeated. This name generator used the following question:

> *Now let's think about people you know in another way. Looking back over the last six months, who are the people especially significant in your life?* ***[IF NECESSARY:*** *By significant, I mean just those who are MOST important to you.****]*** *If you could, just tell me their first name or even the initials of their first AND last names. These may be some of the same people you just mentioned or it may be other people.*

[14] Appendix D of Hampton et al. (2009) provides an extensive description of the survey methodology.

As with the first name generator, if less than five new names are given, the interviewer probed for more names. In total, up to ten names are possible for this name generator.

3.2.3 Position Generator

A position generator was also used to collect occupational information on the respondent's larger social network (relatives, friends, and acquaintances). The position generator used the following question:

> *Next, I am going to ask about types of jobs and whether people you know hold such jobs. These people include your relatives, friends and acquaintances. Do you happen to know someone who is…* ***[INSERT ITEM; RANDOMIZE]****? What about…****[INSERT]****?* ***[IF NECESSARY:*** *Do you know someone who is* ***[INSERT]****?****]***

Respondents notified the interviewer whether they knew or did not know someone with the given occupation. Twenty-two occupations were asked about:

1. a nurse
2. a farmer
3. a lawyer
4. a middle school teacher
5. a full-time babysitter
6. a janitor
7. a personnel manager
8. a hair dresser
9. a bookkeeper
10. a production manager
11. an operator in a factory

12. a computer programmer
13. a taxi driver
14. a professor
15. a policeman
16. a Chief Executive Officer (C-E-O) of a Large Company
17. a writer
18. an administrative assistant in a large company
19. a security guard
20. a receptionist
21. a Congressman
22. a hotel bell boy

3.2.4 Descriptive Statistics

For the analysis in this chapter, the initial dataset of 2,512 respondents was accessed and cleaned. Specifically, individuals were removed who did not disclose political party, race, age, education, marital status, employment, home type, and neighborhood residency length. Additionally, any respondents without core network data was also excluded (i.e. no responses to the name generator). Table 4 summarizes the characteristics of the final sample of 1,895 respondents.

Table 4. Descriptive Statistics from Analysis of Personal Networks and Community Study (After Data Cleaning)

Race /Ethnicity	White	82%	Household Income	< $10k	6%
	Black	11%		$10k-$19k	8%
	Hispanic	3%		$20k-$29k	11%
	Asian	2%		$30k-$39k	10%
	American Indian	1%		$40k-$49k	9%
	Other Race/ Ethnicity	2%		$50k-$74k	14%
Home Type	Detached House	73%		$75k-$99k	12%
	Townhouse / Duplex	5%		≥ $100k	16%
	Apartment / Condo	13%		Missing	13%
	Other House	8%	Age	Mean	49.2
				Median	49.0
Marital Status	Married	52%		Standard Deviation	17.6
	Living with Partner	6%	Education	Less than High School	2%
	Divorced	11%		Grades 9-11	57%
	Separated	2%		High School Graduate or GED	31%
	Widowed	10%			
	Never Married	17%		Technical / Trade / Vocational School	2%
	Single	2%			
Political Party	Republican	29%		Some College / Associate Degree	24%
	Democrat	37%		College Graduate	20%
	Independent	28%		Post-graduate / Professional School	15%
	No Preference / Indifferent	6%	Employment	Full-time	48%
				Part-time	12%
	Other Party	1%		Retired	23%
HH Kids	None	66%		Unemployed for pay	14%
	One	14%		Disabled	3%
	Two	13%		Student	1%
	Three or More	8%		Other	1%
Adults	One	25%	Gender	Female	53%
	Two	54%			
	Three or More	21%		Male	47%

3.3 Modeling Methodology and Formulations

Activity participation is modeled using individual and household characteristics in additions to measures of social capital from individual social networks. These measures will be described, and then the model formulation will be provided.

3.3.1 Ego Network Measures from the Name Generator

The features of social networks can create value for individuals. In this case study, the following egocentric network measures were analyzed from the name generator for their impact on leisure activity generation (Borgatti et al. 2013):

1. Core Network Size (Degree)
2. Homophily
3. Spatial proximity
4. Alter Attributes (Central Tendency)
5. Tie Dispersion

A description of these different network measures are provided below:

Network Size. The network size is the total number of alters that the ego reported in his core network.

Homophily. Homophily is a measure of how similar alters are to an ego. It is measured via the EI index[15] (Krackhardt and Stern 1988):

$$EI = \frac{ties_{a \leftrightarrow b} - ties_{a \leftrightarrow a}}{ties_{a \leftrightarrow b} + ties_{a \leftrightarrow a}} \tag{7}$$

where:

$ties_{a \leftrightarrow b}$ $\equiv$ the number of ties between dissimilar pairs

$ties_{a \leftrightarrow a}$ $\equiv$ the number of ties between similar pairs

[15] An alternative measure is the Yule's Q, but that measure requires complete network data and this study only has ego network data.

The negative of the EI index is used so that a positive value corresponds to greater homophily. Homophily is measured separately for (1) gender, (2) race, and (3) political affiliation.

Spatial Proximity. The distance between an ego and an alter may impact the frequency of activities and the types of activities undertaken. In this study, spatial proximity is measured by proportion of contacts in the core network that live at different levels of physical distance (e.g. in same home, within 1 mile, more than 100 miles).

Alter Attributes. The central tendency of alter attributes impacts the resources, expectations, and experiences of the ego. This is measured for (1) relationship type and (2) social media friendship status. For each type, proportions of alters in different categories (e.g. proportion of alters who are family members) is used in the analysis.

Tie Dispersion. Tie dispersion refers to the variation in the category of different ties in an ego's network. This is measured as (1) the proportion of alters that the ego reported to have discussions with on important matters and (2) the proportion of alters who are especially significant to the ego.

3.3.2 Ego Network Measure from the Positon Generator

From the position generator, the only egocentric network measure that is obtained is the *social network diversity* or *extensity* via occupational relations. The variety of occupational groups among the ego's social contacts is measured by the total number of different groups in the ego's network.

3.3.3 Activity Participation Model Formulation

The survey provides self-reported data on the frequency of different activities. Eight activity locations are provided and respondents provide frequency of visits for the last month. The number of visits recorded is an integer between 0 and 6 (inclusive) – reported values greater than 6 are recorded as 6. Because of the hypothesis that larger social networks would induce activity participation, binary choice models are used to model activity participation where anyone who participates in an activity at least once in the last month is considered a participant, $y_{ni} = 1$. For the explanatory analysis, this hypothesis is tested on two different network specifications: (1) core network from the name generators and (2) alter dispersion from the position generator. The relevancy of social networks to activity participation is tested by hypothesis testing on the corresponding network variables.

For the predictive analysis, five specifications are tested: (1) core network from the name generators, (2) alter dispersion from the position generator, (3) combined network from both the name and position generators, (4) individual and household characteristics without social networks, and (5) naïve model based on average frequency. Using repeated hold-out validation over a sample of 250 training and testing pairings, the models are compared using two-sample t-tests of correct classifications. Additionally, the models fit is also computed using adjusted count-R^2 which is defined as follows (Freese and Long 2006):

$$R^2 = \frac{count\ of\ correct\ classifications - count\ of\ most\ frequent\ outcome}{sample\ size - count\ of\ most\ frequent\ outcome} \quad (8)$$

A binary probit specification was chosen. The latent variable model specification is as follows:

$$\begin{gathered} y_{na}^* = \beta_a x_n + \gamma_a z(G_n) + \varepsilon_{na} \\ y_{na} = \begin{cases} 1 & if\ y_{na}^* \geq 0 \\ 0 & if\ y_{na}^* < 0 \end{cases} \end{gathered} \tag{9}$$

where:

$z(G_n)$ ≡ network indicators for the ego network G_n from individual *n*

x_n ≡ individual and household-level characteristics for individual *n*

y_{na} ≡ a choice indicator corresponding to the participation in activity type *a* for individual *n*

β_a, γ_a ≡ model parameters for activity type *a*

$\varepsilon_{na} \sim N(0,1)$ ≡ normally distributed error term of unobserved factors, IID by activity type and individual

3.4 Social Network Descriptive Statistics

The social network indicators from section 3.3.1 are described in this section for the cleaned dataset. For the core network using data from the name generators, the distribution of core network size is shown in Figure 3. The median core network size was 3 alters with a mean of 3.34 alters. The 5th percentile core network size was 1 alter and the 95th percentile core network size was 7 alters.

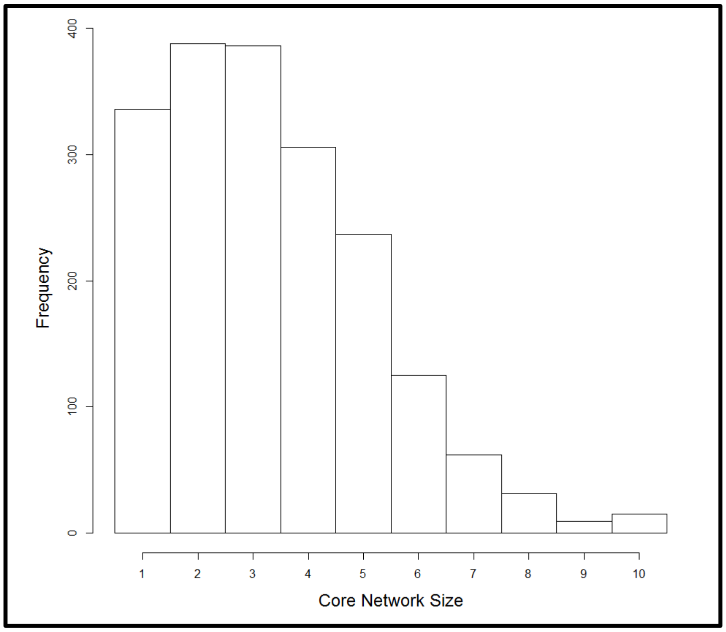

Figure 3. Histogram of Core Network Size as Measured by Name Generator

Data from the name interpreter are summarized in Table 5.

Table 5. Descriptive Statistics on Core Networks from Name Interpreter

Gender Homophily	Median	0.00
	Mean	-0.02
	25th Percentile	-0.50
	75th Percentile	0.30
	90th Percentile	0.30
	95th Percentile	0.50
	99th Percentile	1.00
Racial Homophily	Median	-1.00
	Mean	-0.74
	25th Percentile	-1.00
	75th Percentile	0.00
	90th Percentile	0.00
	95th Percentile	0.20
	99th Percentile	1.00
Political Homophily	Median	0.00
	Mean	0.06
	25th Percentile	-1.00
	75th Percentile	1.00
Spatial proximity	Live within Home	27.7%
	Less than 5 Miles	19.1%
	5-49 Miles	28.8%
	50-99 Miles	4.4%
	Greater than 100 Miles	19.9%
Proportion: Especially Significant Contact	Median	0.75
	Mean	0.67
	25th Percentile	0.50
	75th Percentile	1.00
Proportion: Discuss Important Matters	Median	0.67
	Mean	0.64
	25th Percentile	0.40
	75th Percentile	1.00
Quantity: Social Media Friends	Median	0.00
	Mean	0.10
	75th Percentile	0.00
	90th Percentile	0.50
	95th Percentile	0.75
	97th Percentile	1.00
Proportion: Family/Kin Relation	Median	0.75
	Mean	0.69
	25th Percentile	0.50
	75th Percentile	1.00

The distribution of extensity / social network occupational diversity as measured by number of occupations from the position generator is shown in Figure 4. The median network diversity was 10 occupations with a mean of 9.71 occupations. The 5th percentile network diversity was 1 occupation and the 95th percentile network diversity was 22 occupations.

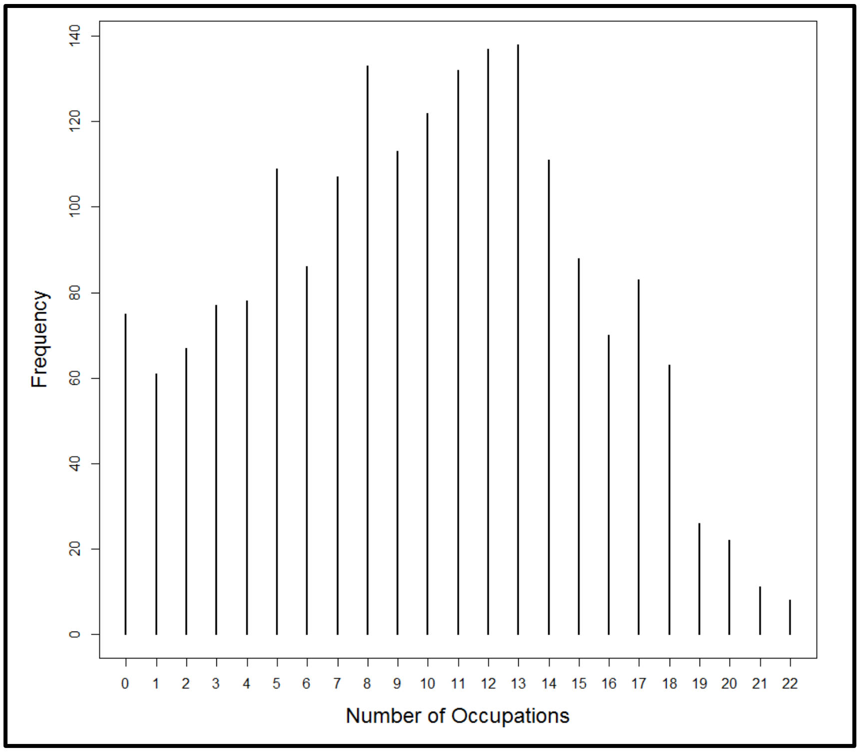

Figure 4. Histogram of Network Diversity as Measured by Position Generator

3.5 Activity Participation Model

In this section, results from the exploratory analysis of activity participation using social capital indicators from a position generator and name generator are compared. Also, a predictive analysis is performed in section 3.5.3 on various model specifications.

3.5.1 Exploratory Analysis: Position Generator

Eight independent probit models were estimated for each activity type / destination using data on respondents and their households as well as the position generator. For all eight models, social network diversity was found to have a significant and positive impact on activity participation. This confirms the hypothesis that network diversity is correlated with activity participation under the assumption that the position generator in this survey is a valid indicator of network diversity. Likelihood ratio test results show that for all eight activity types, the model with network diversity variables fitted the model significantly better than the more restricted model with only non-social variables. Estimation results are provided in Table 6 and an analysis of the results are provided below.

Food and Beverage Establishments. Individual and household characteristics have significant effects on coffee shop, restaurant (fast-food and other), and bar activity participation. Specifically, income tends to increase the probability of eating and drinking out of the home, except in respect to fast-food dining. Fast-food dining and income are negatively correlated when household earn $100,000 or more. Education level was positively associated with dining and drinking outside of one's home for all four activity types. Race tends to not have a significant impact on dining except among African-Americans. Black respondents were less likely to go to coffee shops and non-fast-food restaurants.

Apartment living was positively correlated with trips to coffee shops and non-fast-food restaurants. For marital status, divorced respondents and "never been married" respondents were more likely that married respondents to visit coffee shops and bars. In terms of employment, working part-time was significantly correlated with coffee shop visits but negatively correlated with bar visits. Additionally, disabled respondents tended to avoid coffee shops and bars but there was no significant effect on their tendency to attend other dining establishments.

Network diversity was positively correlated with eating and dining outside of one's home in all four cases. Additionally, individuals with the greatest network diversity (19 occupations or more) saw an additional boost to their probability to visit a coffee shop. In contrast, individuals with the lowest network diversity (2 occupations or more) had an additional deduction in their likelihood to travel to non-fast-food restaurants.

Community Center Activities. Community center activity participation over the last month was observed in 18% of respondents. Community center participation was positively correlated with respondent characteristics of being Hispanic, widowed, retired, and student. Additionally, community center participation was positively correlated with education length and social network diversity.

Place of worship Activities. Place of worship activity participation was observed in 57% of respondents. Specifically, it was positively correlated and statistically significant with female respondents and black respondents. Additionally, place of worship attendance was positively correlated with respondent age and the number of adults in the household. There was no statistically significant log-linear relationship between income and place of worship attendance, but respondents in high income households were less likely to attend a place of worship.

Home type had no effect on place of worship attendance nor did education level. Individuals who lived with a partner, were divorced, or were "single" were less likely to attend a place of worship than married respondents. As compared to fully employed respondents, respondents with part-time employment were more likely to attend a place of worship and disabled respondents were less likely. Politically, all non-Republican political affiliations were negatively correlated with place of worship attendance as compared to Republican respondents. Network diversity was positively correlated with place of worship attendance.

Park Activities. Park visits were observed in 62% of respondents. Using a probit regression model, park visits were found to be positively correlated with education length, income, and political affiliation with the Democratic Party. Park visits were negatively correlated with age and the minority groups of black and other (as compared to white). Additionally, respondents who lived in townhouses were more likely to visit parks than those who lived in detached houses. Students and retirees were more likely to visit parks than full-time employed individuals. In comparison to married respondents, respondents who were "living with partner" were more likely to participate in park activities and "single" respondents were less likely. Social network diversity was found to be positively correlated with park activity participation.

Library Activities. Library visits were observed in 35% of respondents. These visits were found to be positively correlated with length of education and the number of children in a respondent's household. Asians were more likely to attend libraries than whites. Additionally, income and age was found to be negatively correlated with library visits. Compared to being employed full-time, respondents who were employed part-time, retired, unemployed or disabled were more likely to visit libraries. Social network diversity was positively correlated with participation in library activities.

Table 6. Activity Participation Models Using Position Generator Data

Parameter	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Constant	-1.33*	-2.05*	-0.25*	1.51*		-1.40*	1.48*	
Female		0.12**		-0.20*				-0.22*
Log(Respondent's Age)		0.45*	-0.38*	-0.41*			-0.65*	-0.44*
Household Kids	-0.10*	0.13*	0.10*	0.07**	-0.11*			
Household Adults				0.17*				
White	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Black	-0.25*	0.29*			-0.49*		-0.24*	
Asian			0.70*					
Hispanic						0.55*		
Native-American					-0.63*			-0.71**
Other Race							-0.50*	
Log(Household Income)	0.15*		-0.09**		0.21*		0.09**	0.11**
Income Data Unknown	0.50*		-0.42**		0.74*			
Income > $100k		-0.20*		-0.26*	0.23**			0.18**
Detached House	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Apartment	0.16**				0.29*			
Townhouse							0.28**	
Other Home Type								-0.40*
Married	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Living with Partner		-0.24**					0.36*	0.75*
Divorced	0.30*	-0.32*						0.41*
Separated			-0.48**					
Widowed						0.30*		
Never been Married	0.35*							0.36*
Single		-0.64*		-0.48*	-0.63*		-0.50*	
Employed Full-time	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Employed Part-time	0.28*	0.22*	0.47*					-0.27*
Retired			0.36*	0.17**		0.32*		-0.44*
Not Employed			0.31*					-0.30*
Disabled	-0.46*	-0.56*	0.36**					-0.75*
Student						1.03*		
Other Employ. Status								
Republican	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Democrat		-0.39*					0.22*	0.17**
Independent	-0.15**	-0.60*						0.24*
Indifferent to Party		-0.52*			-0.40*			
Other Political Party		-1.55*		-1.13*	-0.74**			
Education in Years	0.08*		0.22*		0.11*	0.05*	0.10*	0.09*
Network Diversity	0.04*	0.07*	0.04*	0.03*	0.04*	0.07*	0.06*	0.03*
Network Diversity > 18	0.47*							
Network Diversity < 3					-0.28*		-0.23**	
Model Statistics: Log-likelihood	-1192	-1132	-1101	-1109	-972	-825	-1075	-951
Nonsocial Log-likelihood	-1224	-1188	-1124	-1127	-1003	-871	-1132	-962

Note: Blank cells are parameter that were estimated but not significant at 95% level (*) or 90% level (**)

3.5.2 Exploratory Analysis: Name Generator

Eight independent probit models were estimated for each activity type / destination using data on respondents and their households as well as the name generator and interpreter. For six out of eight models, core social network size was found to have a significant and positive impact on activity participation. Core network size was not a significant factor in participation in fast-food restaurant and bar activity generation. Homophily was influential in half of the activity types, but for different alter attributes[16]. Spatial proximity was found to be statistically significant only in the likelihood to visit non-fast-food restaurants. Alter attributes were important in half of the activity types and tie dispersion was significant in three models. This confirms the hypothesis that core network size, homophily, spatial proximity, alter attributes, and tie dispersion are correlated with some forms of activity participation.

Likelihood ratio test results show that for all eight activity types, the model with all variables from the name generator / interpreter fitted the model similarly to the more restricted model with only non-social variables. This is due to the large number of additional parameters in the unrestricted model. Estimation results are provided in Table 7 and an analysis of the results in comparison to the results from the position generator models are provided below.

Food and Beverage Establishments. Results for nonsocial parameters were similar to those obtained in the position generator models. Core social network size was positively

[16] If small-network homophily is included (core network size of 1 or 2 alters), then homophily was relevant for all types. Because of the small network sizes, homophily and kin proportion parameters were estimated separately to reduce bias due to the small number of values that were possible.

correlated with coffee shop and non-fast-food restaurant visits. Gender homophily among the core network was negatively correlated with visiting bars, while political homophily was positively correlated with visiting non-fast-food restaurants. Spatial proximity was found to only be significant for non-fast-food restaurant trips. When the core network included a higher proportion of alters who live with the respondent, non-fast-food restaurant trips were more likely. Additionally, having a higher proportion of alters whom the respondent discussed important matters with was positively correlated with fast-food restaurant visits. But, having a higher proportion of alters whom the respondent found to be “especially significant” was negatively correlated with bar activity participation.

Community Center Activities. Core social network size was positively correlated with community center activity participation. Having a higher proportion of alters as friends on social media websites was positively correlated with community center participation. Additionally, for respondents with small core networks, racial homophily was negatively correlated with trips to community centers.

Place of worship Activities. Core social network size was positively correlated with place of worship activity participation. Having a higher proportion of alters whom the respondent discussed important matters with was positively correlated with place of worship participation. Additionally, political homophily was positively correlated with place of worship participation.

Park Activities. Core social network size was positively correlated with park activity participation. Additionally, having a higher proportion of alters whom the respondent discussed important matters with was positively correlated with park visits. Higher proportions of social media friends were found to be positively correlated with park visitation. For respondents with small core networks, gender and racial homophily were negatively correlated with park activity participation.

Library Activities. Core social network size was positively correlated with library activity participation. Higher proportions of social media friends and family ties were found to be positively correlated with library visitation. Racial homophily was found to be negatively correlated with library visits.

Table 7. Activity Participation Model Using Name Generator Data

Parameter	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Constant	-2.06*	-2.25*	-0.71	1.03	-1.47**	-1.71**	0.61	-0.34
Female	-0.13**	0.11**		-0.18*				-0.18*
Log(Respondent's Age)		0.43*	-0.28*	-0.40*			-0.46*	-0.35*
Household Kids	-0.09**	0.14*	0.12*	0.07**	-0.09*		0.08**	
Household Adults				0.18*				
White	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Black	-0.26*	0.35*			-0.38*	0.22**		
Asian			0.59*					
Hispanic						0.37**		
Native-American					-0.52**			
Other Race							-0.56*	
Log(Household Income)	0.15*				0.22*		0.11*	0.13*
Income Data Unknown	0.50*				0.75*			
Income > $100k		-0.20*	-0.35*	-0.27*	0.21**		-0.20*	
Detached House	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Apartment					0.24*			
Townhouse								
Other Home Type	-0.21**							-0.39*
Married	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Living with Partner		-0.26**					0.32*	0.77*
Divorced	0.25*	-0.37*						0.38*
Separated			-0.57*					
Widowed								
Never been Married	0.24*	-0.26*						0.29*
Single		-0.69*		-0.51*	-0.72*		-0.68**	
Employed Full-time	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Employed Part-time	0.18**		0.37*					-0.33*
Retired			0.26*				-0.23*	-0.54*
Not Employed	-0.22*		0.20*				-0.19**	-0.35*
Disabled	-0.54*	-0.69*					-0.42*	-0.78*

Student					-0.67**	0.82*		
Other Employ. Status							-0.68**	
Republican	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Democrat		-0.40*					0.19*	0.17**
Independent		-0.53*						0.24*
Indifferent to Party		-0.50*		-0.26**	-0.46*	-0.31**		
Other Political Party		-1.38*		-1.15*	-0.78**			
Education in Years	0.10*	0.05*	0.23*		0.12*	0.08*	0.12*	0.11*
Homophily: Gender					0.16**			-0.18**
Homophily: Politics		0.11*			-0.12**			
Homophily: Race			-0.16*					
Homophily: Gender (Small Net)							-0.14**	
Homophily: Politics (Small Net)								
Homophily: Race (Small Net)			-0.13**			-0.17**	-0.17**	
Proportion: Family			0.18**					
Proportion: Social Media	0.25**		0.27**			0.58*	0.40*	
Proportion: In Home					0.63**			
Proximity: < 5 Miles								
Proximity: 5-50 Miles								
Proximity: 50-100 Miles								
Proximity: > 100 miles								
Number of Alters	0.06**	0.06*	0.07*		0.09*	0.05*	0.07*	
Proportion: Significant								-0.31*
Proportion: Important		0.27*		0.21*			0.27*	
Model Statistics: Log-Likelihood	-1209	-1172	-1104	-1120	-981	-854	-1101	-950
Nonsocial Log-likelihood	-1224	-1188	-1124	-1127	-1003	-871	-1132	-962
Note: Blank cells are parameter that were estimated but not significant at 95% level (*) or 90% level (**) Homophily parameter estimated on the negative of the EI index								

3.5.3 Predictive Analysis

A predictive analysis was also performed simultaneously with the explanatory analysis[17]. Repeated hold-out validation was used due to the speed of binary probit estimation. The predictive analysis uses six specifications for each activity type tested:

1. naïve *model* based on most frequently chosen,

[17] The model formulations were decided on first from hypotheses about the relationships between network diversity and core network attribute. Then, the explanatory and predictive analyses were performed from using these formulations. This is why the explanatory and predictive models do not pursue any reductions in the number of parameter as this could bias the results for the other analysis.

2. *nonsocial model* with individual and household characteristics and without social networks,
3. *simple core network model* with data from the name generators,
4. *core network with attributes model* with data from the name generators and interpreter,
5. *network diversity model* with data from the position generator, and
6. *combination model* with data from both the name and position generators and name interpreter.

Hold-out validation was repeated 250 times for each model specification and the number of correct predictions was compared. In each validation trial, 80% of the sample (1516 respondents) was used for validation while the remaining 20% of the sample (379 respondents) was used for testing. A concern with repeated hold-out validation is that "some data may be included in the test set multiple times while others are not included at all, or conversely some data may always fall in the test set and never get a chance to contribute to the learning phase" (Refaeilzadeh et al. 2009). Because of the large number of repetitions used, this is not a concern. This can be tested by representing the problem as a coupon collector's problem with group drawings (Stadje 1990). For the drawing size and population size for this dataset, simulation results showed that the 99th percentile number of repetitions needed for full coverage was six.

Table 8 shows the mean prediction count and adjusted count-R^2 for each model type and activity type. Compared to the *naïve* model, all model types predicted better on average for activity types coffee shop, place of worship, library, other restaurant, park, and bar. The *network diversity* model tends to work best in these cases, except for the bar.

No models predicted fast-food restaurant and community center trips better than the *naïve* model. The adjusted count-R^2, which represents a measure of the proportion of predictive accuracy beyond choosing the most frequent option, caps out around 20% additional accuracy for the activity types best predicted by social capital indicators.

Table 8. Mean Prediction Rate and Count-R^2 for Activity Participation Models

Model	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Naïve Model	199.8 [0.00]	215.2 [0.00]	243.8 [0.00]	259.0 [0.00]	273.3 [0.00]	309.4 [0.00]	233.5 [0.00]	273.7 [0.00]
Non-social Model	226.5 [0.15]	236.3 [0.13]	253.4 [0.07]	256.8 [-0.02]	279.6 [0.06]	309.0 [-0.01]	252.9 [0.13]	278.3 [0.04]
Simple Core Network (Name Generator)	229.4 [0.16]	238.9 [0.14]	254.2 [0.07]	255.7 [-0.02]	280.1 [0.06]	309.4 [0.00]	256.2 [0.15]	278.2 [0.04]
Core Network + Attributes (Name Interpreter)	229.1 [0.16]	236.5 [0.13]	253.1 [0.07]	255.4 [-0.03]	279.6 [0.06]	309.2 [0.00]	258.2 [0.17]	276.2 [0.02]
Network Diversity (Position Generator)	234.7 [0.19]	249.4 [0.21]	255.1 [0.08]	259.2 [0.00]	283.4 [0.09]	308.2 [-0.02]	260.4 [0.18]	278.1 [0.04]
Combination: Core Network + Diversity	232.0 [0.18]	248.2 [0.20]	254.7 [0.08]	257.9 [-0.01]	283.2 [0.09]	308.0 [-0.02]	263.3 [0.20]	276.7 [0.03]

Note: In each cell, top number is the mean number of correct predictions on the test set.
In each cell, the number in square brackets is the Adjusted Count-R^2 for the correct predictions versus the naïve model.
Shaded cells represent models in which the mean Adjusted Count-$R^2 \geq 0$ at the 99% level.

To formally test the predictive accuracy of the learning models, two-sample statistical hypothesis tests are used (Refaeilzadeh et al. 2009). Model comparison began by testing for equivalence of variances before proceeding to testing for differences in mean prediction counts. A Bartlett Test (Table 9) was used to determine if the variance of the prediction distributions were similar. For each activity type, each model had a similar variance in its prediction count. Next, the models are compared using pooled two-sample t-tests of correct classifications.

Table 9. Bartlett Test of Variance Equality Results for Cross-validation of Correct Predictions for Activity Participation

Test Statistic	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Bartlett's K^2	3.656	1.031	5.323	0.184	1.371	1.559	0.467	1.436
p-value	0.600	0.960	0.378	0.999	0.928	0.906	0.993	0.920
Reject Null Hypothesis of Equal Variances?	Fail to Reject	Fail to Reject	Fail to Reject	Fail to Reject	Fail to Reject	Fail to Reject	Fail to Reject	Fail to Reject

The t-test results are shown in Table 10. Generally, the predictive power of models using the position generator data was greatest. As shown in the ninth and tenth rows of Table 10, the *network diversity* models tended to be a better predictor than the *simple core network* and *core network with attributes* models. Additionally, the *network diversity* models were better predictive models than the *simple core network* models except for activity types: library, community center, and bar. In only one case – for park activities – was the *combination* model significant better at predicting activity participation than the network diversity model. But the *combination* models tended to be better than the *core network with attributes* models in most cases.

For all activity types except fast-food restaurant and community center, models with non-social and social network data performed better than the *naïve* model. Additionally, the *core network with attributes* models were found to be equally predictive as the *nonsocial* model in all cases except park and coffee shop activity participation. This effect may be attributed to overfitting as the *core network with attributes* models uses a large number of non-significant parameters.

Table 10. Two-Sample T-test Results for Cross-validation of Correct Predictions for Activity Participation

Hypothesis Test Conditions	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
$H_0: \mu_{nonsocial} = \mu_{naive}$ $H_1: \mu_{nonsocial} > \mu_{naive}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (0.999)	Reject (<0.01)	Accept (0.725)	Reject (<0.01)	Reject (<0.01)
$H_0: \mu_{ngni} = \mu_{naive}$ $H_1: \mu_{ngni} > \mu_{naive}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (1.000)	Reject (<0.01)	Accept (0.584)	Reject (<0.01)	Reject (<0.01)
$H_0: \mu_{pg} = \mu_{naive}$ $H_1: \mu_{pg} > \mu_{naive}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (0.379)	Reject (<0.01)	Accept (0.963)	Reject (<0.01)	Reject (<0.01)
$H_0: \mu_{combo} = \mu_{naive}$ $H_1: \mu_{combo} > \mu_{naive}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (0.933)	Reject (<0.01)	Accept (0.984)	Reject (<0.01)	Reject (<0.01)
$H_0: \mu_{ngni} = \mu_{nonsocial}$ $H_1: \mu_{ngni} > \mu_{nonsocial}$	Reject (<0.01)	Accept (0.417)	Accept (0.636)	Accept (0.976)	Accept (0.508)	Accept (0.349)	Reject (<0.01)	Accept (0.998)
$H_0: \mu_{pg} = \mu_{nonsocial}$ $H_1: \mu_{pg} > \mu_{nonsocial}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (0.881)	Reject (<0.01)	Accept (0.631)
$H_0: \mu_{combo} = \mu_{nonsocial}$ $H_1: \mu_{combo} > \mu_{nonsocial}$	Reject (<0.01)	Reject (<0.01)	Reject (0.037)	Accept (0.064)	Reject (<0.01)	Accept (0.940)	Reject (<0.01)	Accept (0.986)
$H_0: \mu_{ng} = \mu_{ngni}$ $H_1: \mu_{ng} > \mu_{ngni}$	Accept (0.354)	Reject (<0.01)	Accept (0.065)	Accept (0.326)	Accept (0.258)	Accept (0.604)	Accept (0.997)	Reject (0.003)
$H_0: \mu_{pg} = \mu_{ngni}$ $H_1: \mu_{pg} > \mu_{ngni}$	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Reject (<0.01)	Accept (0.942)	Reject (<0.01)	Reject (<0.01)
$H_0: \mu_{pg} = \mu_{ng}$ $H_1: \mu_{pg} > \mu_{ng}$	Reject (<0.01)	Reject (<0.01)	Accept (0.109)	Reject (<0.01)	Reject (<0.01)	Accept (0.905)	Reject (<0.01)	Accept (0.593)
$H_0: \mu_{combo} = \mu_{ngni}$ $H_1: \mu_{combo} > \mu_{ngni}$	Reject (<0.01)	Reject (<0.01)	Reject (0.017)	Reject (<0.01)	Reject (<0.01)	Accept (0.974)	Reject (<0.01)	Accept (0.240)
$H_0: \mu_{combo} = \mu_{pg}$ $H_1: \mu_{combo} > \mu_{pg}$	Accept (1.000)	Accept (0.952)	Accept (0.551)	Accept (0.965)	Accept (0.624)	Accept (0.655)	Reject (<0.01)	Accept (0.969)

Note: Hypothesis tests performed at 95% confidence level with "Reject" meaning to reject the null hypothesis H_0 and "Accept" meaning to fail to reject H_0.
The number in parentheses is the p-value for the corresponding test.
Cells in which the null hypothesis is rejected are shaded in gray.
ng = simple core network with only name generator data
ngi = core network with attributes model with only name generator and interpreter data
pg = network diversity model with only position generator data
combo = combination model with both name & position generator and name interpreter data

3.6 Summary

This study found that social capital is correlated with activity participation similarly to research done in Sadri et al. (2015) and Tilahun and Li (2015). Thus this research contributes by increasing knowledge on the applicability of social capital in activity-travel modeling. The name generator was found to be a valid instrument for determining activity participation for most activity types studied as shown in previous research. Additionally, some measures of social capital including homophily and heterogeneity were found to be correlated with activity participation. Similarly to Tilahun et al. (2015), spatial proximity was found to typically not impact activity participation for contacts located within 50 miles.

Network diversity as measured by a position generator was found to be a reasonable explanatory covariate and predictive variable for activity participation. Explanatory models showed that social network diversity was positively correlated with all activity types. Although this was not an exhaustive study of leisure activity generation, these results suggest that network diversity and activity choice are related, but the direction of influence is unclear. Future research with full activity diaries and measurement or modeling designs to account for directionality of influence is suggested. Additionally, because this is not an exhaustive list of activities, activity constraints could not be imposed but should be considered. Future work should use more activity types – and possibly all activity types such as in Sadri et al. (2014).

This conception of social capital and activity participation requires further discussion. The position generator provides a new tool for activity diary & travel survey that is content-free and structure-centric. But, does it need to be tailored to transport and activity? What is meant by “tailored” here is: how valid is Lin’s theory of job prestige linking to social resources (Lin 1982) in this context? It is possible that network diversity enables greater access to opportunities to socialize and share experiences. These opportunities are social resources which are associated with specific activities. The individual then chooses to partake in these activities to access social resources. More activity-travel research using position generators could serve to empirically explore this theory.

Network diversity has not been applied in existing travel and activity behavior research but this study suggests that it can be considered in data collection efforts. Data from a position generator was found to be more predictive than name generator and interpreter data for some activity types. This may be a promising technique for incorporating social capital indicators in travel and activity diaries with less respondent burden. Name interpreter are burdened by recall concerns and the number of questions needed is exponentially related to the size of a person’s network. By contrast, the position generator question count is constant in size relative to a person’s network size. Concerns are warranted with position generators as the professions chosen will impact its validity. Although this study did not analyze the impact of social class and prestige on activity participation, analysis of these would warrant additional design considerations for the position generator. Additionally, it is unknown whether the number of occupations used in a position generator would impact model estimation for activity participation.

Chapter 4: Sensitivity of Discrete Choice Models to Social Diversity Indicators from Position Generators: A Case Study

Table 11. Chapter 4 Summary

Background & Brief Summary	Using a network diversity indicator from a position generator is new to activity-travel analysis as posited in Chapter 2. The sensitivity of choice models to the design of position generator questions is unknown. Thus, understanding the robustness of length of the questionnaire list may provide insight into position generator list length and its effect on model fit and parameter estimation in choice models using network diversity indicators. A case study using the Pew Internet Personal Networks and Community survey is performed to explore the effect.
Motivation	Existing research into the robustness of position generators is limited. Results tend towards evidence that for measuring the extensity of one's social network and access to resources, that occupational list design in position generator is robust to changing the listed occupational types. Its limited use in travel behavior analysis also motivates understanding its robustness in designing activity surveys to incorporate position generator questions.
Results	1. Model fit was found to significantly improve for all sizes (2 - 21 occupations) of the occupation list in the position generator 2. Bias and variablitity in parameter estimates were found to be robust as MSE results were low and steady until occupational list size decreased to about 10 to 15 occupations 3. Similarly, bias and variablitity in parameter ratios were found to be robust as MSE results were low and steady until occupational list size decreased to about 5 to 10 occupations
Limitations	1. The dataset was not designed specifically for understanding the activity process as thoroughly as focused activity diaries 2. The results cannot generate theoretical insight on its own, but over time it could become part of a portfolio of work showing practical proof of the robustness of network diversity indicators from position generators

The analysis in Chapter 3 showed the relevance of social network diversity in a discrete choice model of activity participation. Network diversity was not measured directly, as the whole network of the respondents was not questioned. Network diversity was indicated from data obtained with a position generator question. This position generator entailed providing a list of different occupations and asking respondents if they knew any person, on a first name basis, with any of these occupation. Respondents who had network connections among more occupations are assumed to have greater network diversity and greater access to networked resources. But how does the design of the position generator affect this indicator of network diversity? And how does the this indicator's design and measurement affect models estimated with indicators from position generator questions? Few studies in the existing literature have attempted to answer this question.

Verhaeghe et al. (2013) studied the use of different occupational list in measuring network diversity indicators. Using a parallel test experiment, they test thirteen different position generator measures on two different occupational lists of equal size. They found that the total number of accessed occupations was equally measured between each list, but that the other measures of social class and prestige/status were less reliable.

Hällsten et al. (2015) analyzes the impact of each occupation from a 40-occupation position generator on a composite measure of social capital. Their study analyzes data from a study on social capital and labor market outcomes using a jackknife procedure to determine each occupation's effect on factors related to social capital such as upper-secondary school grades, number of daily contacts with others, and employment. Results showed that some occupations such as medical doctor, engineer, and university student had stronger impacts on these factors than other occupations. In addition, some occupations reduced the strength of the composite social capital measure. Additionally, Hällsten et al. (2015) also repeat the jackknife procedure by removing sets of ten occupations. They found that “some combinations of occupations contribute mainly statistical noise to the measure” (p. 60).

These studies present some promising results on the robustness and limits of position generation, but have some limitations. The specific results of each study are not generalizable. Position generators use different occupation lists and samples are gathered in different locations. The prestige, reputation, and social class of occupations change between and within countries and cultures. But, the results on the extensity or total network occupational diversity show some promise of being more context-free (Hällsten et al. 2015).

In this chapter, the effect of the position generator’s occupation list length is analyzed for discrete choice models that use a social capital extensity indicator (total network diversity by occupation). Using a case study on activity participation, a 22-occupation position generator is reduced in size. At each level of list reduction, model fit and parameter bias is assessed. Results provide additional support for the robustness of extensity to occupation list size.

4.1 Methodology

The sensitivity analysis is based on the case study in Chapter 3. Using the Pew Internet Personal Networks and Community study (Hampton et al. 2009) as the dataset, the position generator from this study is used in models of activity participation using binary probit models. The effect of occupation list length on model fit and parameter estimation are assessed by comparing likelihood values and the mean squared error (MSE) of parameter estimates.

4.1.1 Model Formulation

The model follows similarly to the activity participation models using indicators of social network diversity from the position generator. The model varies by representing the network diversity as the proportion of occupations known to an individual from the position generator. This model is equivalent to the model in section 3.3.2 since the number of occupation known is proportional to the proportion of occupations. But since in the sensitivity analysis, the size of the position generator changes, the parameter estimates cannot be compared unless they are normalized. Using the proportion of known occupations rather than the total number of occupations accomplishes this normalization. A binary probit specification was chosen with the following specification:

$$y_{na}^{*} = \beta_a x_n + \gamma_a v_n(g_n) + \varepsilon_{na}$$
$$y_{na} = \begin{cases} 1 & if\ y_{na}^{*} \geq 0 \\ 0 & if\ y_{na}^{*} < 0 \end{cases} \tag{10}$$

where:

x_n ≡ individual and household-level characteristics for individual *n*

$v_n(g_n)$ ≡ social network diversity indicator for individual *n* and position generator g_n, equals the number of occupations that the individual knows divided by the number of occupations in the position generator.

y_{na} ≡ a choice indicator corresponding to the participation in activity type *a* for individual *n*

β_a, γ_a ≡ model parameters for activity type *a*

$\varepsilon_{na} \sim N(0,1)$ ≡ normally distributed error term of unobserved factors, IID by activity type and individual

4.1.2 Sensitivity Analysis Design

The sensitivity analysis is performed by reducing the size of the position generator in section 3.2.3. This is done through two techniques: (1) complete enumeration of occupation combinations and (2) sampling of occupation combinations. Complete enumeration entails estimating the model from section 4.1.1 on each combination of the 22 occupations in the positon generator for a given position generator size. For example, for a position generator of size 20, $\binom{22}{20}$ or 231 combinations of the occupations are enumerated over. Complete enumeration is practical for position generators of size close to 22 or close to 1 due to the definition of a combination. Sampling of combinations is performed for position generator sizes where more than 10,000 combinations are needed. For these cases, as shown in Table 12, a sampling of 10,000 combinations was used.

Table 12. Combinations Used in Analysis at Different Position Generator Sizes

Position Generator Size, $\lvert g \rvert$	Combinations: $\binom{22}{\lvert g \rvert}$	Combinations Used in Analysis
21	22	22
20	231	231
19	1540	1540
18	7315	7315
16	74613	10000
14	319770	10000
12	646646	10000
10	646646	10000
8	319770	10000
6	74613	10000
4	7315	7315
2	231	231

The sensitivity of the network diversity indicators on model estimation is measured in three areas: (1) model fit, (2) network diversity parameter estimates, and (3) parameter ratios. The model fit is correlated with the likelihood function as this measures how well the model and its parameters correspond to outcomes from the dataset. The likelihood for each model of activity type a is measured based on the standard likelihood function for a traditional binary probit model:

$$\mathcal{LL}_a(\beta_a, \gamma_a; Y_a) = \sum_{n \in N} \Big[y_n \log \Phi\big(\beta_a x_n + \gamma_a v_n(g_n)\big) + (1 - y_n) \log \Big(1 - \Phi\big(\beta_a x_n + \gamma_a v_n(g_n)\big)\Big)\Big] \tag{11}$$

Additionally, the likelihood ratio test is used to test for model selection by comparing the fit of model with a social diversity indicator to a non-social model without an indicator.

The social network diversity parameter estimate describes the strength of the impact of social capital on activity participation. Additionally, the parameter ratios are also another measure often used in discrete choice model because it is not sensitive to changes in model scale. Parameter ratios are useful for describing the relative impact of network diversity on activity participation in relation to non-social factors. To understand if the size of a position generator impacts these measures, normalized mean squared error is used to determine how position generator design affects the bias and variability of estimates of network diversity measures.

4.2 Model Estimation Results

Eight independent probit models as specified in section 4.1.1 were estimated for each activity type / destination using data on respondents and their households as well as the full position generator data (i.e. size 22). The model results shown in Table 13 were nearly identical to results from the analysis in Chapter 3.

Table 13. Base Estimation Results with Position Generator of Size 22

Parameter	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Constant	-1.55*	-2.05*	-0.25*	1.46*		-1.42*	1.40*	
Female	-0.12**	0.13**	0.12**	-0.20*				-0.21*
Log(Respondent's Age)		0.44*	-0.38*	-0.42*			-0.66*	-0.44*
Household Kids	-0.11*	0.13*	0.10*	0.07**	-0.10*			
Household Adults				0.17*				
White	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Black	-0.32*	0.28*			-0.50*		-0.25*	
Asian			0.70*					
Hispanic						0.55*		
Native-American					-0.62*			-0.71**
Other Race							-0.52*	
Log(Household Income)	0.13*				0.22*		0.09**	0.12**
Income Data Unknown	0.45*		-0.40**		0.74*			
Income > $100k		-0.21*	-0.32*	-0.26*	0.23**		-0.17**	0.18**
Detached House	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Apartment	0.16**				0.29*			
Townhouse							0.29**	
Other Home Type								-0.40*
Married	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Living with Partner		-0.24**					0.35*	0.75*
Divorced	0.31*	-0.33*						0.41*
Separated			-0.48**					
Widowed						0.30*		
Never been Married	0.35*				0.21**			0.36*
Single		-0.65*		-0.49*	-0.65*		-0.50*	
Employed Full-time	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Employed Part-time	0.28*	0.22*	0.47*					-0.27*
Retired			0.36*	0.17**		0.32*		-0.44*
Not Employed			0.31*					-0.31*
Disabled	-0.44*	-0.57*	0.36**					-0.76*
Student						1.03*		
Other Employ. Status								
Republican	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Democrat		-0.39*					0.22*	0.17**
Independent	-0.15**	-0.60*						0.24*
Indifferent to Party		-0.52*			-0.41*			
Other Political Party		-1.54*		-1.13*	-0.74**			
Education in Years	0.08*		0.22*		0.11*	0.05*	0.10*	0.09*
Network Diversity	1.06*	1.49*	0.95*	0.81*	1.11*	1.55*	1.52*	0.70*
Model Statistics: Log-likelihood	-1195	-1133	-1102	-1111	-976	-825	-1076	-951
Nonsocial Log-likelihood	-1224	-1188	-1124	-1127	-1003	-871	-1132	-962
Note: Blank cells are parameter that were estimated but not significant at 95% level (*) or 90% level (**)								

The *network diversity* parameter estimates and parameter ratio between the network diversity and the *education in years*, *female, number of kids*, and *income* variables are assumed to represent the true model and are used as the base model for the sensitivity analysis.

4.3 Sensitivity Analysis

The sensitivity analysis compares models estimated on position generators of size smaller than the full size of 22 occupations. This is accomplished through observing:

1. *Changes in Model Fit* as measured through changes in the likelihood function and likelihood ratio tests
2. *Bias and Sensitivity in Network Diversity Parameter Estimates* as measured by the normalized MSE
3. *Bias and Sensitivity in Parameter Ratios* as measured by the normalized MSE

4.3.1 Changes in Model Fit

Model fit as assessed using the log-likelihood and likelihood ratio test. In all activity types, the network diversity indicator significantly increased the model fit as assessed via the likelihood ratio test for one degree of freedom. Model fit was expected to worsen as the occupational list length of the position generator decreased. Results from Figure 5 show that the general trend of model fit worsened as the occupational list decreased. The mean log-likelihood values decrease at an increasing rate as list size approaches zero. The largest rate decreases tend to occur for less than 10 occupations. Although this occurs, even when models are estimated on position generators with an occupation list of two, the model fit is still significantly better than the non-social model. The mean log- and 5^{th}-

percentile log-likelihoods correspond to likelihood ratio tests that reject the null hypothesis of the non-social model having equivalent fit.

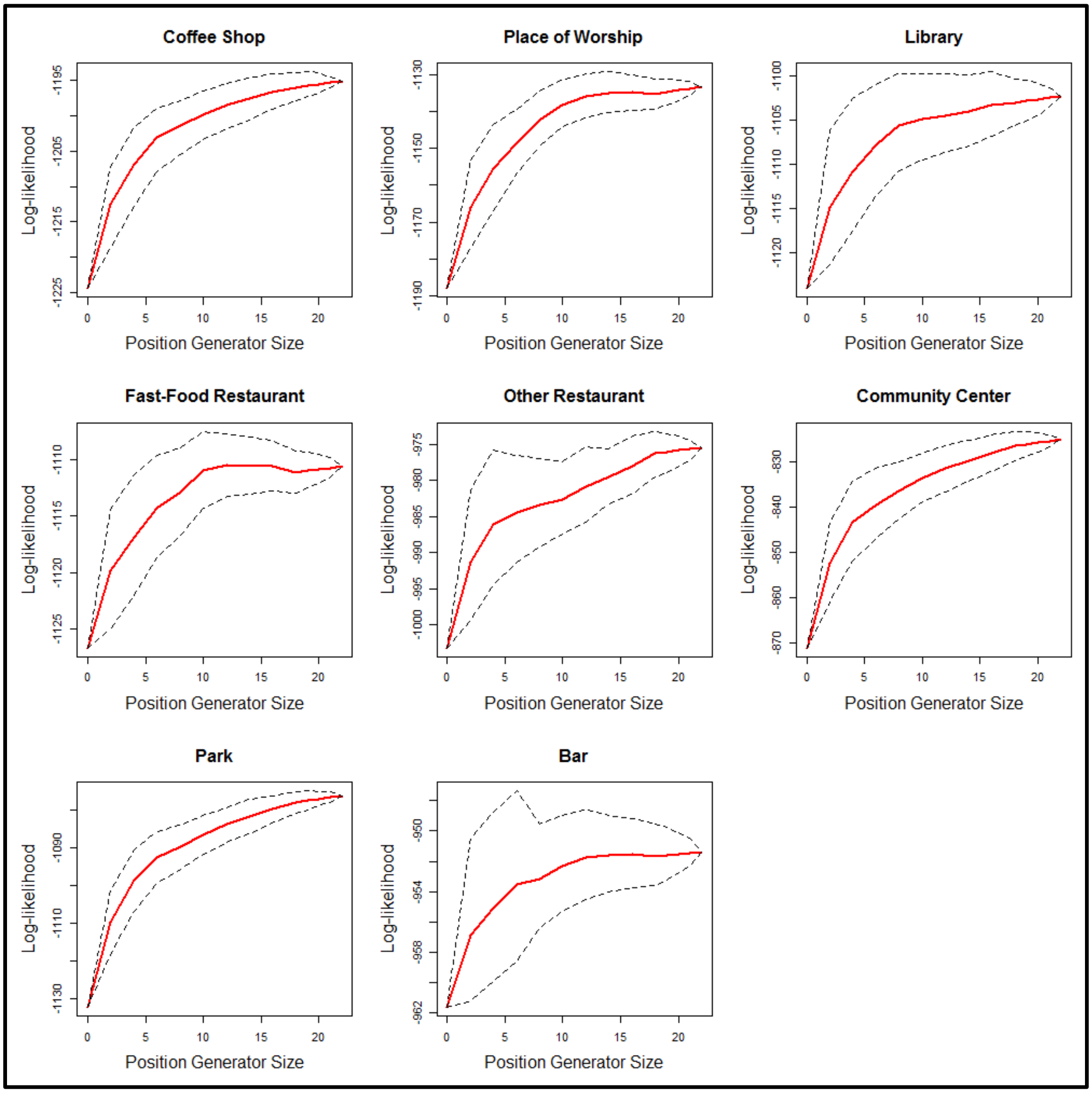

Figure 5. Log-likelihood by Position Generator Size and Activity Type

Additionally, looking at the 95th-percentile log-likelihood values (denoted by the upper dashed line in each plot), some combinations of occupations provided better fit than the full 22-occupation position generator. This may be attested to statistical noise which may provide support for prior claims of statistical noise that some irrelevant occupations may add (Hällsten et al. 2015). This could be tested in future work by regressing log-likelihood percentile on the inclusion of specific occupations.

4.3.2 Bias and Variability in the Diversity Parameter

The MSE is used to assess the bias and variability in the social network diversity parameter γ_a. To allow comparisons between activity types, this measure is normalized by dividing it by the γ_a obtained from the model estimation with the 22-occupation position generator. Results are shown in Figure 6.

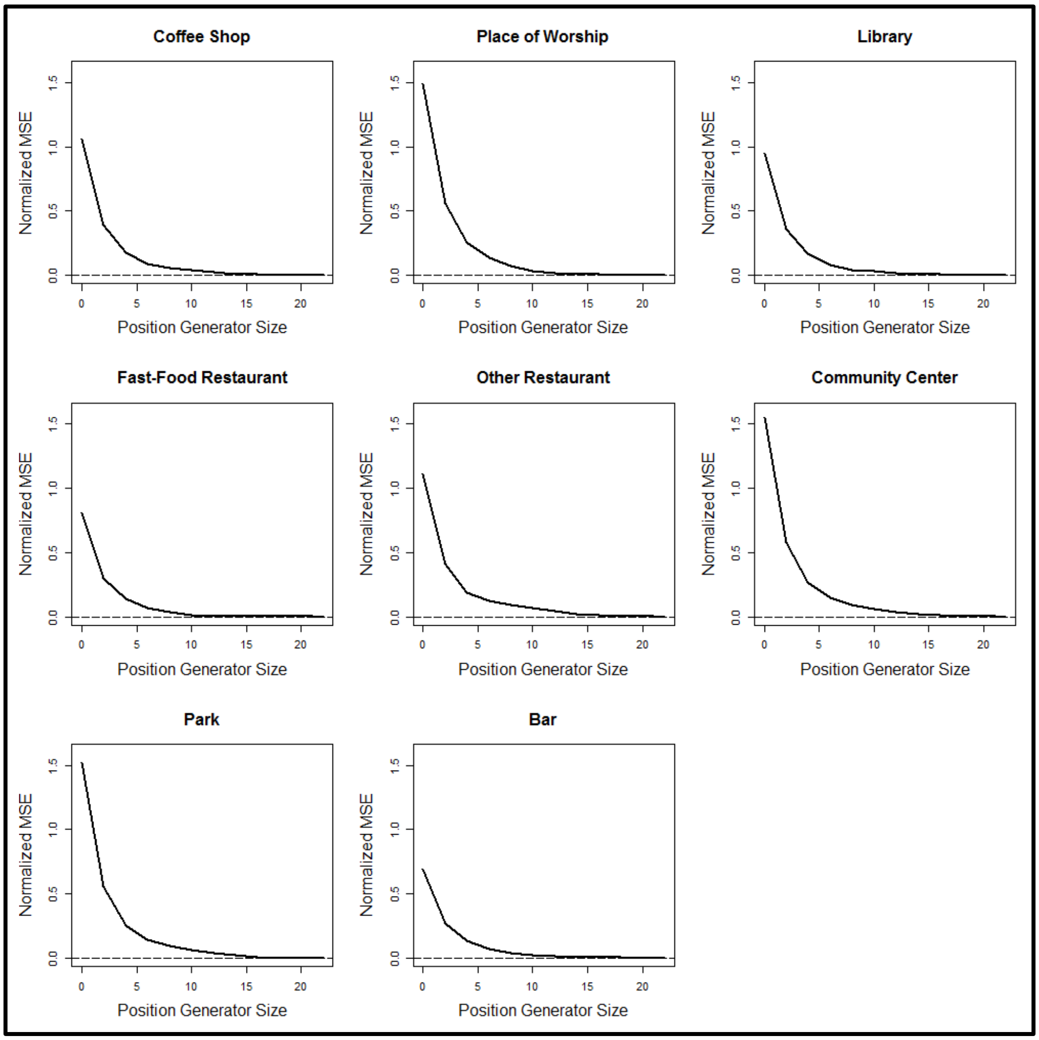

Figure 6. Normalized Mean-Squared Error of Diversity Parameter

For most activity types, the normalized MSE remains steady near zero as the occupation list initially decreases in size. This shows that if the 22-occupation list is believed to faithfully indicate network diversity, then using smaller lists will create similar results. The similarity in estimates tends to occur until the list reduces in size to between 10 and 15 occupations. This lower bound of similarity correlates with the MSE of the non-social model (i.e. zero occupation model) as the models with the largest bias and variability have the greatest lower bound. This effect is seen in the community center and park models where the MSE noticeably increases above zero around 16 occupations. Additionally, the other restaurant model exhibits a similar change but its non-social MSE is not as great.

After reaching the lower bound in similarity, the MSE of the diversity parameter estimate increases. The general trend in the bias and variability of estimates is growth at an increasing rate as occupational list size decreases. As expected, maximum bias and variability are obtained with the non-social model. For all activity types, the rate of increase in MSE accelerates between a list size of two occupations and no occupations.

4.3.3 Bias and Variability in the Ratio of Parameter Pairs

Since the variability of a model can change as covariates change, using parameter ratios is another popular approach to determine the validity of a model and to use its estimates in applications. Parameter ratio estimates were found to be more robust to occupational list length changes than diversity parameter estimates (see Figure 7). The MSE, normalized by the respective parameter ratio estimated with a 22-occupation list (Table 14), remain steady near zero until the list reduces in size to about 5 to 10 occupations.

Table 14. Parameter Ratio Results for Positon Generator of Size 22

Parameter Ratio	Coffee Shop	Place of worship	Library	Fast-Food Restaurant	Other Restaurant	Community Center	Park	Bar
Education / Diversity	0.0772	0.0182	0.2298	-0.0265	0.0969	0.0343	0.0677	0.1322
Female / Diversity	-0.1094	0.0855	0.1220	-0.2426	-0.0177	-0.0441	-0.0424	-0.3059
Kids / Diversity	-0.1009	0.0892	0.1085	0.0829	-0.0943	0.0278	0.0378	-0.0672
Log(Income) / Diversity	0.1205	-0.0118	-0.0906	0.0580	0.1967	-0.0311	0.0622	0.1663
Note: Shaded cells denote ratios where both parameters are significant to at least a 90% level.								

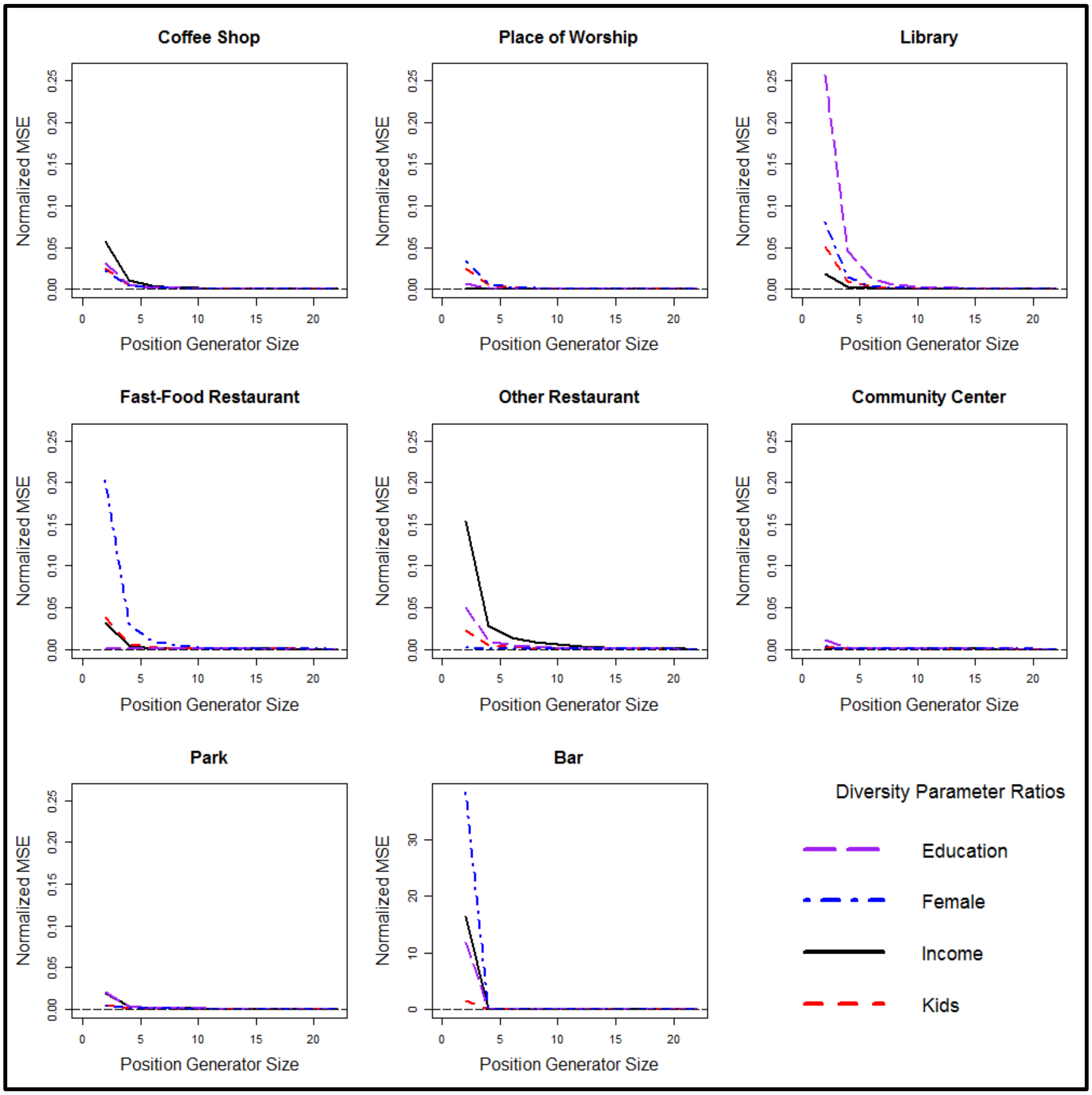

Figure 7. Normalized Mean-Squared Error of Diversity Parameter Ratios (Note: The NMSE axis is the same for all plots except "Bar")

Some caution is warranted with these results since parameter ratios are not guaranteed to have finite moments. This is because the parameter values are not fixed such that the absolute value of the parameter is in $\mathbb{R}^{+}$. A value of zero (or nearly zero) for the diversity parameter would cause an undefined (or very large) value thus making moment calculations impossible. This is what happens with the results from the bar activity models where the absolute value of parameter ratios for some estimates was over 50.

4.4 Summary

Using a network diversity indicator from a position generator is new to activity-travel analysis as posited in Chapter 3. The sensitivity of choice models to the design of position generator questions is unknown. With existing research into the robustness of position generators being limited, the study in this chapter presented additional empirical evidence in support of the claim that position generators are robust in the measurement of extensity / social network occupational diversity.

Results showed that model fit was found to be significantly better than a non-social model for all sizes (2 - 21 occupations) of the occupation list in the position generator. Bias and variability in parameter estimates were found to be robust as MSE results were low and steady until occupational list size decreased to about 10 to 15 occupations for most activity types. Similarly, bias and variability in parameter ratios were found to be robust as MSE results were low and steady until occupational list size decreased to about 5 to 10 occupations.

Future work could test the strength different occupations contribution to representing extensity. Additionally, other travel and activity contexts could be analyzed to contribute additional empirical evidence to support the claim of robustness. Lastly, other indicators of social capital from position generators (e.g. upper reachability of network access, prestige and social class of occupational relations) could be tested for their robustness to the design and length of position generator questions.

Chapter 5: Misspecification of Social Networks in Discrete Choice Models of Social Influence

Table 15. Chapter 5 Summary

Background & Brief Summary	This chapter details a simulation study design to analyze the effect of misspecified social networks on social influence choice model estimation. A simulation study of random additions and omissions of social ties is performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks.
Motivation	Social network data collection can be affected by missing data and measurement error; yet applied social influence choice models often ignore this. Gaining an understanding of the estimator properties for these models under data misspecification can guide methodologists on ways to handle the misspecification as well as aid applied modelers in understanding potential pitfalls in their analyses.
Results	1. Network shape was found to not impact estimator bias and variability nor did it impact model selection 2. Social influence parameter estimates began to lose accuracy and model fit reduced after about 15% to 30% of network ties were changed 3. Individual-level effects parameter estimates remained unbiased as social influence parameter estimates became downwardly biased.
Limitations	1. The results are specific to the network density in the simulations. An exhaustive analysis of the effect of density was not undertaken

5.1 Motivation and Problem Description

Assume that we have a target population of individuals N that are connected via a social network G – with the social connections of an individual n denoted by $g(n)$. The modeler acquires data from the population of interest, specifically some social network data is acquired through a collection process Λ. This network collection process transforms real social networks into a mathematical representation, $\Lambda(G, N) \rightarrow G'$ -- where G' is the acquired network data. This network collection process often is imperfect due to:

- *Representation*. Properly representing the true social network structure can be difficult due to data collection constraints such as survey administration costs and time, privacy, and determination of network/population boundaries. Additionally, the task of choosing an appropriate mathematical representation can be troublesome such as choosing between directed and undirected networks or binary or valued edge weights.
- *Missing Data*. There can be issues with missing network ties or nodes in social network data. An (2011) summarizes reasons for missing data by Handcock and Gile (2007). Prominent examples of missing data include: nonrandom sampling of subjects, missing network ties, not fully traced network ties, absence or attrition of subjects, and missing nodal or tie covariates.
- *Measurement Error*. Measurement error in social network data occurs when the data collector is unable to correctly collect all social ties and nodes in a network. An (2011) gives several reasons for this such as ambiguous survey questions (e.g. varying definitions of friendship), misreporting of social contacts, respondent sensitivity (e.g. income), and inaccurate data input.

These imperfections lead to a situation where $G \neq G'$ and thus models which depend on G but use G' may lead to biased results and conclusions. In the travel behavior field, sources of network data errors may include:

- *Respondent Recall*
- *Fixed-Recall Surveys / Truncated Network*
- *Indirect Sources of Network Data*

The simulation study in Section 5.2 covers examples of missing data and measurement error. This work's purpose is to provide guidance for survey designers and modelers in the amount of tie-based error which is tolerable for determining if a social influence effect is present as well as the bias in measuring that effect's magnitude.

5.1.1 The Name Generator's Data Representation in a Model

To formalize this in a binary choice context, start with a choice modeling exercise where the model is assumed to have a similar specification to the true data generating process. For a binary choice[18], an individual n must choose between performing an action $y_n = +1$ or not performing that action $y_n = -1$. For a binary logit model of conformity and utility maximization, this data generating process may have the following form:

$$\mathcal{U}_n^{(t)} = \beta x_n^{(t)} + \delta \sum_{q \in g(n)} \frac{y_q^{(t-1)}}{\|g(n)\|} + \varepsilon_n^{(t)}, \qquad \forall n \in N$$
$$y_n^{(t)} = \begin{cases} +1 & if\ \mathcal{U}_n^{(t)} \geq 0 \\ -1 & if\ \mathcal{U}_n^{(t)} < 0 \end{cases} \tag{12}$$

Note that the union of all individual-centric social contacts $g(n), \forall n$ is equivalent to the entire social network for the population:

$$\bigcup_{n \in N} g(n) = G \tag{13}$$

[18] The decisions are mapped to {-1,+1} rather than the more familiar {0,1} to simplify the model specification. In this way, the utility difference between doing and not doing an action can take a simpler form in the specification of the endogenous social influence term.

When the modeler obtains the social network data G' (with corresponding individual-level contacts $g'(n)$), the modeler estimates a model on this observed network data, and the following model is estimated:

$$\begin{gathered}\mathcal{U}_n^{(t)} = \hat{\beta} x_n^{(t)} + \hat{\delta} \sum_{q \in g'(n)} \frac{y_q^{(t-1)}}{\|g'(n)\|} + \varepsilon_n^{(t)}, \qquad \forall n \in N \\ y_n^{(t)} = \begin{cases} +1 & if\ \mathcal{U}_n^{(t)} \geq 0 \\ -1 & if\ \mathcal{U}_n^{(t)} < 0 \end{cases}\end{gathered} \tag{14}$$

where:

$g'(n)$	$\equiv$	the observed/measured social contacts of individual *n* (may be different than that individual's true social contacts)
$\hat{\beta}, \hat{\delta}$	$\equiv$	estimates of the corresponding model parameters β, δ

Since $g'(n)$ may not be equivalent to $g(n)$, then changes in the model estimation properties will likely be due to the difference in the social influence mechanism term between equations (13) and (14):

$$\sum_{q \in g'(n)} \frac{y_q^{(t-1)}}{\|g'(n)\|} \neq \sum_{q \in g(n)} \frac{y_q^{(t-1)}}{\|g(n)\|} \tag{15}$$

When the estimators are unbiased, then the expected value of said estimator equals the true value, and accordingly, as the sample size increases, the sample mean of the estimator should converge to the true value. For the binary choice model above, this corresponds to:

$$E[\hat{\delta}] = \delta, \qquad E[\hat{\beta}] = \beta \tag{16}$$

When analyzing if social influence occurs in a population and a decision process, the social influence parameter in the choice model has importance in model application. In particular, it has major influence on the equilibrium properties of the model and on the speed by which behavior permeates through the population (or whether it permeates at all). There is a gap in knowledge about the size of the sample needed to accurately determine the magnitude of the social influence parameter. Specifically, the magnitude of this parameter has important implications on long-run behavior and the equilibrium properties of specific simplified models have been analyzed in the econometrics field.

5.1.2 Equilibrium of Binary Choice Models of Social Influence

In traditional discrete choice models, behavior is determined by exogenous and static attributes of the individual and alternative. But in contrast, endogenous social influence models include endogenous feedback effects. With social spillovers, individuals are adopting new alternatives which induce others to change to that alternative which induces even more people and so on. Under these assumptions, equilibrium may be achieved at some point and individuals will stop switching between alternatives. Equilibrium analysis is important for two primary reasons: (1) it is a long-run behavioral outcome and (2) it provides a metric to compare results from different model specifications.

Blume et al (2011) shows that for binary choice models of conformity with networks of non-overlapping large cliques, the equilibrium field effect for a group m_g is:

$$m_g = \int F_\varepsilon\big(\beta X + \gamma k_g(\cdot) + \delta m_g\big) dF_{X|g} \tag{17}$$

where:

F $\equiv$ the cumulative distribution function of the unobservables ε_n

X $\equiv$ the distribution of individual-level characteristics over the population

Typically, solving this fixed-point problem is analytically difficult since equation (17) often does not have a closed form. For a random sample of group g with S_g members, m_g can be approximated by:

$$m_g \cong \frac{1}{S_g}\sum_{n\in g} F_\varepsilon(\beta x_n + \gamma k_g + \delta m_g) \tag{18}$$

Brock and Durlauf (2001) derive equilibrium properties for a simplified binary choice model[19] where all individuals have the same individual-level characteristics $x_n = x,\ \forall n$, and the only source of heterogeneity is via an individual's error term ε_n. Assuming a scale parameter equal to one, the equilibrium market share for the binary choice field effect model is:

$$m_g^* = \tanh(x + \delta m_g^*) \tag{19}$$

Using the properties of the hyperbolic tangent function, Brock and Durlauf (2001) observe the following conditions for multiple equilibria:

1. For $\delta > 1$ and $x = 0$, multiple equilibria exists with three roots (one positive, one negative, and one zero).
2. For $\delta > 1$ and $x \neq 0$, either single equilibrium or multiple equilibria may exist, depending on whether x is greater than or less than a threshold x'.

Although models in practice will not have homogeneous agents, it is important to realize that multiple equilibria are possible. Since models with heterogeneous agents are

[19] Brock and Durlauf use choice set $J \in \{-1, +1\}$ to simplify their mathematical formulations. Their formulation allows them to multiply each coefficient by +1 or -1 (depending on the alternative) thus making the difference in utility equal to $2(\beta x_n + \gamma k_g + \delta m_{ng} + \mu_g + \varepsilon_n)$.

analytical intractable, numerical analysis is needed to determine equilibrium properties. Using equation (18), Fukada and Morichi (2007) use numerical fixed-point methods to determine equilibria for a binary choice model of bicycle parking. In contrast, Dugundji and Gulyás (2008) analyze temporal dynamics for a nested logit model of mode choice using agent-based simulation. By updating individuals' expectations of group-level market shares over many iterations, their simulated market shares converge to stable equilibria.

It is also important to note that analytical results have not been derived for general non-reflexive social network structures, which are the focus of the work in this chapter. Ioannides (2006) provides analytical results for simple non-reflexive structures including the *star*, *circular*, and *linear path* networks. The details for determining the number of equilibrium[20], which are more complex, are not directly applicable to the work in this chapter, but it shows the importance of accurate measurements of the social influence parameter for equilibrium-based policy analysis.

5.1.3 Policy Analysis

For equilibrium-based policy analysis, policy intervention involves two major strategies:

1. Changing the social choice process from a multiple equilibria system to a single equilibrium system
2. Increasing the probability of achieving a favorable equilibrium in a multiple equilibrium system

[20] Interested readers are encouraged to read Ioannides (2006) for more details.

Strategy #1 shifts the equilibrium curve so that only one intersection occurs. Fukuda and Morichi (2007) provide a good exposition of how this could work. In their analysis of illegal bicycle parking behavior, they observed that increasing the frequency of security patrols leads to lower illegal parking rates. They suggest that increasing patrol frequency could translate the $tanh$ function in equation (19) upwards which would ensure a single equilibrium. Another strategy involves decreasing the degree of social influence. Since only one equilibrium is possible for $\delta < 1$, decreasing individuals' urge to conform will increase the significance of individual-level effects, which tend to be more static. Note however that statistical modeling provides no guidance on how to decrease δ.

Strategy #2 takes into account the fact that the existence of multiple equilibria does not guarantee an equal probability of each equilibrium occurring. Dugundji and Gulyás (2008) observe this with nested logit models of social influence. Their nested logit model theoretically could exhibit five equilibria (three stable and two unstable) but their numerical simulations only exhibited two equilibria. Dugundji and Gulyás (2012b) also found that initial conditions affected the likelihood of equilibrium selection. More formal procedures for studying equilibrium selection are needed in travel demand.

5.2 Simulation Study – MCMC Perturbation of ERGMs

A simulation study will be described in this section to look at the impact of random omissions and additions of social ties in social network data on binary logit model estimation of a conformity model with small-world networks. This section will begin with the intuition that drives the hypotheses. Then, a description of the methodology used for the simulation study is provided followed by a description of how small-world networks were generated. This section concludes with the results of this simulation study.

5.2.1 Problem Exploration

To gain some intuition into this problem, begin by looking at two simplified cases involving omission of network ties:

1. Each individual is only connected to others who share the same behavior
2. Each individual is connected to a mix of individuals with different behaviors

Assuming a binary choice model with linear-in-parameter terms, this section will look at each case and its effect on estimation (at a general concept level).

For case 1, take Figure 8 as an example:

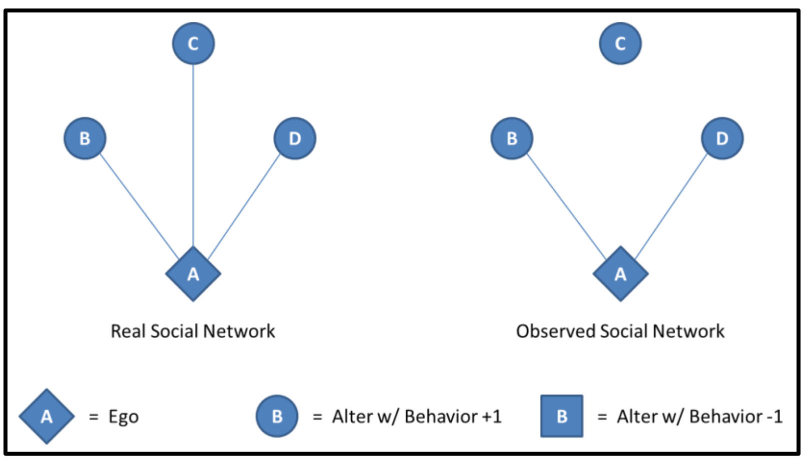

Figure 8. Case 1 Example

The ego under examination is individual A who is connected to three other individuals B, C, and D. Individuals B, C, and D all perform the same behavior, +1, Therefore, the utility for individual A's choice will be as follows:

$$
\begin{aligned}
\mathcal{U}_A^{(t)} &= \beta_0 + \beta_1 x_A + \delta \frac{y_B^{(t-1)} + y_C^{(t-1)} + y_D^{(t-1)}}{3} + \varepsilon_A \\
&= \beta_0 + \beta_1 x_A + \delta \frac{1+1+1}{3} + \varepsilon_A \\
&= \beta_0 + \beta_1 x_A + \delta + \varepsilon_A
\end{aligned} \tag{20}
$$

where:

β_0 $\equiv$ a constant term

β_1 $\equiv$ a parameter weighting the individual-level characteristics

y_B $\equiv$ the behavior of individual B

On the right side of Figure 8 is an example of an observed social network obtained from some network data collection process. Writing down the utility using this observed network, individual A's estimated utility will be:

$$
\begin{aligned}
\hat{\mathcal{U}}_A^{(t)} &= \hat{\beta}_0 + \hat{\beta}_1 x_A + \hat{\delta} \frac{y_B^{(t-1)} + y_D^{(t-1)}}{3} + \hat{\varepsilon}_A \\
&= \hat{\beta}_0 + \hat{\beta}_1 x_A + \hat{\delta} \frac{1+1}{3} + \hat{\varepsilon}_A \\
&= \hat{\beta}_0 + \hat{\beta}_1 x_A + \hat{\delta} + \hat{\varepsilon}_A
\end{aligned} \tag{21}
$$

Individual A's true utility and estimated utility are equal in this case. This is due to the condition that all of individual A's social contacts share the same behavior. Therefore, no matter which combination of social ties is omitted, the total market share of contacts (i.e. the ratio of competing behavior among the contacts) engaging in a given behavior will remain the same. Case 2 will show that this is a very special and specific case.

For case 2, take Figure 9 as an example:

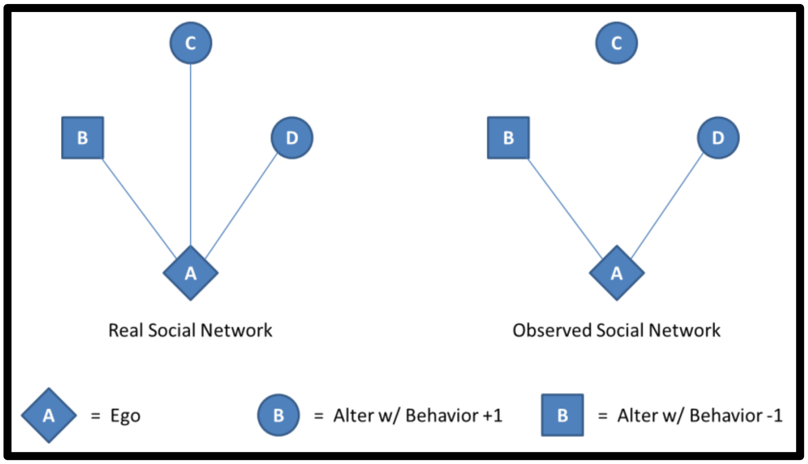

Figure 9. Case 2 Example

The ego under examination is individual A who is connected to three other individuals B, C, and D. Individuals B, C, and D perform various behaviors, with individual B performing behavior -1 and the others performing behavior +1. Therefore, the utility for individual A's choice will be as follows:

$$\begin{aligned}\mathcal{U}_A^{(t)} &= \beta_0 + \beta_1 x_A + \delta \frac{y_B^{(t-1)} + y_C^{(t-1)} + y_D^{(t-1)}}{3} + \varepsilon_A \\ &= \beta_0 + \beta_1 x_A + \delta \frac{-1+1+1}{3} + \varepsilon_A \\ &= \beta_0 + \beta_1 x_A + \delta \frac{1}{3} + \varepsilon_A\end{aligned} \tag{22}$$

Now when the observed social network data is obtained, as shown on the right side of Figure 9, individual A's estimated utility becomes:

$$\begin{aligned}\hat{\mathcal{U}}_A^{(t)} &= \hat{\beta}_0 + \hat{\beta}_1 x_A + \delta \frac{y_B^{(t-1)} + y_D^{(t-1)}}{2} + \hat{\varepsilon}_A \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_A + \hat{\delta} \frac{-1+1}{2} + \hat{\varepsilon}_A \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_A + \hat{\delta} \frac{0}{2} + \hat{\varepsilon}_A\end{aligned} \tag{23}$$

Individual A's true utility and estimated utility are now different in this case. Two parts of the utility equation are changing – with both parts in the social influence term. First, the number of social contacts for individual A decreases from $g(n) = 3$ to $g'(n) = 2$. Second, the market share (ratio) of the observed behavior among the observed contacts changes from $y_B + y_C + y_D = 1$ to $y_B + y_D = 0$. Thus, the estimated social influence term $\hat{\delta}$ will likely have bias. This bias will be transferred to the error $\hat{\varepsilon}$ and constant $\hat{\beta}_0$ terms, likely biasing these estimates in the same direction as the direction of the true social influence term δ.

The main issue complicating analytical analysis of the bias generated by this misspecification of the social network is that the two changing parts are contained in the utility function as a ratio[21]:

$$\frac{\sum_{q \in g(n)} y_q}{\|g(n)\|} \tag{24}$$

So the bias in the estimated social influence parameter is dependent on the change between this ratio in the true network and the observed network. For the example in Figure 9 specifically, note that these ratio can take on a finite set of values and thus the social influence term from the true network can have the following values[22]:

$$\delta \frac{\sum_{q \in g(n)} y_q}{\|g(n)\|} = \{-\frac{3}{3}\delta, -\frac{1}{3}\delta, +\frac{1}{3}\delta, +\frac{3}{3}\delta\} \tag{25}$$

For the estimated social influence term with the observed social network, the following values are possible:

$$\hat{\delta} \frac{\sum_{q \in g\prime(n)} y_q}{\|g'(n)\|} = \{-\frac{2}{2}\hat{\delta}, \frac{0}{2}\hat{\delta}, +\frac{2}{2}\hat{\delta}\} \tag{26}$$

[21] The ratio is rewritten slightly for clarity.

[22] The fractions are not simplified for clarity.

Comparing equations (25) and (26) confirms the results from case 1 and case 2. When the ratio of the sum of others' behavior over the number of social ties is -1 or +1, no bias results since the observed social network contains an equivalent ratio. Otherwise, as shown in case 2, the estimated term value $\frac{0}{2}\hat{\delta}$ can correspond to true term values of $\left\{-\frac{1}{3}\delta, +\frac{1}{3}\delta\right\}$. Additionally, it is also possible for:

- the true term value of $-\frac{1}{3}\delta$ to have a corresponding estimated term value of $-\frac{2}{2}\hat{\delta}$
- the true term value of $+\frac{1}{3}\delta$ to have a corresponding estimated term value of $+\frac{2}{2}\hat{\delta}$

The complexity of this result – even for this simple case – makes generalizing this for large graphs difficult and therefore a simulation design was chosen to study this phenomenon[23].

5.2.2 Simulation Methodology

In this simulation study, an egocentric sampling strategy will be analyzed where the sampled individuals are asked about their direct social contacts and the behavior of those contacts. Individuals will choose between cycling and not cycling by using utility maximizing behavior. In this Monte Carlo simulation, it will be assumed that the data collectors incorrectly identify individual's relevant social ties but accurately know the behavior of all contacts (correct and incorrect contacts). Therefore there are random omission and additions of social ties to the social network data collected.

The steps undertaken for this simulation study are as follows:

[23] A closed form result from theoretical econometrics/statistics is suggested as an area for future research.

1. Initialize a population (100) with individual-specific characteristics, $x_n \sim Normal(0,1)$
2. Give the population some initial cycling choices, $y_n^{(t=0)} \sim B(0.4)$
3. Generate a social network to connect the population with influential others
4. Individuals make new cycling choices depending on social influence, $\mathcal{U}_n^{(t=1)} = -0.2 + 1.2x_n + \delta\bar{y}_m^{(t=0)} + \varepsilon_n$
5. Collect social network data from the population by randomly omitting and adding social ties while maintaining a similar graph density via MCMC graph simulations of ERGMs
6. Estimate a model on this "observed" social network and behavioral data using maximum likelihood estimation
7. Perform a Likelihood Ratio Test and calculate the mean squared error (MSE) of the parameter estimates

5.2.3 Why Not Use a Simultaneous Specification?

The advantages of using this sequential specification include:

1. This specification can be extended to other social network related measures that are incorporated into choice models of social interactions. This includes the work on social capital in Chapter 3 that used a name generator and specifications with contextual effects.
2. The simultaneous equations specification is currently uncommon in travel behavior studies. The purpose of this simulation work is to provide researchers

with guidance on designing surveys and conducting analysis, so a common specification was chosen.

5.2.4 Data – Small-World Network Generation

Small-world networks are sparse networks that exhibit high clustering and short average path lengths. Thus, individuals tend to form relationships such that

1. an individual's friends tend to be friends with each other (i.e. homophily), but
2. "social network [also] tend to have very short paths between essentially arbitrary pairs of people" (Easley and Kleinberg 2010) (i.e. effectance)

To determine if networks are small-world networks, a quantitative measure was sought to determine this subjectively. Humphries and Gurney (2008) provide the *small-world-ness* measure which compares a particular graph instance to an average Bernoulli random graph. Essentially, if there is higher clustering and shorter average path lengths in a graph as compared to a Bernoulli random graph with a similar graph density, then that graph is a small-world graph.

Clustering can be measured with the social network measure called the clustering coefficient:

$$C^{\Delta} = \frac{3 * number\ of\ triangles}{number\ of\ paths\ of\ length\ 2} \tag{27}$$

The average shortest path length measures how easily it is to traverse the graph between any two nodes in a component. It is measured as follows:

$$L_g = \frac{1}{|G|} \sum_{i,j\ :\ i \neq j} L_{ij} \tag{28}$$

The *small-world-ness* of a graph (Humphries & Gurney 2008) is then the ratio of the clustering of a graph and Bernoulli random graphs with similar density divided by the ratio of the average shortest path length between that graph and Bernoulli random graphs with similar density:

$$S^{WS} = \frac{C_g^{\Delta}/C_{rand}^{\Delta}}{L_g/L_{rand}} \tag{29}$$

If $S^{WS} > 1$, then the network is a small-world network. Additionally, the value S^{WS} can give an indication of how "small-world" a graph is such that higher values of S^{WS} correspond to graphs with higher clustering and/or smaller path lengths.

Since small-world networks are a function of the structure of the network and can generated by different processes, the stochastic network generation model chosen is the exponential-family random graph model (ERGM). ERGMs are a family of statistical models for representing networks/graphs by the likelihood of observing counts of certain network configuration terms such as edges, triangles, and k-stars. ERGMs assume that networks are formed by bottom-up processes that work between nodes. For example, transitivity can be modeled by counts of triangles. A graph instance, upon which a model is estimated on, is considered to consist of a dependent series of local processes which are correlated in the local area around any given node but uncorrelated outside of the local. This can lead to macro-level graph behavior emerging, such as small-world networks.

To generate small-world graphs, social circuit dependence (Robins et al. 2007) was assumed. In social circuit dependence, the set of graph configurations are functions of the edges, k-triangles, k-stars, and k-twopaths. Appendix A provides details on some of these functions. Social circuit dependence was assumed rather than Markov dependence because the k-triangles and k-twopaths configurations have a direct impact on the clustering coefficient since the clustering coefficient is affected by the number of triangles and paths of length 2. By contrast, the Markov dependence assumption only includes edges, triangle, and k-star configurations and thus misses the possible importance of the k-twopaths. Additionally, ERGM simulations using social circuit dependence assumptions are less prone to degeneration than Markov dependence assumptions (Robins et al. 2007).

The purpose of this work was to concentrate on multiple triangulation and popularity (degree distribution) while penalizing isolated nodes[24]. The model formulation used to generate small-world graphs based on triangulation and popularity was as follows:

$$P(G = g|\theta) = \frac{1}{\kappa(\theta)} \exp\{\theta_v z_v(g; \theta_\lambda^v) + \theta_d z_d(g; \theta_\lambda^d) + \theta_i z_i(g)\} \tag{30}$$

where:

$z_v(g; \theta_\lambda^v)$ ≡ the geometrically weighted edgewise shared partner (*gwesp*) statistic, a measure of *multiple triangulation* in the graph

$z_d(g; \theta_\lambda^d)$ ≡ the geometrically weighted dyadic shared partner (*gwdegree*) statistic, a measure of *popularity* in the graph

$z_i(g)$ ≡ the graph isolates statistic, a count of the number of isolated nodes in the graph

$\theta_v, \theta_v, \theta_i$ ≡ model parameters corresponding to the given graph statistics

[24] Simulation studies with multiple connectivity were also performed and had similar results. Results from that work are available upon request.

$\theta_\lambda^v, \theta_\lambda^d$ $\equiv$ a scale parameter (assumed to be $\log(2)$ in all simulations)

The isolates parameter was fixed to -4.00 in order to strongly discourage the formation of graphs with isolated nodes. This would reduce the efficiency of each simulation since isolated nodes contribute nothing towards increasing the information on the effect of social influence in the sample. Additionally, the number of edges in the graphs was constrained to always have a density of 0.05. This decision was made to remove variation due to varying graph densities, thus reducing the dimensionality of the problem.

The parameter space was explored to find parameter pairings that would generate small-world networks as given by Humphries and Gurney's definition of a small-world network (see equation (29)). The procedure for finding small-world networks was as follows:

1. *Define the parameter space.* The *multiple triangulation* parameter (*gwesp*) was allowed to vary from 0.0 to +1.6 in steps of size 0.4, while the *popularity* parameter (*gwdegree*) was allowed to vary from -1.0 to +1.0 in steps of size 0.5. The isolates parameter was held constant at -4.0.
2. *Generate simple random graphs.* One thousand simple random graphs with a 0.05 probability of tie formation were generated and average clustering coefficient and average shortest path length were stored.
3. *Generate networks.* At each parameter pairing, 1000 graphs were generated and network statistics for each graph were stored.
4. *Calculate Small-world-ness.* For each graph generated at a given parameter pairing, that graph's *small-world-ness* was calculated from its clustering coefficient and average shortest path length and from the simple random graphs.

5. *Select Small-World Parameter Pairings.* For the graphs generated at each parameter pairing, selection criteria were applied.

The following selection criteria were applied:

- At least 95% of the graphs generated must have a small-world-ness factor greater than 1.0. For chosen parameter pairings, all graphs generated were small-world.
- The mean size of the graphs' giant components must be greater than 94.00. This criterion was chosen to reduce instances of multiple components, particularly small cliques or isolated dyads.

Fifteen graphs were chosen according to these rules and the selected parameter pairings are listed in Table 16. Additionally, Figure 10 provides samples of graph generated for each parameter pairing. The grid in Table 16 corresponds to the grids in Figure 10.

Table 16. Chosen Parameter Pairings (gwdegree, gwesp, isolates) for Small-World Networks[25]

(-1.0,0.4,-4.0)	(-1.0,0.8,-4.0)	(-0.5,0.4,-4.0)
(-0.5,0.8,-4.0)	(-0.5,1.2,-4.0)	(0.0,0.4,-4.0)
(0.0,0.8,-4.0)	(0.0,1.2,-4.0)	(0.5,0.4,-4.0)
(0.5,0.8,-4.0)	(0.5,1.2,-4.0)	(1.0,0.4,-4.0)
(1.0,0.8,-4.0)	(1.0,1.2,-4.0)	(1.0,1.6,-4.0)

[25] The ordering of the pairing in this grid is repeated in all figures showcasing results for all the parameter pairings (triangulation and popularity).

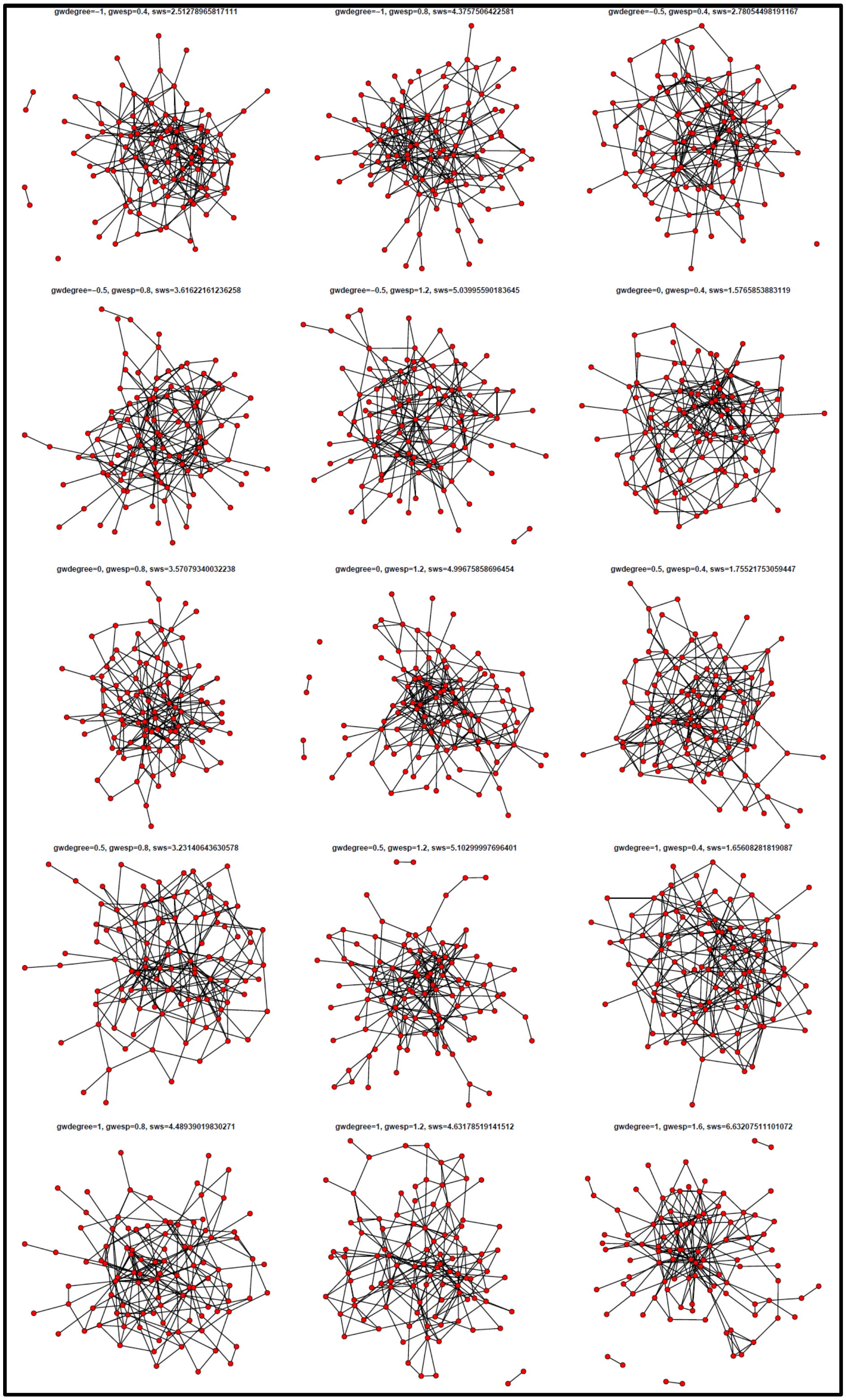

Figure 10. Example Graphs Generated from the Triangulation and Popularity Parameter Pairing

The graphs generated by each parameter pairing had the following statistics:

Table 17. Graph Statistics for Simulated Networks for Triangulation and Popularity Parameter Pairings

gwdegree	gwesp	SWS	Clustering Coefficient	Mean Path Length	Isolates	Size of Giant Component
-1.00	0.40	2.59 (0.38)	0.06 (0.0070)	2.23 (0.17)	0.14 (0.37)	99.55 (0.87)
-1.00	0.80	4.29 (0.46)	0.10 (0.0071)	2.17 (0.16)	0.37 (0.59)	97.99 (1.85)
-0.50	0.40	2.42 (0.37)	0.06 (0.0072)	2.26 (0.17)	0.07 (0.27)	99.73 (0.73)
-0.50	0.80	4.16 (0.46)	0.09 (0.0074)	2.18 (0.17)	0.22 (0.46)	98.67 (1.61)
-0.50	1.20	5.65 (0.54)	0.12 (0.0065)	2.10 (0.15)	0.58 (0.75)	94.79 (3.17)
0.00	0.40	2.37 (0.36)	0.05 (0.0070)	2.26 (0.18)	0.03 (0.17)	99.88 (0.45)
0.00	0.80	4.01 (0.45)	0.09 (0.0068)	2.21 (0.17)	0.10 (0.31)	99.23 (1.24)
0.00	1.20	5.58 (0.55)	0.12 (0.0072)	2.10 (0.17)	0.28 (0.53)	96.80 (2.58)
0.50	0.40	2.31 (0.36)	0.05 (0.0069)	2.25 (0.18)	0.02 (0.12)	99.90 (0.43)
0.50	0.80	3.94 (0.47)	0.09 (0.0071)	2.20 (0.19)	0.05 (0.22)	99.57 (0.91)
0.50	1.20	5.47 (0.57)	0.12 (0.0074)	2.11 (0.16)	0.15 (0.39)	97.95 (2.09)
1.00	0.40	2.27 (0.37)	0.05 (0.0070)	2.25 (0.18)	0.00 (0.04)	99.96 (0.30)
1.00	0.80	3.96 (0.48)	0.09 (0.0074)	2.19 (0.18)	0.01 (0.12)	99.73 (0.78)
1.00	1.20	5.53 (0.64)	0.12 (0.0079)	2.10 (0.18)	0.06 (0.24)	98.84 (1.64)
1.00	1.60	6.60 (0.70)	0.14 (0.0078)	2.07 (0.16)	0.17 (0.40)	95.50 (3.22)

Note: All statistics given are the mean with the standard deviation in parentheses.

5.2.5 Data – Generating "Observed" Networks

To generate social networks with random omissions and additions of social ties, the true social network is modified by using a ERGM MCMC process to generate graphs that are close to the initial graph but different. For example, if we assume that additions and omissions are random but that the density of the graph should remain similar (but not fixed), then the following ERGM is used:

$$
\begin{aligned}
P(G = g|\theta_L) &= \frac{1}{\kappa(\theta_L)}\exp\{\theta_L\, z_L(g)\} \\
\theta_L &= -\log\left(\frac{1}{\text{density}(g)} - 1\right)
\end{aligned} \tag{31}
$$

where:

$z_L(g)$ ≡ the number of edges in the graph g

θ_L ≡ model parameter corresponding to the edge count

Figure 11 shows an example of this process for a network with 16 nodes and an edge density of 0.20. By using the "real" network as a basis for the ERGM simulation process, after 20 intervals of randomly turning ties "on and off" and then accepting or rejecting changes according to evaluations of the new network's likelihood, an "observed" network is generated. This new network no longer has the very dense pocket of blue, green, and purple nodes and the density has reduced to 0.18.

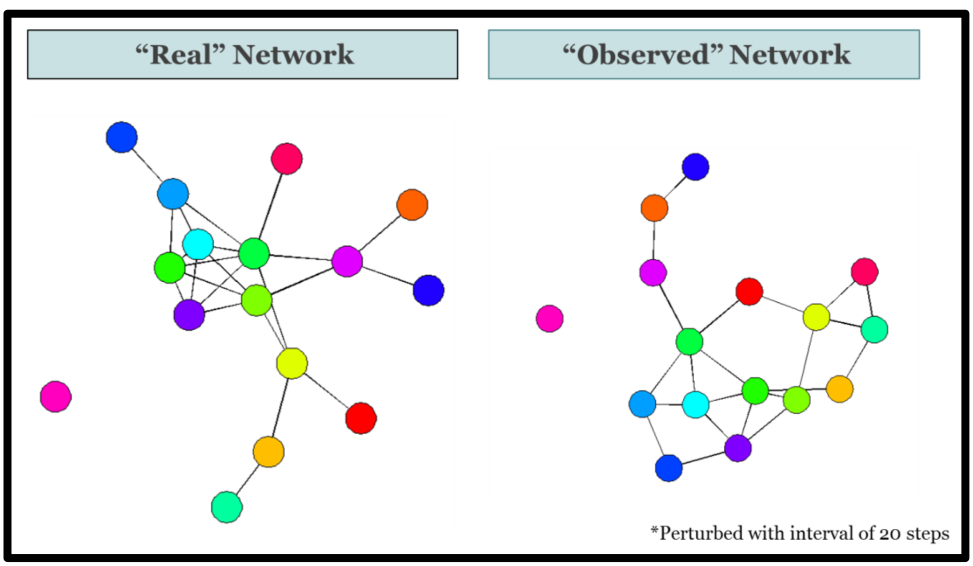

Figure 11. Observed Network Generation Example

5.2.6 Study Hypothesis

In this study, the following hypotheses are made:

1. Network structure will have an impact on the model estimation. Since the observed networks are perturbed by an ERGM likelihood function for Bernoulli random graphs, the networks will higher triangulation will likely experience a change in the degree distribution. The triangulation and average path lengths will trend towards an average simple graph with similar edge density. This may affect the mapping of possible states for the social influence term.
2. As the number of graph changes increases, the accuracy of the likelihood ratio test will decrease and the bias and variability in the social influence parameter estimates will increase.
3. The social influence parameter estimates will become biased as the number of graph changes increases. It is expected that the incorrect contacts will add more statistical noise to this estimate, thus making the social influence parameter estimates less accurate. This noise will be significant and the social influence parameter estimates will trend downwards (towards 0) due to less correlation between the actual and observed contacts as the number of graph changes increases.
4. The constant parameter estimates will become biased in response to the downward bias in the social influence parameter. This is because in binary logit model, the constant insures that the market share predicted by the estimation across the population becomes equivalent to the actual observed market share.

5. The individual-level effects parameter will be unbiased, because there is no correlation between the network structure and individual-level characteristics.

5.2.7 Results

Results for the procedures described in Section 5.2.1 indicate that the hypothesis is correct for all graph configurations covered for the different multiple triangulation and popularity pairing. In particular, for simulations where social influence occurs, as the observed network is more greatly perturbed away from the true network, the likelihood ratio test increasingly rejects the null hypothesis less often than expected. For simulations where social influence does not occur, the likelihood ratio test exhibits similar behavior. Figure 12 shows the typical results that are observed and Figure 13 shows the average results for each parameter pairing.

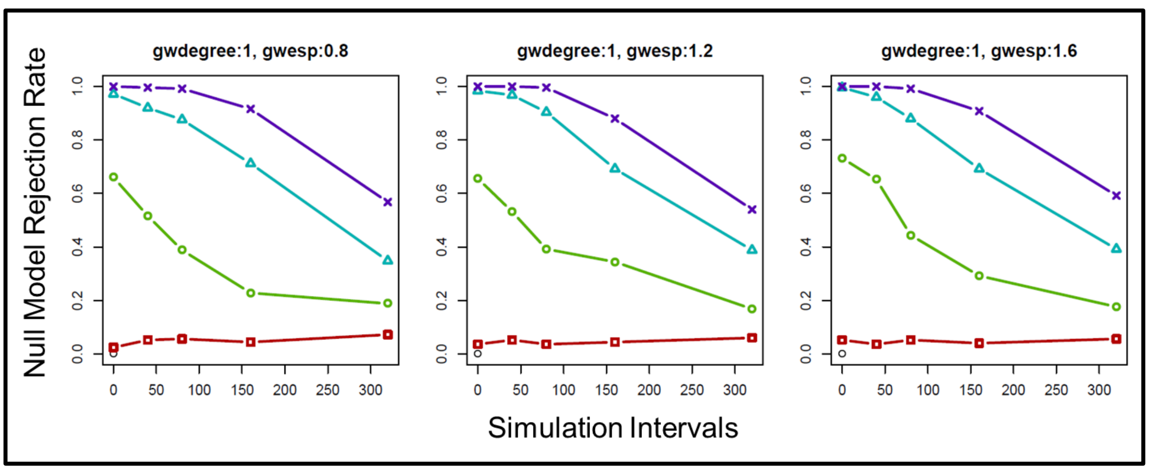

Figure 12. Typical Results for Likelihood Ratio Tests for Misspecified Networks

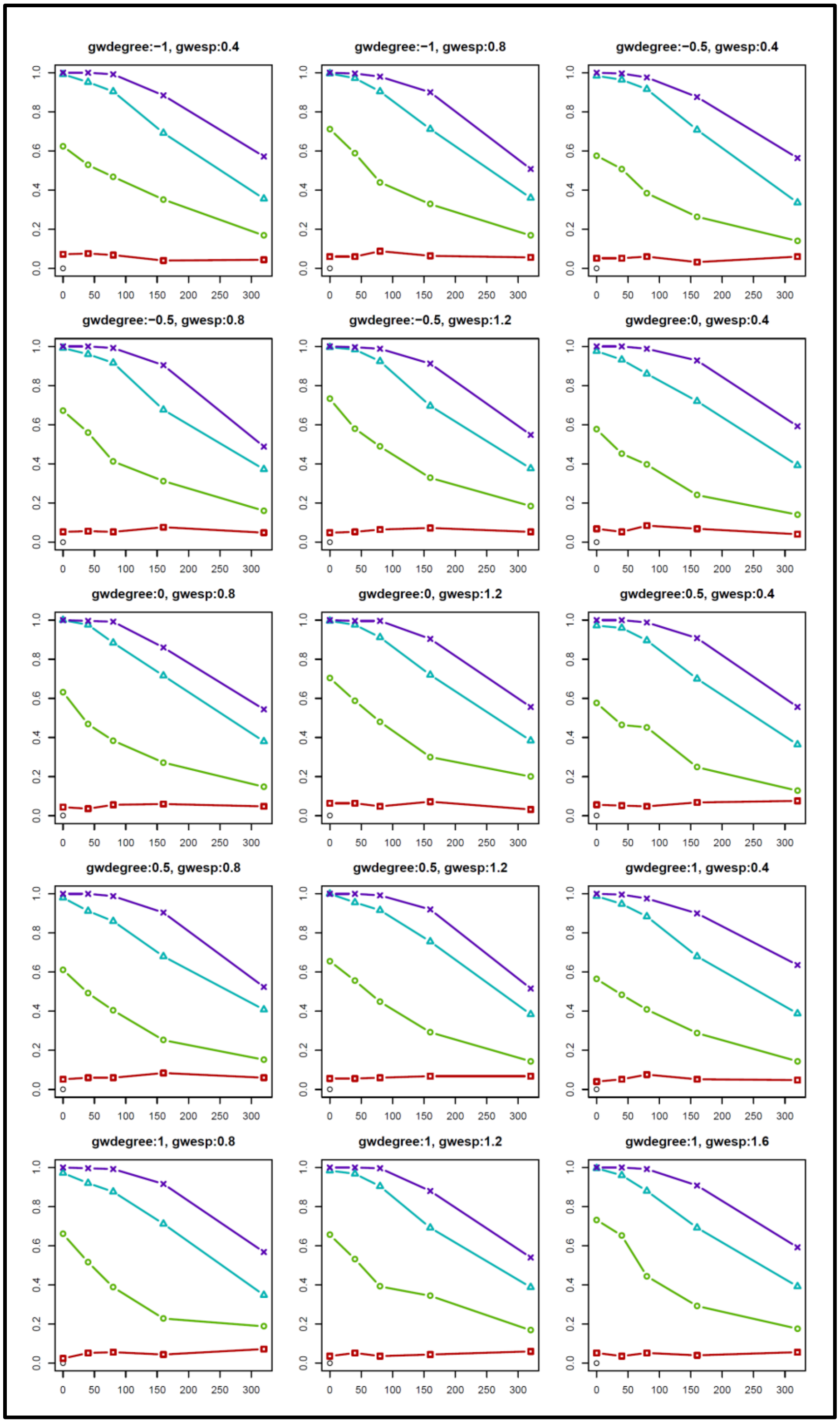

Figure 13. Average Likelihood Ratio Test Results for Triangulation and Popularity Parameter Pairings (Misspecified Networks)

Results indicate that the hypothesis that bias and variability in the estimation of the social influence parameter would increase with greater deviation from the true network is correct for all graph configurations covered. The results for all the network parameter pairings are shown in Figure 14. Most networks exhibit a similar pattern to that shown in the typical results. As expected, MSE increased with increasing deviation from the original network. This increase tends to occur at the same time that there is a drop in model fit. This occurs after about 50 to 100 iterations of the MCMC sampler. Observations of the parameter estimates confirmed the hypothesis of a tendency towards underestimation of the social influence parameter as graph perturbation increased. This tended to not lead to bias in the individual-level parameter estimates (Figure 15), except in the case where $\delta = 3$.

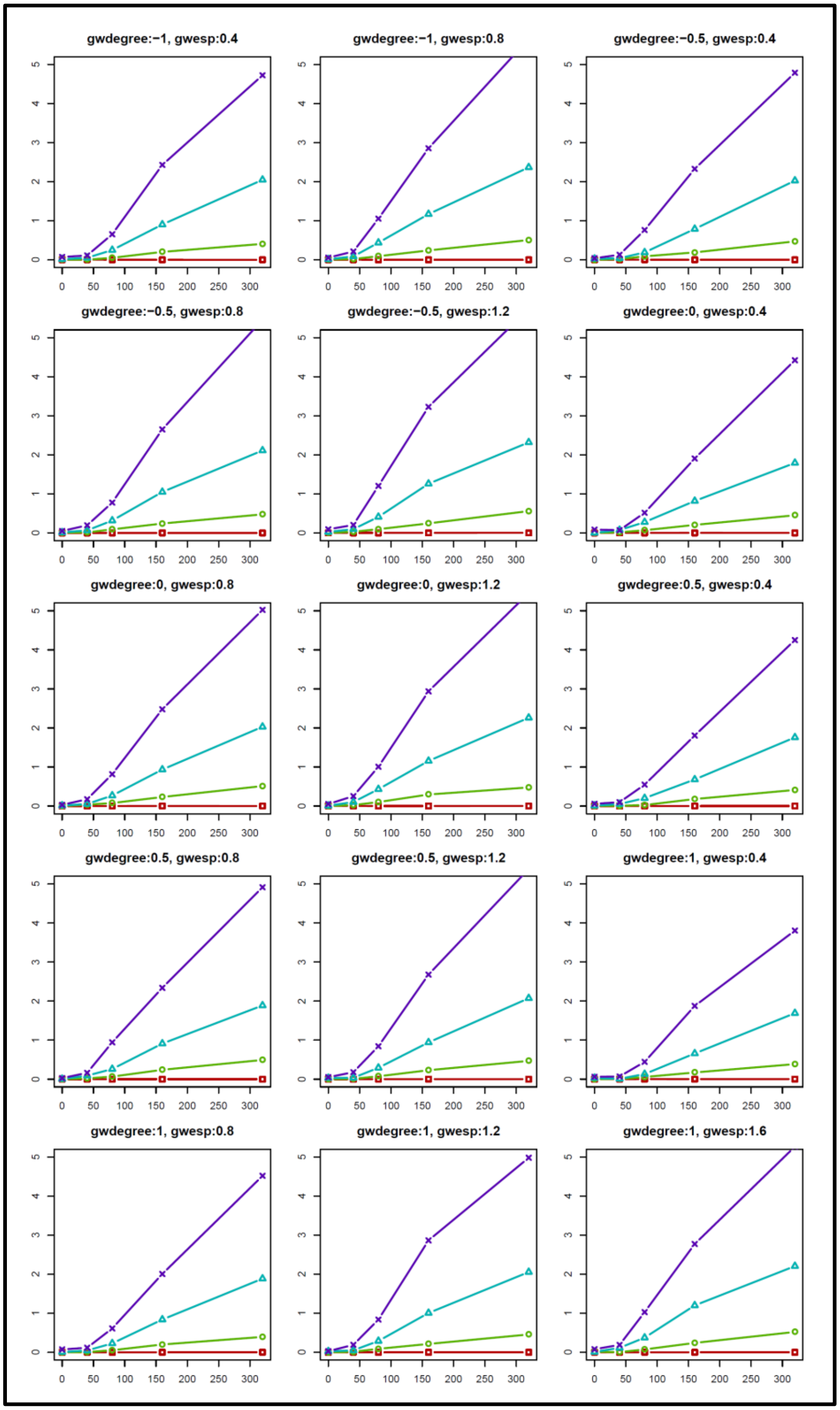

Figure 14. MSE Results for Social Influence Parameter for Multiple Triangulation and Popularity Parameter Pairings (Misspecified Networks)

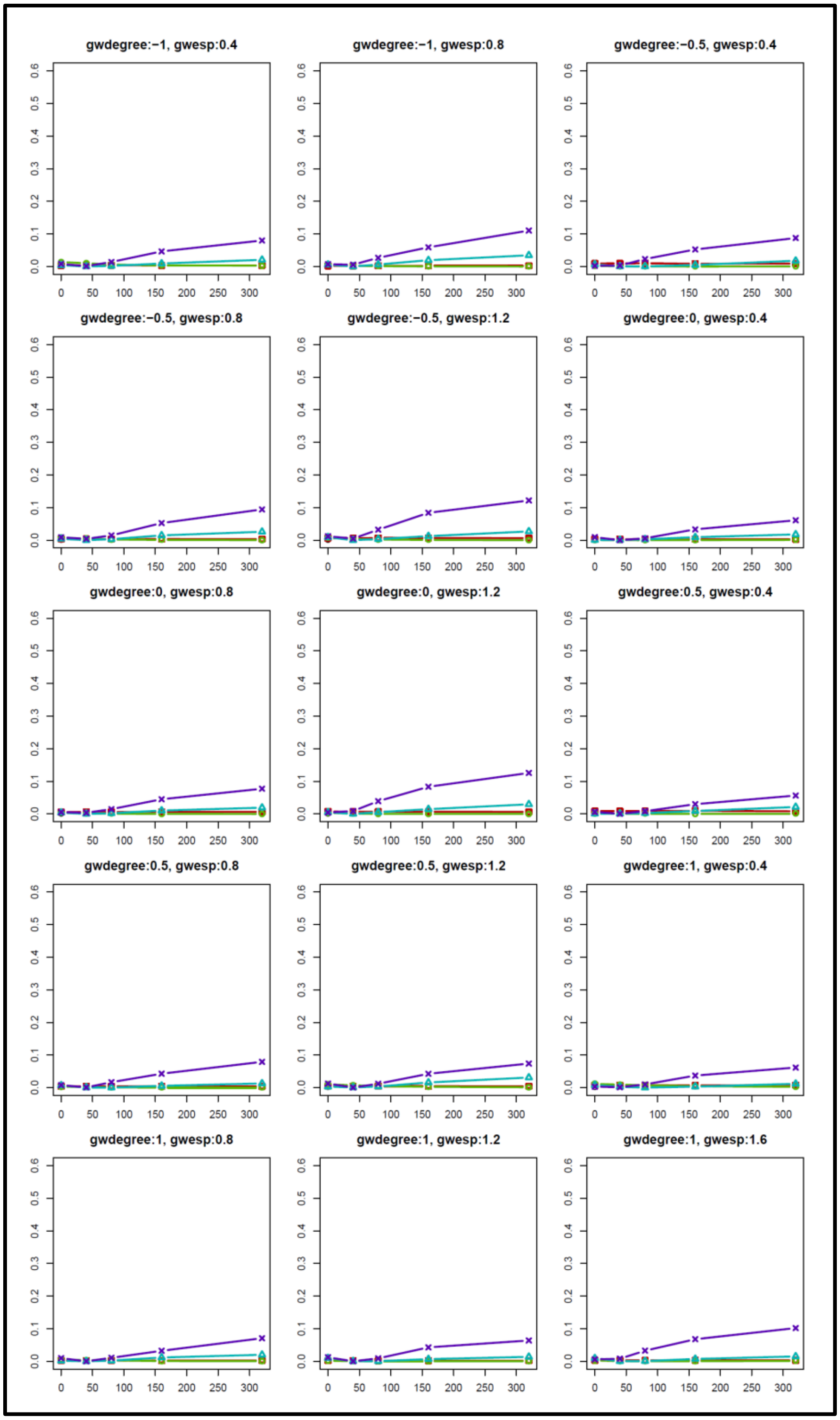

Figure 15. MSE Results for Individual-Effects Parameter for Multiple Triangulation and Popularity Parameter Pairings (Misspecified Networks)

5.3 Summary

This chapter details a simulation study design to analyze the effect of misspecified social networks on social influence choice model estimation. A simulation study of random additions and omissions of social ties is performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks. The use of small-world networks in social influence choice models was studied using ERGMs. Additionally, a perturbation protocol was used where ERGM simulations generated new graphs in which to test model estimation.

Results showed that network shape had no impact on estimator bias and variability nor did it impact model selection. Social influence parameter estimates began to lose accuracy and model fit reduced after about 15% to 30% of network ties were changed. Individual-level effects parameter estimates remained unbiased as social influence parameter estimates became biased towards zero.

The results in this simulation study are specific to the network density in the simulations. Derivation of analytical results would strengthen such analysis. Additionally, future work should perform more extensive analysis of the effect of density on model estimation.

Chapter 6: Sampling Bias in Discrete Choice Models of Social Influence

Table 18. Chapter 6 Summary

Background & Brief Summary	This chapter details a simulation study design to analyze the effect of network shape and sample size on social influence choice model estimation. A simulation study of egocentric sampling is performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks.
Motivation	Social network data is often collected by various sampling techniques in the travel behavior field; yet there is limited information on how these sampling techniques affect social influence choice model estimation. Gaining an understanding of the estimator properties for these models can guide methodologist on ways to handle possible sampling bias as well as aid applied modelers in understanding potential pitfalls in their analyses.
Results	1. Network shape was found to not impact estimator bias and variability nor did it impact model selection 2. The strength of social influence had the largest impact for a given sample size with smaller strength of social influence needing larger samples before model fit tests worked consistently 3. Parameter estimates exhibited similar patterns in regards to bias and variability
Limitations	1. Although egocentric sampling is the primary technique used in transportation, snowball sampling is also used in other social interaction work 2. The results are specific to the network density in the simulations. An exhaustive analysis of the effect of density was not undertaken.

6.1 Problem Description

Assume that we have a target population of individuals N and we have a sampling strategy, $\mathcal{S}(N) \rightarrow S$ such that $S \subseteq N$, that samples individuals from the target population. In a choice modeling exercise, the model is assumed to have a similar specification to the true data generating process (or decision process). For a binary choice, an individual *n* must choose between performing an action $y_n = +1$ or not performing that action $y_n = -1$. For a binary logit model of conformity and utility maximization, this data generating process may have the following form[26]:

$$\mathcal{U}_n^{(t)} = \beta x_n^{(t)} + \delta \sum_{q \in g(n)} \frac{y_q^{(t-1)}}{\|g(n)\|} + \varepsilon_n^{(t)}, \qquad \forall n \in N$$
$$y_n^{(t)} = \begin{cases} +1 & if\ \mathcal{U}_n^{(t)} \geq 0 \\ -1 & if\ \mathcal{U}_n^{(t)} < 0 \end{cases} \tag{32}$$

When the modeler estimates a model on the sample population, the following model if estimated:

$$\hat{\mathcal{U}}_s^{(t)} = \hat{\beta} x_s^{(t)} + \hat{\delta} \sum_{q \in \tilde{g}(s)} \frac{y_q^{(t-1)}}{\|\tilde{g}(s)\|} + \hat{\varepsilon}_s^{(t)}, \qquad \forall s \in S$$
$$y_s^{(t)} = \begin{cases} +1 & if\ \mathcal{U}_s^{(t)} \geq 0 \\ -1 & if\ \mathcal{U}_s^{(t)} < 0 \end{cases} \tag{33}$$

where:

$\tilde{g}(s)$	$\equiv$	the observed/measured social contacts of sampled individual *s* (may be different than that individual's true social contacts)
$\hat{\beta}, \hat{\delta}$	$\equiv$	estimates of the corresponding model parameters β, δ

6.2 Simulation Study – Egocentric Data

A simulation study will be described in this section to look at the impact of egocentric sampling on binary logit model estimation of a conformity model with small-world

[26] The decisions are mapped to {-1,+1} rather than the more familiar {0,1} to simplify the model specification. In this way, the utility difference between doing and not doing an action can take a simpler form in the specification of the endogenous social influence term.

networks. This section will begin with the intuition which drives the hypotheses. It also includes a description of the methodology used for the simulation study and a description of how small-world networks were generated. It concludes with the results of this simulation study.

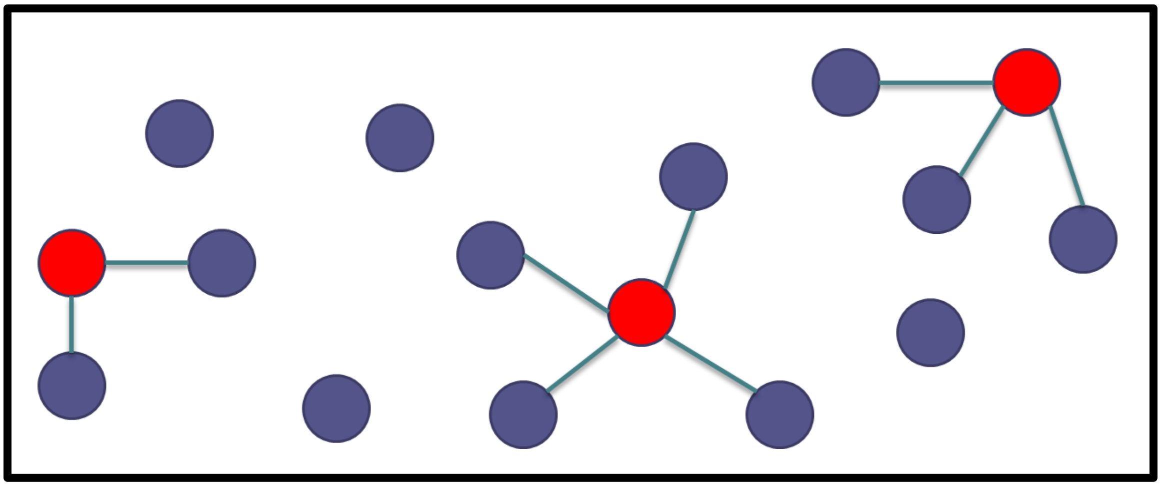

Figure 16. An Example of an Egocentric Sample

6.2.1 Problem Exploration

To gain some intuition into this problem, this section will work thorough a simplified example of a population and sample. Assuming a population N of homogeneous agents with individual-level characteristics $x_n = x, \forall n \in N$ and a social network G that connects the individuals in the population, the choice generation process is a utility maximizing framework with each individual's utility as follows:

$$\mathcal{U}_n^{(t)} = \beta_0 + \beta_1 x + \delta \sum_{q \in g(n)} \frac{y_q^{(t-1)}}{\|g(n)\|} + \varepsilon_n^{(t)}, \qquad \forall n \in N$$
$$y_n^{(t)} = \begin{cases} +1 & if\ \mathcal{U}_n^{(t)} \geq 0 \\ -1 & if\ \mathcal{U}_n^{(t)} < 0 \end{cases} \tag{34}$$

For an egocentric sampling technique, the population is sample by individual with each sampled individual's social contacts known as well as their corresponding behaviors. Therefore for each individual in the sample, their estimated utility is as follows:

$$\hat{\mathcal{U}}_s^{(t)} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\delta} \sum_{q \in \tilde{g}(s)} \frac{y_q^{(t-1)}}{\|\tilde{g}(s)\|} + \hat{\varepsilon}_s^{(t)}, \qquad \forall s \in S \subseteq N \tag{35}$$

For any given individual in the population, their estimated utility has an equivalent form and equivalent explanatory variables (due to the assumption that all contacts are observed fully and accurately). Due to this, the properties of the estimator likely are dependent on the estimation methodology used and less on the properties of the social network. The intuition for this simple example motivates the general methodology which centers on changing the sample size for estimation. As the sample size decreases, the properties of the maximum likelihood estimator are less well-defined[27] so the estimators should become less accurate (i.e. generates more bias or variability/inefficiency).

6.2.2 Simulation Methodology

In this simulation study, an egocentric sampling strategy will be analyzed where the sampled individuals are asked about their direct social contacts and the behavior of those contacts. In this Monte Carlo simulation, it will be assumed that individuals can correctly recall all relevant social ties and accurately know the behavior of their contacts.

The steps undertaken for this simulation study are as follows:

1. Initialize a population (100) with individual-specific characteristics, $x_n \sim Normal(0,1)$

[27] Maximum likelihood estimation properties are generally valid as the sample size increases to infinity.

2. Give the population some initial cycling choices, $y_n^{(t=0)} \sim B(0.4)$
3. Generate a social network to connect the population with influential others
4. Individuals make new cycling choices depending on social influence, $\mathcal{U}_n^{(t=1)} = -0.2 + 1.2x_n + \delta \bar{y}_m^{(t=0)} + \varepsilon_n$
5. Apply the sampling technique $\mathcal{S}$ & estimate a model on this "observed" data using maximum likelihood estimation
6. Perform a Likelihood Ratio Test and calculate the mean squared error (MSE) of the parameter estimates
7. Repeat steps (1) through (6) 250 times

6.2.3 Data – Generating "Real Networks"

The same small-world network generation process is used as in section 5.2.4 as given by equation (30).

6.2.4 Study Hypothesis

In this study, the following hypotheses are made:

1. Network structure will have no impact on the model estimation. This will be due to the design of the data generating process which only considers the behavior of direct contacts from the preceding time period[28]. Since it is assumed that the contacts of an individual and their behaviors are properly recorded, all relevant

[28] This may be an area for future research into the misspecification of the time period. For example, if we estimate the model at $t = 2$ rather than $t = 1$, but still use the behavior of direct contacts in time period $t = 0$. Then the utility and possibly behavior of an individual in $t = 2$ will be impacted by the behavior in $t = 1$ of a direct contact. This contact's behavior was affected by that contact's contacts' behaviors in $t = 0$. Thus the behavior of the individual could be affected by behavior of his contacts' contacts.

explanatory variables will be the same between the data generating process and the estimated model.

2. Sampling rate (sample size-to-population size) will have a negative impact on the accuracy of the likelihood ratio test and increase the variability in the social influence parameter estimates. As the sampling rate decreases, the accuracy will decrease and the variability will increase. Because a maximum likelihood estimator is used, the small sample properties of this estimator are not equivalent to its large sample properties. The sample sizes in this study will be less than 100 individuals.
3. The social influence parameter estimates will be unbiased but will increase in variability as the sample rate decreasing.
4. The individual-level effects parameter and constant parameter will be unbiased. There is no correlation between the network structure and individual-level characteristics.

6.2.5 Results – Triangulation and Popularity

The likelihood ratio test is used to test a restricted model (without social influence) against an unrestricted model (with social influence). The likelihood ratio statistic has a chi-squared distribution and the assumed confidence level is 95 percent ($\alpha = 5$). Using the definition of a confidence level for a likelihood ratio statistic, when social influence truly occurs in the population, then in $(100 - \alpha)\%$ confidence level for a likelihood ratio statistic, when social influence truly occurs in the population, then the likelihood ratio test should accept the null hypothesis in α out of 100 sample runs.

Thus in simulations where the true data generation process includes social influence, it should be expected that $(100 - \alpha)\%$ of the simulation runs should include a likelihood ratio test indicating the rejection of the null hypothesis. For the simulation runs where there is no social influence in the true data generation process, only $\alpha\%$ of the runs should indicate rejection of the null hypothesis according to the likelihood ratio test.

Figure 17 shows expected results for an experiment with a sample of social networks with (gwdsp, gwesp) parameters equal to (-0.4, 0.9) and varying strengths of the social influence parameter, $\delta = \{0,1,2,3\}$.

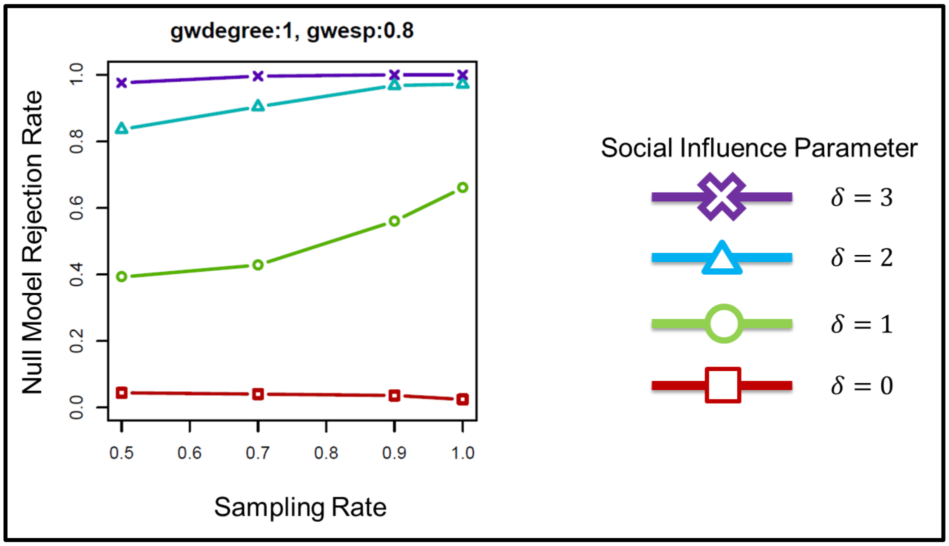

Figure 17. Expected Results for Likelihood Ratio Test on Egocentric Sampling

The hypothesized results occur for all parameter pairing shown in Figure 18. For estimation when $\delta = 0$, the likelihood ratio test was expected to remain steady at about 0.05 null model rejections. The results show that this expectation is maintained as triangulation and popularity varies.

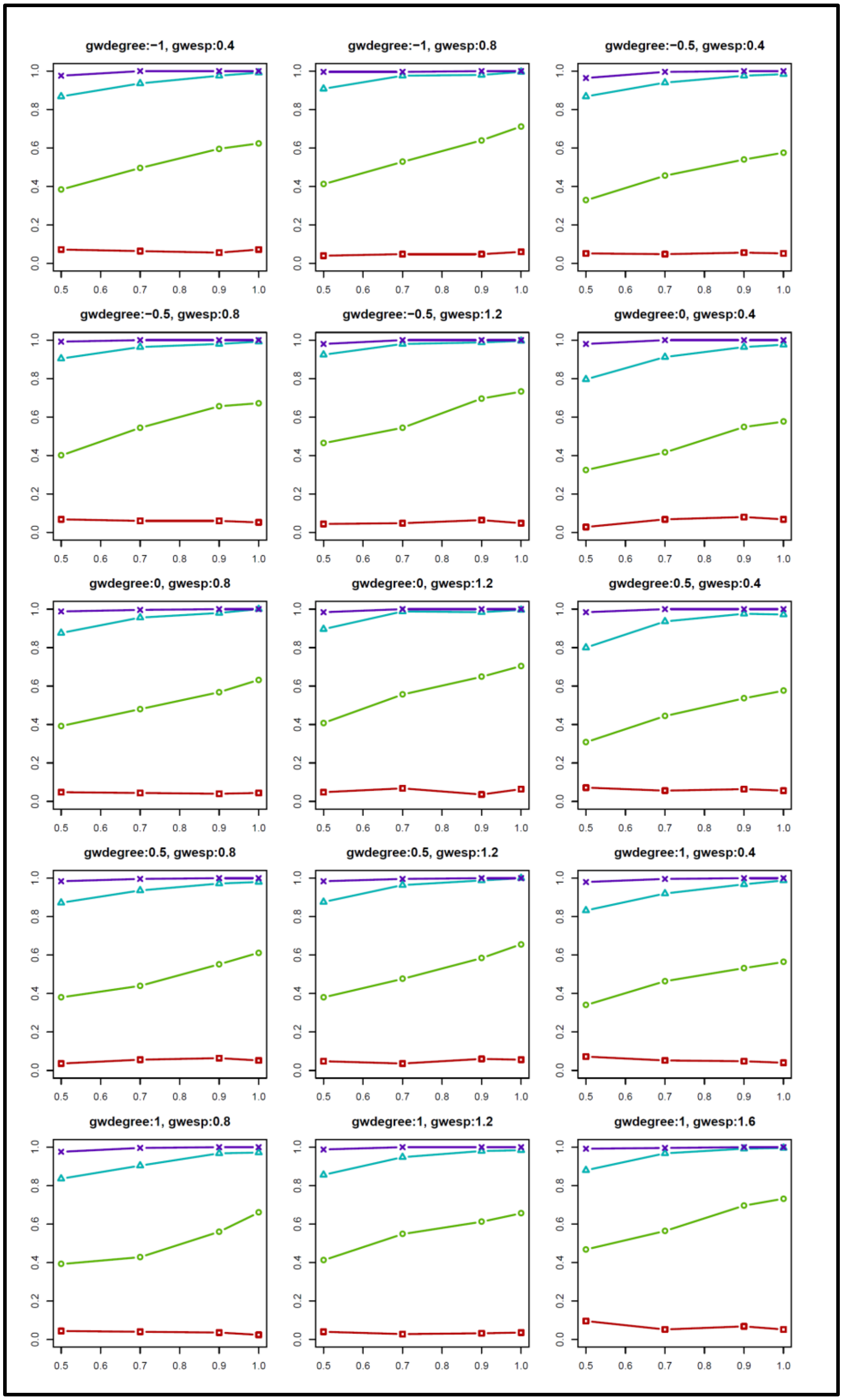

Figure 18. Average Likelihood Ratio Test Results for Triangulation and Popularity Parameter Pairings (Egocentric Sampling)

Additional analysis explored if this behavior is experienced while varying two additional factors in the simulation study design: (1) Initial Conditions and (2) Individual-level and constant parameters. There were concerned that these two factors may be correlated in this proof-of-concept study. Specifically, the initial choice tends towards behavior $y_{ni} = -1$ as well as the negative constant parameter. It was found that varying the initial conditions for population-wide average behavior from 10% to 90% did not produce an appreciable difference in estimator behavior.

To check the bias and variability in the social influence parameter estimates, the mean squared error of the estimates were calculated and plotted (Figure 20). Typical results are included in Figure 19.

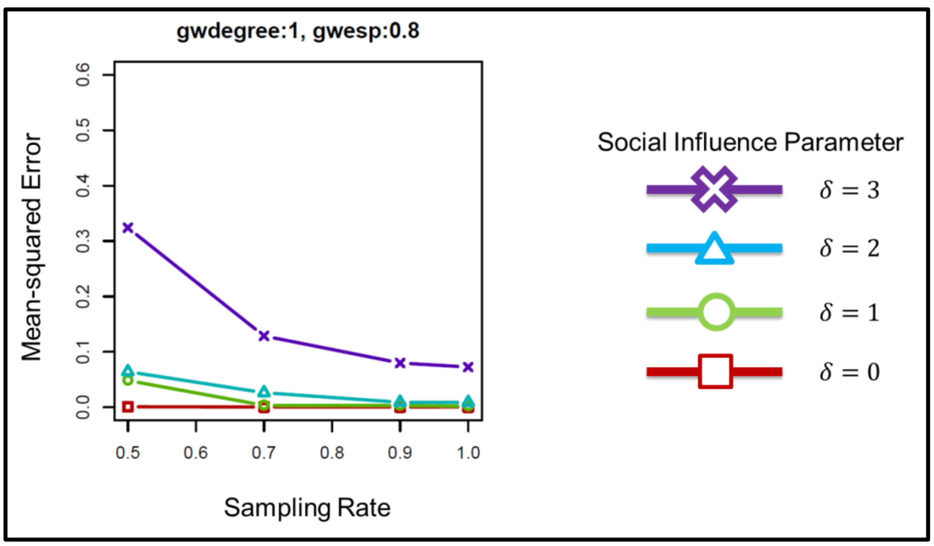

Figure 19. Expected Results for MSE (Egocentric Sampling)

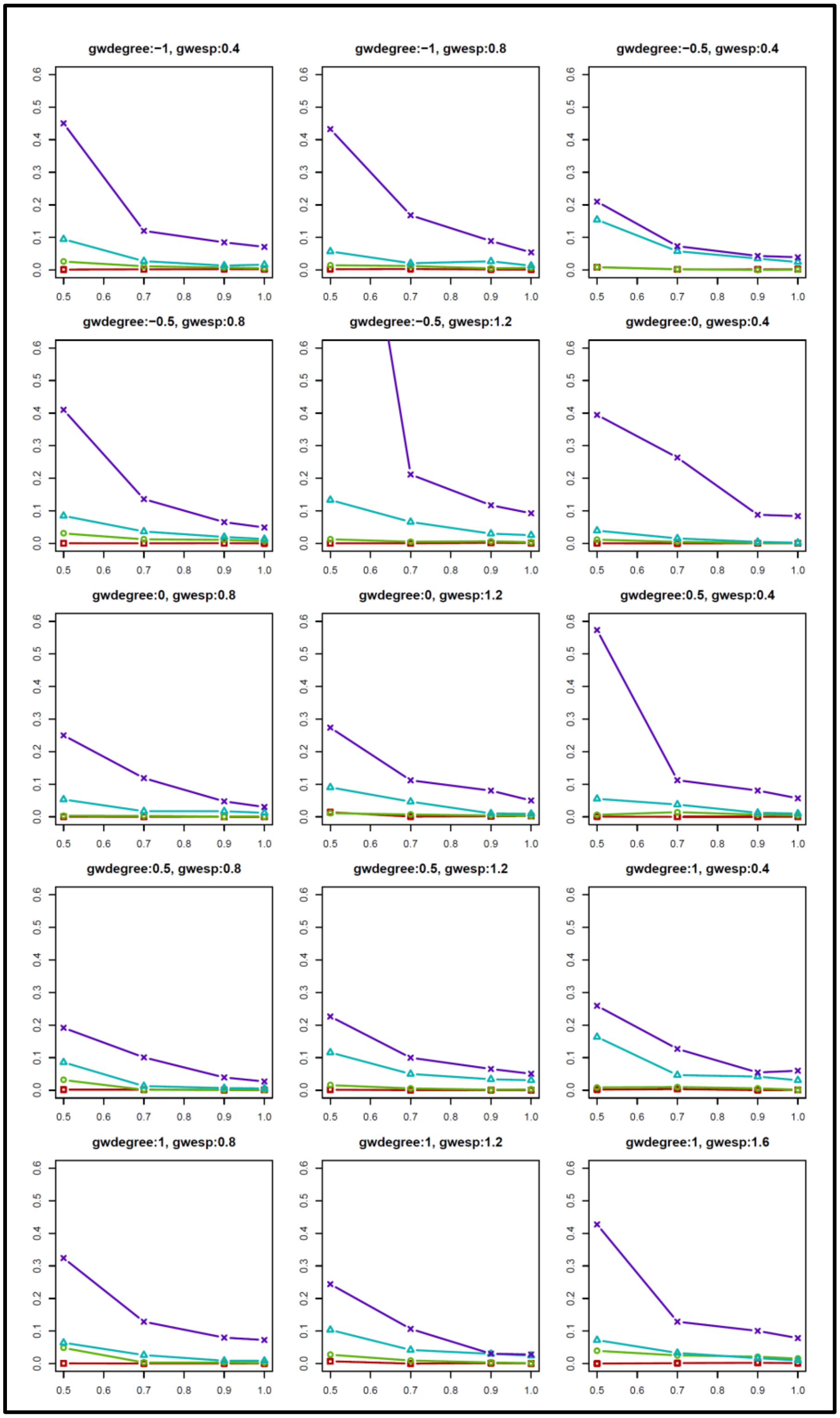

Figure 20. Social Influence Parameter MSE Results for Triangulation and Popularity Parameter Pairings (Egocentric Sampling)

As expected, MSE increased with decreasing sample size. There tended to be a disproportionate increase in the MSE for simulations where the true social influence parameter were highest, $\delta = 3$. There were a small number of estimates that were an order of magnitude higher than 3. This skewed the MSE heavily since the MSE is calculated on the square of the difference between the observed and true parameter values. This is due to the small sample sizes.

6.3 Summary

This chapter details a simulation study design to analyze the effect of network shape and sample size on social influence choice model estimation. A simulation study of egocentric sampling is performed to measure the accuracy of the likelihood ratio test and the biasness and variance in social influence parameter estimates for binary choice models of conformity with small-world social networks.

Results showed that network shape did not impact estimator bias and variability nor did it impact model selection. The strength of social influence had the largest impact for a given sample size. For smaller strengths of social influence, larger samples were needed before model fit tests worked consistently. At a density level of 0.05, estimation with social influence parameter strengths of 1.00 were not achieving their expected rate of model acceptance for samples sizes of 100 individuals or less.

Parameter estimates were found to increase in biasness and variability as sample sizes decreased. This effect was strongest – even after normalizing MSE values – for simulations with high levels of social influence.

The results in this simulation study are specific to the network density in the simulations. Derivation of analytical results would strengthen such analysis. Additionally, future work should perform more extensive analysis of the effect of density on model estimation. Until more extensive and generalizable knowledge is available, it is suggested that modelers perform similar tests of the sensitivity of their models to changes in social influence parameter strength and sample size before design data collection efforts and during their model building process.

Chapter 7: Generalized Behavioral Framework for Social Influence Choice Models

Table 19. Chapter 7 Summary

Background & Brief Summary	This chapter showcases a generalized behavioral framework for choice models of social influence. This framework stresses the important interconnection between the social influence mechanism and associated social networks and influence sources. The current state-of-practice in travel behavior studies of social influence is described as well as the limitations of current approaches. Then theories of social influence via conformity and compliance are described, plus theories of social network formation and structural properties of social networks are also described.
Motivation	Social influence choice models incorporate theories and terminology from a variety of social science fields. Additionally, various model specifications using during social network specifications, influence sources, and social influence types and processes have been developed. Understanding how social factors are incorporating into choice modeling can be a confusing and daunting task for new users. The behavioral framework developed in this chapter seeks to consolidate the current state-of-the-art in a clear and easy to understand format. This will clarify areas for improvement and future research topics.
Results	1. Developed a behavioral framework to describe social influence choice models 2. Summarized the existing travel behavior literature and suggests gaps in behavioral realism and model applicability 3. Suggested behavioral enhancements to social influence processes using individuals' motivations as an example 4. Described the importance of social networks in the social influence choice modeling context, specifically emphasizing the importance of network formation and structure 5. Recommended new research for model development and application
Limitations	1. An exhaustive search of the social influence literature in the social sciences was not undertaken, but the framework is flexible enough to handle other social influence theories

7.1 Framework

7.1.1 Historical Background

Conceptually, Manski (1993, 1995) outlines three different ways in which similarities in group behavior can be explained in a model, namely[29]:

- **Endogenous Social Influence Effects**, "*wherein the propensity of an individual to behave in some way varies with the prevalence of that behavior in the group*";
- **Contextual Social Influence Effects**, "*wherein the propensity of an individual to behave in some way varies with the distribution of exogenous background characteristics in the group"*; and
- **Correlated Individual-level and Correlated Environmental Effects**, "*wherein individuals in the same group tend to behave similarly because they face similar institutional environments [(environmental)] or have similar unobserved individual characteristics [(individual-level)]"*

Endogenous and contextual social influence effects characterize the relevance of group level behavior and group level characteristics respectively for individual behavior. An important distinction between these two specifications however, is that endogenous social influence effects allow for the possibility of direct feedback between individual behavior and group level behavior, which can potentially be reinforcing over the course of time. Contextual social influence effects, while social, are presumed (at least short-term) not to involve direct behavioral feedback between the individual and others. In contrast,

[29] Manski refers to these effects respectively as endogenous, contextual, and correlated effects, but they are renamed here to maintain consistency with the rest of the text.

correlated individual-level and correlated environmental effects are presumed to be entirely non-social. This distinction between the different sources of similar group behavior are delineated by Manski in a measurement context as Blume et al. (2011) explains: “the current generation of social interactions models focuses on a relatively crude division of social interactions between factors that are predetermined [(contextual / exogenous) and those that are contemporaneous [(endogenous)]” (p. 941).

7.1.2 Framework Description

The framework described in this paper approaches the subject from a behavioral and microfoundations perspective rather than through this measurement and macro-level perspective. Remaining within the discrete choice modeling framework is still useful in this context as these models lend themselves well to linking behavioral theory and statistical modeling (Durlauf and Ioannides 2010). This is due to the latent variable derivation of the payoff which can represent a theoretical quantity that can be minimized or maximized.

The framework, shown in Figure 21, rests on the assumption that the individual makes choices according to a decision rule that depends on some aspects specific to the individual and a social component which depends on the social systems (actual or perceived) surrounding the individual. The backbone of the framework is the traditional discrete choice model with its focus on individual-level effects generated from individual characteristics and properties of the individual’s environment. Individual-level characteristics, environmental factors, and non-choice related social factors (i.e.

exogenous influence sources) are assumed to be exogenous to the individual’s decision process (denoted by gray in Figure 21).

Individuals are connected to one another through the social networks in their lives. These networks, which may have structures formed by self-selection due to individual characteristics, provide a reference to society through which social influence occurs. Social influence is a function of an individual’s social networks. This is an important part of this framework, as it is an explicit acknowledgment of the importance of the social network. Different social networks may imply the use of different social influence mechanisms as well as different influence sources – endogenous or exogenous. Different choice contexts may imply the use of different social networks, e.g. mode choice may imply the use of co-worker networks whereas social trips would use friendship networks. Additionally, social networks can vary between individuals in their structure and the relationships between individuals.

When environmental factors and individual characteristics are correlated with an individual’s social network (i.e. an individual’s social contacts share the same environment or have similar unobserved characteristics), they become correlated environmental and individual-level effects, respectively, which can seem social when measured but truly are behaviorally non-social. A similar correlation can occur between the influence sources and the social networks. This can manifest in homophily of behaviors, attitudes, and values where individuals are connected to each other because they prefer to be around similar others.

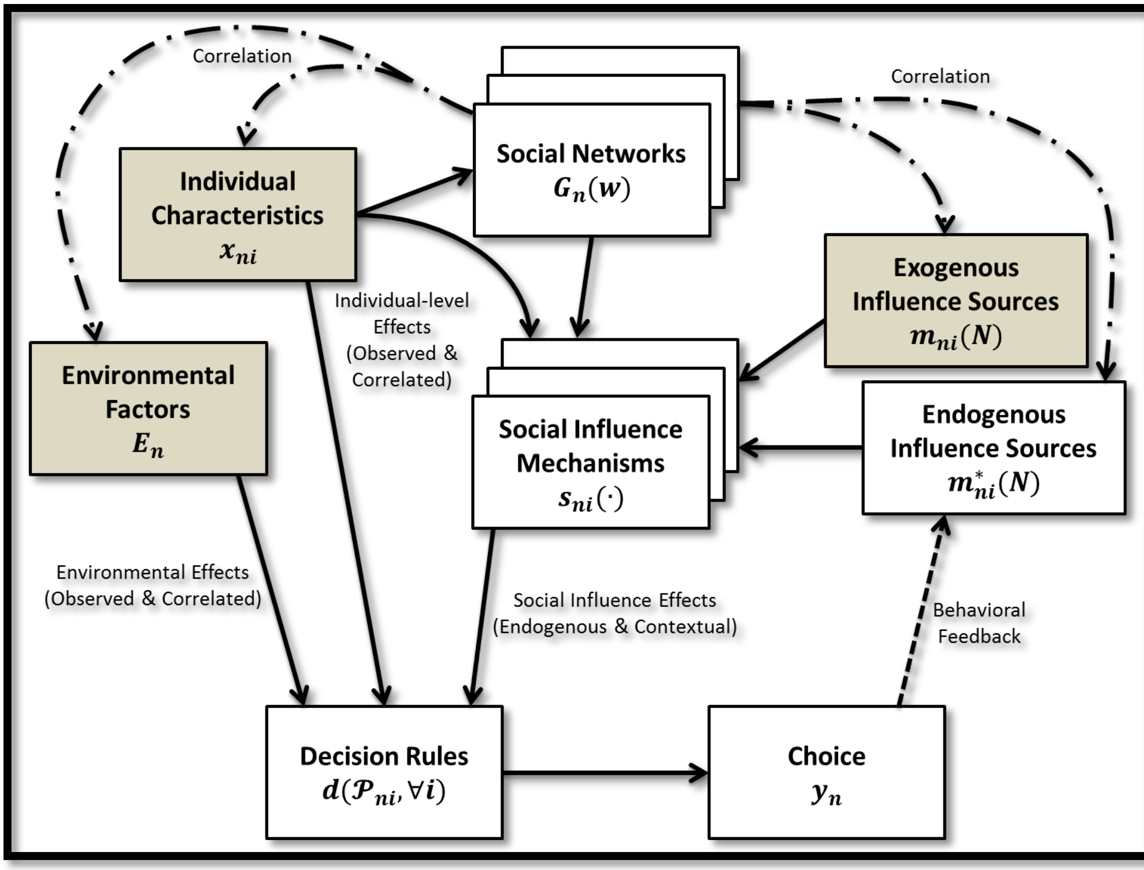

Figure 21. Generalized Framework for Choice Models of Social Influence

From the combination of individual-level, environmental-level, and social influence effects, individual *n* obtains some payoff $\mathcal{P}_{ni}$ when choosing an alternative *i*. This payoff is a function of various effects at the individual, environmental, and social influence level. Assuming a linear-in-parameter form, the payoff function takes the following form:

$$\mathcal{P}_{ni} = \beta_i x_{ni} + \psi_i s_{ni}\big(G_n(w), m_{ni}(N), m^*_{ni}(N)\big) + \mu_i E_n + \varepsilon_{ni} \tag{36}$$

where:

x_{ni} ≡ individual-level characteristics of individual *n* for alternative *i*

$s_{ni}(\cdot)$ ≡ social influence mechanisms for individual *n* for alternative *i* due to endogenous and contextual factors

$G_n(w)$ ≡ individual *n*'s social contacts and the strength of these relationships (modeled through a weighting function *w*)

$m_{ni}(N)$ ≡ exogenous social influence sources of the population on individual *n* for alternative *i*

$m_{ni}^*(N)$ ≡ endogenous social influence sources of the population on individual *n* for alternative *i*

N ≡ the population of all individuals

E_n ≡ environmental factors on individual *n* (may include correlated environmental factors)

ε_{ni} ≡ unobserved effects on individual *n* for alternative *i* (includes correlated individual-level effects and alternative-specific unobservables)

β_i, ψ_i, μ_i ≡ model parameters (these can be alternative-specific)

This formulation can be expanded to separate the endogenous and contextual social influence effects as follows:

$$\mathcal{P}_{ni} = \beta_i x_{ni} + \gamma_i k_{ni}\big(G_n(w), m_{ni}(N)\big) + \delta_i l_{ni}\big(G_n(w), m_{ni}^*(N)\big) + \mu_i E_n + \varepsilon_{ni} \tag{37}$$

where:

$k_n(\cdot)$ ≡ contextual social influence mechanisms for individual *n* for alternative *i* due to contextual factors

$l_n(\cdot)$ ≡ endogenous social influence mechanisms for individual *n* for alternative *i* due to endogenous factors

γ_i, δ_i ≡ model parameters (these can be alternative-specific)

The individual chooses an alternative by evaluating the payoffs from each alternative according to a decision rule, $d(\mathcal{P}_{ni}, \forall i) \to y_n$.

7.1.3 Comparison to Prior Work

The framework in this paper provides a behavioral basis for social influence choice models. Previous work defined endogenous and contextual effects only "in terms of [the] types of variables rather than via particular mechanisms" (Blume et al. 2011, p. 941). The framework contrasts with previous works which classified on structural terms such as Manski's initial work (1993, 1995) on linear models which was primarily concerned with conformity based on actual behavior and Brock and Durlauf's (2001, 2003, 2006, 2007) extension to binary and multinomial choice models of conformity based on perceptions of behavior and rational expectations with complete information. The framework in this paper emphasizes this behavioral rather than measurement focus by:

1. Explicitly mentioning the importance of social networks and its part as a function of social influence processes,
2. Consolidating endogenous and contextual social effects into a single concept of a social influence mechanism which depends on endogenous and exogenous influence sources respectively,
3. Generalizing influence sources beyond observed or perceived choices,
4. Allowing for heterogeneity in social influence and social networks (Roy et al. 2012) since both may vary depending on characteristics of the individual, and
5. Generalizing the decision rule space beyond utility maximization.

7.2 State of Practice in Transportation

Travel behavior research analyzes social influence through applied inferential analyses, agent-based simulations, and experiments. The primary behavioral paradigm in discrete choice models of transportation is random utility maximization where an individual chooses the alternative which gives that individual the most utility. Two forms of social influence mechanisms have been used in travel behavior models: conformity (an endogenous social influence mechanism) and compliance (a contextual social influence mechanism). These models have the following form for the utility $\mathcal{U}_{ni}$ an individual *n* obtains from choosing alternative *i* and a utility maximizing decision rule[30]:

$$\mathcal{U}_{ni} = \beta_i x_{ni} + \gamma_i k_{ni}(G_n(w), m_{ni}(N)) + \delta_i l_{ni}\big(G_n(w), m_{ni}^{*}(N)\big) + \varepsilon_{ni}$$
$$y_{ni} = \begin{cases} 1 & if\ U_{ni} = \max_{j \in C} U_{nj} \\ 0 & otherwise \end{cases} \tag{38}$$

7.2.1 Specific Works

Applying the generalized framework to prior work provides a sample taxonomy for describing social influence models of discrete choice. Table 20 summarizes many social influence models used in studies of travel behavior by classifying each according to the social networks, social influence mechanism and sources, and decision rule used in the study. From this classification, we can see certain patterns emerge in prior research. Social influence models in transportation are primarily models of conformity rooted in utility maximization. The modelers tend to use social network structures that are either:

[30] Environmental effects are not included since models in transportation generally ignore correlated environmental effects. Also, identification issues arise in cross-sectional models with correlated environmental (Brock and Durlauf 2007).

(1) large cliques of individuals joined by similar demographics or spatial proximity or (2) sparse networks of intimate social connections. Since data collection tends to be cross-sectional, influence sources generally are based on current behavior and are from in-sample connections. Few studies elicit data from the respondent on their social networks and the behavior of these social contacts nor do they get information directly from their social contacts.

The following examples describe three unique models using the framework's taxonomy:

- Dugundji and Walker (2005) analyze mode choice using utility maximization via mixed logit models. Social influence occurs through conformity based endogenously on current behavior in multiple large cliques that are based socially on income and spatially on postal code. This was the first major paper in the social influence and travel field and its use of cliques and conformity was emulated in other work.

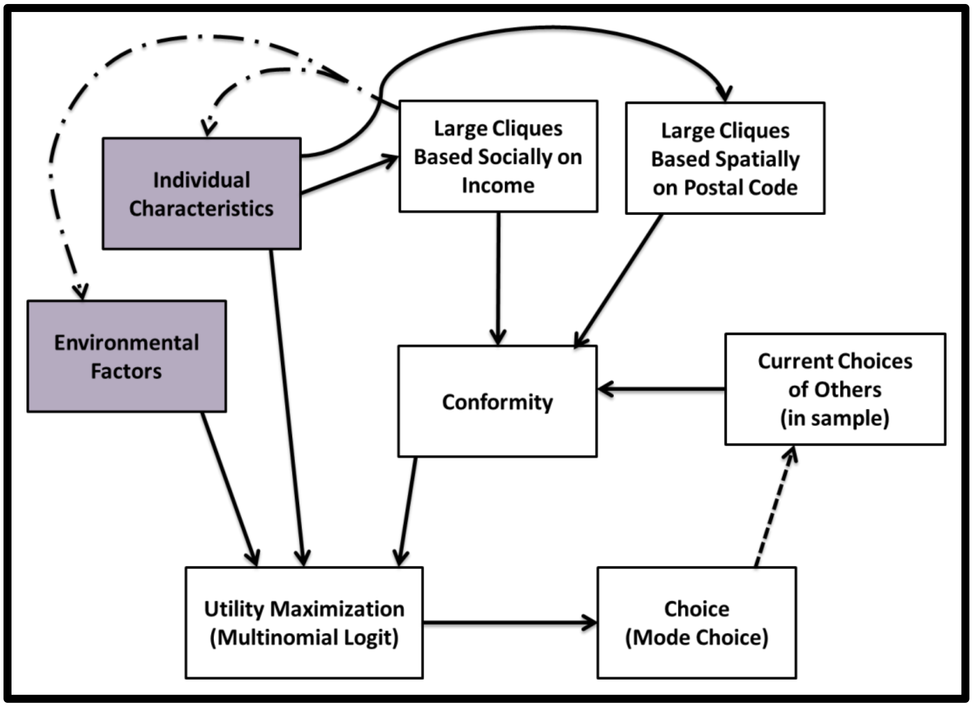

Figure 22. Dugundji and Walker (2005) Model Specification

- Abou-Zeid and Ben-Akiva (2012) analyze mode choice using utility maximization via hybrid choice models (with latent variables). Social influence occurs through conformity based endogenously on social comparison. The individual compares their commute trip to the commute of an acquaintance with a similar commute. This paper provided a unique social influence source that contrasts with the common sources of past and current behavior.
- Goetzke and Weinberger (2012) analyze mode choice utility maximization via a binary probit model. Social influence occurs through conformity based endogenously on current behavior of large cliques spatially based on census tracts. Additionally, the authors theorize that social influence occurs contextually through social norms based on measures of the aggregated demographics of

individuals in the census tract. This paper was unique in its use of contextual variables from an outside dataset.

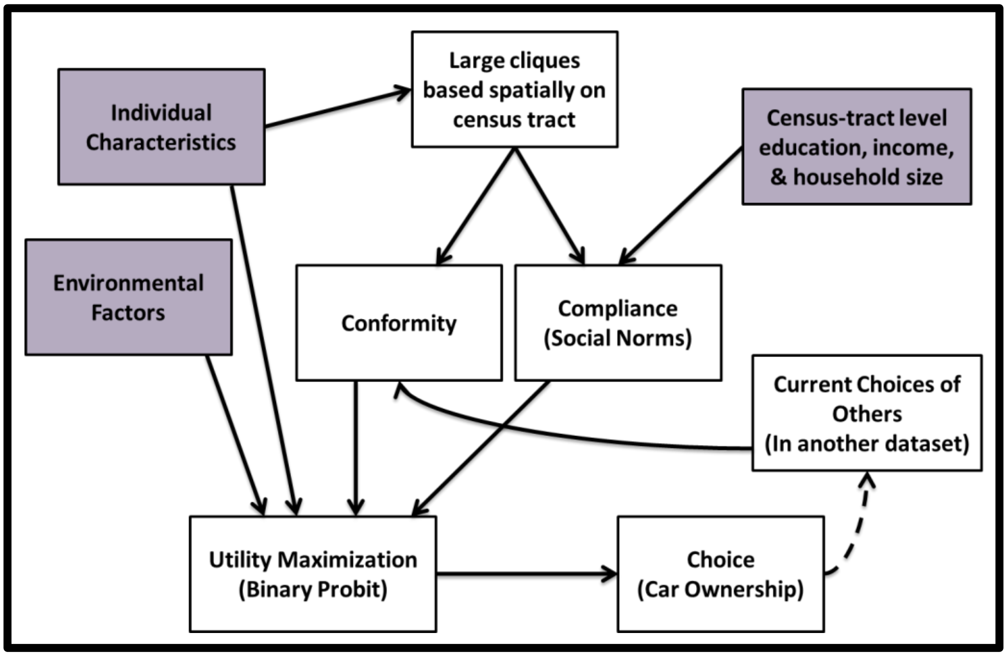

Figure 23. Goetzke and Weinberger (2012) Model Specification

Table 20. Discrete Choice Models of Social Influence in Travel Behavior Research

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
Abou-Zeid & Ben-Akiva (2011)	Mode Choice	An acquaintance (respondent-reported) with a similar commute	Conformity (Social Comparison)	Current Behavior and Social Comparison Indicators	Utility Maximization (Hybrid Choice Model with Latent Variables)
Adjemian, Lin, & Williams (2010)	Vehicle Ownership	Nearest neighbors based spatially within 8 km	Conformity	Past Behavior / Lagged Behavior (in sample)	Utility Maximization (Binary Logit)
Baltas & Saridakis (2013)	Vehicle Type Choice	Egocentric network of friends, relatives, & acquaintances; not explicitly measured	Compliance or Conformity (Informational)	Whether Information was Sought before Purchase	Utility Maximization (Multinomial Logit)
		Access to car TV shows, car literature, car salesmen, & brochures	Compliance (Affect & Arousal)	Whether Information was Sought before Purchase	
Dugundji (2012)	Mode Choice	A large clique of everyone	Conformity	Current Behavior (in sample)	Utility Maximization (Nested Logit)
Dugundji & Gulyas (2003)	Theoretical	Random sparse networks (Poisson/Erdos-Renyi & Watts-Strogatz graphs)	Conformity (Dynamical Systems)	Past Behavior (from previous simulation timestep)	Utility Maximization (Binary Logit)
Dugundji & Gulyas (2008)	Mode Choice	[1] Large cliques based spatially on residential district	Conformity (Dynamical Systems)	Current Behavior (in sample) & Past Behavior (from previous simulation timestep)	Utility Maximization (Multinomial Logit, Nested Logit)
		[2] Large cliques based socially on age, income, & education	Conformity (Dynamical Systems)		

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
Dugundji & Gulyas (2012a)	Mode Choice	Random sparse networks (Poisson/Erdos-Renyi graphs), equal probability of linking between all individuals	Conformity (Dynamical Systems)	Past Behavior (from previous simulation timestep)	Utility Maximization (Binary Logit)
Dugundji & Gulyas (2012b)	Mode Choice	[1] Large cliques based spatially on residential district	Conformity (Dynamical Systems)	Current Behavior (in sample) & Past Behavior (from previous simulation timestep)	Utility Maximization (Nested Logit)
		[2] Large cliques based socially on age, income, & education	Conformity (Dynamical Systems)		
Dugundji & Gulyas (2013)	Mode Choice	[1] Large cliques based spatially on residential district or postal code	Conformity (Dynamical Systems)	Current Behavior (in sample) & Past Behavior (from previous simulation timestep)	Utility Maximization (Multinomial Logit & Nested Logit)
		[2] Large cliques based socially on age, income, & education	Conformity (Dynamical Systems)		
Dugundji & Walker (2005)	Mode Choice	[1] Large cliques based spatially on residential district or postal code	Conformity	Current Behavior (in sample)	Utility Maximization (Mixed Cross-Nested Logit)
		[2] Large cliques based socially on age, income, & education	Conformity	Current Behavior (in sample)	
Fukada & Morichi (2007)	Bicycle Parking	Large clique based spatially on bicyclists who share a railway station	Conformity (Normative, Dynamical Systems)	Current Behavior (in sample)	Utility Maximization (Binary Logit)
[1] Gaker, Zhang, & Walker (2011)	Vehicle Ownership	A directed four-mode network of temporally-lagged participants in an economics experiment	Conformity (Information Cascade)	Past Behavior (in experiment)	Utility Maximization (Multinomial Logit)

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
[2] Gaker, Zhang, & Walker (2011)	Pedestrian Crossing Behavior	[1] Hypothetical source of State Law	Compliance (Authority & Obedience)	Shown current law (stated in choice experiment)	Utility Maximization (Multinomial Logit)
		[2] Hypothetical source of Citation Rates and Fine Penalties	Compliance (Authority & Obedience)	Past Actions (citation rates and fine amount for Jan. 2009, stated in choice experiment)	
		[3] Large clique of all University students & staff	Conformity (Informational)	Past behavior (out of sample, stated in choice experiment)	
		[4] Hypothetical source of Accident Statistics	Compliance (Affect and Arousal)	Shown statistics (stated in choice experiment)	
Goetzke (2008)	Transit Mode Choice	Nearest sampled neighbors based spatially (≤ 40 individuals within 1.2 km); weighted equally	Conformity	Current behavior (in sample)	Utility Maximization (Binary Logit, Conditional Autoregressive)
Goetzke & Andrade (2010)	Walking Mode Choice	Nearest three sampled neighbors based spatially; weighted by spatial distance	Conformity	Current behavior (in sample)	Utility Maximization (Binary Logit)
Goetzke & Rave (2011)	Bicycle Mode Choice	Large cliques based spatially on municipality	Conformity	Current behavior (in sample, instrumented)	Utility Maximization (Binary Logit)
Goetzke & Weinberger (2012)	Vehicle Ownership	[1] Large cliques based spatially on census tract	Conformity	Current Behavior (from another dataset)	Utility Maximization (Binary Probit)
		[2] Large cliques based spatially on census tract	Compliance (Social Norms)	Census-tract level education, income, & household size	

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
Grinblatt, Keloharju, & Ikaheimo (2008)	Vehicle Ownership	Nearest sampled neighbors based spatially; weighted by distance ranking (1-500 neighbors placed in rings of varying sizes)	Conformity (Informational)	Previous Behavior (in sample)	Utility Maximization (Binary Logit)
Kuwano, Chikaraishi, & Fujiwara (2013)	Personal Mobility Ownership	[1] Egocentric network of friends	Conformity	Current Behavior (stated in choice experiment)	Utility Maximization (Binary Logit)
		[2] Hypothetical large cliques based on regions	Conformity	Current Behavior (stated in choice experiment)	
Kuwano, Tsuaki, & Matsubara (2012)	Vehicle Ownership	A hypothetical single large clique of vehicle buyers	Conformity (affected classes of individuals)	Current Behavior (stated in choice experiment)	Utility Maximization (Latent Class RUM)
Kuwano, Zhang, & Fujiwara (2011)	Vehicle Ownership	[1] Large cliques based socially on income	Conformity	Current Behavior (in sample)	Utility Maximization (Dynamic GEV)
		[2] Large cliques based spatially on neighborhood	Conformity	Current Behavior (in sample)	
		[3] A large clique representing a nation	Conformity	Current Behavior (in sample)	
Páez & Scott (2007)	Teleworking Choice	Random sparse networks with links probabilistically based on homophily from a two-dimensional social lattice (9 networks generated)	Conformity (Dynamical Systems)	Past Behavior (from previous simulation timestep)	Utility Maximization (Binary Logit)
Páez, Scott, & Volz (2008)	Residential Choice	Random sparse network with links probabilistically based on varying degree distributions & clustering (24 networks generated)	Conformity (Dynamical Systems)	Past Behavior (from previous simulation timestep)	Utility Maximization (Multinomial Logit)

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
[1] Pike (2014)	Mode Choice	Egocentric network of up to five contacts (contacted within last six months)	Conformity	Current Behavior (respondent reported)	Utility Maximization (Multinomial Logit)
[2] Pike (2014)	Mode Choice	Nearest neighbors based spatially ranging from within 250 to 25,000 feet	Conformity	Current Behavior (in sample)	Utility Maximization (Multinomial Logit)
Scott, Dam, Páez, & Wilton (2012)	Teleworking Choice	[1] Egocentric network of co-workers with advice-seeking contact	Conformity	Current Behavior (respondent reported)	Utility Maximization (Multinomial Probit)
		[2] Egocentric network of co-workers without advice-seeking contact	Conformity	Current Behavior (respondent reported)	
Sidhartan, Bhat, Pendyala, & Goulias (2011)	Mode Choice	A large clique of everyone; weighted inversely to spatial distance	Conformity	Current Behavior (in sample)	Utility Maximization (Multinomial Probit)
Walker, Ehlers, Banerjee, & Dugundji (2011)	Mode Choice	[1] Large cliques based spatially on residential postal code	Conformity	Current Behavior (in sample)	Utility Maximization (Multinomial Logit)
		[2] Large cliques based socially on income	Conformity	Current Behavior (in sample)	

Paper Authors	Application Area	Social Network	Social Influence	Influence Sources	Decision Rule (Model)
Wu, Zhang, & Chikaraishi (2013)	Tourism Participation	[1] Large cliques based spatially on prefecture	[a] Conformity	Current Behavior (in sample)	Utility Maximization (Mixed Multinomial Logit)
			[b] Compliance (Social Norms)	Prefecture-level education, household size, & household income	
		[2] Large cliques based socially on occupation	Conformity	Current Behavior (in sample)	
		[3] Large cliques based socially on income	Conformity	Current Behavior (in sample)	

Note: Social networks are assumed to be undirected and weighted evenly between contacts unless otherwise stated.
Multiple entries from a single paper represent distinctively different models.

7.2.2 Discrete Choice Models of Conformity

As shown in Table 20, conformity is the primary social influence mechanism modeled in transportation. An individual conforms when that individual desires to change their own behavior to that of another person or persons (Cialdini and Goldstein 2004). Since the behavior of others may be correlated through unobservable factors and simultaneity, conformity is considered an endogenous social influence mechanism. The use of conformity in travel studies is a simple extension of previous methods because it requires no additional data collection since choice data is often collected in travel studies. Individuals also tend to be able to observe the choices of those closest to them and, sometimes, the publicly made choices of people outside of their social contacts. Therefore, an individual may have perceptions of the choices of people with similar social standing although they may not have intimate relationships with these individuals. This is also a weakness as these perceptions may be biased since individuals have limited information (e.g. limited exposure, extrapolations from limited observations, media bias) and limited cognitive abilities (Kadushin 2012).

Due to the cross-sectional nature of most travel surveys, most conformity models use an endogenous influence source based on the current behavior of peers. Most of these models use the following form to represent the utility observed for an alternative *i* during the current time period *t*:

$$\mathcal{U}_{ni}^{(t)} = \beta x_{ni}^{(t)} + \delta \sum_{q \in g(n)} \frac{y_{qi}^{(t)}}{\|g(n)\|} + \varepsilon_{ni}^{(t)} \tag{39}$$

where:

$\mathcal{U}_{ni}^{(t)}$ ≡ the utility an individual *n* obtains from choosing alternative *i* in the current time period *t*

$y_{qi}^{(t)}$ ≡ 1 if individual *q* chose alternative *i* in time period *t*; 0 otherwise

$q \in g(n)$ ≡ an individual *q* in individual *n*'s social contacts

$\|g(n)\|$ ≡ the number of individuals in individual n's group of social contacts

$\varepsilon_{ni}^{(t)}$ ≡ unobserved individual-level effects in time period *t* for alternative *i* (can include individual-level correlated effects)

This form assumes a direct-benefit effect is generated from conforming to the behavior of others (i.e. utility itself is directly increased by conforming). But conformity is not always generated by the same motivation; the question of why are people conforming often is not being answered. Are individuals transferring information? Are people envious of others and aspiring to obtain a similar status? Is this just a fad and people are just following the crowd? These motivations have important implications when the dynamical processes of behavior are analyzed – to determine the decision process between time periods and thus long-run behavior. Dynamical models in the literature[31] use the past behavior of peers as an endogenous influence source and typically use the following form:

$$\mathcal{U}_{ni}^{(t)} = \beta x_{ni}^{(t)} + \delta \sum_{q \in g(n)} \frac{y_{qi}^{(t-1)}}{|g(n)|} + \varepsilon_{ni}^{(t)} \tag{40}$$

where:

$y_{qi}^{(t-1)}$ ≡ 1 if individual *q* chose alternative *i* in time period *t-1;* 0 otherwise

[31] See Table 20 for examples where social influence occurs through conformity via dynamical systems.

But this model specification is most relevant for behavior where imitating others provides direct benefits such as in popularity and status seeking. In contrast, if the conformity is informational, then perhaps the individual's choice set should change to include this new option or the attributes of the new alternatives should increase in attractiveness[32].

The level of detail to determine the factors motivating the social influence process are lacking in the travel behavior field. With proper data and modeling techniques, a better understanding of the social influence processes may be inferred. Grinblatt et al. (2008) presents an example with their thorough analysis of Finnish vehicle ownership study involving state-provided data on location and vehicle purchasing behavior. With an extensive dataset and varying model specifications and descriptive statistics, they suggest that transfer of information is the most likely method of influence in their study. Additionally, they found that their results could possibly support conformity or status signaling but likely refutes hypotheses about individuals feeling envy towards other car owners.

The travel behavior field needs to place greater focus on the behavioral and societal motivations behind the social influence process. In the next section, some examples of the motivations for social influence are explored.

7.3 Social Influence: Types and Motivations

Social influence has been studied extensively in the social sciences and a comprehensive introduction and review is beyond the scope of this paper. Kelman (1958) provides an early taxonomy to describe social influence through the types of compliance,

[32] A model of informational conformity is developed in Chapter 8.

identification, and internalization. In Kelman's work, compliance is the recognition of a request, implicit or explicit, and a desire to fulfill that request[33]. Compliance is a public act and may involve no change in one's personal beliefs. "Thus the satisfaction derived from compliance is due to the *social effect* of accepting influence" (Kelman 1958, p. 53). Identification involves association with an individual or group and a desire to partake of behavior that is expected of said individual or group. Identification is a public act with the purpose of relationship-building with the influential individual or group. "Thus the satisfaction derived from identification is due to the *act* of performing as such" (Kelman 1958, p. 53). Internalization involves the recognition of one's value system and partaking of behavior that one believes in. Internalization is a public and private act that involves changing one's beliefs to those of the influencing entity (individual or group). "Thus the satisfaction derived from internalization is due to the *content* of the new behavior" (Kelman 1958, p.53). However, conformity, the most commonly modeled influence process in travel behavior, can be a component of each of these influence processes.

Due to this limitation, this paper instead will emphasize a specific view that closely parallels the generalized framework for social influence models of discrete choice and the majority of work in the travel behavior field. The types of social influence are simplified along the lines of Cialdini and Goldstein (2004). Their review concentrates on late 1990s and early 2000s social influence literature which tended to look at "subtle, indirect, and nonconscious" sources of social influence. These are the processes most likely to be present at the data scales relevant for travel behavior research that uses

[33] Please note that a different definition of compliance (see section 7.3.2) was used in Table 20.

discrete choice modeling. Cialdini and Goldstein (2004) separate social influence into the two types of conformity and compliance. Individuals are influenced by others when it serves their motivations for *accuracy*, *affiliation*, and/or *maintenance of a positive self-concept*.

7.3.1 Conformity

Individuals conform when they attempt to match the behavior of others. Thus, conformity parallels the discussion on endogenous social influence effects since the influence medium is a function of the choices of others. Conformity can be informational, where the goal is “to form an accurate interpretation of reality and behave correctly,” or normative, with “the goal of obtaining social approval from others” (p. 606). Cialdini and Goldstein (2004) frame different research areas in conformity through the motivations of accuracy, affiliation, and maintenance of a positive self-concept. These research areas are listed below:

- Accuracy
    - *Perceived Consensus*
    - *Dynamical Systems*
    - *Automatic Activation*
- Affiliation
    - *Behavioral Mimicry*
    - *Gaining Social Approval*
- Maintaining a Positive Self-Concept
    - *Majority & Minority Influence*

- *Deindividuation Effects*

The model forms generally used for conformity can fit many motivations, but these motivations need different interventions to generate changes in the strength of social influence effects. For example, if influence is motivated by accuracy due to perceived consensus, then changing behavior may involve exposing individuals to alternative behaviors in order to break the consensus perception.

The conformity modeled in the common model forms in equations (39) and (40) introduces ambiguity in the identification of the influence mechanism. Often these models can be explained by each motivation – accuracy, affiliation, and maintenance of a positive self-concept. For example, *perceived consensus* parallels the Brock and Durlauf (2001, 2003) model where social influence occurs through perceptions of the behavior of others, but they assume rational expectations which correspond with the actual average behavior[34]. Applied work in transportation has not measured behavioral perceptions. In *individual activation*, individuals minimize cognitive effort by imitating the actions of others. This technique may possibly be measured in social influence models if individual-level effects $\beta_i x_{ni}$ are approximately zero. For *gaining social approval*, individuals may imitate the actions of others in order to "restore their sense of belonging and their self-esteem" (Cialdini and Goldstein 2004, p. 611). In *majority influence*, group members "[identify] with a message source" (p. 612) and may desire to signal to themselves and others that they are a member of said groups by exhibiting similar behavior. *Deindividuation effects* parallels research on the social identity approach (Reicher et al.

[34] Li and Lee (2009) counter the rational expectations assumption by using data that measured behavioral perceptions and Manski (2004) encourages the measurement of expectations.

1995) and may present as a social norms-based influence where the norm is conveyed through the observed actions of similar others.

7.3.2 Compliance

In contrast with conformity, compliance draws parallels to contextual social influence. Influence is not from the individual seeing or perceiving the behavior of other but by advice, commands, and norms[35] that trigger specific behaviors. These triggers can be explicit (e.g. direct request from a supervisor) or implicit (e.g. an advertisement). For social influence through compliance, Cialdini and Goldstein (2004) mention a number of different research areas for motivating compliance including:

- Accuracy
    - *Affect and Arousal*
    - *That's-not-all Technique*
    - *Resistance*
    - *Authority and Obedience*
    - *Social Norms*
- Affiliation
    - *Liking*
    - *Reciprocation*
    - *Door-in-the-face Technique*

[35] Norms may affect individuals through both conformity and compliance. In normative conformity, the norm is conveyed directly through the behavior of others. In compliance, other avenues of influence – such as advertisements, advice, commands, policies, laws, and ideal types – are used to convey the norm to an individual.

- Maintaining a Positive Self-Concept
    - *Foot-in-the-door Technique*
    - *Consistency and Commitment*

Compliance motivations have been limited in travel models of discrete choice with the only examples involving *social norms* (Wu et al. 2013, Goetzke and Weinberger 2012). Social norms can be classified into injunctive – “what is typically approved/disapproved” – or descriptive norms – “what is typically done” (Cialdini and Goldstein 2004, p. 597). *Authority and obedience* may be pertinent to work on the influence of authority figures at work and home and counter-culture elements. The *foot-in-the-door technique* is used often by individuals, groups, and institutions to encourage compliance by removing barriers to an option for a limited time such as free transit days or bike-to-work days. In *consistency and commitment*, an individual may be motivated to perform behavior in accordance with a prior promise they made. The individual will attempt to maintain consistency with their self-concept. Cialdini and Goldstein (2004) note that this is particularly effective in individualistic societies compared to collectivist societies. *Affect and arousal* also has relevance due to advertising techniques. Advertisements attempt to entice favorable emotions in their ads to compel individuals to change behavior.

Cialdini and Goldstein (2004) is not the only appropriate taxonomy for describing social influence, but their work is showcased here to show that the motivational patterns of social influence work through different processes and serve various motivations. Additionally, these motivation patterns work at different levels of social interactions and access different types of people that an individual may come into contact. Thus, the social network of the individual and the processes that form and shape that network will have important implications on the effect of social influence in the decision process.

7.4 Social Influence and Social Networks

In social influence processes, it is critically important to understand who transfers influence to an individual. In social network terminology[36], nodes represent individuals and edges indicate the connections between individuals. These linkages between individuals form a comprehensive social network, and the synergies between social networks and social influence need to be taken into account when modeling social influence.

In studies of social influence and diffusion, varying strains of research support and refute the hypotheses that influence occurs primarily due to: (1) personal influence between the direct contacts of an individual, (2) the influence of social groups, social circles, and social position, and (3) the influence of marketing and the media (Kadushin 2012). Since each of these sources entails different social interactions, this translates into a critical connection between the social influence mechanism, the underlying social network, and the sources of influence. Thus, social influence hypotheses require different social networks to explain their behavioral processes correctly, such as:

- *Minority Influence*. In minority influence, individuals in a smaller group (the minority) may influence the behavior of members of the majority by appealing to

[36] Extensive definitions of social network terms will not be covered here. Please consult a textbook on social networks such as Jackson (2008).

a shared identity. Because of the importance of overlapping social circles to create shared identities, a network of close contacts as well as acquaintances would be an appropriate social network for studying minority influence.

- *Comparative Happiness*. In comparative happiness, individuals compare their current situation with that of a target peer. If there is a discrepancy, the individual may emulate the target peer to gain a more favorable condition. Because the cognitive costs of making comparisons are high, a social network with small, intimate connections would likely be most appropriate.
- *Authority and Obedience.* In authority and obedience (i.e. social power), an individual emulates the behavior of those with higher social position. Thus, a hierarchical social network showing roles in an organization and the directions of social power would be helpful.
- *Affect and Arousal.* In affect and arousal, a source attempts to appeal to the emotions of the individual in order to trigger favorable behavior. A possible network structure for this influence mechanism may include a bipartite network showing connections between individuals and advertising sources.

7.5 Summary and Areas for Future Research

In this chapter, a generalized framework to behaviorally describe choice models of social influence was presented. The framework focuses on the microfoundations of social influence, but also emphasizes the similarities in different forms of social influence models previously presented in the literature and focuses on the role of social networks in these models. Understanding the motivations for social influence has a critical role in determining the social network and influence sources to use in modeling various choice behaviors. Interdependence between these aspects affects the behavioral explanation of choice decisions which will have impacts on the effectiveness of different policy prescriptions. The complexity of social influence and the various and conflicting motivations for social influence make it critically important to understand the behavioral process rather than solely comparing competing model specifications for statistical significance alone. As Kadushin (2012) explains, identifying influence is difficult due to:

- "the practical problems of finding the influencers"
- "the theoretical problems of modeling the source and nature of the influence," and
- "distinguishing between the effect of media and the social environment and specific individuals who might inform or persuade (or both)" (p. 140).

The differences in social influence types, sources, and motivations have implications in applying these models for short-run and long-run policy analysis due to varying outcomes[37].

The flexibility of the framework presented can be used as a taxonomy for describing social influence choice models as well as a springboard for further research and application. In particular, new focus can be applied to:

- New decision rules such as regret minimization (Chorus 2010, Zeelenberg and Pieters 2007), prospect theory, and elimination by aspects (Hess et al. 2011)
- Heterogeneity in the social influence mechanism depending on classes of individuals[38] (Kuwano et al. 2012)

[37] An example of this is shown in Chapter 8 where the equilibrium properties of a model of informational conformity vary from those of a model with direct-benefit conformity such as Brock and Durlauf (2001).

- Incorporating new social influence mechanisms and motivations[39]
- Exploratory work to find new influence sources that affect social influence aside from the choices of others such as attitudes, perceptions, past experiences, ideal types, and the salience of social identities
- Mixing social network types and structures when using multiple social influence mechanisms
- Deriving and analyzing dynamical and equilibrium behavior beyond reflexive large cliques and mean-effect conformity due to the greater variety of social influence processes, network configurations, and decision rules possible
- Incorporating cognitive and spatial limitations on network formation
- Developing and applying dynamic models of network formation and discrete choice (Gulyás and Dugundji 2006, Snijders et al. 2010)
- Assuming more complex payoff forms beyond the linear-in-parameter formulation such as multiplicative combinations of factors (e.g. cross effects)
- Incorporating network statistics (e.g. centrality, closeness, diameter) into the modeling process as explanatory variables (Dugundji et al. 2011) or to trigger changes in social network mechanisms and influence sources

---

[38] A sample model formulation using a confirmatory latent class model is shown in Appendix E.

[39] A model of informational conformity is formulated in Chapter 8.

Chapter 8: An Informational Conformity Binary Choice Model

Table 21. Chapter 8 Summary

Background & Brief Summary	This chapter describes a formulation of a choice model of informational conformity which uses a latent class structure. The class membership model depends on the proportion of group members exhibiting a particular behavior. Membership into the "more informed" class will cause a change in the preferences of those individuals, thus making the behavior more attractive. These "informed" individuals are motivated to conform due to the goal of accuracy.
Motivation	As stated in section 7.2.2, the current state-of-the-art is conformity models with direct benefit social influence effects. Specifically, indirect effects have seen limited development and one example of this is the modeling of an informational conformity hypothesis. The informational conformity model developed in this chapter will demonstrate the additional behavioral realism possible with choice models and serve as an example of an enhancement to current models.
Results	1. A model was formulated for incorporating informational conformity into a choice model via a latent class discrete choice model form 2. Equilibrium properties were derived for this model which showed the possibility of multiple equilibria 3. The direct-benefit conformity model was found to be similar to the informational conformity model with different behavioral assumptions which resulted in different equilibrium properties 4. A Bayesian inference procedure was proposed to handle hypothesis testing and to derive predictive distributions 5. A two-stage control function approach was proposed to handle endogeneity in the social influence effect was handle
Limitations	1. The model cannot prove the existence of informational conformity without data obtained through an appropriate causal experimental design 2. Actual informational levels are latent and cannot be observed and are thus inferred through model fit, hypothesis testing, and modeler experience

8.1 Introduction

Informational conformity is a social influence process where the goal is "to form an accurate interpretation of reality and behave correctly" (Cialdini and Goldstein 2004, p.606). Informational conformity is a form of social influence that is mostly motivated by the *goal of accuracy*. To model social influence via informational conformity, a confirmatory latent class approach (Hess 2014) is used. In this approach, an a priori behavioral hypothesis is made that informational conformity occurs through a stratification of the population into different informed classes. This stratification is posited to be correlated with individual-level and environmental factors as well as social influence effects. Individuals in each informed class have differing choice perspectives which are represented by class-specific choice models. A generalized model of informational conformity using the confirmatory latent class approach is formulated in Appendix D.

In this chapter, a binary logit informational conformity model is formulated with two information classes. It is hypothesized that a choice depends on information derived from individual factors and social influence. This information causes changes in the preferences (i.e. choice model parameters) of individuals. Figure 24 summarizes the approach where two classes will be modeled with class $\mathscr{a}$ representing individuals who have been informed of some preferable features of a particular type of behavior and class $\mathscr{b}$ representing individuals who are not socially influenced.

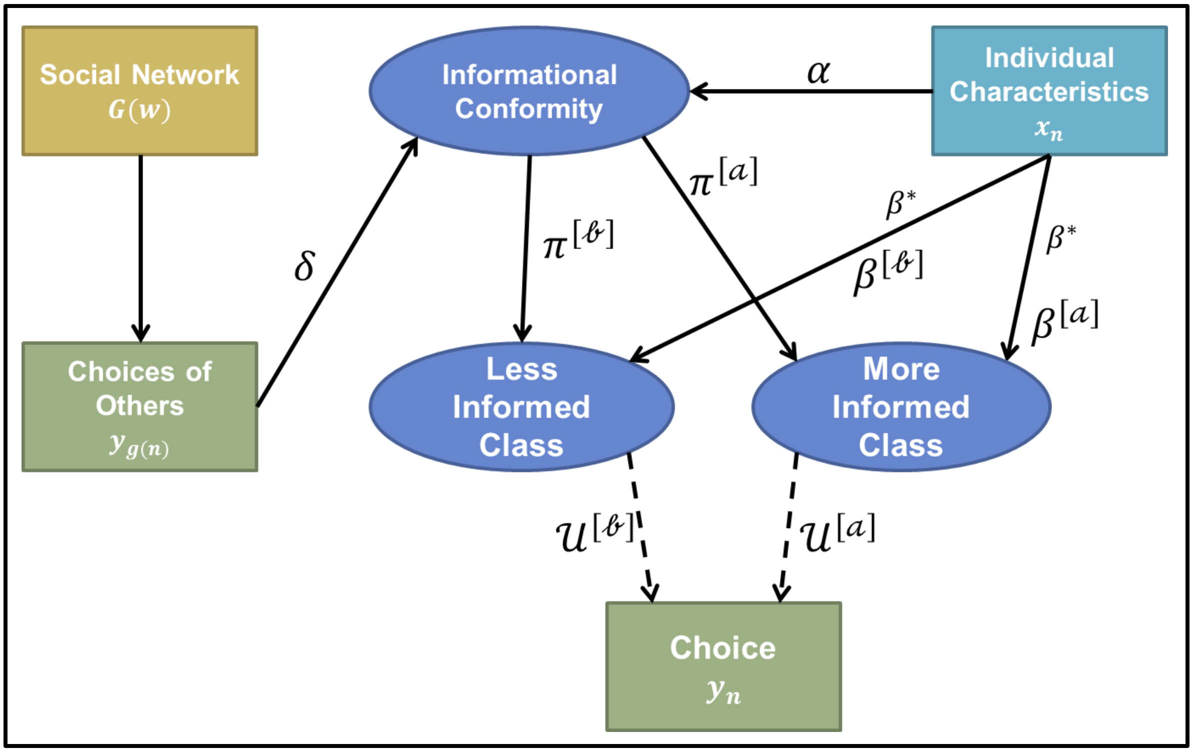

Figure 24. An Informational Conformity Model Specification

8.2 Binary Choice Formulation

This formulation will be written under a binary choice decision. Begin by assuming a population *N* of decision makers where individuals are connected in a social network *G*. Each individual *n* is faced with a choice task where the individual must choose between two alternatives $y_n = \{0,1\}$. In this population, individuals may be influenced via informational conformity (class a) or not influenced (class b). This process is unobserved and will be modeled latently with discrete classes. Class membership is affected by the prevalence of a behavior in a population. Specifically, this could represent knowledge of the behavior being transferred to individuals through seeing the behavior in the population. This will be denoted by the information function $\mathcal{F}$, which will take the following linear-in-parameter form[40]:

$$\mathcal{F}_n = \alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} + \varepsilon_n^{\mathcal{F}} \tag{41}$$

where:

z_n $\equiv$ individual-level characteristics of individual *n*

α, δ $\equiv$ information (class-membership) model parameters

$\varepsilon_n^{\mathcal{F}}$ $\equiv$ error term for individual *n*

Assuming that the error term $\varepsilon_n^{\mathcal{F}}$ is IID logistic (with location 0 and scale 1), then the probability for an individual to be in the "more informed" class takes the familiar binary logit (logistic regression) form as follows:

$$\pi_n^{[a]} = \text{Prob}(c_n = a) = \frac{\exp(\mathcal{F}_n)}{1 + \exp(\mathcal{F}_n)} = \frac{\exp(\alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|})}{1 + \exp(\alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|})} \tag{42}$$

where:

c_n $\equiv$ the class of individual *n*, which can be either informed a or uninformed b

Accordingly, the probability of being in class b follows:

$$\pi_n^{[b]} = \text{Prob}(c_n = b) = 1 - \pi_n^{[a]} \tag{43}$$

[40] No time superscripts are used in this formulation as it is assumed that the modeler will choose an appropriate formulation. The model can be formulated for both the cross-sectional and dynamic cases. The equilibrium properties for the model (derived in the next section) will assume simultaneity.

It is expected that individuals surrounded by others who perform the action, $y = 1$, may be able to reevaluate their preferences for the alternative under the new information they receive from being exposed to the behavior more often than other people. Thus, the preferences of these "more informed" individuals may vary compared to the "less informed" individuals. Assuming utility maximizing behavior for individuals, the utility differences between behaviors $y_n = \{1,0\}$ for an individual *n* for each class $\{\mathcal{a}, \mathcal{b}\}$ is given as follows:

$$\Delta\mathcal{U}_n^{[\mathcal{a}]} = \beta^{[\mathcal{a}]} x_n^{[\mathcal{a}]} + \beta^* x_n^* + \varepsilon_n^{[\mathcal{a}]} \tag{44}$$

$$\Delta\mathcal{U}_n^{[\mathcal{b}]} = \beta^{[\mathcal{b}]} x_n^{[\mathcal{b}]} + \beta^* x_n^* + \varepsilon_n^{[\mathcal{b}]} \tag{45}$$

where:

$x_n^{[\mathcal{a}]}, x_n^{[\mathcal{b}]} \equiv$ individual-level characteristics of individual *n* that are specific to the choice models for class $\mathcal{a}$ and class $\mathcal{b}$ (may be the same as those in the class-membership model)

x_n^* $\equiv$ individual-level characteristics of individual *n* that are shared between both class $\mathcal{a}$ and class $\mathcal{b}$ (may be the same as those in the class-membership model)

$\beta^{[\mathcal{a}]}, \beta^{[\mathcal{b}]} \equiv$ model parameters specific to class $\mathcal{a}$ and class $\mathcal{b}$

β^* $\equiv$ model parameters shared by both class $\mathcal{a}$ and class $\mathcal{b}$

β $= \left\{\beta^{[\mathcal{a}]}, \beta^{[\mathcal{b}]}, \beta^*\right\}$

$\varepsilon_n^{[\mathcal{a}]}, \varepsilon_n^{[\mathcal{b}]} \equiv$ unobserved effects on individual *n* for classes $\mathcal{a}$ and $\mathcal{b}$, distributed IID Logistic(0,1)

Thus, the probability of observing a choice $y_n = 1$ for individual *n* given that *n*'s class is as follows:

$$P_n^{[a]} = P_n(y_n = 1 | c_n = a) = \frac{\exp(\beta^{[a]} x_n + \beta^* x_n)}{1 + \exp(\beta^{[a]} x_n + \beta^* x_n)} \tag{46}$$

$$P_n^{[b]} = P_n(y_n = 1 | c_n = b) = \frac{\exp(\beta^{[b]} x_n + \beta^* x_n)}{1 + \exp(\beta^{[b]} x_n + \beta^* x_n)} \tag{47}$$

Taken together, the probability of observing a choice $y_n = 1$ for individual *n* is as follows:

$$P_n = \pi_n^{[a]} P_n^{[a]} + \pi_n^{[b]} P_n^{[b]} \tag{48}$$

8.2.1 Likelihood Function

Using equation (48) and substituting values from equations (42), (43), (46), and (47), the likelihood of an observation for individual *n* can be written as follows:

$$\begin{aligned}
\mathcal{L}_n &= \mathcal{L}_n(\alpha, \beta, \delta; y_n) = (P_n)^{y_n}(1 - P_n)^{1-y_n} \\
\mathcal{L}_n &= \pi_n^{[a]} \cdot \left[\left(P_n^{[a]}\right)^{y_n}\left(1 - P_n^{[a]}\right)^{1-y_n}\right] + \pi_n^{[b]} \cdot \left[\left(P_n^{[b]}\right)^{y_n}\left(1 - P_n^{[b]}\right)^{1-y_n}\right] \\
&= \frac{\exp(\mathcal{F}_n)}{1 + \exp(\mathcal{F}_n)}\left[\left(\frac{\exp\left(\Delta\mathcal{U}_n^{[a]}\right)}{1 + \exp\left(\Delta\mathcal{U}_n^{[a]}\right)}\right)^{y_n} \cdot \left(\frac{1}{1 + \exp\left(\Delta\mathcal{U}_n^{[a]}\right)}\right)^{1-y_n}\right] \\
&+ \frac{1}{1 + \exp(\mathcal{F}_n)}\left[\left(\frac{\exp\left(\Delta\mathcal{U}_n^{[b]}\right)}{1 + \exp\left(\Delta\mathcal{U}_n^{[b]}\right)}\right)^{y_n} \cdot \left(\frac{1}{1 + \exp\left(\Delta\mathcal{U}_n^{[b]}\right)}\right)^{1-y_n}\right]
\end{aligned} \tag{49}$$

The likelihood and log-likelihood for a sample drawn randomly from a population under a simple random sample for a set of parameters are given as follows:

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}(\alpha, \beta, \delta; y_n) = \prod_{n \in N} \mathcal{L}_n \\
\mathcal{LL} &= \mathcal{LL}(\alpha, \beta, \delta; y_n) = \sum_{n \in N} \log(\mathcal{L}_n)
\end{aligned} \tag{50}$$

8.2.2 Equilibrium Properties

To understand the equilibrium properties of this model, this section will consider the case where all individuals have the same individual-level characteristics and are all connected in a single large clique. Therefore the only heterogeneity between individuals is from the random utility ($\varepsilon_n^{[a]}, \varepsilon_n^{[b]}$) and random information ($\varepsilon_n^{\mathcal{F}}$) terms. Therefore each individual has the same probability of choosing alternative $y_n = 1$:

$$
\begin{aligned}
&P_n = \pi_n^{[a]} P_n^{[a]} + \pi_n^{[b]} P_n^{[b]} = \frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})} P^{[a]} + \frac{1}{1+\exp(h+\delta\bar{y})} P^{[b]}, \\
&where\ h = \alpha z_n\ and\ z_n = z_m, \qquad \forall n, m \in N \\
&and\ P_n^{[a]} = P_m^{[a]} = P^{[a]}, \qquad P_n^{[b]} = P_m^{[b]} = P^{[b]}, \qquad \forall n, m \in N \\
&and\ \bar{y}_n = \bar{y} = \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} = \sum_{q \in N, q \neq n} \frac{y_q}{\|N\| - 1}, \qquad \forall n \in N
\end{aligned} \tag{51}
$$

From these three assumptions of heterogeneity in individual characteristics, heterogeneity in class membership probability, and a large clique social network, the market share of behavior $y = 1$ across the population is as follows[41,42]:

$$
\bar{y} = \frac{1}{|N|} \sum_{n \in N} P_n = \frac{1}{|N|} \sum_{n \in N} \left[\frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})} P_n^{[a]} + \frac{1}{1+\exp(h+\delta\bar{y})} P_n^{[b]} \right] \tag{52}
$$

$$
\bar{y} = f(\bar{y}) = \frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})} P^{[a]} + \frac{1}{1+\exp(h+\delta\bar{y})} P^{[b]} \tag{53}
$$

[41] As mentioned in footnote 40, this formulation assumes simultaneity in the formulation because these are the equilibrium properties. As $t \to \infty$, the dynamic and simultaneous formulations become equivalent.

[42] It is assumed that $\bar{y} = \frac{1}{|N|} \sum_{q \in N} y_q \approx \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} = \sum_{q \in N, q \neq n} \frac{y_q}{\|N\|-1}$. This assumption is most valid for large groups as excluding a single individual will have little effect on the overall group market share.

This is a fixed point problem since $\bar{y} = f(\bar{y})$. Finding the solutions to this problem is non-trivial, but a fixed point does exist in this case.

Proposition 7.1. (Equilibrium Existence for Informational Conformity Model). For the model specified by the likelihood function (49) with heterogeneous agents in a large clique, there exists at least one equilibrium $\bar{y}^*$ such that:

$$\bar{y}^* = \frac{P^{[\mathscr{b}]} + P^{[a]} \exp(h + \delta \bar{y}^*)}{1 + \exp(h + \delta \bar{y}^*)} \tag{54}$$

Proof:

The function $f(\bar{y})$ in equation (53) maps an interval unto itself, explicitly $\{0,1\} \rightarrow \{0,1\}$. Since $\frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})} + \frac{1}{1+\exp(h+\delta\bar{y})} = 1$ and $0 \leq P^{[a]} \leq 1$ as well as $0 \leq P^{[\mathscr{b}]} \leq 1$, then $f(\bar{y}) = \frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})} P^{[a]} + \frac{1}{1+\exp(h+\delta\bar{y})} P^{[\mathscr{b}]}$ must be between 0 and 1. By Brouwer's fixed point theorem and the intermediate value theorem, there must exist at least one fixed point such that $\bar{y}^* = f(\bar{y}^*)$.

Multiple equilibria are also possible with this model formulation when the following assumptions are made:

Assumption 7.1. (Behavioral Properties of Informational Conformity Model)

a) $P^{[a]} \geq P^{[\mathscr{b}]}$ [Superiority of Preferences for More Informed Class over Less Informed Class]

b) $\delta \geq 0$ [Increasing Market Share cannot Decrease Information Transfer]

Specifically, the properties of the first and second derivatives can be used to observe when it is possible for multiple equilibria to exist. The existence of multiple equilibria is explained in the following proposition:

Proposition 7.2. (Existence of Multiple Equilibria in Informational Conformity Model)

On the interval of $\bar{y} = \{0,1\}$ the number of equilibria for equation (54) can be determined through the following conditions and properties:

a) When $P^{[a]} = P^{[b]}$, there exists a unique fixed point to equation (54) at $\bar{y} = P^{[a]} = P^{[b]}$.

b) When $\delta = 0$, there exists a unique fixed point to equation (54) at $\bar{y} = \frac{\exp(h)}{1+\exp(h)} P^{[a]} + \frac{1}{1+\exp(h)} P^{[b]}$.

c) When $-{}^{h}/_{\delta} < 0$ or $-{}^{h}/_{\delta} > 1$, there exists a unique fixed point to equation (54).

d) When $0 \leq -{}^{h}/_{\delta} \leq 1$, there can exist 1, 2, or 3 fixed points to equation (54).

Proof:

Starting from equations (53) and (54):

$$\frac{\partial f(\bar{y})}{\partial \bar{y}} = \frac{\delta\left(P^{[a]} - P^{[b]}\right)\exp(h + \delta\bar{y})}{[\exp(h + \delta\bar{y}) + 1]^2} \tag{55}$$

$$\frac{\partial^2 f(\bar{y})}{\partial \bar{y}^2} = \frac{\delta^2 \left(P^{[a]} - P^{[b]}\right) \exp(h + \delta\bar{y})(\exp(h + \delta\bar{y}) - 1)}{[\exp(h + \delta\bar{y}) + 1]^3} \tag{56}$$

a) For Proposition 7.2(a), when $P^{[a]} = P^{[b]}$, $\partial f(\bar{y})/\partial \bar{y} = 0$. Thus the function is constant and takes a single value $f(\bar{y}) = P^{[a]} = P^{[b]}$. Since this value is between 0 and 1, there is a single fixed point at $\bar{y}^* = P^{[a]} = P^{[b]}$.

b) Proof of Proposition 7.2(b) follows similarly to the proof Proposition 7.2(a). When $\delta = 0$, $\partial f(\bar{y})/\partial \bar{y} = 0$ and $f(\bar{y})$ is thus a constant function that exists between values 0 and 1.

c) Equation (55) and Assumption 7.1 shows that $f(\bar{y})$ is a continuously differentiable and monotonically increasing function since the first derivative of $f(\bar{y}) \geq 0$ for $\bar{y} \in \{0,1\}$. Thus, the only way for the function to cross the line $\bar{y} = f(\bar{y})$ multiple times is via an inflection point. Using equation (56) to find inflection points by setting the second derivative equal to zero, an inflection point will occur at $h + \delta\bar{y} = 0$ or equivalently at $\bar{y} = -h/\delta$. Thus if this inflection point occurs outside of the region $\bar{y} \in \{0,1\}$, then inside the region $\bar{y} \in \{0,1\}$, $f(\bar{y})$ will cross the line $\bar{y} = f(\bar{y})$ only once.

d) Continuing from the proof for Proposition 7.2(c), if the inflection occurs inside the region $\bar{y} \in \{0,1\}$, then it is possible for the curve $f(\bar{y})$ to have a root located before the inflection point at a value $0 \leq \bar{y} < -h/\delta$ and/or after the inflection point at a value

$-h/\delta < \bar{y} \leq 1$. If a root occurs both before and after the inflection point, then there could possibly be one additional root before, at, or after the inflection point.

An additional rare possibility is for the curve $f(\bar{y})$ to have a root at the inflection point and to be tangent to the line $\bar{y} = f(\bar{y})$ at this inflection point. If this occurs, then a root may also occur before or after the inflection point as well, thus causing two roots.

Additionally the following corollary was derived for understanding conditions pertaining to characteristics of the equilibrium system as the individual-level and endogenous social influence effects vary:

Corollary 7.1. (Limits on individual-level effects and endogenous social influence effects)

a) As $h \rightarrow -\infty$, there will exist a fixed point to equation (54) with a root at $\bar{y} = P^{[b]}$.

b) As $h \rightarrow \infty$, there will exist a fixed point to equation (54) with a root at $\bar{y} = P^{[a]}$.

c) As $d \rightarrow \infty$, there will exist a fixed point to equation (54) with a root at $\bar{y} = P^{[a]}$.

Following from Corollary 7.1 the following corollary was derived for understanding conditions pertaining to complete adoption of an alternative:

Corollary 7.2. (Complete Adoption Conditions)

a) When $P^{[a]} = 0$ and as $h \to -\infty$, there will exist a fixed point to equation (54) with a root at $\bar{y} = 0$, and thus, all individuals will choose option $y_n = 0, \forall n$.

b) When $P^{[\mathscr{b}]} = 1$ and as $h \to \infty$, there will exist a fixed point to equation (54) with a root at $\bar{y} = 1$, and thus, all individuals will choose option $y_n = 1, \forall n$.

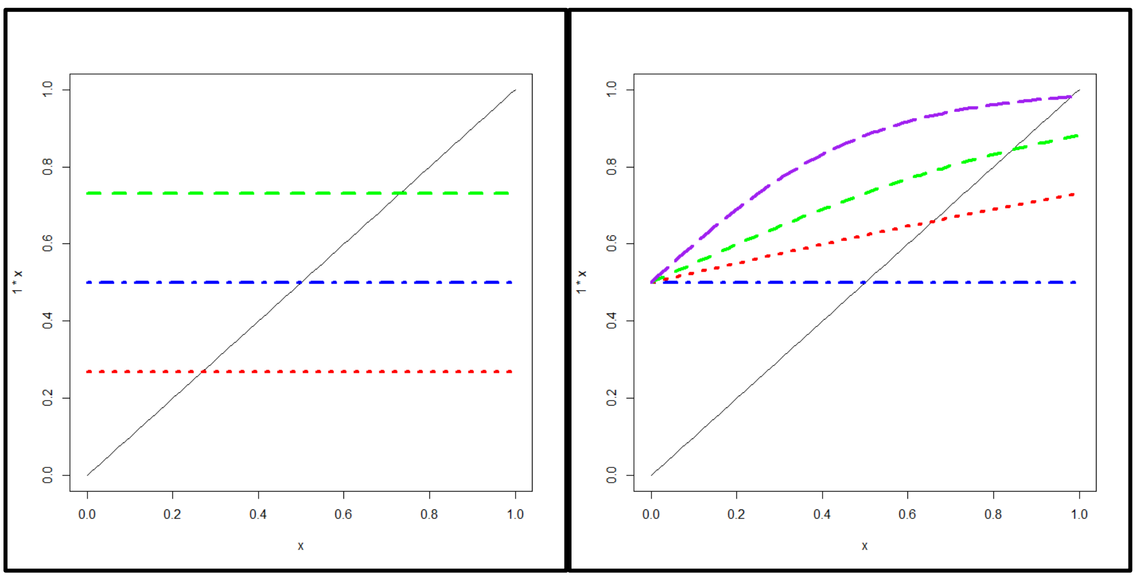

Figure 25. Equilibrium Plots for Varying h and δ Parmeters

To aid the reader, the figures that follow are various plots of equation (54) for different values of the parameters. For Figure 25 through Figure 27, $P^{[a]} = 0$ and $P^{[\mathscr{b}]} = 1$.

In Figure 25, the left plot shows equilibrium conditions in which $\delta = 0$. The red short-dash line corresponds to $h = -1$, the blue dot-dash line corresponds to $h = 0$, and the green long-dash line corresponds to $h = 1$. The right plot shows equilibrium conditions when h is fixed to 0. The lines (in order from bottommost to topmost) corresponds to values of $\delta = 0,1,2,4$.

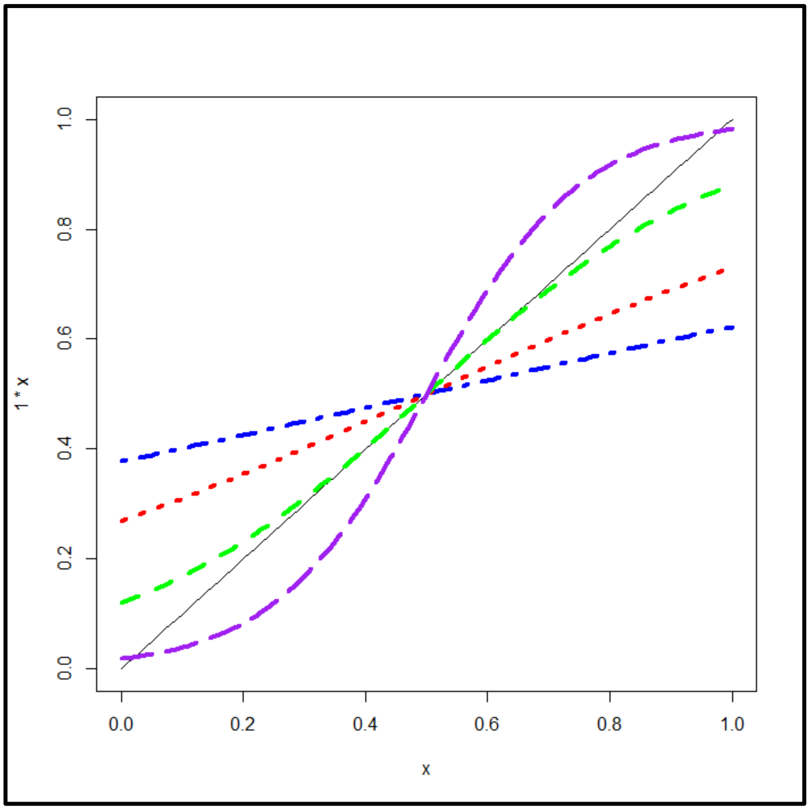

Figure 26. Equilibrium Plots with h/δ fixed at -0.5

In Figure 26, the ratio of h/δ is fixed at 0.5 which corresponds to a condition where multiple equilibria may exist. All four (h, δ)-tuples have a root at $\bar{y}^* = 0$. Three tuples has a single, unique fixed point: the blue dot-dash $(-0.5,1)$ line, the red dotted $(-1,2)$ line, and the green short-dash $(-2,4)$. The purple long-dash $(-4,8)$ line has three equilibria with the additional equilibria located near 0 and 1.

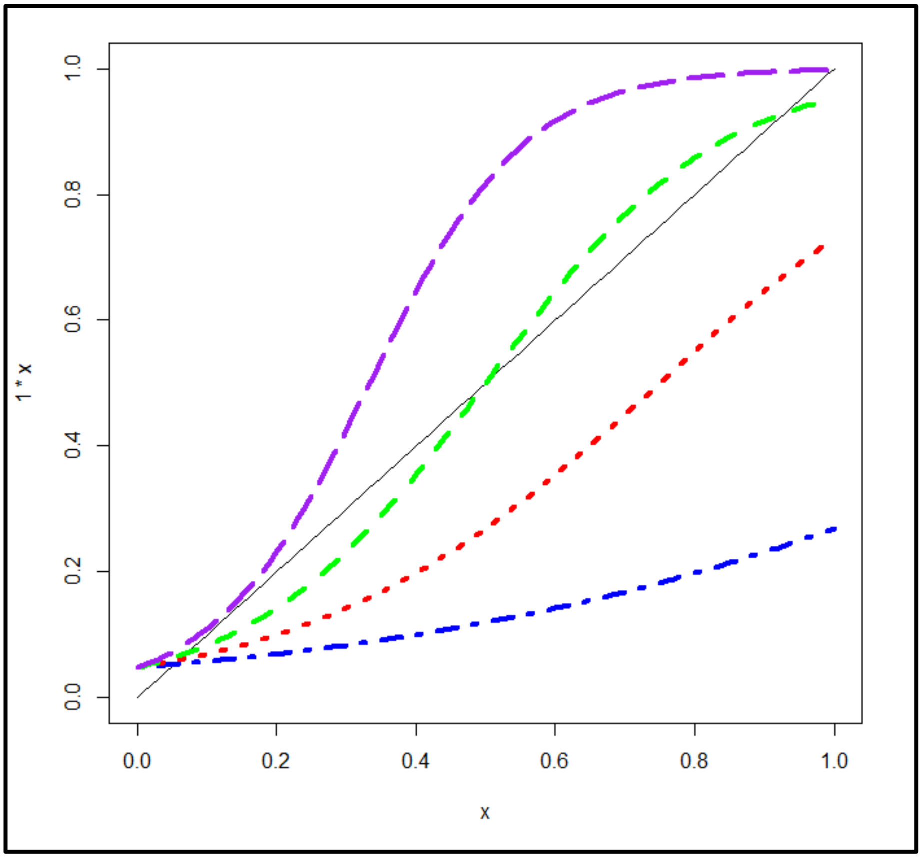

Figure 27. Equilibrium Plots for Varying h/δ

Figure 27 shows how the change in slope varies as the ratio h/δ changes. The lines (in order from bottommost to topmost) corresponds to the (h, δ)-tuples values of $\{(-3,2), (-3,4), (-3,6), (-3,9)\}$.

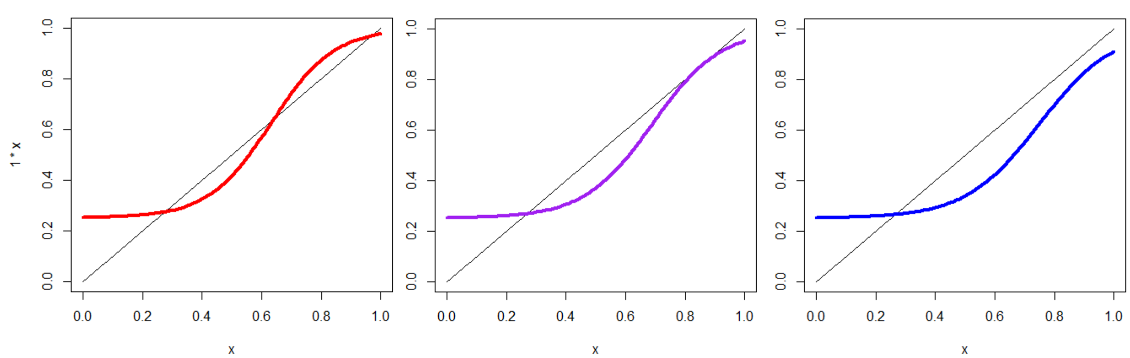

Figure 28. Examples of 3, 2, and 1 Equilibria Conditions

Figure 28 shows examples of the number of equilibria possible from the informational conformity model formulation. For the plot in order from left to right, the (h, δ)-tuples values correspond to the following $\{(-6,9.5), (-6,6.6882), (-6,8)\}$. The leftmost plot is an example of three equilibria, while the middle plot shows a rare two equilibria solution.

8.2.3 Comparison to Statistical Mechanics Field-Effect Formulation

The statistical mechanics formulation of Brock and Durlauf (2001) is a special case of the informational conformity model, but it requires differences in behavioral assumptions. Specifically, it requires that individuals in the "less informed" class only make choice $y_n = 0$ and individuals in the "more informed" class only make choice $y_n = 1$. Additionally, the endogenous social influence effect in the information model of equation (41) must allow for both positive and negative influence[43]. To show this, the formulation will be eased by making the binary choice map to $y_n = \{-1, +1\}$. Accordingly, the average behavior of social contacts now ranges from -1 to +1 as explained in footnote 43. The formulation is as follows:

$$\begin{gathered}
\mathcal{F}_n = h + \delta \bar{y}_n + \varepsilon_n^{\mathcal{F}} \\
c_n = \begin{cases} a & if\ \mathcal{F}_n \geq 0 \\ \mathscr{b} & otherwise \end{cases} \\
\Delta \mathcal{U}_n^{[a]} = \infty + \varepsilon_n^{[a]} \\
\Delta \mathcal{U}_n^{[\mathscr{b}]} = -\infty + \varepsilon_n^{[\mathscr{b}]} \\
y_{ni} = \begin{cases} +1 & if\ \Delta \mathcal{U}_n^{[c_n]} \geq 0 \\ -1 & otherwise \end{cases}
\end{gathered} \tag{57}$$

[43] In other words, if less than 50% of one's social contacts choose $y = 1$, then individuals will experience negative information and thus be even more likely to be members of the "less informed" class. The network and influence term is remapped from {0,1} to {-1,+1} by $\bar{y}' = 2\bar{y} - 1$.

The equilibrium conditions are as follows when homogeneity of individual-level effects is assumed:

$$\bar{y} = \frac{1}{|N|}\sum_{n\in N} P_n = \frac{1}{|N|}\sum_{n\in N}\left[\frac{\exp(h+\delta\bar{y})}{1+\exp(h+\delta\bar{y})}\cdot(1) + \frac{1}{1+\exp(h+\delta\bar{y})}\cdot(-1)\right] \quad (58)$$

$$\bar{y} = f(\bar{y}) = \frac{\exp(h+\delta\bar{y})-1}{1+\exp(h+\delta\bar{y})} = \tanh\left(\frac{1}{2}(h+\delta\bar{y})\right) \quad (59)$$

This reduction to the hyperbolic tangent form makes the model's equilibrium conditions equivalent to the Brock and Durlauf (2001) formulation when h and δ are doubled. The existence of equilibrium follows from Brock and Durlauf (2001) Proposition 1 due to Browuer's fixed point theorem. The multiplicity of equilibria follows similar to Proposition 2 such that:

1. When $h = 0$ and $\delta > 2$, three equilibria exist.
2. When $h \neq 0$ and $\delta > 2$, then the number of equilibria (1, 2, or 3) depends on a threshold. This threshold is a function of h and δ.

8.3 Bayesian Inference and Application Procedure

Latent class discrete choice models are known for difficulties in estimation via maximum likelihood estimation (Hess 2014, Bhat 1997). This is primarily due to two concerns:

1. *Local Optima.* The likelihood function for a latent class discrete choice model is not globally concave as in the common multinomial logit model. Because of this, traditional optimization approaches may converge to local optima with no guarantee of it being a global optimum.
2. *Non-invertible Hessian Matrix.* Because numerical approximations of the Hessian matrix are used, near-flat likelihood functions and estimate values near the

boundaries of the parameter space (Chung et al. 2006) can cause issues with properly deriving the Hessian. This can cause the optimization models to converge to an optimum but not be able to invert the Hessian matrix to obtain an asymptotic variance-covariance matrix.

3. *Model Identification.* This is a limitation of the class membership model. The classes do not have natural orderings, so swapping of coefficients between classes can cause two different model specifications to have the same likelihood value. In a simple two class, binary choice example, assume that the probability of membership in class 1 is constant at 0.25 with a corresponding choice probability of 0.75 for its class 1's choice model and class 2 has a choice probability of 0.35. An equivalent model is obtained where class 1 has membership probability 0.75 with corresponding class 1 choice probability of 0.35 and class 2 choice probability of 0.75.

Model identification can be handled by normalization of class-specific parameters or by proper interpretation by the modeler. For the informational conformity model formulated in section 8.2, this can be handled by normalizing the social influence parameter to be positive. The most common methods for handling local optima and Hessian inversion are to use traditional optimization routine at multiple starting values or by using the Expectation-Maximization (EM) algorithm (Bhat 1997). The EM algorithm is a popular approach but suffers from issues with the speed of convergence and a limitation to models without shared coefficients. Since the informational conformity model may have shared coefficients and may suffer from optimization problems, another approach is warranted.

8.3.1 Motivations for Using Bayesian Inference

An alternative approach for estimating latent class models is to use Bayesian inference such as in Hoijtink (1998), Garrett and Zeger (2000), Garrett et al. (2002), and Chung et al. (2006). A Bayesian inference approach has a number of advantages including:

1. Bayesian inference procedures do not require the optimization of a non-linear function. This strength is most relevant when new specifications of choice models are used with multiple optima such as the latent class DCMs used in Chapter 8 and Appendix E.
2. Draws from the posterior can be used to test an almost limitless number of statistical hypotheses. Functions of parameters can be tested using MCMC draws of the parameter estimates in a straightforward fashion (Lenk 2014).
3. Predictive distributions for forecasting can be constructed by simulating the model for each draw from the posterior. This provides a means to describe the uncertainty in model predictions.
4. With proper choice of proper, calculating the mean of the posterior distribution via Bayesian inference is asymptotically equivalent to maximum likelihood estimation due to the Bernstein-von Mises theorem (Train 2009). As Train (2009) says: "The researcher can therefore use Bayesian procedures to obtain parameter estimates and then interpret them the same as if they were maximum likelihood estimates" (p. 284).

These four motivations make the use of Bayesian inference viable for informational conformity model estimation. Specifically, it has strengths in guaranteeing that the modeler obtains variance estimates for parameters and to do other hypothesis testing as well as it allows for describing the uncertainty in forecasting. The latter point is important as social influence is an imperfect process and can be difficult to measure. Uncertainty in parameter estimations can cause large swings in the predicted equilibria in some social influence models.

8.3.2 Bayesian Inference for Fixed Coefficient DCMs

This section describes a procedure for Bayesian inference on fixed coefficient discrete choice models such as the informational conformity model described in section 8.2. Using a MNL model as an example, the model and sampling procedure can be described as follows:

Utility:

$$\mathcal{U}_{\mathrm{ni}} = \beta_{\mathrm{n}} x_{ni} + \varepsilon_{ni}$$

$$\varepsilon_{ni} \sim \text{IID extreme value}$$

Observed Choice:

$$y_{ni} = 1 \Leftrightarrow \mathcal{U}_{ni} \geq \mathcal{U}_{nj} \;\; \forall j \in C \tag{60}$$

Priors:

$$\pi(\beta) \text{ is } N(\beta_0, s_0) \text{ where } s_0 \text{ is very large}$$

Conditional Posteriors:

$$\pi(\beta|X,Y) \propto \prod_{i} \prod_{n} \left\{ \exp(\beta_{\mathrm{n}} x_{ni}) \middle/ \sum_{j \in C} \exp(\beta_{\mathrm{n}} x_{nj}) \right\}$$

To calculate the posterior density for this model, the Metropolis-Hastings algorithm is used. A prior for β is assumed to be normal with a very large variance (i.e. β is almost flat). The steps of the process are as follows:

1. Initialize the chain with initial parameter values , $\beta^{\{0\}}$
2. Set d = 1, repeat until burn-in is achieved ($d = b$) and until a sufficient number of draws from the posterior are achieved ($d = b + D$):
    a. Draw |β| independent values from a standard normal distribution, $\phi \sim Normal(0, \mathbb{I}(|\beta|))$
    b. For each parameter, generate a new value, $\beta^{\{d\}} = \beta^{\{d-1\}} + \phi$
    c. Calculate the likelihood given the new parameter values: $l^{\{d\}} = Prob(y = Y|\beta^{\{d\}})$
    d. Draw a value from a uniform random distribution $\zeta \sim Uniform(0,1)$.
    e. Set $l^{\{d\}} = l^{\{d-1\}}$ if $\zeta > \min\left\{1, \frac{l^{\{d\}}}{l^{\{d-1\}}}\right\}$. Otherwise, retain $l^{\{d\}}$.[44]
    f. Increment d by 1
3. Use the draws $\beta^{\{d\}}$ for $d > b$ to calculate any posterior measures required (e.g. mean, standard deviation, confidence intervals)

In addition to deriving posterior measures corresponding to the parameters themselves, the parameter draws can be used for constructing predictive distribution as described in the next section.

[44] Note that the posterior does not include the prior in this calculation because the prior is assumed to be essentially flat (i.e. normally distributed with very large variance). Thus, since the prior probability for a value of β, $\pi(\beta)$, is the same for all values of β, $\zeta > \min\left\{1, \frac{l^{\{d\}}\pi(\beta^{\{d\}})}{l^{\{d-1\}}\pi(\beta^{\{d-1\}})}\right\} = \min\left\{1, \frac{l^{\{d\}}}{l^{\{d-1\}}}\right\}$.

8.3.3 Constructing Predictive Distributions and Performing Hypothesis Tests

The application of choice models often involves using model parameters to make predictions. For example, in the analysis of latent class models, class memberships in a population are often calculated using a point estimate – the mean parameter values. Because only one set of parameter values is used, the class membership is presented as a singular value with no degree of uncertainty. The uncertainty in class membership could be derived because the parameter estimates from classical MLE have an asymptotic distributional assumption, but this derivation can be tedious. Hess et al. (2011) show that parameters exhibit correlation in latent class logit models, which also increases the complexity of deriving uncertainty analytically.

The MCMC procedures used for obtaining posterior distributions of the parameters provide a useful technique for describing the uncertainty of a model and thus that model's predictions as well. The draws from the posterior used for model estimation can be reused in an analysis that depends on the parameter estimates. This has the advantage that the technique is straightforward to implement, asymptotic assumptions are not needed, and distributions do not need to be derived. As long as the draws from the estimation are retained, they can be used to run the analysis on each draw of the parameter set. Additionally, this makes the procedure easy to parallelize in computer applications. To construct a predictive distribution for some predictor of interest, the following procedure can be used:

1. Obtain draws of parameter values from the model's posterior density, $\beta^{\{d\}}$, where $d = b + 1, b + 2, b + 3, \ldots, D$

2. Create a function $\mathcal{A}(N, \beta) \to \mathcal{Y}$ which takes input from a population N and a set of parameter values β and creates a prediction $\mathcal{Y}$,
3. Iterate through the set of draws $d = \{b + 1, b + 2, b + 3, \dots, D\}$:
    a. Run the analysis on $\beta^{\{d\}}$ and store the prediction, $\mathcal{A}\left(N, \beta^{\{d\}}\right) \to \mathcal{Y}^{\{d\}}$
4. Use the predictions $\mathcal{Y}^{\{d\}}$ for $d = b + 1, b + 2, b + 3, \dots, D$ to describe the uncertainty in the prediction by calculating any measures required (e.g. mean, confidence intervals)

Examples of analysis functions include the calculation of the location and number of equilibria (see section 8.2.3), class membership probabilities (as performed in section 9.6.1), and parameter elasticity (as performed in section 9.6.3). Hypothesis tests can also be performed by altering the analysis function to output a hypothesis test result. The confidence interval for a test is obtained simply by obtaining "the fraction of times the MCMC draws are in the hypothesized region" (Lenk 2014, p. 488).

8.4 Endogeneity in the Binary Choice Informational Conformity Model

Social influence choice models are susceptible to endogeneity bias in model estimation. This is primarily due to three common sources of correlation in these models: (1) correlated individual-level and environmental-level effects; (2) social network link formation due to homophily of behavior, opinion, and values; and (3) behavioral feedback between an individual's behavior and the behavior of others[45]. In this section,

[45] See Appendix B for a more in-depth overview of endogeneity and methods of handling it that have been used in previous research on social influence choice models.

we will consider the case of correlated unobservables in the latent class binary choice model of informational conformity derived in the previous section.

8.4.1 Two-step Control Function Approach

The two-step control function approach is chosen due to flexibility compared to the BLP approach for social influence choice models and due to being simpler to code in commercial software[46] than both the BLP and simultaneous control function approach. Specifically, the BLP approach is useful when a study has social networks comprised of large reflexive groups (Walker et al. 2011) – which are analogous to market-level effects. When non-reflexive networks are used, each individual may have a different value for the social influence term. Thus the control function approach is more flexible in this respect.

To simplify the explanation, the average behavior of individual *n*'s social contacts will be denoted as follows:

$$\bar{y}_n = \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} \tag{61}$$

Using equation (61), the informational conformity model can be rewritten as follows:

$$\begin{gathered}
\mathcal{F}_n = \alpha z_n + \delta \bar{y}_n + \varepsilon_n^{\mathcal{F}} \\
c_n = \begin{cases} a & if\ \mathcal{F}_n \geq 0 \\ \mathscr{b} & otherwise \end{cases} \\
\Delta\mathcal{U}_n^{[a]} = \beta^{[a]} x_n^{[a]} + \beta^* x_n^* + \varepsilon_n^{[a]} \\
\Delta\mathcal{U}_n^{[\mathscr{b}]} = \beta^{[\mathscr{b}]} x_n^{[\mathscr{b}]} + \beta^* x_n^* + \varepsilon_n^{[\mathscr{b}]} \\
y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[c_n]} \geq 0 \\ 0 & otherwise \end{cases}
\end{gathered} \tag{62}$$

[46] This trade-off is pertinent for use in forecasting where practitioners often do not write custom code for estimating choice models.

The main issue is that due to correlated unobservables between an individual and the individual's social contacts, $\bar{y}_n$ may be correlated with the error terms $\varepsilon_n^{[a]}, \varepsilon_n^{[b]}$ in the choice model portion. In this latent class model context, the endogenous variable is located in the class membership model while the correlated error term is in the choice model. Unfortunately, this separation does not prevent endogeneity bias as shown in the simulation study in section 8.3.2. Thus, we still may have bias in the estimation of parameters $\alpha, \beta,$ and δ.

To handle this endogeneity in the social influence term, the two-stage control function approach attempts to find an appropriate function for the endogenous variable that uses instrumented variables w_n:

$$\bar{y}_n = \theta_n w_n + \nu_n \tag{63}$$

$$E(w_n \nu_n) = 0, \qquad E\left(\nu_n \varepsilon_n^{[a]}\right) \neq 0, \qquad E\left(\nu_n \varepsilon_n^{[b]}\right) \neq 0 \tag{64}$$

where:

w_n ≡ instrumented variables of individual *n*

θ ≡ social influence measure model parameters

ν_n ≡ error term for individual *n*

For the first step in the control function approach, $\bar{y}_n$ is regressed on the instruments w_n and residuals $\hat{\nu}_n$ are obtained using ordinary least squares regression (OLS). Then, in the second step, these residuals are inserted into the choice model and parameter estimates are obtained using the original explanatory variables and the residuals:

$$\begin{gathered} \mathcal{F}_n = \alpha z_n + \delta \bar{y}_n + \varepsilon_n^{\mathcal{F}} \\ c_n = \begin{cases} a & if\ \mathcal{F}_n \geq 0 \\ b & otherwise \end{cases} \end{gathered} \tag{65}$$

$$\Delta\mathcal{U}_n^{[a]} = \beta^{[a]} x_n^{[a]} + \beta^* x_n^* + \beta_v^{[a]} \hat{v}_n + \varepsilon_n^{[a]}$$
$$\Delta\mathcal{U}_n^{[b]} = \beta^{[b]} x_n^{[b]} + \beta^* x_n^* + \beta_v^{[b]} \hat{v}_n + \varepsilon_n^{[b]}$$
$$y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[c_n]} \geq 0 \\ 0 & otherwise \end{cases}$$

where:

$\beta_v^{[a]}, \beta_v^{[b]} \equiv$ parameters for the OLS residuals from the first step of the control function approach (these can be class-specific)

8.4.2 Simulation Study

To test the control function approach, a Monte Carlo experiment was set up with a synthetic population of 10,000 agents. Each agent is placed into a class according to a class membership generation process and then makes a binary choice $y_n = \{0,1\}$. The data generation and model estimation were both performed in the R programming language. Model estimation was performed using the maxLik package (Henninsen and Toomet 2011). In this experiment, the true data generating process is as follows[47]:

$$\begin{gathered} \bar{y}_n = 1.5 \cdot w_n + 0.3 \cdot v_n \\ \mathcal{F}_n = -1 + 1 \cdot \bar{y}_n + \varepsilon_n^{\mathcal{F}} \\ c_n = \begin{cases} a & if\ \mathcal{F}_n \geq 0 \\ b & otherwise \end{cases} \\ \Delta\mathcal{U}_n^{[a]} = -1 + (-1) \cdot x_n + v_n + \varepsilon_n^{[a]} \\ \Delta\mathcal{U}_n^{[b]} = 1 + 0.2 \cdot x_n + v_n + \varepsilon_n^{[b]} \\ y_n = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[c_n]} \geq 0 \\ 0 & otherwise \end{cases} \end{gathered} \tag{66}$$

where:

[47] Simulations were also run by assuming a uniform and lognormal distribution for v_n. Model estimates were still recoverable under those distributional assumptions.

$w_n \sim iid\ Normal(-0.5,0.5)$

$v_n \sim iid\ Normal(0.4,1)$

$x_n \sim iid\ Normal(0,1)$

$\varepsilon_n^{[a]}, \varepsilon_n^{[b]}, \varepsilon_n^{\mathcal{F}} \sim iid\ Logistic(0,1)$

The experiment begins by generating the population of agents and generating class and choice assignments for each. Second, a naïve model is estimated that does not account for endogeneity between $\bar{y}_n$ and $\varepsilon_n^{[a]}, \varepsilon_n^{[b]}$ due to v_n. The naïve model will take the following form:

$$\mathcal{F}_n = \alpha_0 + \delta \cdot \bar{y}_n + \varepsilon_n^{\mathcal{F}}$$

$$c_n = \begin{cases} a & if\ \mathcal{F}_n \geq 0 \\ b & otherwise \end{cases}$$

$$\Delta\mathcal{U}_n^{[a]} = \beta_0^{[a]} + \beta_1^{[a]} \cdot x_n + \varepsilon_n^{[a]} \tag{67}$$

$$\Delta\mathcal{U}_n^{[b]} = \beta_0^{[b]} + \beta_1^{[b]} \cdot x_n + \varepsilon_n^{[b]}$$

$$y_n = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[c_n]} \geq 0 \\ 0 & otherwise \end{cases}$$

Due to omitted variable bias, the expectation is that the estimated model's $\varepsilon_n^{[a]}$ will attempt to represent the sum of the true $\varepsilon_n^{[a]}$ and v_n (i.e. a sum of a logistic and normal random variable). The results of this naïve estimation are shown in Table 22.

Table 22. Naïve Model Estimates for LC Endogeneity Experiment

Parameter	True Value	Estimate	Std. Error	T-value
α_0	-1.00	-1.71	0.31	-5.59
δ	1.00	-1.45	0.38	-3.84
$\beta_0^{[a]}$	-1.00	0.55	0.08	6.98
$\beta_1^{[a]}$	1.00	0.27	0.07	3.91
$\beta_0^{[b]}$	1.00	1.01	0.03	29.78

$\beta_1^{[b]}$	0.20	-0.23	0.04	-6.56

The naïve results show bias in all variables with a reversal of sign for the δ, $\beta_0^{[a]}$, and $\beta_1^{[a]}$ parameters. Additionally, the ratio of $\beta_1^{[a]}/\beta_1^{[b]}$ also varies from the true model.

In the third part of the experiment, the two-stage control function approach is applied. For the first stage, $\bar{y}_n$ is regressed on w_n as follows:

Table 23. OLS Regression Results for 2SCF Approach (LC Endogeneity Experiment)

Parameter	Estimate	Std. Error	T-value
intercept	0.12	0.0013	90.5
θ	1.50	0.0019	795.2
Model Statistics: Observations	100000		
Adjusted R^2	0.863		
F-statistic	632000		

Additionally the distribution of the residuals is as follows:

Table 24. OLS Residual Statistics for 2SCF Approach (LC Endogeneity Experiment)

Statistic	Value
Minimum	-1.33
1st Quartile	-0.20
Median	0.00
3rd Quartile	0.20
Maximum	1.27

In the second stage, the OLS residuals $\hat{v}_n$ are placed in the latent class model and the corrected model is estimated. The results of the estimation for the correct model are shown in Table 25.

Table 25. Endogeneity Corrected Model Estimates for LC Endogeneity Experiment

Parameter	True Value	Estimate	Std. Error	T-value
α_0	-1.00	-1.05	0.08	-12.46
δ	1.00	0.98	0.06	17.15
$\beta_0^{[a]}$	-1.00	-0.63	0.14	-4.49
$\beta_1^{[a]}$	1.00	-1.03	0.09	-11.86
$\beta_\nu^{[a]}$	--	3.19	0.21	15.35
$\beta_0^{[b]}$	1.00	1.43	0.03	43.13
$\beta_1^{[b]}$	0.20	0.21	0.02	12.03
$\beta_\nu^{[b]}$	--	3.41	0.06	55.30

In this corrected model, the class membership parameters are now all statistically equal to their actual values. Additionally, $\beta_1^{[a]}$ and $\beta_1^{[b]}$ are now statistically equal to their actual values as well. The constants for the two classes are not equivalent, but this is due to the constants absorbing the mean of the omitted variable ν_n. Each constant is biased multiplicatively[48] by the mean of ν_n which is 0.4. Additionally the parameter estimates for the residuals are significant thus confirming that endogeneity is present.

8.5 Summary

In this chapter, an informational conformity model of discrete choice is proposed using a confirmatory latent class choice model framework. A binary choice formulation is proposed using two classes: a "more informed" and "less informed" class. Class membership depends on the decision maker and environmental characteristics as well as endogenous social influence. Once separated into separate classes, the choice model which depends only on decision maker and environmental characteristics, exhibits

[48] This result was confirmed in another experiment in which the mean of ν_n was changed to 0.2 and similarly the constants were also biased by a multiplicative factor of 0.2.

differing coefficients between decision-makers in each class. Because of these features, it is a type of indirect social influence model which is unique in the travel behavior literature. From a behavioral side, the model does not exhibit properties similar to a direct-benefit formulation as the behavior of others does not in itself increase the favorability of an alternative.

For this formulation, the equilibrium properties of the model are derived. Under homogeneity and mild behavioral assumptions, the model exhibits behavior similar to other social influence / adoption models – namely, the familiar "s-shape curve." Multiple equilibria are possible in this formulation, particularly determined by the ratio of endogenous social influence effect as compared to other non-endogenous effects in the class membership model.

A Bayesian inference procedure is suggested due to optimization concerns, statistical hypothesis testing, and construction of forecast distributions. The procedure's flexibility makes it applicable to other social influence choice models since it provides a technique to construct distributions of equilibria market shares as well as distributions of the quantity of equilibria. Additionally, Bayesian procedures' flexibility allows for the use of prior information if available.

Additionally, this chapter covered endogeneity issues with a two-step control function approach. Although not the only procedure available (BLP can also be used), the control function approach is flexible since it works for differing social network structures – whereas the BLP approach requires reflexive cliques. A simulation study was performed and model estimates were recovered.

In conclusion, this chapter presented a tractable model of indirect social influence based on a motivation for accuracy. A binary choice formulation was presented in this chapter and will be used in a case study on bicycle ownership in Chapter 9. Additionally, Appendix D provides a generalized model formulation that expands the model to handle not only preference difference, but also expectation and constraint differences. This generalized model could be applied in future research.

Chapter 9: Informational Conformity Case Study – Bicycle Ownership in the United States

Table 26. Chapter 9 Summary

Background & Brief Summary	Qualitative studies have noted the importance of social influence in cycling behavior. This can start a process in which the individual may begin to research the suitability of cycling and adjust their opinions and behaviors. To test for this informational conformity statistically, bicycle ownership is modeled on a national scale by using data from the National Household Travel Survey. Specifically, a latent class discrete choice model is formulated which places individuals into classes based on information exposure. Preferences for bicycle ownership vary between these classes. Information is signaled by city-level bicycle usage where greater usage may induce households to change their preferences.
Motivation	Existing research in cycling behavior has found the existence of social influence, along with household factors and built environment factors. There have been limited studies of social influence in cycling behavior and none have explored informational conformity.
Results	1. Informational conformity model showed that "more informed" households had a greater chance of owning a bike due to preference changes, instead of direct benefits from others' behaviors, with less sensitivity to smaller home footprints and limited incomes 2. The Bayesian inference and application procedure shown in section 8.3 was successfully demonstrated 3. The behavioral hypothesis of positive preference change due to information transfer was confirmed 4. Observed ownership share matched predicted local-level equilibrium in many MSAs and the model did not show a strong bias towards overpredicting or underpredicting 5. The elasticity of social influence was found to range locally from about 0.50% to 0.80%
Limitations	1. The dataset has no information on individual's social contacts or their social groups 2. The analysis uses choice-data approach to determine a latent social influence process 3. This is not a controlled experiment so the model can only determine that this is a possible behavioral process explanation that may exist 4. Without time-series or panel data sources, equilibrium cannot be confirmed to have been achieved in the observed ownership shares

There is much interest in bicycling for promoting livable communities, improving public health and the environment, and reducing energy usage. Bicycling has the benefit of being a cheap, efficient mode of travel. According to the National Bicycling and Walking Study (Federal Highway Administration 2010), bicycle funding has grown from less than 0.5% of funding in 1993 to about 2.0% of funding in 2009. Additionally, the study stated that bike trips have grown to about four billion trips as reported in the 2009 National Household Travel Survey (NHTS). But even with this growth, cycling is heavily underutilized by American households. Approximately one percent of all trips are performed by bicycle even though a large number of American household trips are within biking distance.

Wigan (1984) proposed that understanding which households owned bicycles was important for determining bicycle usage. Measuring a person’s access to a bicycle directly (e.g. a bicycle belongs to Person A and Person B is not allowed to ride it) is generally not done in travel surveys. Therefore, household bicycle ownership is perceived to be a proxy in its place. But bicycle travel surveys tend to be inconsistent, with some asking about bicycle ownership while others do not. Analyses that have included bicycle ownership have found it to have a significant effect on many aspects of cyclist behavior.

9.1 Prior Research

9.1.1 Studies of Bicycle Ownership

For example, Sener et al. (2009) analyzed perceptions of bicycle facilities and safety as well as trip frequency for both commute and non-commute trips. This research found that bicycle ownership (quantity of bicycles) was a determinant of commuting and non-commuting bicycle usage. Bicycle ownership more heavily influenced non-commute trips, but the effect was statistically significant for both trip types. Sener et al. (2009) summarized some major findings from the field of bicycle research:

- Men bike more often than women
- Individuals between 25 and 45 years of age are more likely to bicycle
- Caucasians and part-time workers are more likely to bike
- Higher income households are more likely to have cyclists
- Bicycle trip frequency decreases with vehicle ownership and increasing household size
- Good land use mix encourages cycling

Vij et al. (2013) attempted to understand and quantify different modality styles. The study assumed that individuals have lifestyles which affect their mobility style, or factors which determine how often a person can travel. One's mobility style then has an effect on his modality style which is his tendency to use a mode type when traveling. Bicycle ownership, a mobility style, had a positive effect on individuals who had a tendency to be multimodal and individuals who were multimodal and environmental conscious. Pinjari et al. (2009) modeled residential location and activity time-use choice using a joint mixed logit and multiple discrete-continuous extreme value (MDCEV) model. This study proposed that people who own a bicycle may be more physically active. It found that an increase in household bicycle ownership of one bike decreased in-home relaxation time by 27%. The joint model showed that bicycle facility design mostly affected those with bicycles. Additionally the study claimed that households with more bicycles tended to pursue physically active recreational travel and self-selected into neighborhoods with good cycling infrastructure. This effect was also found in Pinjari et al. (2011), a study that examined self-selection effects and the built environment's impact on travel. Additional finding from this study were that bicycle ownership was positively correlated with bicycle mode usage and that higher population and employment densities encouraged non-motorized transport. Bhat et al. (2006) stated that the number of bikes was positively correlated with out-of-home recreation but negatively correlated with social activities and in-home relaxation and recreation.

More research on the connection between activities and bike ownership includes Bhat and Srinivasan (2005) which examined the frequency of weekend activities. The quantity of adult and children bikes was positively correlated with physical recreation activities – with children bike ownership having a greater effect. The study's authors were unsure if this effect was casual or just spurious. Bhat and Lockwood (2004) focused on out-of-home weekend recreation. This research found that young adults (16-17 years old) are less likely to participate in physically active recreation, such as bicycling, compared to older adults. This study again cautioned that the relationship between bicycle ownership and physical activity may not be causal, but it still advocates that policies which encourage non-motorized mode ownership may foster physically active recreational pursuits. Additionally, Bhat and Gossen (2004) modeled recreation activity choice using a mixed logit model and found that bicycle ownership may have encouraged out-of-home recreation activities.

Additional studies about the influences of bicycle ownership on cyclist travel behavior include:

- Waller (1971) analyzed bicycle ownership among young children (3-12 years old) in Vermont. A survey was conducted which asked about bicycle style, ownership duration, age of first bike ride, and injuries. This study found that bicycle style affected injury rates and that the activity level of children affected bicycle style.
- Owen et al. (2010) analyzed differences in bicycle ownership in Australia and Belgium. This study found no difference in bicycle ownership between men and women and that bicycle usage was more likely in areas with high walkability.
- Primerano (2006) analyzed travel mode choice but restricted the bicycle mode choice to only households with bicycles, thus making bicycle ownership a determinant of mode availability. This exclusion allowed for less biased alternative specific constant estimates for the bicycle mode.
- Handy et al. (2005) found that young bicycle owners with higher levels of education were more likely to travel by bicycle.
- Ryley (2008) collected bicycle ownership data in the United Kingdom and found that bicycle ownership was lower in areas with shorter travel distance. This was attributed to bicycle storage problems in some urban neighborhood. The lack of a bicycle was also cited as a major reason for not riding a bicycle.

As shown above, bicycle ownership affects many aspects of bicycle use. Therefore directly modeling bicycle ownership may provide insight into bicyclist travel behavior. The author has found only three examples in the literature of estimated bicycle ownership models.

Pinjari et al. (2011) proposed to use a joint modeling approach that looked at bicycle ownership and neighborhood type choice. Bicycle ownership was modeled using an ordered logit formulation. Variables which positively correlated with bicycle ownership included number of active adults in the household, number of children, male householder, Caucasian households, household income, home ownership, living in a single-family dwelling, and living in bicycle-friendly neighborhoods. In contrast, single-person households and having a householder over 60 years of age were negatively correlated. The research found that ignoring self-selection effects in residence choice can lead to underestimation of the effects of the built environment on bike ownership. It also found bicycle ownership was influenced more by sociodemographics characteristics than neighborhood type.

Yamamoto (2009) modeled bicycle, automobile, and motorcycle ownership in Osaka, Japan and Kuala Lumpur, Malaysia. The study found that bicycle ownership was affected by the number of workers, retirees, and children, population density, public transit accessibility, income, and land use mix. Multinomial logit and trivariate probit models were used to study the relationship.

Handy et al. (2010) examined bicycle ownership in six US cities. A nested logit model of bicycle ownership, cycling frequency, and trip purpose was estimated. The study found that income and being Caucasian had positive effects on cycling while aging had a negative effect. They also found that the perceptions of cyclists had negative effects such as the perception that cyclists are poor or spend too much on bicycles.

9.1.2 Social Influence Studies of Cycling Behavior

Recently, there have been a number of studies finding that social influence affects cycling behavior. Using a survey in Portugal and Belgium, Bourdeauhuij et al. (2005) found that utilitarian and recreational biking trips were both impacted by levels of social support and social norms. Among both Portuguese and Belgian adults, social support from friends was significantly correlated with frequency of cycling. Additionally, social norms impacted cycling and varied between the two locations. Sherwin et al. (2014) used semi-structured interviews and thematic analysis to analyze cycling behavior in the UK. Their research found that individuals experienced direct social influence from family, friends, co-workers, and government programs. Additionally, individuals also experienced indirect social influence from seeing strangers cycle, varying cycling culture between towns, and gender norms. Their paper specifically mentions two quotes that support a theme of information social influence in cycling:

> *"I've encouraged others actually, cause lots of the children said to their parents 'Oh I want to come to school on the bikes', so it kind of started a few people doing it."* (p. 41)

> *"You see people on their bikes, you see all ages from young to really old people on their bikes(..) I quickly worked out that I could get to the shops on cycle lanes without going on a road, so I started going out on my bike"* (p. 42)

Both quotes show that individuals can influence others who they are not in direct contact with through their actions. In the first quote, children changed their perceptions of bicycling. And in the second quote, a woman noticed the cycling patterns of others to infer the availability of bicycle routes.

Goetzke and Rave (2011) use a binary logit model to study social influence for bicycle trips in 20 German municipalities. Their work found that social influence effects were correlated with shopping and recreational bicycle trip generation but not for work/school or errand trips. Fukada and Morichi (2007) studied illegal bicycle parking behavior and social influence in Tokyo. Their models found that social influence was a determinant in parking behavior. Using an equilibrium analysis, they suggested that police intervention could be used to shift aggregate parking behavior to more legal parking. Additional studies that have found an effect between social influence and cycling use through the use of discrete choice models include Dugundji and Walker (2005), Walker et al. (2011), and Pike (2014, 2015).

9.2 Objective and Contributions

Although social influence has been identified as a factor in bicycle behavior including mode choice and illegal parking behavior, no study has studied the effect of social influence on bicycle ownership. This chapter proposes to contribute to knowledge about the role of social influence in travel behavior – in particular, bicycle ownership – through the use of social influence choice models. Using data from the 2001 National Household Travel Survey, an informational conformity model (Figure 29) is estimated and compared to a direct-benefit conformity choice model and a non-social choice model.

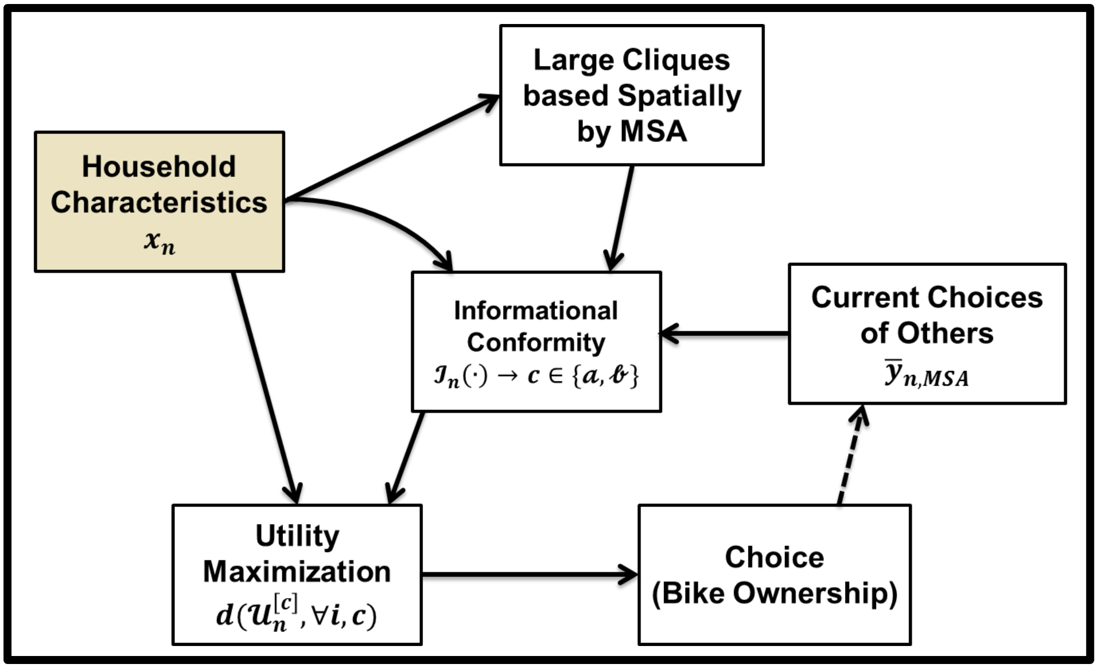

Figure 29. Informational Conformity Model of Bike Ownership

This study aims (1) to determine if social influence and bicycle ownership are correlated, (2) to understand the differences in behavioral explanations of social and non-social factors, and (3) to use the informational conformity model to analyze the effect of informational conformity on levels of information class membership, local-area ownership equilibrium, and social influence elasticity. This chapter makes the following contributions:

1. Confirms the hypothesis of correlation between social influence and bicycle ownership in the United States
2. Provides a behavioral explanation to account for some of the regional and local variations in bicycle ownership
3. Confirms the behavioral hypothesis that "more informed" households experience a preference change that induces higher bicycle ownership probability

4. Derives measures of the uncertainty in the effects of social influence in respect to information class membership, ownership equilibrium, and social influence elasticity

9.3 Data

For this case study, the 2001 National Household Travel Survey (NHTS) was used (Federal Highway Administration 2001). Although the 2009 NHTS dataset is more recent, bicycle ownership was not measured in that survey. The NHTS is a national travel survey funded by the US Federal Highway Administration, Bureau of Transportation Statistics, and National Highway Transportation Safety Administration. The 2001 NHTS collected data about households and their travel habits. The analysis in this chapter will only consider household level data from this dataset. The survey collected information about households directly through telephone interviews and travel diaries and some built environment variables were also included in the dataset (e.g. population density).

The 2001 NHTS consisted of a total sample of 69,817 interviewed households with 26,038 households from the national sample and 43,779 households from the nine add-on areas. The analysis in this chapter will draw from the total sample excluding household not in one of the 50 largest metropolitan statistical areas (MSAs). From these households in MSAs, household without educational, age, bike quantity, home ownership, home type, or race data were excluded. Additionally, households that were college dorms or owned by a respondent's job or the military were also excluded. Thus, the analysis sample size is 25,563 households. Because of oversampling in the add-on areas and limitations in survey recruitment, the analysis will use the sample weights provided in the NHTS dataset to reduce sampling bias for MSA-wide bicycle ownership and for scenario analysis.

Table 27 summarizes some characteristics about the households in the sample. About 53% of households owned at least one bicycle. Taking account of weighting, about 54% of American households in the 50 largest MSAs owned a bicycle. Most homes that owned a bicycle have one or two bicycles, while a small percentage owned three or more bicycles. The average household size was 2.54 persons with about 26% of households having children between 6 and 17 years of age[49]. Most households had at least one vehicle with a median of two vehicles.

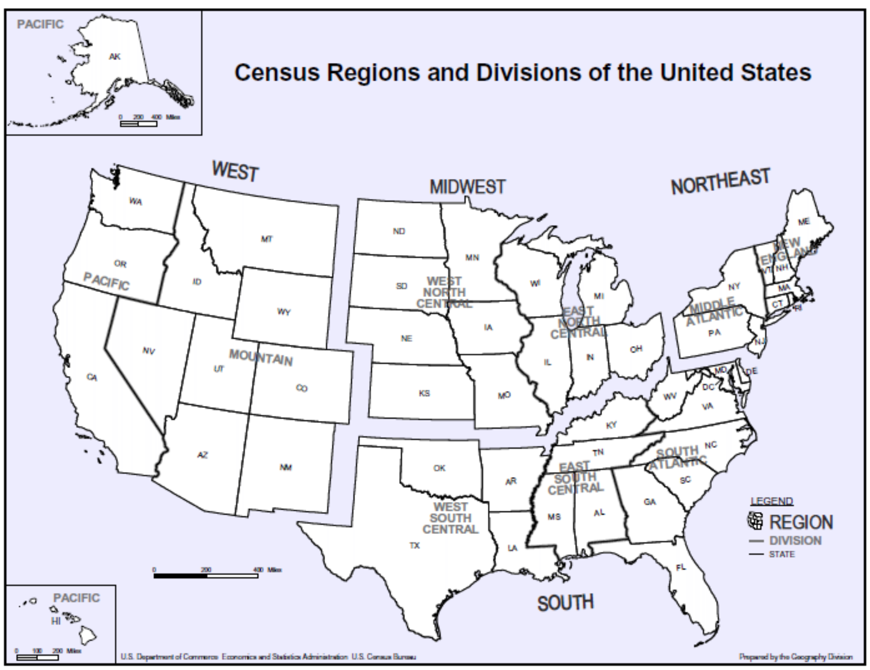

Figure 30. US Census Divisions (US Census Bureau, n.d.)

[49] The 6-17 year age group was used as this is the age at which children may be able to ride adult-sized bicycles.

Table 27. Descriptive Statistics from 2001 NHTS (Only Households in MSAs)

Variable	Value Label	Value	Variable	Value Label	Value
Sample Size	Households	25,563	Age	6-17 years old	18.3%
Bike Ownership	0 Bicycles	53.0%		18-54 years old	54.4%
	1 Bicycle	18.0%		55 years or older	27.3%
	2 Bicycles	19.0%	Home Tenure	Own	74.4%
	3 Bicycles	5.5%		Rent	25.6%
	4 or More Bicycles	4.5%	Home Type	Detached Home	63.7%
Bike Ownership (Weighted)	0 Bicycles	54.4%		Apartment	15.3%
	1 Bicycle	18.4%		Townhouse	8.5%
	2 Bicycles	17.6%		Duplex	4.1%
	3 Bicycles	5.3%		Mobile Home	1.9%
	4 or More Bicycles	4.3%	Census Division	New England	3.4%
Household Size	Mean	2.54		Mid Atlantic	30.0%
	Median	2.00		East North Central	12.6%
Number of Adults	Mean	1.89		West North Central	2.6%
	Median	2.00		South Atlantic	23.2%
Gender (Adults)	Male	45.2%		East South Central	1.3%
	Female	54.8%		West South Central	7.2%
Number of Children (6-17 years)	Mean	0.43		Mountain	3.2%
	Median	0.00		Pacific	16.4%
	Households with Children	26.1%			
Number of Vehicles	0 Vehicles	9.6%			
	1 Vehicle	29.6%			
	2 Vehicles	39.6%			
	3 Vehicles	14.2%			
	4 or More Vehicles	7.0%			

The distribution of households was geographically skewed towards the areas with add-on samples. Figure 30 (US Census Bureau, n.d.) details which states are in each division. About one-third of the sample was in the Middle Atlantic division and approximately one-quarter of households were in the South Atlantic division. The smallest samples were found in the Mountain, East South Central, West North Central, and New England divisions.

9.4 Model Development

Three models were compared in this case study as follows:

1. *Informational Conformity Latent Class Model*
2. *Direct-Benefit Conformity Logit Model (All Regressors)*
3. *Non-social Logit Model (All Regressors)*

These models will be described in sections 9.4.2 through 9.4.4. In section 9.4.1, the choice of social network for the analysis is described.

9.4.1 Social Network Choice and Justification

Due to a lack of spatial data, the chosen social network for this analysis is a large clique based on MSA. This choice of network is justified for the informational conformity case because information can be conveyed through observation as well as direct contact. It is hypothesized that greater bicycle ownership in a MSA would correlated with a greater chance of seeing other individuals' bicycle through means such as observing others ride, seeing bicycles in a neighbor's garage, or seeing bicycles parked. From these observations, the household may reevaluate their preferences for bicycle ownership.

9.4.2 Informational Conformity Latent Class Model Formulation

This model follows similarly from the specification in section 8.2. As a reminder, the class membership model takes the following form:

$$\mathcal{F}_n = \alpha z_n + \exp(\delta) \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} + \varepsilon_n^{\mathcal{F}} \tag{68}$$

The regressors used in the class membership model, z_n, include the following:

- *Respondent Education* (Base: High School Education)

- *Respondent Race and Ethnicity* (Base: White, Non-Hispanic)
- *Household Census Division*
- *Household Vehicle Ownership*

These regressors were assumed to affect the transfer of information between individuals. Specifically, it was hypothesized that individuals with higher education would be more likely to be "more informed" about cycling. Additionally, it was hypothesized that minority groups would transfer information about cycling less than white households. Due to data limitations, it was assumed that the householder who answered the questionnaire was representative of the household's educational level and racial composition. For census division, it was assumed that some regions may have different cycling tendencies because prior research showed lower bike ownership in the South compared to other regions (Maness 2012). There was no clear hypothesis for the vehicle ownership parameter as greater vehicle ownership has been found to decrease bicycle usage but greater vehicle ownership may be correlated with greater income.

The endogenous social influence source chosen for this model was the MSA-level bicycle ownership average. The social influence effect parameter δ is exponentiated in the estimation to aid in model identification and to make each group fit the behavioral prediction of increasing ownership leading to more membership in class $\mathcal{a}$ (i.e. "more informed" class).

The choice model component involves a binary choice between a household owning a bicycle ($y_n = 1$) and not owning a bicycle ($y_n = 0$). The choice model for each class $\{\mathcal{a}, \mathcal{b}\}$ is defined through a utility function difference with a logistically distributed error term, thus making each class choice model an independent binary logit model:

$$
\begin{aligned}
&\Delta\mathcal{U}_n^{[a]} = \beta^{[a]} x_n^{[a]} + \beta^* x_n^* + \varepsilon_n^{[a]} \\
&y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[a]} \geq 0 \\ 0 & if\ \Delta\mathcal{U}_n^{[a]} < 0 \end{cases}
\end{aligned}
\tag{69}
$$

$$
\begin{aligned}
&\Delta\mathcal{U}_n^{[b]} = \beta^{[b]} x_n^{[b]} + \beta^* x_n^* + \varepsilon_n^{[b]} \\
&y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n^{[b]} \geq 0 \\ 0 & if\ \Delta\mathcal{U}_n^{[b]} < 0 \end{cases}
\end{aligned}
\tag{70}
$$

The regressors that are shared between both classes' models are:

- *Household Number of Children (aged 6-17)*
- *MSA Level Bicycle Ownership Residual*

The number of children in the household was placed in the shared regressors list because children were assumed to have limited direct influence on information dissemination's effect on travel preferences among the adults in the household. So while children likely increased the likelihood of owning a bike, they are modeled here as contributing equally between "more informed" and "less informed" households.

Class-specific regressors were chosen such that it was hypothesized that being informed about the properties of bicycle ownership could change a household's preferences. The regressors that are specific to either class' choice model include:

- *Household Number of Adults (aged 18-54)*
- *Household Number of Women (aged 18-54)*
- *Household Number of Adults (aged 55 and over)*
- *Household Number of Women (aged 55 and over)*
- *Home Rent Status*
- *Home Type* (Base: Detached House)

- *Household Income* (Base: Middle Income, $25,000 - $75,000)

Prior research has shown that the size of a household impacts the number of bicycles owned. Since larger households tend to own more bicycles, it was hypothesized that having more individuals would induce more cycling. Age is also a factor in bicycle ownership and usage and it was hypothesized that older adults would be less likely to own bicycles. Additionally, gender has also been found to be a factor in bicycle ownership as shown in the section 9.1. This prior research has found that women are less likely to own and use bicycles in the United States and that hypothesis is tested in Model 1 as well.

Rent status was included in the specification to indicate the likelihood of moving. Bicycles are oddly shaped devices which are difficult to transport, aside from being ridden. Therefore, it was hypothesized that renting a home would discourage bicycle ownership. For home type, it was assumed that home type was a proxy for the available space for bicycle storage and the ease of access and egress between storage and the street. For detached single-family homes, these homes are more likely to be larger (in terms of available floorspace) or have additional facilities for storage such as garages and sheds. Even though townhouses, duplexes, and mobile homes are also single-family dwelling, these home types are likely to have smaller footprint and fewer storage opportunities. Additionally, apartments also have small footprints and fewer storage opportunities as well as their access and egress may be hampered by stairways and elevators.

Although bicycles are priced such that many American households can afford at least one bicycle (prices can range from about $100 to thousands of dollars), past research has shown that higher income household are still more to cycle than lower income household (Sener et al. 2009). Thus it was hypothesized that lower income households would be less likely to own bicycles as compared to higher income households. For households that did not disclose their income, there was no hypothesis for the direction of this effect. If it is assumed that these incomes are missing at random, then the direction and strength of this effect could reveal the average income of households in this group among the sample.

The likelihood function for this model follows from that shown in section 8.2.1 in equation (49).

9.4.3 Direct-Benefit Conformity Logit Models

The informational conformity model will be compared to a traditional direct-benefit conformity model specification. In this model, the decisions of other individuals in the same MSA (average behavior or mean-effect) are used as a covariate in the regression. This is the common model described in section 7.2.2 as the current state-of-practice. It has its origins in the statistical mechanics specification of Brock and Durlauf (2001). The direct-benefit conformity logit model has the following specification:

$$\begin{gathered} \Delta\mathcal{U}_n = \beta x_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} + \varepsilon_n \\ y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n \geq 0 \\ 0 & if\ \Delta\mathcal{U}_n < 0 \end{cases} \end{gathered} \tag{71}$$

The household-level characteristics x_n include all the variables in the class-membership and choice models of the informational conformity model. It is expected that the direction of the effect of each variable in this model will be similar to the hypothesized direction in the informational conformity described in section 8.4.3. The likelihood function for this model is the same as the common binary logit likelihood.

9.4.4 Non-Social Logit Models

Lastly, the model will also be compared to a non-social logit model. In this model, no social influence is incorporated. The non-social logit model has the following formulation:

$$\Delta\mathcal{U}_n = \beta x_n + \varepsilon_n$$
$$y_{ni} = \begin{cases} 1 & if\ \Delta\mathcal{U}_n \geq 0 \\ 0 & if\ \Delta\mathcal{U}_n < 0 \end{cases} \tag{72}$$

The household-level characteristics x_n include all the variables in the class-membership and choice models of the informational conformity model. It is expected that the direction of the effect of each variable in this model will be similar to the hypothesized direction in the informational conformity class-membership and choice models described in section 8.4.3. The likelihood function for this model is the same as for common logit models.

9.5 Estimation Results

Three model formulations were estimated corresponding to the formulations discussed in section 9.4. Because of correlations in environmental effects among individuals in the same MSA (e.g. similar bicycle infrastructure, recreational facilities, bike shops, public bicycle funding), the mean MSA-level bicycle ownership term was tested for endogeneity. The mean bike ownership of the closest MSA is used as an instrument for a MSA's mean bike ownership in a two-stage control function approach (2SCF). Table 28 shows regression results from the first stage of the 2SCF approach.

Table 28. OLS Regression Results for 2SCF Approach (Bike Ownership)

Parameter Name	Estimate	Std. Error	T-value
Intercept	0.160	0.057	2.80
Closest MSA Bike Ownership	0.658	0.123	5.35
Model Statistics: Observations	50		
Adjusted R^2	0.36		
F-statistic	28.6		

Additionally the distribution of the residuals is as follows:

Table 29. OLS Residual Statistics for 2SCF Approach (Bike Ownership)

Statistic	Value
Minimum	-0.12
1^{st} Quartile	-0.05
Median	0.00
3^{rd} Quartile	0.05
Maximum	0.14

As shown in the regression, this instrument is significant and the R^2 for the model is 0.36 with a F-statistic of 28.6. Although there are no strict tests for weak instruments in discrete choice models (Guevara-Cue 2010), this regression does pass the F-stat test for weak instruments for linear models (Stock et al 2002). It does not pass the R^2 test for weak instruments (Hanh and Hausman 2002), but is close (the threshold is 0.40).

Table 30. Class Membership Model Estimation Results for Informational Conformity Latent Class Model

Parameter	Informational Conformity LC
Class Constant	-3.38*
Mean MSA Bicycle Ownership˜	5.03*
Less than HS Diploma or GED	-0.56*
Associate Degree	0.41*
Bachelor Degree or Higher	0.53*
African-American or Black	-0.62*
Asian-American or Asian	-1.18*
Native American/Pacific Islander	-0.32
Hispanic	-0.34*
Other Race, Non-Hispanic	-0.05
Vehicles per Person in HH	0.61*
HH with No Vehicles	0.27*
New England Census Division	-0.08
Middle Atlantic Census Division	0.15**
South Atlantic Census Division	-0.05
East North Central Division	0.11
West North Central Division	-0.13
East South Central Division	-0.40
West South Central Division	-0.34*
Mountain Census Division	0.09
HH Located in Hawaii	0.09
Note: * denotes estimate p-value ≤ 0.05. ** denotes estimate p-value > 0.10 and < 0.05. ˜ denotes estimate is the natural exponential function of the model estimated value	

Table 31. Choice Model Estimation Results for Binary Choice of Bike Ownership

Parameter	Informational Conformity LC		Direct-Benefit Conformity Logit	Non-Social Logit
	More Informed Class	Less Informed Class		
Constant	1.22*	-0.96*	-2.17*	-0.40*
Mean MSA Bicycle Ownership			3.60*	
Mean MSA Bike Own Residual	1.38*		0.80**	
Number of Adults (aged 18-54)	2.62*	0.16*	0.29*	0.29*
Number of Women (aged 18-54)	-1.30*	-0.12	-0.20*	-0.20*
Number of Adults (aged 55+)	0.11	-0.17	-0.11*	-0.12*
Number of Women (aged 55+)	-0.70*	-0.46*	-0.41*	-0.41*
Number of Children (aged 6-17)	0.74*		0.43*	0.55*
Rent Home	-0.48*	-0.61*	-0.26*	-0.26*
Duplex	0.04	-0.38*	-0.10	-0.10
Townhouse / Rowhouse	-0.38**	-0.43*	-0.21*	-0.19*
Apartment	-0.59*	-0.84*	-0.40*	-0.41*
Mobile Home	0.46	-1.05*	-0.34*	-0.36*
Single Person HH	-0.77*	-0.98*	-0.53*	-0.53*
Low Income HH (< $25k)	-0.68*	-0.68*	-0.38*	-0.37*
High Income HH (> $75k)	0.24	0.49*	0.25*	0.22*
HH Income Unknown	-0.43*	-0.28*	-0.22*	-0.22*
Less than HS Diploma or GED			-0.40*	-0.39*
Associate Degree			0.28*	0.29*
Bachelor Degree or Higher			0.36*	0.36*
African-American or Black			-0.37*	-0.39*
Asian-American or Asian			-0.69*	-0.71*
Native American/Pacific Islander			-0.17	-0.20
Hispanic			-0.28*	-0.33*
Other Race, Non-Hispanic			-0.07	-0.09
Vehicles per Person in HH			0.48*	0.48*
HH with No Vehicles			0.16*	0.13**
New England Census Division			-0.02	-0.08
Middle Atlantic Census Division			0.16*	0.16*
South Atlantic Census Division			0.02	0.22*
East North Central Division			0.14*	0.33*
West North Central Division			-0.06	-0.02
East South Central Division			-0.16	-0.56*
West South Central Division			-0.19*	-0.36*
Mountain Census Division			0.08	0.11
HH Located in Hawaii			0.08	0.29*
Model Statistics: Log-Likelihood	-14709[a]		-14745	-14838
AIC	29519[a]		29578	29743
BIC	29935[a]		29871	30020
Number of Parameters	51 (30 choice + 21 class model)		36	34

Note: * denotes estimate p-value ≤ 0.05. ** denotes estimate p-value > 0.10 and < 0.05.
[a] note that these are estimates of the log-likelihood, AIC, and BIC from the MLE estimation using the Bayesian inference as the starting values. MLE estimation results are shown in Appendix G.
Blank cells denote parameters that were not included in that specific model.

Table 31 shows the choice model estimation results for models 1 through 3 when endogeneity from omitted variables is accounted for with a 2SCF approach. Endogeneity in the latent class model (model 1) was found to exist as a t-test of the control function residuals statistic was accepted at the 5% level with a p-value of 0.024. In models 2 and 3, the t-tests of the control function residuals statistics were rejected at the 5% level with p-values of 0.083 and 0.838 respectively.

The class-membership model estimates (Table 30) describe the relative influence of household characteristics and location on the information state of a household. The class-membership model parameter estimates show the following:

- *Mean MSA-level Bicycle Ownership.* The social influence effect was found to strongly influence class membership, but this is strongly countered by the large negative constant term. When considering that average national bicycle ownership is about 45%, social influence plus the constant would account for the first 25% of class membership in the "more informed" class.
- *Respondent Education Level.* Education level was found to be proportional to the likelihood of being in the "more informed" class. Using a high school diploma / GED as the reference group, respondents with an education level below high school diploma belonged to households that were less likely to be in the "more informed" class. In contrast, household with a respondent with a college level degree were more likely to be in the "more informed" class.
- *Respondent Race and Ethnicity.* Households with minority respondents were less likely to be in the "more informed" class as compared to the households of white,

non-Hispanic respondents. This effect was found to be statistically insignificant for Native American and Other Minority Race, Non-Hispanic respondents.

- *Household Vehicle Composition.* Households with more vehicles per person were more likely to be in the "more informed" class. Income was found to be positively correlated with bicycle ownership in both choice models, so this effect may be attributed to higher income households being able to receive or being more receptive to information about bicycle ownership.
- *Household Regional Location.* Region generally had an insignificant impact on the likelihood of being "more informed." Households in the Middle Atlantic census division were more likely to be in the "more informed" class, while households in the West South Central division were less likely.

Overall, these results show that college-educated households and white households are more likely to be influenced by informational conformity in choosing to own a bicycle. But, a significant amount of influence is due to the overall bicycle ownership among other households in a household's MSA.

When looking at the choice model estimates for each class in the informational conformity model, the parameters for the more informed class tend to be greater than the same parameters in the less informed class. There were no model restrictions to enforce this "greater than" relationship which is an encouraging sign since it shows that the data supports this assumption rather than the model forcing said assumption. The class-specific parameter estimates show the following:

- *Household Size and Composition.* For more informed households, adults under the age of 55 in the household have a large impact on the probability of the household owning at least one bicycle. Male householders contribute more than female householders at all adult age levels. For less informed household, the size and composition has little effect on bicycle ownership. Specifically, only men younger than 55 and women over 54 years old have an effect on ownership (positive and negative effect respectively). Men 55 years of age and older have no effect on bicycle ownership while older women have a negative effect on bicycle ownership. Children have a positive effect on bicycle ownership.

- *Home Tenure and Type.* More informed households are less sensitive to household type and they rent their home. This may suggest that these households are willing to accommodate their bicycles in home storage/parking and moving decisions.
- *Single-person Households*. In both classes, single-person households were less likely to own a bicycle. In more informed households, the net effect of household size, composition, and being a single-person household for an adult of age less than 55 was still positive. But, in the less informed households, the effect was negative. For older adult single-person households, the net effect in negative in both classes.
- *Household Income.* For more informed households, low income (less than $25,000) had a negative impact on bike ownership with a similar effect felt by less informed households. Less informed household were more likely to own a bicycle when they had higher incomes (greater than $75,000). Households who withheld income information were less likely to own a bicycle in both information classes of households.

The non-social and direct-benefit conformity models have similar directionality of estimates as compared to the estimates in the informational conformity model. When the non-social and direct-benefit conformity models are compared to each other, most estimates are similar except for the regional fixed-effects. Specifically, for the non-social model, five out of eight census divisions plus the state of Hawaii have significantly different (at the 10% level) bicycle ownership propensity as compared to the Pacific census division. Comparatively, the direct-benefit conformity model has only three census divisions that are significantly different. Additionally, the magnitude of the fixed effect is larger or equivalent for each census division and Hawaii in the non-social model compared to the direct-benefit conformity model. This result shows that the conformity term may account for some of the fixed effect observed between the different regions. Thus the direct-benefit conformity model increases the explanatory power of the choice model by accounting for these fixed effects as due to conformity rather than being unobserved.

9.6 Analysis of Informational Conformity Model

This section analyzes the properties of the informational conformity model estimated in section 9.5. The section begins by presenting the class membership allocations nationally and locally and the uncertainty in class membership allocations. Then, the behavioral hypothesis of a positive change in preference from a shift in information class is tested. The section concludes with distributional analyses of local-level ownership equilibrium and the elasticity of the social influence effect on ownership probability.

9.6.1 City-level Class Membership Allocations

The national average class membership in the "more informed" class for households located in MSAs was 0.369. The posterior distribution of average class membership exhibited a central tendency with median ownership 0.369 and 5th-percentile membership of 0.325 and 95th-percentile membership of 0.418. This national measure masks the vast differences in local-level class membership between MSAs.

Figure 31 shows the distributions of the posterior local-level class membership by MSA. Each graph is scaled on the x-axis by the proportion of individuals in the "more informed" class (class a) from 0% on the left to 70% on the right. The y-axis denotes density of the distribution and its ranges from 0 at the bottom to 14.5 at the top. Most cities experience class membership allocations that are heavily peaked with small spreads of about 20% membership share. Class membership was directly correlated with local bike ownership shares, and the ordering of mean class membership share closely follows the ordering of actual bike ownership share. The cities with the least membership included Memphis and Nashville, with statistically significant portions of posterior class membership of less than 10%. These two cities also distributions with wider spreads of membership share. Grand Rapids exhibited the "most informed" population.

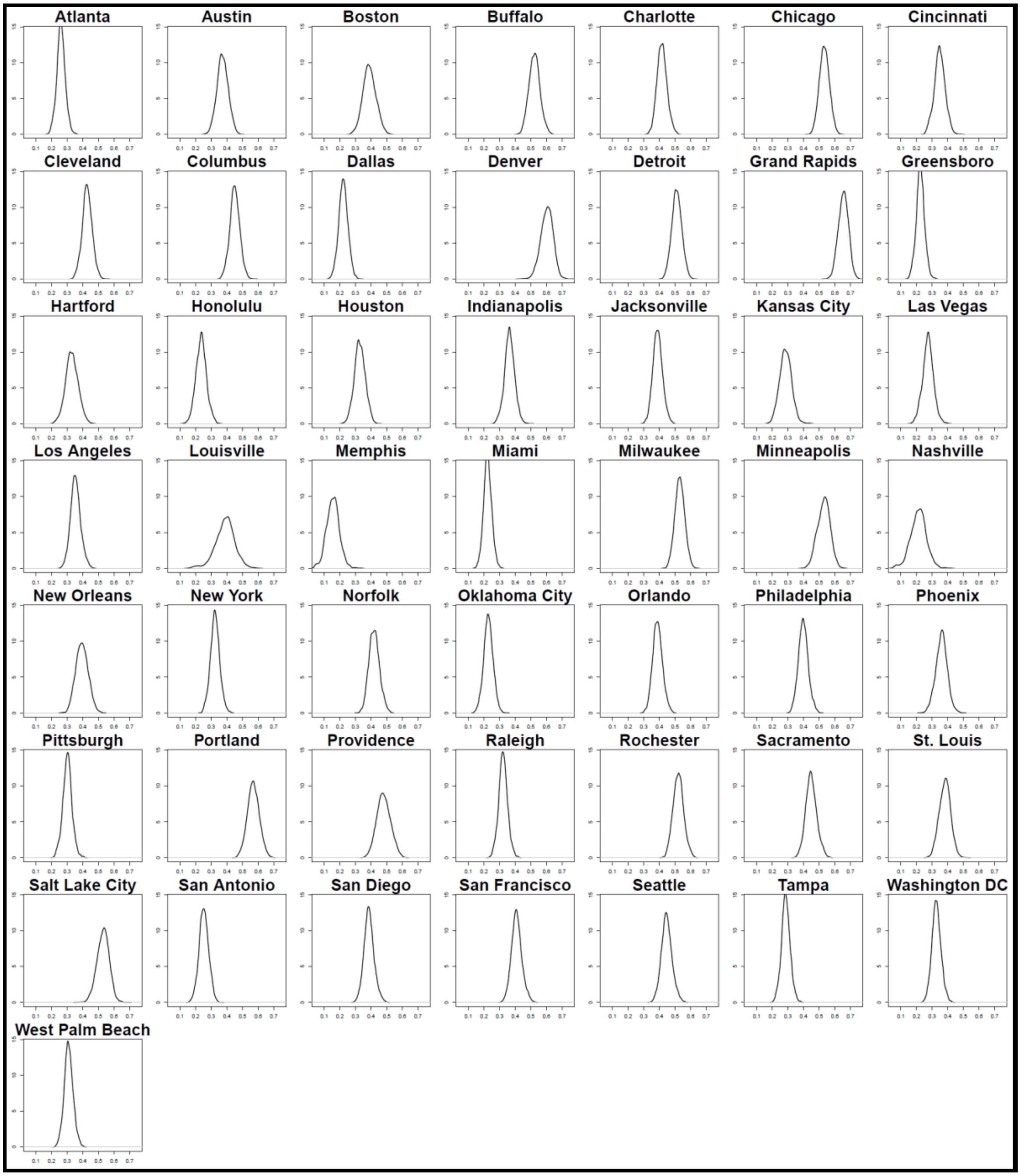

Figure 31. Posterior Class Membership by MSA (Weighted and Same Density Scale)

9.6.2 Testing the Hypothesis of Preference Differential

The major behavioral assumption of the informational conformity model is that the probability of performing an action when "more informed" is greater than when someone is "less informed." This is attributed to a change in individual preferences, $P_n^{[a]} > P_n^{[b]}$. The Bayesian inference framework allows this to be tested in a straightforward fashion by calculating for each individual and each draw, the difference in probability of ownership between class a and b:

$$Prob\left(P^{[a]} > P^{[b]}\right) = \frac{1}{|N| \cdot |D|} \sum_{n \in N} \sum_{d \in D} 1 \left\{ P_n^{[a]}(d) > P_n^{[b]}(d) \right\} \tag{73}$$

where:

$1\{\cdot\}$ $\equiv$ the indicator function; evaluates to 1 if the expression in the curly brackets is true and evaluates to 0 otherwise

Figure 32 and Table 32 describe the distribution of the difference in probabilities between the two classes, $P_n^{[a]} - P_n^{[b]}$. Over the 5000 parameter draws used in the Bayesian estimation of the parameter posterior, none of the draws[50] encountered an instance in which a preference increase did not occur – i.e. $P_n^{[a]} \ngtr P_n^{[b]}$. This confirms the hypothesis that there is a preference change between the classes and that preference changes induces an increased probability in bike ownership among individuals in the "more informed" class.

[50] The minimum over the posterior density was 0.00234 which is very close to zero. Thus, a negative value may be possible, but the hypothesis is confirmed with 99% confidence.

Table 32. Summary Measures of Probability Difference between Informed Classes

Probability Difference at:	Mean	Percentile						
		1%	5%	25%	50%	75%	95%	99%
Posterior Mean	0.563	0.113	0.200	0.423	0.551	0.721	0.872	0.915
Posterior	0.559	0.106	0.193	0.412	0.566	0.721	0.873	0.915

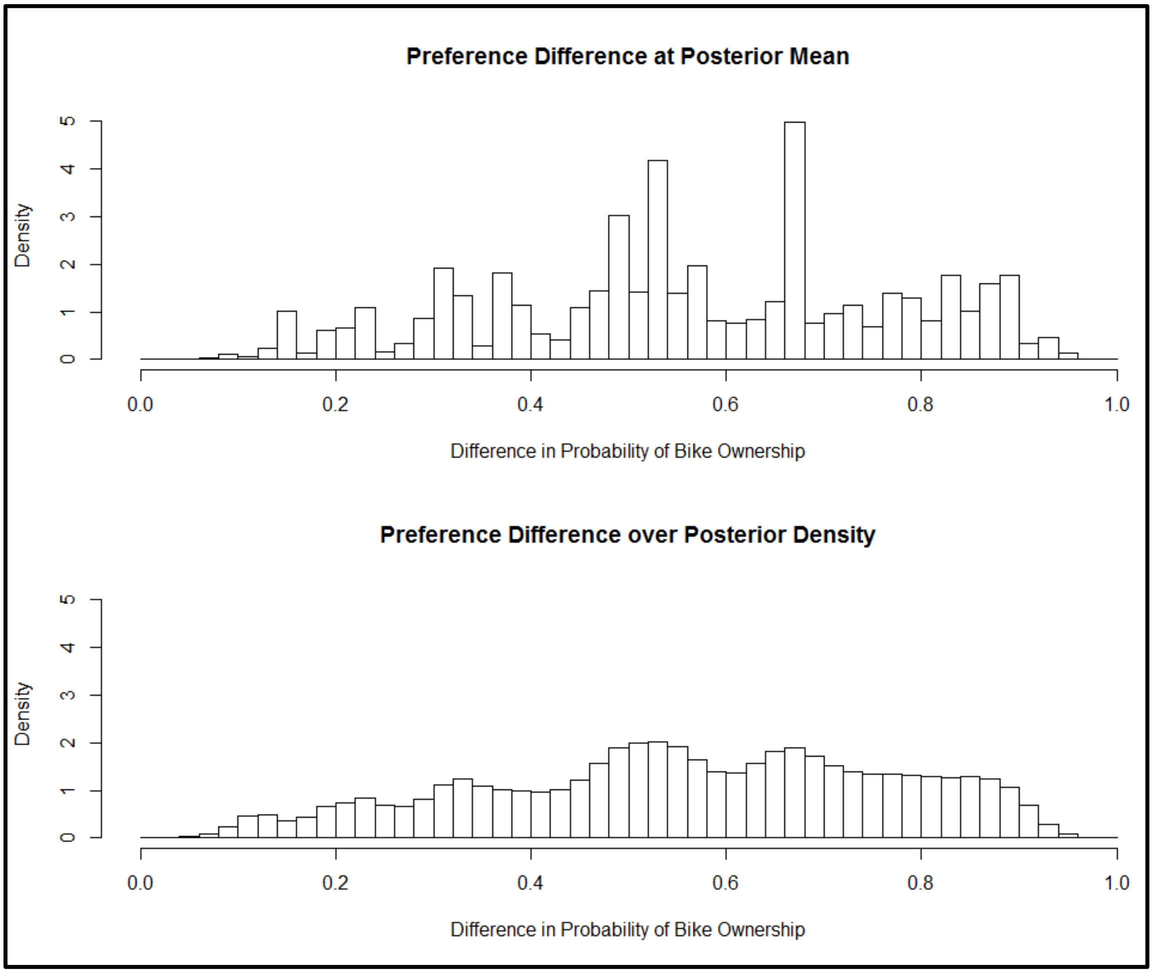

Figure 32. Difference in the Probability of Bike Ownership

9.6.3 Equilibrium Analysis for Different Cities

Predicted equilibria were calculated by approximating MSA-wide behavior:

$$\bar{y} = \frac{1}{|N|}\sum_{n\in N} P_n = \frac{1}{|N|}\sum_{n\in N}\left[\frac{\exp(\alpha z_n + \delta \bar{y}_N)}{1+\exp(\alpha z_n + \delta \bar{y}_N)} P_n^{[a]} + \frac{1}{1+\exp(\alpha z_n + \delta \bar{y}_N)} P_n^{[b]}\right] \qquad (74)$$

The population N in equation (74) corresponds to all sampled individuals in a MSA. Predicted equilibria corresponded to fixed-points of equation (74). Table 33 (ordered by difference between observed ownership and equilibrium predicted ownership) shows the predicted equilibrium ownership at parameter values corresponding to the posterior mean from the informational conformity model.

A linear regression analysis showed that there was no statistically significant relationship between the absolute value of the difference and neither the true ownership, predicted equilibrium ownership, nor the number of observations. The distribution of the difference in true ownership versus predicted equilibrium ownership is shown in the left plot of Figure 33. The mean difference is -1.49% and median difference is 0.45%. The 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles are -10.7%, -5.45%, 3.10%, and 12.5% respectively.

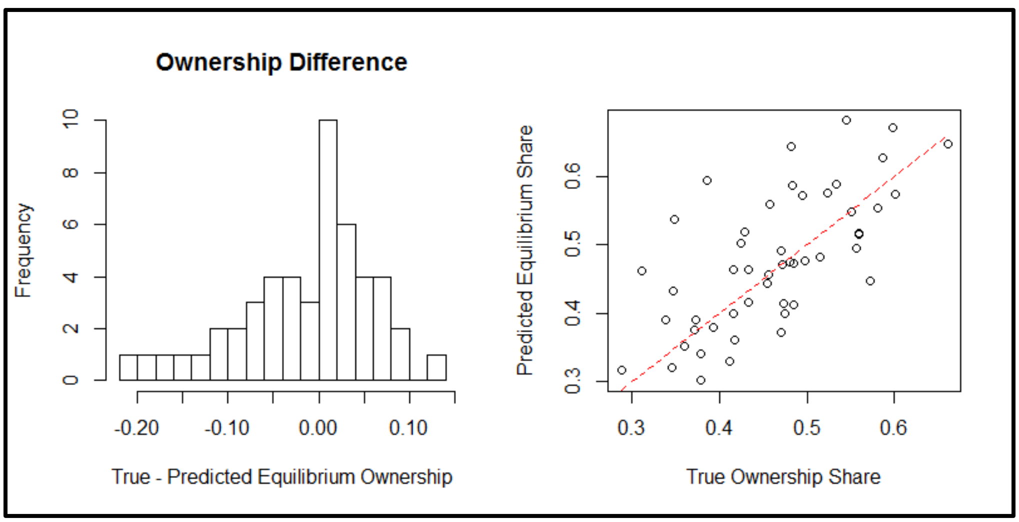

Figure 33. Difference between Actual and Predicted Equilibrium Ownership Shares

The equilibrium analysis was also performed using the posterior distribution of parameter estimates. This allowed for analysis of both the number and location of equilibria. Results found that all MSAs tended towards singular equilibrium conditions – none of the parameter draws resulted in multiple equilibria. Although the model was unbiased mostly as stated previously, the model appeared to be unable to mimic the observed market share in similar proportion to the equilibrium distributions' confidence intervals. For a 95% confidence interval, 26 out of 50 MSAs had an observed ownership share within the confidence interval. At the 99% confidence interval, 37 out of 50 MSAs' observed ownership shares were within the confidence interval. Whether this is due to unobservable factors or if these areas have not achieved equilibrium bicycle ownership is unknown, but could be tested with dynamic bicycle ownership data.

Table 33. MSA-level Predicted Equilibrium Ownership (at Posterior Mean)

MSA	Actual Rate	Posterior Mean Equilibrium	Difference	Percentile Equilibrium				Obs.
				0.5%	2.5%	97.5%	99.5%	
New Orleans	57.2%	44.7%	-12.5%	36.6%	38.4%	52.2%	54.2%	106
Louisville	47.0%	37.1%	-9.9%	29.9%	31.6%	46.4%	49.6%	91
West Palm Beach	41.2%	32.9%	-8.3%	27.8%	29.5%	36.9%	37.8%	97
Tampa	37.8%	30.1%	-7.7%	26.7%	27.8%	33.4%	34.3%	241
Orlando	47.5%	39.9%	-7.6%	31.8%	34.9%	46.3%	48.2%	131
Houston	48.5%	41.2%	-7.3%	36.5%	37.7%	46.9%	48.7%	546
Buffalo	55.6%	49.5%	-6.2%	45.5%	46.7%	54.7%	56.4%	582
San Francisco	47.3%	41.5%	-5.8%	36.5%	37.8%	47.1%	49.1%	556
San Antonio	41.7%	36.1%	-5.6%	32.5%	33.4%	40.1%	41.3%	254
Milwaukee	55.9%	51.5%	-4.5%	45.6%	47.3%	56.5%	57.8%	1087
Providence	55.9%	51.7%	-4.3%	41.0%	43.8%	61.5%	64.9%	96
Honolulu	37.8%	34.2%	-3.7%	30.2%	31.3%	39.1%	41.0%	1593
Norfolk	51.5%	48.3%	-3.2%	40.5%	43.0%	56.4%	58.9%	146
Portland	60.2%	57.5%	-2.7%	48.5%	50.9%	65.7%	68.4%	226
Chicago	58.1%	55.4%	-2.7%	49.4%	51.3%	60.9%	62.4%	713
Miami	34.6%	32.1%	-2.5%	28.4%	29.6%	35.7%	36.7%	228
Austin	49.8%	47.7%	-2.1%	39.9%	41.8%	56.7%	59.7%	274
Cincinnati	43.4%	41.5%	-1.8%	35.8%	37.4%	48.2%	50.5%	183
New York	41.6%	39.9%	-1.7%	37.2%	38.1%	43.1%	44.3%	5425
Grand Rapids	66.1%	64.7%	-1.4%	59.8%	61.0%	69.2%	70.5%	109
Kansas City	39.2%	37.9%	-1.3%	32.0%	33.4%	46.3%	49.3%	181
Cleveland	48.5%	47.4%	-1.1%	42.1%	43.6%	52.7%	54.0%	279
Los Angeles	45.4%	44.4%	-1.0%	38.9%	40.5%	50.9%	53.3%	1023
Las Vegas	36.0%	35.2%	-0.8%	27.3%	29.4%	45.6%	50.4%	131
Boston	48.1%	47.4%	-0.6%	39.2%	41.0%	59.0%	63.2%	528
Detroit	55.1%	54.8%	-0.3%	48.9%	50.7%	59.3%	60.5%	451
Jacksonville	47.3%	47.1%	-0.1%	42.9%	44.0%	54.4%	57.7%	102
St. Louis	45.7%	45.7%	0.1%	38.7%	40.7%	53.2%	56.3%	241
Dallas	37.2%	37.5%	0.3%	33.2%	34.3%	42.8%	45.0%	569
Oklahoma City	37.3%	39.0%	1.7%	34.0%	35.3%	44.7%	46.8%	66
Philadelphia	47.0%	49.2%	2.2%	46.2%	47.0%	54.8%	56.5%	483
Memphis	28.7%	31.6%	2.9%	26.4%	27.4%	40.9%	47.7%	90
Washington DC	43.3%	46.4%	3.2%	43.1%	44.0%	52.1%	54.8%	3948
Minneapolis	58.8%	62.7%	4.0%	49.9%	53.3%	72.4%	75.0%	381
Hartford	41.6%	46.5%	4.8%	35.9%	38.1%	58.8%	62.7%	106
Nashville	33.8%	39.0%	5.3%	30.2%	32.1%	57.9%	66.0%	111
Sacramento	52.3%	57.7%	5.3%	51.1%	52.6%	66.4%	69.1%	188
Rochester	53.4%	58.9%	5.5%	53.2%	54.3%	66.3%	68.0%	880
Denver	59.8%	67.1%	7.3%	51.9%	57.0%	75.3%	77.4%	267
Seattle	49.5%	57.3%	7.9%	48.8%	50.5%	67.3%	69.1%	378
Indianapolis	42.4%	50.3%	7.9%	43.2%	45.1%	57.7%	59.6%	162
Pittsburgh	34.6%	43.3%	8.7%	38.2%	39.3%	50.5%	54.0%	252
Phoenix	42.9%	51.9%	9.0%	41.6%	44.0%	61.0%	64.5%	295
San Diego	45.8%	55.9%	10.1%	49.5%	51.0%	64.8%	67.4%	209
Columbus	48.3%	58.7%	10.4%	51.6%	53.6%	65.8%	67.7%	135
Salt Lake City	54.5%	68.2%	13.7%	55.1%	58.3%	76.2%	78.4%	116
Greensboro	31.0%	46.2%	15.1%	41.0%	42.2%	54.1%	57.5%	135
Charlotte	48.2%	64.4%	16.2%	57.9%	59.3%	71.3%	73.1%	120
Atlanta	34.9%	53.7%	18.8%	48.6%	49.7%	62.7%	65.9%	317

Raleigh	38.5%	59.5%	20.9%	52.5%	54.1%	69.0%	71.2%	135

The remainder of this section will show plots of the distribution of equilibria as well as the equilibrium at the posterior mean from a sample of MSAs across different regions of the United States.

Washington-Baltimore

The Washington-Baltimore MSA has a predicted equilibrium ownership share close to its actual ownership share. The actual ownership share is not located within the 95% confidence interval of the equilibrium predictions, but it is within the 99% confidence interval. The spread of equilibria is also right skewed and smaller than most of the MSAs sampled in this section.

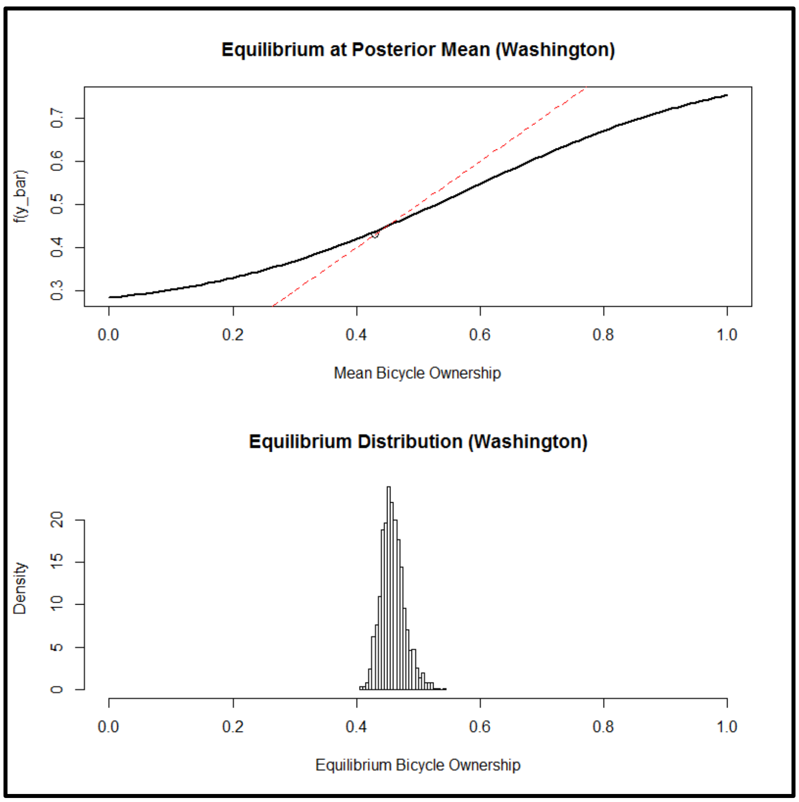

Figure 34. Bicycle Ownership Equilibrium Distribution for Washington MSA

New York-Northern New Jersey-Long Island

Similar to the Washington-Baltimore MSA, the New York MSA has a predicted equilibrium ownership share close to its actual ownership share. The actual ownership share is located within the 95% confidence interval of the equilibrium predictions and the spread of equilibria is relatively tight compared to other featured MSAs.

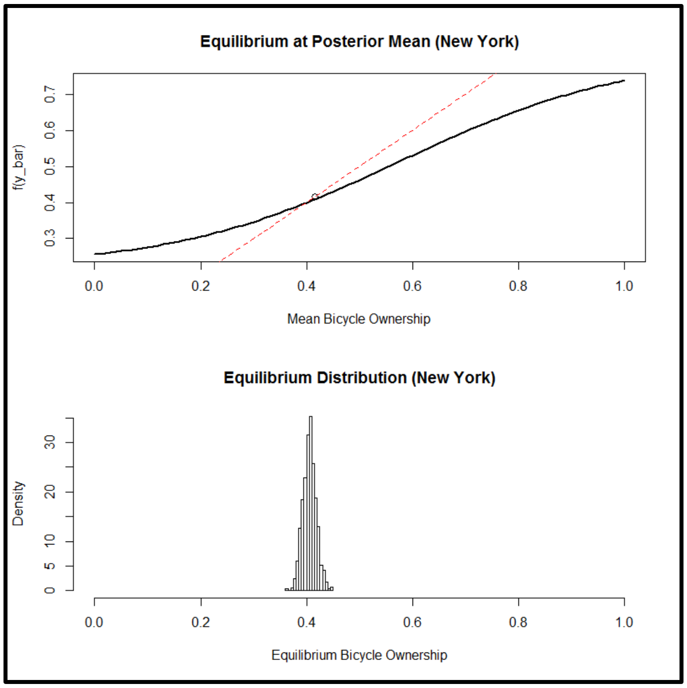

Figure 35. Bicycle Ownership Equilibrium Distribution for New York MSA

Grand Rapids-Muskegon-Holland

The Grand Rapids MSA had the highest actual mean ownership share with 66.1%. At the posterior mean, the model slightly underpredicted its equilibrium ownership. Compared to Washington and New York MSAs, Grand Rapids has a similar spread of predicted equilibria.

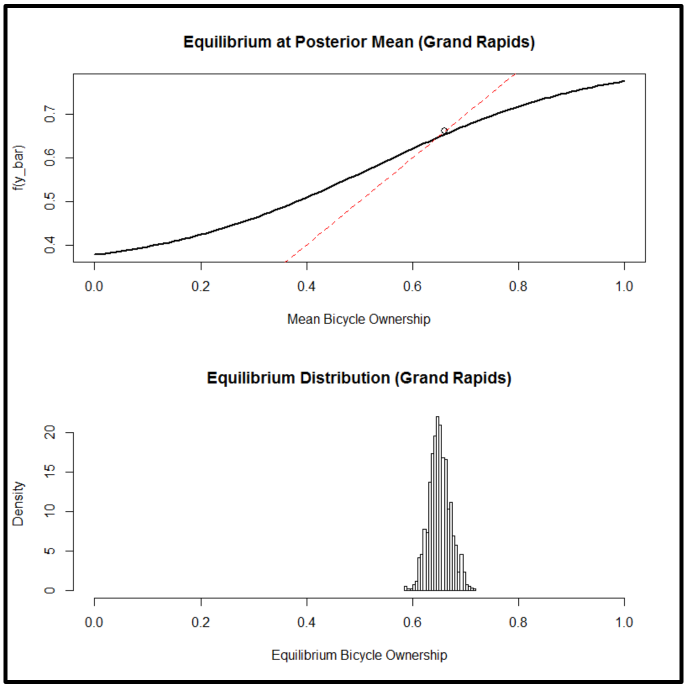

Figure 36. Bicycle Ownership Equilibrium Distribution for Grand Rapids MSA

Minneapolis-St. Paul

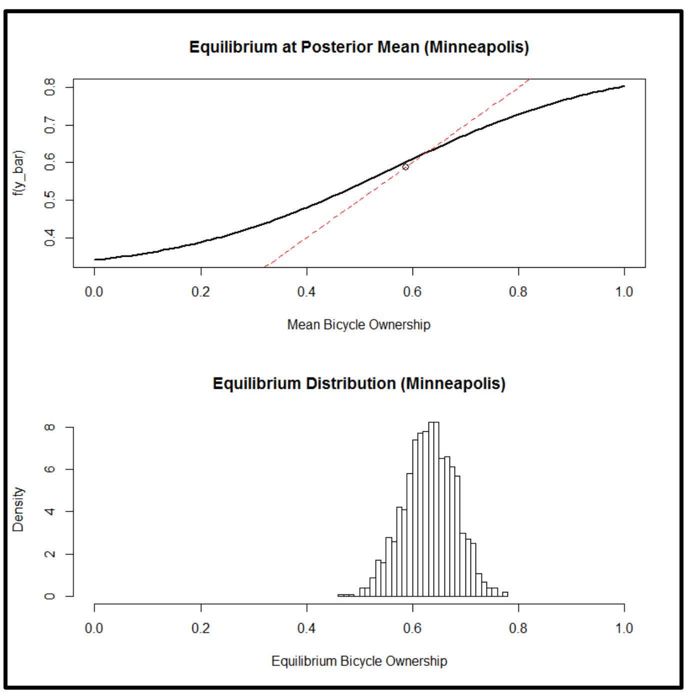

Figure 37. Bicycle Ownership Equilibrium Distribution for Minneapolis MSA

The Minneapolis MSA's ownership was slightly overpredicted at the posterior mean with a difference of 4.0%. There is a wide spread of equilibria predicted representing large uncertainty in the predicted equilibria.

Portland-Salem

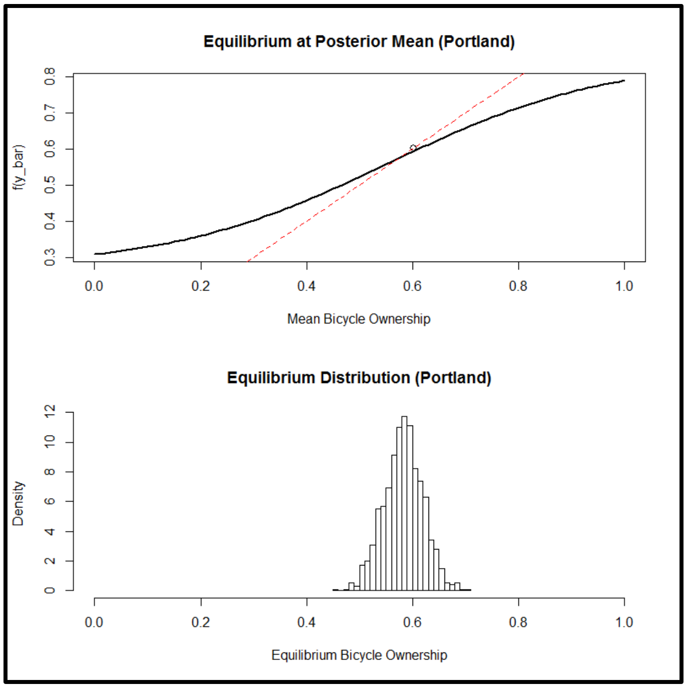

Figure 38. Bicycle Ownership Equilibrium Distribution for Portland MSA

Portland is an often-cited example of an American city that emphasizes cycling. The model underestimated the population share in Portland at the posterior mean parameter values.

Denver-Boulder-Greeley

Located in the Mountain census division, Denver's predicted equilibria at the posterior mean overestimated the ownership share. The actual ownership was still located within the wide spread of the equilibrium distribution within the 95% confidence interval.

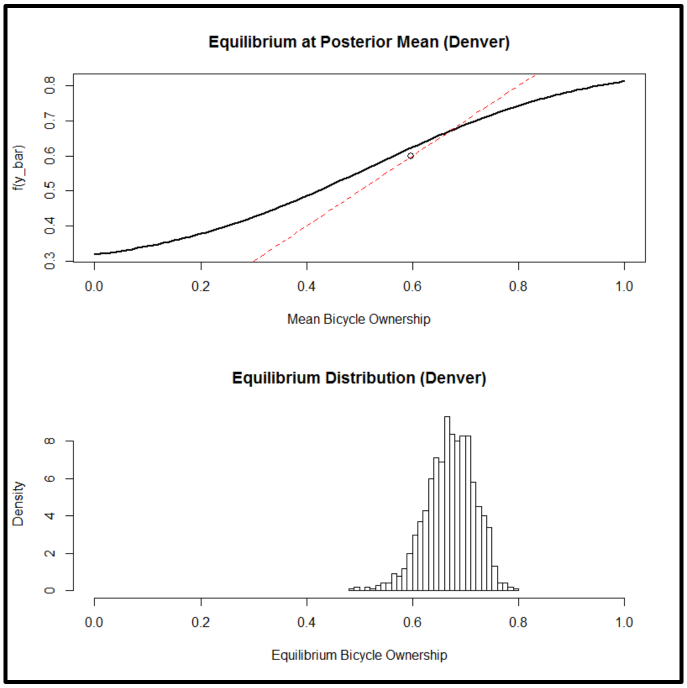

Figure 39. Bicycle Ownership Equilibrium Distribution for Denver MSA

Houston-Galveston-Brazoria

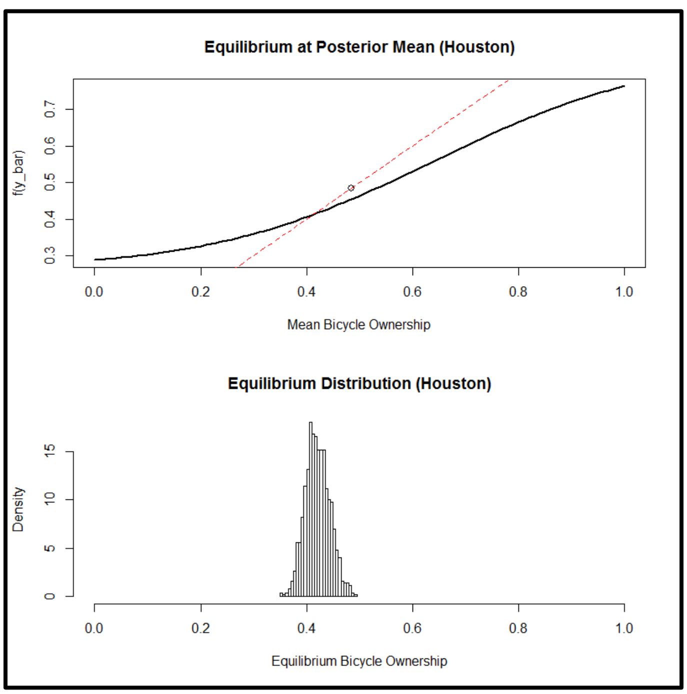

Figure 40. Bicycle Ownership Equilibrium Distribution for Houston MSA

Located in the South West Central census division, Houston's predicted equilibrium was overestimated by 7.3%. The actual ownership share of 48.5% is within the spread of the distribution, but neither within the 95% nor the 99% confidence interval.

Atlanta

The Atlanta MSA is the second most overpredicted MSA with an overestimation difference 18.8%. The equilibrium distribution is right skewed and the actual bicycle ownership share is not within the 95% confidence interval of the distribution. This difference may be due to the disconnection between the model and reality (i.e. models do not predict perfectly and are simplification of reality). Additionally, Atlanta may not have reached equilibrium bicycle ownership or additional unobserved factors are impacting bicycle ownership. Dynamic ownership data and models could be used to determine if these hypotheses are valid.

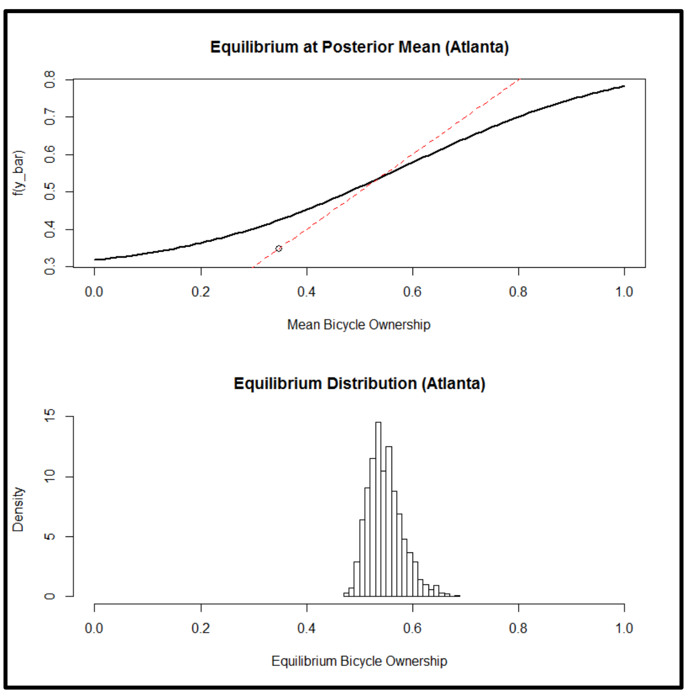

Figure 41. Bicycle Ownership Equilibrium Distribution for Atlanta MSA

New Orleans

In contrast to Atlanta, the New Orleans MSA has much higher bicycle ownership rates than predicted by an equilibrium analysis at the posterior mean. In contrast, the distribution is skewed towards the observed ownership share. But, the observed ownership is not within the 95% confidence interval of the equilibrium distribution.

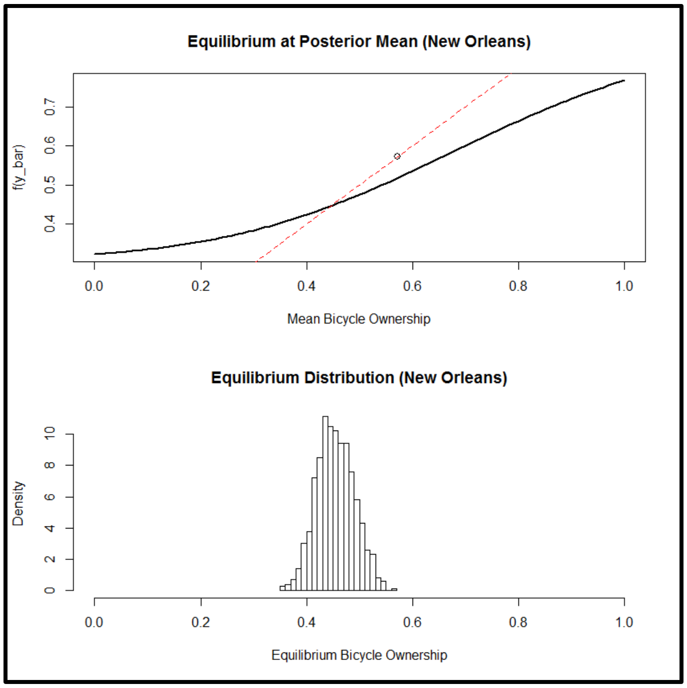

Figure 42. Bicycle Ownership Equilibrium Distribution for New Orleans MSA

9.6.4 Elasticity of Social Influence

Hess et al. (2011) derives the elasticity for a latent class logit model when choice model covariates change. By contrast, in order to understand the elasticity of the social influence covariate in the informational conformity model, the elasticity for changes in class model covariates must be derived. For the informational conformity model, the partial derivative of mean behavior among peers with respect to the probability of choosing $y_n = 1$ is derived as follows:

$$\frac{\partial P_n}{\partial \bar{y}_n} = \frac{\partial}{\partial \bar{y}_n}\left\{\pi_n^{[a]} P_n^{[a]} + \pi_n^{[b]} P_n^{[b]}\right\} = \frac{\partial \pi_n^{[a]}}{\partial \bar{y}_n} P_n^{[a]} + \frac{\partial \pi_n^{[b]}}{\partial \bar{y}_n} P_n^{[b]}$$

$$= P_n^{[a]} \cdot \left\{\frac{\exp(\mathcal{F}_n)}{1+\exp(\mathcal{F}_n)} \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n} - \left(\frac{\exp(\mathcal{F}_n)}{1+\exp(\mathcal{F}_n)}\right)^2 \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\} \quad (75)$$

$$+ P_n^{[b]} \cdot \left\{-\frac{\exp(\mathcal{F}_n)}{(1+\exp(\mathcal{F}_n))^2} \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\}$$

This results in the following simplified partial derivative:

$$\frac{\partial P_n}{\partial \bar{y}_n} = P_n^{[a]} \cdot \left\{\pi_n^{[a]} \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n} - \pi_n^{[a]^2} \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\} - P_n^{[b]} \cdot \left\{\frac{\pi_n^{[a]}}{1+\exp(\mathcal{F}_n)} \cdot \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\} \quad (76)$$

The result in equation (76) is used to derive the elasticity with respect to mean behavior among peers as follows:

$$\mathcal{E}_{\bar{y}_n} = \frac{\partial P_n}{\partial \bar{y}_n} \cdot \frac{\bar{y}_n}{P_n}$$

$$= \frac{\bar{y}_n \left(P_n^{[a]}\left\{\pi_n^{[a]} \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n} - \left(\pi_n^{[a]}\right)^2 \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\} - P_n^{[b]}\left\{\frac{\pi_n^{[a]}}{1+\exp(\mathcal{F}_n)} \frac{\partial \mathcal{F}_n}{\partial \bar{y}_n}\right\}\right)}{\pi_n^{[a]} P_n^{[a]} + \pi_n^{[b]} P_n^{[b]}} \quad (77)$$

$$= \frac{\bar{y}_n \left(P_n^{[a]}\left\{\pi_n^{[a]} \delta - \left(\pi_n^{[a]}\right)^2 \delta\right\} - P_n^{[b]}\left\{\frac{\pi_n^{[a]}}{1+\exp(\mathcal{F}_n)} \delta\right\}\right)}{\pi_n^{[a]} P_n^{[a]} + \pi_n^{[b]} P_n^{[b]}}$$

With the result from equation (77), the elasticity for each individual can be obtained for any set of draws of the parameters from the informational conformity model in Table 30 and Table 31. Figure 43 shows the elasticity distributions for the national sample at the individual- and population-levels. The top left plot shows the elasticity across the national sample when the posterior mean parameter estimates are chosen. At the posterior mean, the average elasticity is 0.699 with a standard deviation of 0.325. The median elasticity is 0.712. This distribution is skewed to the right with a skewness of 0.08.

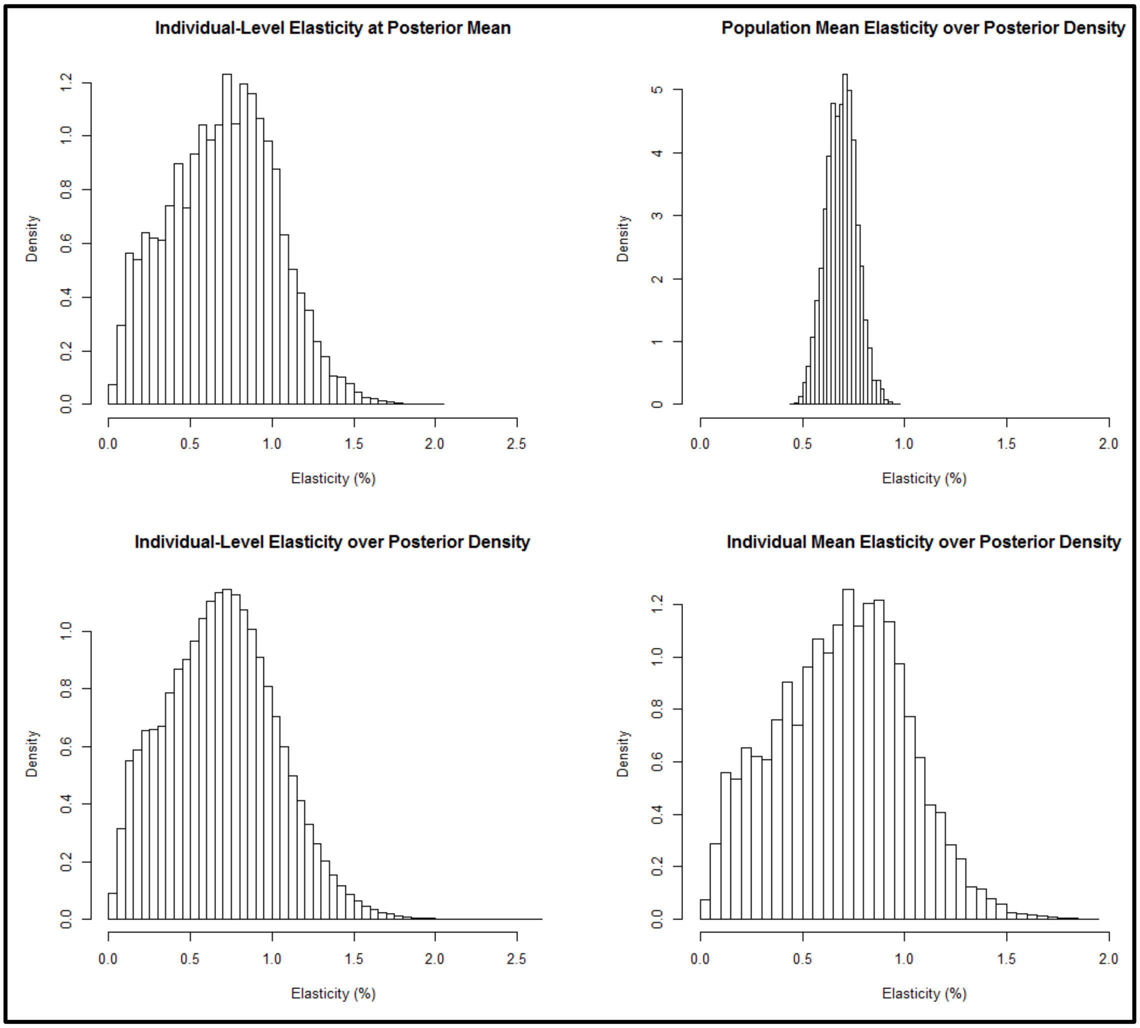

Figure 43. Elasticity Distributions: Bike Ownership Informational Conformity Model

The full distribution of individual-level elasticity is shown in the bottom left plot of Figure 43. It is skewed positively with a skewness of 0.27. The mean individual-level elasticity is 0.689 with standard deviation 0.335 and the median individual-level elasticity is 0.685.

The top right plot of Figure 43 is a histogram of the mean elasticity for the population at each draw among a 5000 parameter draw sample of the posterior density. The population mean elasticity is centered at 0.689% with a standard deviation of 0.075 and median of 0.691. This distribution is skewed positive with a skewness of 0.05. The bottom right plot shows the mean elasticity by individual over the posterior density of parameter estimates. The individual mean elasticity averages 0.689% with a standard deviation of 0.315 and median of 0.705. This distribution is skewed positive with a skewness of 0.05.

On average nationally, a 1.00% increase in MSA-level bicycle ownership will induce an increase in household-level bicycle ownership by 0.69%. The elasticity of social influence was found to range locally from about 0.50% to 0.80%. This result helps to check the reasonableness of the model. On average, bicycle ownership does not grow out of control hyperbolically. This effect combined with the equilibrium conditions shown in section 9.6.2 help to show that social influence occurs but does not overpower the decision process.

The distributions of elasticities at the national-level and local-level exhibit a central tendency. These elasticity distributions are summarized in Table 34.

Table 34. Mean Social Influence Elasticity by MSA

MSA	Mean	Standard Deviation	5% Percentile	25% Percentile	50% Percentile	75% Percentile	95% Percentile
National	0.69	0.34	0.15	0.44	0.69	0.92	1.26
Atlanta	0.48	0.27	0.09	0.28	0.47	0.66	0.96
Austin	0.76	0.37	0.16	0.47	0.77	1.01	1.38
Boston	0.73	0.35	0.18	0.44	0.72	0.98	1.32
Buffalo	0.75	0.34	0.22	0.50	0.74	0.98	1.34
Charlotte	0.65	0.36	0.14	0.36	0.62	0.91	1.27
Chicago	0.75	0.37	0.18	0.48	0.73	0.99	1.43
Cincinnati	0.72	0.32	0.20	0.51	0.72	0.92	1.27
Cleveland	0.73	0.34	0.15	0.47	0.74	0.96	1.30

Columbus	0.71	0.36	0.14	0.42	0.72	0.98	1.31
Dallas	0.56	0.31	0.10	0.32	0.56	0.77	1.10
Denver	0.69	0.34	0.20	0.42	0.67	0.91	1.27
Detroit	0.70	0.35	0.18	0.42	0.68	0.94	1.31
Grand Rapids	0.61	0.36	0.17	0.30	0.54	0.84	1.28
Greensboro	0.49	0.25	0.10	0.30	0.48	0.65	0.92
Hartford	0.68	0.30	0.17	0.47	0.69	0.89	1.18
Honolulu	0.63	0.30	0.13	0.41	0.63	0.84	1.12
Houston	0.73	0.36	0.15	0.48	0.73	0.98	1.34
Indianapolis	0.66	0.30	0.18	0.44	0.65	0.88	1.15
Jacksonville	0.70	0.32	0.20	0.45	0.69	0.92	1.24
Kansas City	0.63	0.31	0.15	0.38	0.62	0.84	1.15
Las Vegas	0.66	0.30	0.15	0.45	0.66	0.86	1.18
Los Angeles	0.72	0.34	0.15	0.47	0.72	0.94	1.30
Louisville	0.74	0.37	0.14	0.51	0.70	0.96	1.44
Memphis	0.40	0.22	0.06	0.22	0.39	0.55	0.78
Miami	0.61	0.30	0.12	0.40	0.61	0.82	1.11
Milwaukee	0.73	0.34	0.21	0.47	0.71	0.95	1.32
Minneapolis	0.72	0.36	0.19	0.44	0.70	0.96	1.37
Nashville	0.50	0.26	0.11	0.31	0.47	0.65	0.99
New Orleans	0.76	0.38	0.22	0.47	0.72	1.01	1.45
New York	0.66	0.32	0.14	0.42	0.67	0.89	1.20
Norfolk	0.74	0.34	0.23	0.50	0.71	0.96	1.35
Oklahoma City	0.58	0.33	0.11	0.31	0.55	0.81	1.18
Orlando	0.79	0.34	0.21	0.57	0.80	1.02	1.34
Philadelphia	0.69	0.33	0.15	0.45	0.70	0.92	1.25
Phoenix	0.66	0.34	0.12	0.41	0.66	0.89	1.25
Pittsburgh	0.60	0.27	0.14	0.42	0.62	0.79	1.04
Portland	0.72	0.38	0.18	0.43	0.68	0.98	1.41
Providence	0.76	0.35	0.23	0.47	0.76	1.02	1.35
Raleigh	0.55	0.30	0.11	0.32	0.54	0.76	1.08
Rochester	0.70	0.34	0.17	0.46	0.68	0.92	1.31
Sacramento	0.72	0.38	0.17	0.43	0.68	0.98	1.39
St. Louis	0.70	0.35	0.15	0.43	0.69	0.93	1.31
Salt Lake City	0.68	0.32	0.14	0.45	0.68	0.89	1.22
San Antonio	0.65	0.33	0.12	0.43	0.66	0.86	1.22
San Diego	0.67	0.37	0.16	0.37	0.63	0.92	1.34
San Francisco	0.73	0.34	0.18	0.48	0.74	0.97	1.30
Seattle	0.72	0.34	0.17	0.47	0.72	0.93	1.32
Tampa	0.71	0.30	0.16	0.52	0.72	0.91	1.22
Washington DC	0.74	0.33	0.19	0.51	0.74	0.96	1.28
West Palm Beach	0.73	0.31	0.15	0.53	0.76	0.94	1.20

9.7 Summary

Observing the bicycle ownership and cycling behavior of others may provide information on the benefits of cycling. This can start a process in which the individual may begin to research the suitability of cycling and adjust their opinions and behaviors.

To test for this effect statistically, this chapter explored bicycle ownership in the United States by using explanatory models of social influence. The more traditional direct-effect conformity model is contrasted with an indirect-effect informational conformity model.

The informational conformity model is a latent class discrete choice model is formulated that places individuals into classes based on information exposure where information is signaled by greater city-level bicycle usage. In contrast to existing work, the informational conformity model shows that “more informed” households have a higher probability of owning a bike due to changes in preferences rather than direct benefits from others’ behaviors – they are less sensitive to smaller home footprints and limited incomes, and being single-person households. But, “more informed” households are sensitive to household membership size and composition.

Additionally, a Bayesian inference procedure is proposed as a tool for hypothesis testing and forecasting the distribution of market equilibria. The behavioral hypothesis of higher preferences for “more informed” household was confirmed. Local-level “more informed” class membership varied across the country but the uncertainty in this membership tended to be similarly shaped. For most MSA areas surveyed, the observed market share falls within middle 90% of the predicted equilibrium distribution. Also, social influence elasticity was found to vary locally from about 0.5% to 0.8%.

Areas for future research include a need to understand why the “more informed” households were less sensitive to home type. Qualitative study into whether home moving patterns (e.g. moving frequency) and home footprints impact bicycle ownership could be useful. Additionally, panel and time-series data could be used to aid in identify social influence. This study is limited by the use of cross-sectional data. Time-series data could also be used to test for equilibrium and could serve as a method for model selection and confirmation. Thus it may allow for identification of social influence motivations.

Chapter 10: Conclusion

Understanding the determinants of activities and travel is critical for transportation policy makers, planners, and engineers to design and manage transportation systems. These systems, and their externalities, are interwoven with social systems in communities, cities, regions, and societies. But discrete choice models – the predominant modeling tool for researching travel behavior and planning transportation systems – are grounded in theories of individual decision-making. Travel behavior analysis' shift to the social is currently underway. The incorporation of social context into models of travel behavior has the potential to enhance the behavioral realism of these models and lead to better understanding of activity and travel behavior. This social context is typically incorporated in three areas: (1) social cooperation, (2) social influence, and (3) social capital. And these areas are linked through the social networks of individuals. Thus far, research efforts have resulted in work showing that all three factors plus social networks may be relevant in travel decision making.

The incorporation of social interactions into discrete choice models is non-trivial since discrete choice models are grounded in theories of and methods for independent decision makers. Modeling efforts have been taken to develop techniques and to test hypotheses. This dissertation continues those efforts by:

- Incorporating network capital indicators from a position generator in model of activity selection
- Developing a behavioral framework of social influence choice modeling to classify existing research and spur new modeling directions that emphasize social influence motivations

- Studying the impact of informational conformity on choice by formulating a latent class choice model
- Applying social interaction modeling techniques in the applied study of activity selection and bicycle ownership
- Analyzing the estimation properties of discrete choice models of social capital and social influence in the presence of misspecified social network data

10.1 Directions for Future Research

Although the work presented in this dissertation has expanded upon the methodology and empirical analysis of social interactions in activity and travel behavior, the work can be extended upon with subsequent research. In the area of social capital in activity and travel the following issues need to be resolved:

- *Access to Resources and Activity Participation:* This linkage between the diversity of networked resource (as indicated by an indicator of network occupational diversity) can be explored more. Expanding work by surveys and models to including a greater variety and exhaustive list of activities would increase understanding of this link and panel data would allow researchers to determine the directionality between resource access and activity participation frequency.
- *Activity Diary Design:* Position generators measure additional components of social capital that are not captured in current survey and diary designs. Incorporating position generators in surveys and diary may be promising in understanding the linkage between social capital, activities, and travel.

- *Combining Position and Name Generators Data*: Results showed that combining position and name generator data was useful for some activity types in improving prediction. More empirical evidence is needed to understand when combining these datatypes is useful and to apply composite measures of social capital from both generators.
- *Additional Indicators:* The previous literature concentrated on name generator measures centered on core networks. Chapter 3 showed the relevance of using position generators in activity selection models but the only measure used was extensity of resource access via social network occupational diversity. Incorporating measures of prestige and social class need to be pursued.

In the area of choice modeling and the effect of misspecification errors due to social network data collection, future research could explore:

- *Robustness of Position Generator Measures:* Other travel and activity contexts could be analyzed to contribute additional empirical evidence to support the robustness results in Chapter 4.
- *Generalizing the Effects of Network Misspecification:* Analytical results on network misspecification are lacking and may be difficult to obtain. Until more extensive and generalizable knowledge is available, it is suggested that data collectors and modelers use methods similar to chapters 4, 5, and 6. Agent-based modeling and simulation can be used before design data collection efforts to guide sampling and survey design and during the model building process to guide model selection and understand model sensitivity.

In the area of social influence in travel, the following directions could be pursued:

- *Applications:* The review of social influence research in the travel behavior field showed that social influence has been applied to many areas of travel behavior. New applied research incorporating new social network data sources, time-series data, and varying model specifications are needed to contribute additional evidence to support efforts to incorporate social interactions in transportation planning and policy analysis.
- *Heterogeneity of Social Influence Motivations:* Using latent class discrete choice models presents an opportunity to test hypotheses about multiple motivations among the population. Applications of models similar to the one in Appendix F on travel datasets could be pursued.
- *Comparing Different Models of Social Influence Motivations:* Models with differing social influence motivations may exhibit varying dynamic properties. Combined with panel data, this offers an additional approach to aid in identifying the existence of social influence as well as its type.
- *Exploring Informational Conformity:* A generalized informational conformity was described in Appendix D. Applying this model could test hypotheses of expectation and constraint changes in addition to preference changes due to social influence. Additionally, the model can be expanded to allow for multiple levels of information acquisition.

10.2 Answers and Contributions

This dissertation began with a set of questions about social interactions in travel and choice modeling. To conclude, this dissertation has contributed to answering these questions in the following ways:

1. *Who cares about social interactions in travel?* The case studies in Chapter 3 and 9 show that travel behavior researchers need to consider social interactions in the study of activity participation and cycling behavior. The social capital work in Chapter 3 shows that diversity of access to resources (extensity) is correlated with activity participation. This will be of value to government organizations that try to evaluate the social value of travel and in attempts to increase social capital access among disadvantaged communities.
2. *How are social interactions incorporated into choice models?* Two new model formulations were proposed for social influence modeling involving informational conformity (Chapter 8 and Appendix D) and multiple social influence motivations (Appendix E).
3. *What are the indicators of social capital in activity-travel and how is it measured?* Extensity, or social network occupational diversity, was found to be correlated with leisure activity participation. Previous researched used name generators and interpreters to study social capital in activity-travel, but results from Chapter 3 show that position generator data was as useful as name generator and often resulted in better explanatory and predictive performance.
4. *How does the measurement of social network indicators impact estimation of social capital and social influence choice models?* The robustness of position

generator data in a social capital choice model was demonstrated in Chapter 4 with network diversity indicators found to still improve model fit even for shortened and misspecified occupational lists. Name generator data was used on social influence choice models where it was found that estimates of social influence parameters and model selection were generally unaffected by network shape for small-world network. Estimates were most affected by the relative strength of social influence, sample size, and the degree of network distortion. Networks with about 15% to 30% change in network ties lost predictive and model selection accuracy.

5. *How is social influence incorporated in choice modeling?* An extensive review of discrete choice models of travel behavior that modeled social influence was undertaken. Using a microfoundations / behavioral framework, the literature was categorized, and it was found that primarily cross-sectional models rooted in conformity with modeler-determined large cliques are used. It is recommended that travel behavior modelers study and incorporate specific motivations for social influence into their models.

6. *How can the motivation for accuracy be incorporated into social influence choice models?* Chapter 8 formulates a model of informational conformity through latent class discrete choice framework. Equilibrium analysis shows that the long-run implications of the model vary from the direct-benefit model that dominates the literature. The model is then demonstrated in the context of bicycle ownership in Chapter 9 where it explains difference in bicycle ownership as due to difference

in the effect of home type and ownership as compared to a direct-benefit conformity model.

Appendix A: ERGM Primer

A.1 Model Specification

Exponential-family random graph models (ERGMs) are a family of statistical models for representing networks/graphs by the likelihood of observing counts of certain network configuration terms such as edges, triangles, and k-stars. ERGMs assume that networks are formed by bottom-up processes that work between nodes. For example, transitivity can be modeled by counts of triangles. A graph instance, upon which a model is estimated on, is considered to consist of a dependent series of local processes which are correlated in the local area around any given node but uncorrelated outside of the local. This can lead to macro-level graph behavior emerging, such as small-world networks.

For a fixed set of individuals N, the probability of observing a graph instance given a set of configuration parameters in an ERGM (Lusher et al. 2013) is:

$$P(G = g|\theta) = \frac{1}{\kappa(\theta)} \exp\left\{\sum_{c \in C} \theta_c z_c(g)\right\} \tag{78}$$

where:

$z_c(g)$ $\equiv$ the count of configurations *c* in graph g

C $\equiv$ the set of graph configuration types in the model's chosen dependence assumption

θ_c $\equiv$ model parameter corresponding to the count of configurations of type *c*

$\kappa(\theta)$ $\equiv$ the model's normalization term

The set of graph configuration types are typically determined by the model's dependence assumption. This determines how particular ties in the network are correlated with one another. For example, the simplest graph configuration type is the Bernoulli dependence assumption where all ties are assumed to be independent of one another. In other words, a tie has a given probability of being observed and this does not change if any other tie is added or removed from the graph. The set of graph configurations for a Bernoulli dependence assumption is $\{z_L\}$ which is just a count of the number of edges (or the graph density).

One strength of the ERGM approach is the ease at which graphs can be simulated. Although the normalization term $\kappa(\theta)$ is intractable for any graph with greater than only a few nodes[51], the models can be simulated via Markov chain Monte Carlo (MCMC) techniques which do not depend on the normalization term. The Metropolis-Hastings algorithm is a popular technique in the ERGM community and is used in the *statnet* R (Handcock et al. 2007) package which is used to simulate the graphs in this dissertation.

A.2 Social Circuit Dependence Functions

As shown in Hunter (2007), the geometrically weighted degree (*gwdegree*) statistic is defined as:

$$z_u(\boldsymbol{y};\theta_s) = e^{\theta_s} \sum_{i=1}^{N-1} \left\{1 - \left(1 - e^{-\theta_s}\right)^i\right\} D_i(\boldsymbol{y}) \qquad (79)$$

When $\theta_s = \log \lambda_s$, the *gwdegree* statistic takes on a similar form to the alternating k-star statistic from Snijders et al. (2006):

$$z_u(\boldsymbol{y};\log \lambda_s) = 2S_1(\boldsymbol{y}) - \frac{S_2(\boldsymbol{y})}{\lambda_s} + \cdots + (-1)^N \frac{S_{N-1}(\boldsymbol{y})}{\lambda_s^{N-2}} \qquad (80)$$

[51] The number of different graph configurations for an undirected graph with *n* nodes is $2^{n(n-1)/2}$.

where:

$\boldsymbol{y}$	$\equiv$	a particular network instance
$S_k(\boldsymbol{y})$	$\equiv$	the number of k-stars in network $\boldsymbol{y}$
$D_i(\boldsymbol{y})$	$\equiv$	the number of nodes in network $\boldsymbol{y}$ with degree i
λ_s	$\equiv$	a scale parameter
θ_s	$\equiv$	a scale parameter

As shown in Hunter (2007), the geometrically weighted edgewise shared partner (*gwesp*) statistic is defined as:

$$z_v(\boldsymbol{y};\theta_t) = e^{\theta_s}\sum_{i=1}^{N-2}\left\{1-\left(1-e^{-\theta_t}\right)^i\right\}EP_i(\boldsymbol{y}) \tag{81}$$

Correspondingly, the Snijders et al (2006), specification for the alternating k-triangle statistic is as follows:

$$\begin{gathered} T_k(\boldsymbol{y}) = \sum_{i=k}^{N-2}\binom{i}{k}EP_i(\boldsymbol{y}), \qquad 2 \le k \le N-2 \\ T_1(\boldsymbol{y}) = \frac{1}{3}\sum_{i=0}^{N-2} i\, EP_i(\boldsymbol{y}) \end{gathered} \tag{82}$$

where:

$EP_k(\boldsymbol{y})$	$\equiv$	the number of k-triangles in network $\boldsymbol{y}$ or "the number of unordered pairs [$\{n,m\}$] such that [$y_{nm} = 1$] and [n and m] have exactly k common neighbors in" network $\boldsymbol{y}$
$T_1(\boldsymbol{y})$	$\equiv$	the number of triangles in network **y**

As shown in Hunter (2007), the geometrically weighted dyadic shared partner (*gwdsp*) statistic is defined as:

$$z_w(\boldsymbol{y};\theta_p) = e^{\theta_p} \sum_{i=1}^{N-2} \left\{1 - \left(1 - e^{-\theta_p}\right)^i\right\} DP_i(\boldsymbol{y}) \tag{83}$$

Correspondingly, the Snijders et al (2006), specification for the alternating k-path statistic is as follows:

$$\begin{aligned} P_k(\boldsymbol{y}) &= \sum_{i=k}^{N-2} \binom{i}{k} DP_i(\boldsymbol{y}), \qquad 1 \le k \le N-2, k \ne 2 \\ P_2(\boldsymbol{y}) &= \frac{1}{2}\sum_{i=2}^{N-2} \binom{i}{2} DP_i(\boldsymbol{y}), \qquad 1 \le k \le N-2, k \ne 2 \end{aligned} \tag{84}$$

where:

$DP_k(\boldsymbol{y})$ ≡ the number of *k*-twopaths in network ***y*** or “the number of pairs [$\{n, m\}$] such that [$y_{nm} = 0$] and [*n* and *m*] share exactly *k* common neighbors in” network ***y***

$P_2(\boldsymbol{y})$ ≡ the number of two-paths in network **y**

Appendix B: Selected Models of Social Influence and Choice

This appendix describes some existing models of social influence and choice from the areas of social influence network theory, social network analysis, social and spatial econometrics, experimental economics, behavioral economics, and travel behavior.

B.1 Social Influence Network Theory

Social influence network theory is a dynamical theory of opinion and attitude change that combines social network formulations with a mathematics formalization of dynamic behavior. In particular, the equilibrium properties of social influence network theory models are emphasized in the literature. In this section, the mathematical origins of social influence network theory as described by Friedkin and Johnsen (2011) are summarized.

B.1.1 French's Formal Theory of Social Power

$$y_n^{(t)} = \left[\sum_{q \in N} w_{nq} y_q^{(t-1)}\right] \Big/ \sum_{m \in N} w_{nm} \tag{85}$$

In this formulation, all weights are equal between individuals who are connected.

B.1.2 Harary and DeGroot's Generalizations

Harary (1959) and DeGroot (1974) develop a similar formulation but allow for weights to varying between individual pairings.

$$y_n^{(t)} = \left[\sum_{q \in N} w_{nq} y_q^{(t-1)}\right] \Big/ \sum_{m \in N} w_{nm} \tag{86}$$

such that: $0 \leq w_{nm} \leq 1, \quad \sum_{m \in N} w_{nm} = 1 \quad \forall n, m \in N$

B.1.3 Friedkin and Johnsen's Generalizations

Friedkin and Johnsen (1990) relax the assumption that attitudes only depend on the attitudes from the time period immediately prior. Specifically, they allow for an individual's initial attitudes to continuously impact their decisions.

$$y_n^{(t)} = \delta \left[\sum_{q \in N} w_{nq} y_q^{(t-1)} \right] \Big/ \sum_{m \in N} w_{nm} + (1-\delta) y_n^{(1)}$$

such that: $0 \leq w_{nm} \leq 1, \quad \sum_{m \in N} w_{nm} = 1 \quad \forall n, m \in N$ (87)

AND $\quad 0 \leq \delta \leq 1$

This model is currently the "standard model" (Friedkin and Johnsen 2011) in the field. Operationalizing the model for discrete choices yields a formulation similar to discrete choice models. Specifically, Friedkin and Johnsen (2011) mention two methods of choice: most preferred and criterion attainment. The most preferred form uses equation (87) for a threshold function ${y^*}_{ni}^{(t)}$ and has individual choose the option among a set category such that:

$$y_n^{(t)} = i \quad if \; y_{ni}^{(t)} = \max\left({y^*}_{n1}^{(t)}, {y^*}_{n2}^{(t)}, \ldots\right) \tag{88}$$

In this context, ${y^*}_{n1}^{(t)}, {y^*}_{n2}^{(t)}, \ldots$ are threshold functions. This formulation draws parallels to random utility models. The criterion attainment formulation has individuals choose an option among a set of categories such that:

$$y_{ni}^{(t)} = \begin{cases} 1 & if\, y^{*(t)}_{ni} \geq \psi \\ 0 & if\, y^{*(t)}_{ni} < \psi \end{cases} \tag{89}$$

This formulation draws parallels with discrete choice models that use satisficing decision rules.

B.2 Statistical Models of Social Networks

B.2.1 Stochastic Actor-Based Network Dynamics

Snijders et al. (2010) provides a description of models for the dynamics of networks and behavior. In these models, both tie selection and behavior selection are modeled simultaneously over time. These are distinguished from other models in the literature by the assumptions described by Snijders et al. (2010):

1. "The underlying time parameter is continuous" (p. 54). This stands in stark contrast to the majority of models showcased in this appendix. Therefore, stochastic actor-based models of network and behavior dynamics are not limited by the time-slices from a data collection design.
2. Both the networks and behavior change by a Markov process.
3. A change in one tie or a one-unit change in behavior occurs at any instance in time. As Snijders et al. explains, this is most limiting when the number of behaviors possible is large. Also, the behavior must be binary or have some ordinal meaning, else a one-unit change would be difficult to explain behaviorally through the model specification.
4. "The actors control their outgoing ties as well as their own behavior" (p. 54).

The model formulation involves processes governed by a rate function, objective function, evaluation function, and endowment function (Snijders et al 2007). The rate function determines how often individuals can change their ties and behavior. This Poisson process is governed two different rate function for tie changes $\lambda_n^{[G]}$ and behavior change $\lambda_n^{[Y]}$ and are given by the following equations over a time period $t_m < t < t_{m+1}$:

$$\begin{aligned} \lambda_n^{[G]} &= \varpi_m^{[G]} \exp\left\{\sum_{c \in C} \theta_c^{[G]} z_c\big(G, Y(t)\big)\right\} \\ \lambda_n^{[Y]} &= \varpi_m^{[Y]} \exp\left\{\sum_{c \in C} \theta_c^{[Y]} z_c\big(G, Y(t)\big)\right\} \end{aligned} \tag{90}$$

where:

$z_c(\cdot)$ $\equiv$ the count of configurations *c* in graph *g* and among ties and the behaviors of others Y

C $\equiv$ the set of configuration types in the model's chosen dependence assumption

$\theta_c \in \theta$ $\equiv$ the model parameter corresponding to the statistics for configurations of type *c*

$\varpi_m^{[G]}, \varpi_m^{[Y]}$ $\equiv$ model parameters corresponding to period-dependence

How decisions are made at times given by the rate function are determined through an objective function. These objective functions (one for tie changes and one for behavior change) depends linearly on an evaluation function, an endowment function, and unobservables.

The details of the evaluation and endowment functions will not be described here[52]. Due to the assumptions on the error form, each decision making step follows a multinomial logit form.

B.2.2 Autologistic Actor-Attribute Models

The autologistic actor-attribute model, or ALAAM, (Robins et al. 2001) is related to the exponential-family random graph model (ERGM). In contrast to an ERGM where the network is a random variable, the ALAAM has the individual's behavior as random variables conditional on an exogenously given network. The model takes the following form:

$$P(Y = y | G = g, \theta) = \frac{1}{\kappa(\theta)} \exp\left\{\sum_{c \in C} \theta_c z_c(g, x, y)\right\} \tag{91}$$

where:

$z_c(\cdot)$	$\equiv$	the count of configurations *c* in graph *g* and among nodal variates *x* and the behaviors of others *y*
C	$\equiv$	the set of graph, behavior, and nodal variate configuration types in the model's chosen dependence assumption
$\theta_c \in \theta$	$\equiv$	the model parameter corresponding to the count of configurations of type *c*
$\kappa(\theta)$	$\equiv$	the model's normalization term

Different choices of count configurations lead to different dependence assumptions. For example, an independent behavior assumption leads to the independence ALAAM model

[52] Interested readers can see an example in Snijders et al. (2007).

which is analogous to a logistic regression model. The network dependence assumption assumes that behaviors y are conditionally dependent on network ties. A major limitation of ALAAMs is their inability to handle heterogeneity.

B.2.3 Exponential-Family Random Network Models

The exponential-family random network model is a cross-sectional model that "[models] the joint relation between the processes of tie selection and nodal variate influence in a cross-sectional network" (Fellows and Handcock 2012). The model takes an exponential-family form similar to an ERGM. The model takes the following form:

$$P(G = g, X = x|\theta) = \frac{1}{\kappa(\theta)} \exp\left\{\sum_{c \in C} \theta_c z_c(g, x)\right\} \tag{92}$$

where:

$z_c(g, x)$ ≡ the count of configurations c in graph g and among nodal variates x

C ≡ the set of graph and nodal variate configuration types in the model's chosen dependence assumption

$\theta_c \in \theta$ ≡ the model parameter corresponding to the count of configurations of type c

$\kappa(\theta)$ ≡ the model's normalization term

B.3 Statistical Mechanics and Social Econometrics

Blume et al. (2011) denotes five different influence types in social influence models[53]:

1. $x_n \equiv$ individual-level observables for individual n
2. $k_g \equiv$ group-level observables for group g (contextual effects)
3. $\mu_{ng}^e(y_{-ng}) \equiv$ an individual's expectations (beliefs) of the behaviours of others in the group (endogenous effects) which are generally unobservable; here it is expressed as the individual's expectation of group members' average behaviour
4. $\varepsilon_n \equiv$ individual-level unobservables (correlated effects)
5. $\zeta_g \equiv$ group-level unobservables (unobserved contextual effects)

Most discrete choice models are based on random utility maximization in which an individual chooses the alternative which gives him[54] the most utility. Traditionally, an individual's utility is based on his characteristics and attributes of each alternative, but discrete choice models of social influence expand this to include contextual and endogenous social effects. Assume the maximization of some payoff, typically denoted as utility, from a set of alternatives. This payoff depends on the expectations of the individual, his preferences for individual-specific factors together with contextual and endogenous social factors, and the constraints imposed by his finite choice set C:

$$y_{ng} = argmax_{j \in C} V(j, x_n, k_g, \mu_{ng}^e(y_{-ng}), \varepsilon_n, \zeta_g) \qquad (93)$$

[53] The convention in this paper will generally refer to observables with Latin letters and unobservables with Greek letters.

[54] We refer to the modeler as *she* and an individual decision maker as *he*.

From this general form, the transportation literature generally uses model formulations based on field effects or network effects.

B.3.1 Field Effect DCMs

The statistical mechanics formulation, also known as the field effect formulation, was imported into social econometrics by Brock and Durlauf (2001). It involves the imposition of groups where all members of a group are influenced by the same individuals. These models are generally closed by assuming self-consistency – that all group members have the same expectations of other group members' behavior. The following general assumptions are made:

- **Static Game**: All individuals choose their actions before they see the actions of others.
- **Incomplete Information**: Individual n of group g knows x_m, k_g, and ζ_g for all individuals m in group g, but does not observe other individuals' random terms ε_{jm}.
- **Rational Expectations:** An individual's belief must equal the expected value of the market share for each alternative.
- **Self-Consistency**: $\mu_{ng}^{e}(y_{-ng}) = E(y_{-ng}|x_m, k_g, \zeta_g, \varepsilon_n \;\; \forall m \in g)$

In the game theory literature, this situation is analogous to finding the Bayes-Nash equilibrium.

Brock and Durlauf (2001) present a binary choice formulation. Brock and Durlauf (2002, 2006) extend the field effect formulation to finite choice sets with three or more alternatives. For conciseness, the multinomial formulation is presented here. Assuming

Bayes-Nash equilibrium with self-consistent expectations, the indirect utility of choosing alternative i for individual n in group g is[55]:

$$\mathcal{U}_{ing} = \alpha_i + \beta_i x_n + \gamma_i k_g + \delta_i r_{ing} + \zeta_{ig} + \varepsilon_{in}$$

$$\begin{aligned} y_{ni} &= 1 \quad if\ U_{ing} = \max_{j \in C} U_{jng} \\ y_{ni} &= 0 \quad otherwise \end{aligned} \tag{94}$$

The self-consistent expectations $\mu_{ng}^{e}(y_{-ng})$ are replaced with the field effect r_{ing}. This field effect represents the expected proportion of group members choosing alternative i. Most empirical studies close the model by setting r_{ing} equal to each alternative's observed market share. Assuming that ε_{in} is IID Gumbel distributed, we obtain a multinomial logit model (MNL) with social influence:

$$P_{ing} = \frac{\exp(\alpha_i + \beta_i x_n + \gamma_i k_g + \delta_i r_{ing} + \zeta_{ig})}{\sum_{j \in C} \exp(\alpha_j + \beta_j x_n + \gamma_j k_g + \delta_j r_{ing} + \zeta_{jg})} \tag{95}$$

B.3.2 Self-Selection Field Effect DCMs via Nested Logit

In travel demand models, group membership is generally assumed to be exogenous and self-selection into groups is not taken into account. While this assumption of exogenous group membership may be valid for studies on ethnic group or gender, other groups often require a (conscious) choice of group such as neighborhood selection, work environment, and income group.

Since there are a finite number of groups in social influence studies, group choice can be modeled as a discrete choice problem. For example, Ioannides and Zabel (2008)

[55] The parameters are subscripted by alternative to show that the same set of individual-level and contextual variables need not be used in each alternative's indirect utility. Standard normalization rules for random utility models still apply.

analyze social effects in housing demand (continuous variable) but account for self-selection into neighborhoods with a multinomial choice model. Brock and Durlauf (2003) suggested that a nested logit model could be used to model group and discrete behaviour choice simultaneously. Zanella (2007) develops the formulation for a nested logit model of endogenous group membership and discrete behaviour choice.

Zanella decomposes the formulation into two logit models: one for the probability of choosing group g and the other for the probability of choosing alternative j conditional on being in group g.

$$P_i(g,j) = P_i(j|g) \cdot P_i(g) \tag{96}$$

The probability of choosing alternative j conditional on being in group g is the basic model given before[56]:

$$P_i(j|g) = \frac{\exp(\alpha_j + \beta_j x_i + \gamma_j k_g + \delta_j p_{igj}^e)}{\sum_{l \in C} \exp(\alpha_l + \beta_l x_i + \gamma_l k_g + \delta_l p_{igl}^e)} \tag{97}$$

Zanella then provides a group choice model which depends on the group-level observables k_g and the inclusive utility W_{ig}. The inclusive utility is the expected utility an individual is expected to obtain from choosing one of the available alternatives. Since individuals are assumed to exhibit utility maximizing behaviour, this log-sum is proportional to the expected maximum value between Gumbel distributed random variables.

$$P_i(g) = \frac{\exp(\lambda \psi k_g + \lambda W_{ig})}{\sum_{h \in G} \exp(\lambda \psi k_h + \lambda W_{ih})} \tag{98}$$

[56] In this derivation, Zanella's work is extended to three or more alternatives (behavior choices).

$$W_{ig} = log \sum_{l \in C} \exp(\alpha_c + \beta_l x_i + \gamma_l k_g + \delta_l p_{igl}^e) \tag{99}$$

The strength of this approach is that it maintains utility maximizing behavior and the model can be estimated using standard nested logit software. However, Zanella's model is limited by its assumptions on the random component of utility.

This random component can be decomposed into components that vary across groups only, alternatives only, and both:

$$\varepsilon_{ing} = \xi_{ng} + \xi_{in} + \xi_{ing} \tag{100}$$

Zanella closes the model (as with nested logit in general) by assuming that "[ξ_{ing} and ξ_{ng}] are independent for all individuals, groups, and behaviors." This assumption can be inappropriate in some contexts, such as the choice of neighborhood and travel mode. Because the decision to live near transit is likely correlated with the decision to choose transit, ξ_{ing} and ξ_{ng} are likely correlated. Possible approaches to relax these assumptions include multinomial probit, mixed logit, or other generalized extreme value models.

B.3.3 Endogenous Spatial Weights

In Conley and Topa (2007), a dynamic model of choice under social influence is described and applied to the application of finding employment. Individuals can exist in one of two states ($y_n^{(t)}$): employed ($y_n^{(t)} = 1$) or unemployed ($y_n^{(t)} = 0$). Conley and Topa assume that social networks only matter when an individual is unemployed to aid in model identification and due to data limitations. Individuals are connected in reflexive graphs (cliques) by census tract. Their model represents a Markov process with transitions occurring with a logit conditional transition function. The probability of transitioning from employed to unemployed is as follows:

$$Prob\left(y_n^{(t+1)} = 0 \middle| y_n^{(t)} = 1\right) = \frac{\exp(\beta x)}{1 + \exp(\beta x)} \tag{101}$$

This function depends only on characteristics of the individuals. The probability of transitioning from unemployed to employed is as follows:

$$\begin{aligned} &Prob\left(y_n^{(t+1)} = 1 \middle| y_n^{(t)} = 0\right) \\ &= \frac{\exp(\beta x + \delta_1 \sum_{q \in g_n} y_q^{(t)} \cdot 1(x_{race,n} = x_{race,q}) + \delta_2 \sum_{q \in g_n} y_q^{(t)} \cdot 1(x_{race,n} \neq x_{race,q}))}{1 + \exp((\beta x + \delta_1 \sum_{q \in g_n} y_q^{(t)} \cdot 1(x_{race,n} = x_{race,q}) + \delta_2 \sum_{q \in g_n} y_q^{(t)} \cdot 1(x_{race,n} \neq x_{race,q})))} \end{aligned} \tag{102}$$

This probability depends on the individuals' characteristics as well as their assumed social network (in this case, census tract-level behavior). The transition back to employment depends on the total number of individuals in one's census tract who are employed. Additionally, there is varying influence occurring between individuals in the census tract of the same race versus individuals in the census tract of another race. Conley and Topa claim that their model is analogous to contact processes in interacting particle systems (Liggett 1985, 1999). To estimate their model, a calibration-based simulation procedure is used.

B.4 Spatial Econometrics

The network effect formulation, with origins in spatial econometrics and social network analysis, emphases that individuals are connected in varying ways and that the heterogeneity generated from varying network structures and influence patterns is important for analysis. The primary features of network effects-based social influence models are:

- **Social Distance:** The modeler must define a measure of distance between individuals. Physical distance is most commonly used but social measures can be used instead, such as education, income, or political views.
- **Influence Transmission:** The modeler must identify a mechanism that transmits the influence. This is often the actual choices of others.

Network effects models use a weighting matrix to represent social distance. A weighting matrix describes the degree of influence between each individual in the population. For example, Sidhartan et al. (2011) estimates an inverse-distance weighting matrix, where people living closer to an individual exert more influence than people living farther away. Most empirical studies in transportation use predetermined (exogenously-formed) weighting matrices with equal influence structures such as:

$$w_{nm} = \begin{cases} 1/S_n & if\ individual\ m\ is\ in\ n's\ social\ network \\ 0 & otherwise \end{cases} \tag{103}$$

where S_n is the number of people in individual n's social network.

B.4.1 Conditional Autoregressive DCMs

The conditional spatially autoregressive discrete choice model is similar to the social econometrics formulation. Begin by assuming that each individual has a personalized social network and that he knows and is influenced by the decisions of all members in his social network. An individual's indirect utility for alternative i is:

$$\mathcal{U}_{ni} = \beta_{0i} + \beta_i x_n + \delta_i \sum_{m=1}^{M} w_{nm} y_m(i) + \varepsilon_{in} \tag{104}$$

where $y_m(i)$ equals one if individual m chose alternative i and zero otherwise. The modeler must be careful with this formulation as the social lag term $y_m(i)$ is likely endogenous due to omitted variable bias and simultaneity. A useful property of this model is that it can be estimated using standard logit or probit software as long as the weighing matrix has no parameters (i.e. the weights are fully known before estimation) and appropriate instruments are used for handling endogeneity in $y_m(i)$.

Goetzke (2008) analyzes transit mode choice using this form but assumes that $y_m(i)$ is exogenous to simplify model estimation. Adjemian et al. (2010) use a similar model to predict auto ownership by class with a series of binary logit models. They justify the exogeneity assumption by stating that automobile purchases are major household purchases therefore influence must be one-directional but do not explicitly account for omitted variables. Páez and Scott (2007) present a similar model but modify equation (104) by having the utility an individual gains from choosing an alternative depend on the past choices of his peers. This breaks the simultaneity issue but the modeler must be careful to choose an appropriate length between time periods.

B.4.2 Simultaneous Autoregressive DCMs

Since conditional autoregressive models have simultaneity issues, some researchers model the decision process as a system of simultaneous equations. Behaviourally, this formulation is different from the conditional autoregressive and field effect formulations as the individual is affected by perceptions of the preferences of others $\mu_{ng}^{e}(U_{-ng})$ rather than their decisions $\mu_{ng}^{e}(y_{-ng})$.

As an introduction, the binary choice formulation[57] of the spatially autoregressive lagged dependent variable model (Fleming 2004) with M individuals is presented where:

$$\mathcal{U}_n = \mathcal{U}_n(1) - \mathcal{U}_n(0) = \alpha + \beta x_n + \delta \sum_{m=1}^{M} w_{nm} \mathcal{U}_m + \varepsilon_n$$

$$y_n = \begin{cases} 1, \ if \ \mathcal{U}_n \geq 0 \\ 0, \ if \ \mathcal{U}_n < 0 \end{cases} \tag{105}$$

The formulation becomes clearer when written in matrix form. Let $\mathcal{U} = [\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_M]$, $W = [w_{11}, w_{21}, \dots, w_{1M}; \dots; w_{M1}, w_{M2}, \dots, w_{MM}]$, $X = [x_1, x_2, \dots, x_M]$, $K = [k_1, k_2, \dots, k_M]$, and $\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M]$, then model can be rewritten as:

$$U = (I - \delta W)^{-1}(\alpha + X\beta + \varepsilon) \tag{106}$$

where I is the identity matrix. In the spatial econometrics literature, Fleming (2004) calls this the spatially autoregressive lagged dependent variable model (SAL)[58]. For a binary probit model, ε is multivariate normal with mean zero and variance-covariance matrix:

$$\Omega = (I - \delta W)^{-1}((I - \delta W)^{-1})' \sigma_\varepsilon^2 \tag{107}$$

Estimation of the probit SAL is computationally difficult since the likelihood function involves a multidimensional integral of the form:

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_M = y_M) = \int_{-\infty}^{a_1} \dots \int_{-\infty}^{a_M} \phi(\varepsilon) d\varepsilon$$

$$where: \phi(\varepsilon) = (2\pi)^{-\frac{M}{2}} |\Omega|^{-\frac{1}{2}} exp^{\left(-\frac{1}{2}\right)(\varepsilon' \Omega^{-1} \varepsilon)} \tag{108}$$

[57] See Sidharthan et al. (2011) for an example of the multinomial probit form.

[58] Fleming mentions the spatially autoregressive error model (SAE) in which the error terms are spatially correlated by weights. This is not covered since the focus is on endogenous social effects.

As the size of social networks increase, the greater the dimensionality of the integral will be. Fleming (2004) surveys the approaches for estimation of binary probit SAL models, including the Expectation-Maximization and Gibbs Sampler methods, and Sidharthan et al. (2011) suggests the maximum composite marginal likelihood approach for multinomial probit.

B.5 Information Cascades Experiments

In many segments of society, collections of people follow similar behavior. In some cases this uniform social behavior is fragile, such as in fashion and fads. In other cases this behavior is not fragile, such as cultural norms and religion. Understanding why people tend to conduct similar behavior is researched in psychology, sociology, and economics. In economics, the information cascade explanation of these behaviors has risen in prominence. In this framework, sequences of individuals make decisions about a task of interest. Each individual can see the decisions that prior decision makers have made (public information) but not necessarily the reasons for those decisions (private information). Information cascade research attempts to explain when people begin to ignore their private information and just follow the direction of others. When this occurs, a information cascade forms. Experimental economists have attempted to analyze this phenomenon in the laboratory and in the field[59] to see if it occurs and why it occurs.

This section begins by looking at the theoretical basis for information cascade experiments. Bikhchandani et al. (1992) proposed a novel approach for understanding the fragility of some forms of uniform social behavior such as fads, fashion, customs, and

[59] No field experiments are featured in this literature review.

culture. To explain their approach, “a specific model” is shown that looks at a sequence of individuals deciding on a binary decision. Each individual receives a private signal and knows the decisions of past respondents but not their signals. The authors state that a Bayesian decision process would be the rational approach for an individual to pick the appropriate action and that this process would lead to cascade behavior as the number of individuals in a sequence increases. Experimental research of information cascades generally uses this model as a starting point.

The experimental setup and interpretation of the decision process is important in studying information cascades through experimental economics. Anderson and Holt (1997) was the first significant experiment to test Bikhchandani et al (1992) theory. Their lab experiment involved urns with varying numbers of balls of different types. A binary signal was given to the subjects and a binary decision was expected of them. Their results showed that cascades occurred in this laboratory setup and their modeling approach seemed to support that these decisions tended to be Bayesian. In other words, individuals were cascading because they reasoned via Bayesian decisions processes, which predicts that cascades should often occur. Hung and Plott (2001) found similar results of rational Bayesian decision making.

Since their novel experiment, criticism of Anderson and Holt’s (1997) interpretation of their cascade results being caused by Bayesian decisions have been prevalent. Noth and Weber (1999) and Huck and Oechssler (2000) found that respondents did not act rationally as expected by Bayes’ rule. Their analysis found inconsistencies in what the expected rational behavior of students should be and their actual decisions. Additionally, self-reported explanations of respondent’s decision reasons showed that few attempted to use Bayes' rule and no one who claimed to use Bayes' rule applied it correctly. The authors proceed to discuss that, in Anderson and Holt's experiment, alternative decision techniques were in play that just led to similar results from Bayesian decision processes. Spiwoks et al. (2008) performed an experiment to explicitly test whether respondents acted according to Anderson and Holt (1997) or Huck and Oechssler (2000). This research found that Huck and Oechssler's findings hold and suggested that a reinterpretation of Anderson and Holt's analysis was needed.

Additional skepticism of the Bayesian decision process in information cascades included the work of Kübler and Weizsäcker (2004). Their study modifies the Anderson and Holt design by forcing respondents to choose whether to buy their private signal or not. They found that respondents were unable to think on a high enough level[60] to decide via Bayes' rule and thus respondents chose to trust their own signals more and to distrust some of the choices of others.

After showing this selected progression of the Anderson and Holt's experimental approach, the literature review concludes by looking at some alternative approaches to cascade experiments. Particular emphasis is placed on Çelen and Kariv (2004), where the change from a discrete signal to a continuous signal and belief elicitation allow for differentiating cascades – where private information is ignored and individuals only follow the latest actions – and herd behavior – where private information and public information are used jointly to decide. Information cascades are also looked at in voting

[60] In regards to thinking about themselves, relying on others decisions, as well as others also relying on others decisions, and so on.

(Hung and Plott, 2001), markets (Cipriani and Guarino, 2005), travel behavior analysis (Gaker et al., 2010) and social networks (Choi et al., 2005).

B.6 Identity Economics

Akerlof and Kranton's (2000) paper “Economics and Identity” was not the first paper to incorporate the concept of identity into an economic framework, but it is the most popular conception. In this conception, social identity theory is combined with the psychodynamics of personality in an economic framework of utility maximization or cognitive dissonance minimization (Davis 2010). Traditional economic models are concerned with utility maximizing individuals who are only affected by their own actions and the actions of others. Akerlof and Kranton propose to make this utility function depend on a person's “identity” or sense-of-self as well.

B.6.1 Theoretical Foundations

Akerlof and Kranton (2010) summarize their approach as containing two parts: standard utility and social context. A person's standard utility includes “a person's tastes for goods, services, or other economic outcomes.” The social context is comprised of three components:

- *Social Group Identity*. These are the *social categories* which differentiate individuals. Akerlof and Kranton refer to this as a person's *identity*.
- *Social Norms/Prescriptions*. These are the expected courses of conduct for individuals in different social categories. Akerlof and Kranton also refer to the concept of *ideals* which individuals strive to achieve.

- *Identity Utility*. Achieving and not achieving one's ideals affect one's utility. When prescriptions are followed (by an individuals or others), then an individual may gain utility. Akerlof and Kranton refer to this as a gain in identity (or self-image). Not following prescriptions has the opposite effect; it can cause losses in one's utility. These loses can induce anxiety or cognitive dissonance, as the self feels a disconnect between its ideals and its realization of itself. This in turn may cause individuals to perform actions to minimize this cognitive dissonance.

To formalize this, their framework begins with the following utility function:

$$U_j = U(a_j, a_{-j}, I_j) \tag{109}$$

where a_j is the actions of individual j , a_{-j} are the actions of other individuals, and I_j is individual *j*'s identity/sense-of-self. The identity (self-image) component, I_j, has the following form:

$$I_j = I(a_j, a_{-j}, c_j, \epsilon_j, P) \tag{110}$$

where c_j is the social category of individual j, ϵ_j is *j*'s characteristics or abilities, and *P* are the prescriptions or social norms associated with each social category. It is not required that all of these elements are dynamic, and often to simplify analysis or to focus on a particular problem, it is assumed that some components of the identity utility are given. Akerlof and Kranton typically assume that the social categories, individual characteristics, and prescriptions are given, which Davis (2011) has criticized as limiting. Akerlof and Kranton (2000) say that their work expands economic analysis in the following ways:

1. “Identity can explain behavior that appears detrimental”
2. “Identity underlies a new type of externality. One person’s actions can have meaning for and evoke responses in others”
3. “Identity reveals a new way that preferences can be changed”
4. “Choice of identity may be the most important ‘economic’ decision people make”

The next section describes an example of their approach applied to education.

B.6.2 Education and Identity

Akerlof and Kranton (2002) presents an identity model of student effort. In classic models of education in economics, students exhibit tradeoffs between effort and opportunity cost and school quality is generally a function of resource expenditures. In their identity model, students “choose” groups to identify with (leading crowd, nerd, burnout) and adjust effort levels to the ideals of the group. They support their specification based on work from social psychology experiments (e.g. Robbers Cave) and behavioral observational studies.

To formalize their model, in this school, students separate into the categories: Leading Crowd, Nerds, and Burnouts. Utility is a function of effort in class and salary returns from this effort. Students choose a social group and level of effort, but are impacted by the status level and social norms of the school. Akerlof and Kranton formulate a model for the utility of student *i* as follows:

$$U_i(L) = p\left[wk_i - \frac{1}{2}e_i^2\right] + (1-p)\left[I_L - t(1-l_i) - \frac{1}{2}\left(e_i - e(L)\right)^2\right]$$

The tradeoff between salary and effort time is represented by $wk_i - \frac{1}{2}e_i^2$. The second term in the utility represents the social status of a student with I_L equal to the status of the student's social category, $t(1 - l_i)$ describes the similarity between individual and ideal group member and its importance (i.e. social fit), and $\frac{1}{2}\left(e_i - e(L)\right)^2$ is the distance between ideal effort level for this group and the individual's effort (i.e. social norms).

B.7 Travel Behavior Modeling

This section briefly describes some of the social influence models used in the travel behavior modeling community.

B.7.1 Borrowed Inspirations

The statistical mechanics formulation of Brock and Durlauf (2001) serves as the most prominent source of inspiration for travel behavior models of social influence. Fukada and Morichi (2007) used this form to study illegal bicycle parking behaviour in Tokyo and looked at the equilibrium properties and policy interventions. Goetzke and Rave (2011) estimate binary logit models of bicycle mode choice in Germany by trip purpose similarly but handle endogenity in the social influence term with instrumental variables. Goetzke and Weinberger (2012) used a similar form to Brock and Durlauf (2001) but estimated a binary probit model instead. Another slight modification includes Kuwano et al. (2011) studies vehicle ownership over time with a dynamic GEV model of diffusion at the national, neighborhood, and income group level.

The spatial econometrics formulations have been the second most prominent inspiration. Goetzke (2008) analyzes transit mode choice using this form but assumes that the behavior of others is exogenous to simplify model estimation. Adjemian et al. (2010) use a conditional spatially autoregressive model to predict auto ownership by class with a series of binary logit models. They justify the exogeneity assumption by stating that automobile purchases are major household purchases therefore influence must be one-directional but this does not account for endogeneity from omitted variable bias. Additionally, they set the neighborhood level auto ownership to the observed ownership at the time of the vehicle purchase. Páez and Scott (2007) present a similar formulation but use a temporally and spatially lagged term. This breaks the simultaneity issue, but the modeler must be careful to choose an appropriate length between time periods. Sidharthan et al. (2011) uses the simultaneous formulation in a multinomial probit model and uses maximum composite marginal likelihood approach to make the estimation tractable.

B.7.2 Spatial and Social Heterogeneity

Dugundji and Walker (2005) describe a mixed GEV model similar to Brock and Durlauf (2000) that incorporates group-level random-effects. Smirnov (2010) terms this the *spatial heterogeneity* model. Smirnov (2010) think this model is likely “meaningful in the context of social interactions, where social group membership effectively channels interactions between individuals to within-the-group interactions” but not for between-group interactions. The model formulation is as follows:

$$\begin{gathered} \mathcal{U}_{ing} = \alpha_i + \beta_i x_n + \gamma_i k_g + \delta_i r_{ing} + \zeta_{ig}(g) + \varepsilon_{in}, \qquad \zeta_{ig} \sim N(0, \sigma_{ig}) \\ y_{ni} = 1 \quad if\ U_{ing} = \max_{j \in C} U_{jng} \\ y_{ni} = 0 \quad otherwise \end{gathered} \tag{111}$$

Appendix C: Endogeneity in Social Influence Choice Models

Econometric modeling concentrates heavily on the properties of expectations. Most models are generally concerned with a dependent variable y – a linear-in-parameters β function of Q variables, $X = [x_1, x_2, \dots, x_R]$. To account for unobservables, an error term ϵ is added, giving the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_Q x_Q + \epsilon \tag{112}$$

Estimating these models generally entails two assumptions:

- **Zero Mean Error:** $E(\epsilon) = 0$
- **Exogenous Variables:** $Cov(\epsilon, x_q) = 0 \;\; \forall q \in \{1,2, \dots, Q\}$

When the covariance between the error term and a variable x_q is non-zero, x_q is called an endogenous variable and the estimation of its corresponding parameter can be biased. Endogeneity, correlation between a variable and the error term, is caused by many factors (Antonakis et al. 2010) and the most common causes in social influence models include:

- *Omitted Variables.* This is a major problem when unobserved group-level factors are correlated with the social influence terms. This is also the origin of social influence model, as excluding social factors can bias traditional travel demand model when social influence is present.
- *Self-Selection.* Individuals may select into particular groups because they share similar preferences. For example, individuals who like transit may choose to live in a neighborhood with good transit connections; a model could overestimate the social multiplier effect for transit mode choice among residents of this neighborhood if it does not correct for self-selection of residential location.

- *Simultaneity.* Social influence models often include the group-level market share for an alternative. Since individuals are making decisions at the same time, individual choice and aggregate market share are determined simultaneously, and it is difficult to differentiate the direction of this influence in single-equation models.
- *Model Misspecification.* This cause is difficult to determine but generally involves making too many assumptions or poor assumptions in the model. In the social influence choice modeling field, this typically entails misspecification of the social influence mechanism and social network. Additionally, some full-information estimators have to make strong assumptions about conditional probabilities, while two-stage approaches may relax those assumptions.

To ensure consistent and unbiased estimates, the modeler must ensure that only exogenous variables are present. This typically entails finding instruments which are correlated with the endogenous variable x_q but uncorrelated with the error term ϵ. The modeler can then estimate an exogenous model and determine the endogenous variable's effect through its impact on its corresponding instruments. In social influence models, endogeneity due to omitted variables and simultaneity has typically been handled through the BLP and control function approaches.[61]

C.1 BLP Approach

The BLP approach, which originates from Berry, Levinsohn, and Pakes (1995), is a three-stage approach to dealing with endogeneity at the market-level. The approach

[61] Train (2009) discusses endogeneity in discrete choice models.

requires the endogenous variable to be shared between multiple decision makers in the same market. First, a choice model is estimated with constants that correspond to each alternative and market. Second, each constant is regressed on the endogenous variable and its corresponding instruments in a linear model using an instrumental variable (IV) estimation technique. Finally, the model is corrected by inserting the effects of the endogenous and instrumental variable from the linear model into the discrete choice model.

Walker et al. (2011) used this approach to deal with endogeneity in the field effect term in a cross-nested logit model of mode choice in the Netherlands. In their research, they place individuals into two groups – a spatial group based on postal code and a social group based on income – and therefore each person has two field effect variables. Since this field effect is the same between individuals in the same spatial or social group, the BLP approach can be used since this each group is interpreted as a "market."

To simplify the presentation of the approach, assume that each individual belongs to only one group, but note that the BLP approach can work for multiple endogenous variables. For this approach, Walker et al. begins with a model with individual-level characteristics and a field effect p_{igj}^{e}:

$$U_{ig}(j) = \alpha_j + \beta x_{in} + \delta p_{igj}^{e} + \varepsilon_{ij} \qquad (113)$$

There is correlation between the field effect and U_{igj} due to omitted variables (similar people may have similar unobservables) and simultaneity. We need the error term to be uncorrelated with the regressor, so Walker et al. break apart ε_{ij} into λ_{igj} (correlated with the field effect) and ε_{ij}^{*} (uncorrelated with the field effect).

$$U_{ig}(j) = \alpha_j + \beta x_{ij} + \delta p^e_{igj} + \lambda_{igj} + \varepsilon_{ij} \tag{114}$$

Assuming that all members have the same expectations of others' behaviour, the endogenous field effect is the same for all group members. Walker et al. then replace the field effect in (114) for all individuals in the same group with a group-specific constant ϕ_{igj}.

$$U_{ig}(j) = \alpha_j + \beta x_{ij} + \phi_{igj} + \varepsilon_{ij} \tag{115}$$

Endogeneity in the linear model, $\phi_{igj} = \delta p^e_{gj} + \lambda_{igj}$, is corrected for with a two-stage IV approach:

$$\begin{aligned} &1.\quad p^e_{igj} = \theta_i + \theta_m I_g + v_{ig} \rightarrow \hat{p}^e_{igj} = \hat{\theta}_i + \hat{\theta}_p I_g \\ &2.\quad \phi_{igj} = \delta_i + \delta_p \hat{p}^e_{igj} + \lambda_{igj} \end{aligned} \tag{116}$$

First a regression is run on the actual group-level field effects, where v_{ig} has zero mean and is uncorrelated with I_g. Then estimates are obtained for the parameters θ_i and θ_p and an estimate for the field-effect. In the second stage, the group-specific constant ϕ_{igj} is regressed on the estimated field effect $\hat{p}^e_{igj}$ to obtain the corrected social multiplier effect δ_p.

The appropriate choice of instruments in IV estimation is critically important. As stated above, the instruments must be correlated with the endogenous variables (the stronger the better) but must be uncorrelated with the error term in the utility functions / latent variable model. For the models in Walker et al. (2011), instruments were needed for the field effect in each spatial and social group. For the spatial groups, the average of the field effects from adjacent postcodes was used. These were deemed as "natural instruments" and the use of adjacent zonal characteristics as instruments is common in spatial models. It is critically important that the zones are genuinely distinct otherwise the instruments will be invalid and endogeneity bias will persist (e.g. breaking up a homogeneous neighborhood into two separate zones does not solve endogenity bias since both zones have similar unobservables).

Instruments for non-spatial social groups are harder to deal with and researchers have had difficulty finding "natural instruments" for their work. Walker et al. initially tried to use the social field effect from adjacent zones as well as the field effect from the next lowest income group, but the correlations were not sufficient. Their final model used income, age, and weekly work hours as instruments for the social group's field effect.

C.2 Control Function Approach

The control function approach, with discrete choice origins in Rivers and Vuong (1988), is a two-stage approach for dealing with endogeneity at the individual-level. First, the endogenous variable for each individual is regressed on some instruments and residuals are calculated. The estimated endogenous effect and the residuals are then used in the estimation of the choice model.

Goetzke and Weinberger (2012) use a control function approach in their study of the influence endogenous and contextual social effects on automobile ownership in New York City. Social groups are delineated by census tracts and the decision to own a vehicle is modeled by a binary probit model with the following latent variable structure:

$$y_{ig}^* = \alpha + \beta x_i + \gamma k_g + \delta p_{ig}^e + \varepsilon_{ig}$$
$$y_i = \begin{cases} 1, \ if \ y_i^* \geq 0 \\ 0, \ if \ y_i^* < 0 \end{cases} \tag{117}$$

where x_{ig} are household-level characteristics, p_{ig}^{e} is the observed mean auto ownership (between 0.0 and 1.0) in census tract g (endogenous social effects), and k_g are observed tract-level built environment characteristics and average socio-economic characteristics of the tract's residents (contextual social effects). Since tract-level mean ownership is endogenous, a control function is used:

$$p_{ig}^{e} = \theta_0 + \theta_1 x_i + \theta_2 k_g + \theta_3 I_p + \upsilon_i \tag{118}$$

where I_p are additional instruments that are not contextual or individual-specific effects. With the probit model and control function, some assumptions must be made about the error terms to make the model tractable. A bivariate normal distribution is assumed for (u_i, ν_i) with zero mean and the following normalization from Wooldridge (2009):

$$var\ u = 1 \text{ AND } u_i = \nu_i \frac{cov(\nu, u)}{var\ \upsilon} = \eta_i \tag{119}$$

With (119), the latent variable model as can be rewritten as:

$$y_{ig}^{*} = \alpha + \beta x_i + \gamma k_g + \delta p_{ig}^{e} + \lambda \nu_i + \eta_i \tag{120}$$

First, the control function regression is run with ordinary least squares (OLS) and the residuals $\hat{\nu}_i$ are calculated. Then the probit model is run with $\hat{\nu}_i$ used as an estimate of υ_i in the latent variable model (120) and the coefficient λ can be used to test if p_{ig}^{e} is exogenous (null hypothesis of $\lambda = 0$). This two-stage conditional maximum likelihood approach leads to consistent estimates of the parameters.

C.3 Instrument Selection

The appropriate choice of instruments is critically important. As stated above, the instruments must be (strongly) correlated with the endogenous variables but uncorrelated with the error terms in the utility functions. In Walker et al. (2011), instruments were needed for the field effect of each spatial and social group. For the spatial groups, the average of the field effects from adjacent postal codes was used. These were deemed as "natural instruments" and the use of adjacent zonal characteristics as instruments is common in spatial models. It is critically important that the zones are genuinely distinct otherwise the instruments will be invalid and endogeneity bias will persist (e.g. breaking up a homogeneous neighborhood into two separate zones does not solve endogenity bias since both zones have similar unobservables).

Instruments for non-spatial social groups are more difficult and researchers have had difficulty finding parsimonious instruments for their work. Walker et al. (2001) initially try to use the social field effect from adjacent zones as well as the field effect from the next lowest income group, but the correlations were insufficient. Their final model uses income, age, and weekly work hours as instruments for the social group's field effect. Goetzke and Weinberger (2012) suggest using the characteristics of the entire group (contextual variables) as instruments for the field effect. But this suffers from the same issues of instrument appropriateness since there is no theoretical foundation for instrument selection in social group problems. Instrument selection is an ongoing problem and requires the modeler to explore different specifications and think clearly about what constitutes the error term in her particular modeling situation.

Appendix D: A Generalized Informational Conformity RUM

The informational conformity model shown in Chapter 8 can be generalized beyond just preference changes. As a reminder, Durlauf and Ioannides (2010) note that social interactions are "direct interdependences in preferences, constraints, and beliefs of individuals, which impose a social structure on individual decisions" (p.452). The informational conformity model can be generalized to incorporate changes in constraints and beliefs as well as preferences. The generalized informational conformity discrete choice model is a latent class discrete choice model that can be formulated in a conformity latent class structure.

D.1 Class-Membership Model

The model begins with an information term that defines class membership. It is defined as follows for individual *n* and class *c*:

$$\mathcal{F}_n^{[c]} = \alpha^{[c]} z_n + \delta^{[c]} s_n^{[c]}\big(G_n(w), m_n(N), m_n^*(N)\big) + \epsilon_n^{[c]} \tag{121}$$

where:

z_n $\equiv$ individual-level characteristics of individual *n*

$s_{ni}^{[c]}(\cdot)$ $\equiv$ social influence mechanisms for individual *n* in class *c* for alternative *i* due to endogenous and contextual factors

$G_n(w)$ $\equiv$ individual *n*'s social contacts and the strength of these relationships (modeled through a weighting function *w*)

$m_{ni}(N)$ $\equiv$ exogenous social influence sources of the population on individual *n* for alternative *i*

$m_{ni}^{*}(N)$ $\equiv$ endogenous social influence sources of the population on individual *n* for alternative *i*

N $\equiv$ the population of all individuals

$\alpha^{[c]}, \delta^{[c]}$ $\equiv$ information (class-membership) model parameters (these can be class-specific)

$\epsilon_{n}^{[c]}$ $\equiv$ error term for individual *n*

For each class, a separate discrete choice model specification can be defined with varying preferences, beliefs, and constraints. In the next four sections, each of these variations will be defined in isolation then combined in the final specification.

D.2 Utility Function for Preference Changes

This specification is the same as shown in Chapter 8. It follows similarly from equations (44) and (45), but generalized to more than two choices. For each alternative and each class, an individual's utility for making a choice is as follows:

$$\mathcal{U}_{ni}^{[c]} = \dot{\beta}_{i}^{[c]} \dot{x}_{ni} + \varepsilon_{ni}^{[c]} \tag{122}$$

where:

$\dot{x}_{ni}$ $\equiv$ individual-level characteristics of individual *n* and alternative *i*

$\dot{\beta}^{[c]}$ $\equiv$ model parameters specific to class *c*

$\varepsilon_{ni}^{[c]}$ $\equiv$ unobserved effects on individual *n* for classes *c* and alternative *i*

D.3 Utility Function for Belief Changes

New information may cause individuals to update their beliefs on the attributes of an alternative. In other words, individuals may change the individual-level characteristics of an alternative in accordance with their beliefs of the characteristics of that alternative. This is represented in the model as a change in x_{ni} for each class. For each alternative and each class, an individual's utility for making a choice is as follows:

$$\mathcal{U}_{ni}^{[c]} = \ddot{\beta}_i \ddot{x}_{ni}^{[c]} + \varepsilon_{ni}^{[c]} \tag{123}$$

where:

$\ddot{x}_{ni}^{[c]}$ $\equiv$ individual-level characteristics of individual *n* and alternative *i*

$\ddot{\beta}_i$ $\equiv$ model parameters

$\varepsilon_{ni}^{[c]}$ $\equiv$ unobserved effects on individual *n* for classes *c* and alternative *i*

D.4 Decision Rule for Constraint Changes

New information may increase the knowledge of available options for an individual. In discrete choice models, constraints are handled by the choice set. Thus to deal with this in a latent class formulation, the corresponding classes will have different choice sets. The formulation for the decision rule is as follows:

$$y_{ni} = \begin{cases} 1 & if\ \mathcal{U}_{ni}^{[c]} = \max_{j \in J^{[c]}} \mathcal{U}_{nj}^{[c]} \\ 0 & otherwise \end{cases} \tag{124}$$

where:

y_{ni} $\equiv$ a choice indicator function for individual *n* for alternative *i*

J $\equiv$ the set of all possible choices for the population / sample

$J^{[c]}$ $\equiv$ the choice set for class *c*, where $J^{[c]} \in J, \forall c$

D.5 Combined DCM Specification

The generalized model combines the preference, expectations, and constraint changes into a single model under a latent class discrete choice framework. The model is specified as follows:

$$\mathcal{F}_n^{[c]} = \alpha^{[c]} z_n + \delta^{[c]} s_n^{[c]}\big(G_n(w), m_n(N), m_n^*(N)\big) + \epsilon_n^{[c]}, \qquad \forall c$$

$$c_n = c, \text{ where } \mathcal{F}_n^{[c]} = \max_{b \,\in\, \forall c} \mathcal{U}_{nj}^{[b]}$$

$$\mathcal{U}_{ni}^{[c]} = \dot{\beta}_i^{[c]} \dot{x}_{ni} + \ddot{\beta}_i \ddot{x}_{ni}^{[c]} + \dddot{\beta}_i \dddot{x}_{ni} + \varepsilon_{ni}^{[c]}, \qquad \forall c \tag{125}$$

$$y_{ni}^{[c]} = \begin{cases} 1 & if\ \mathcal{U}_{ni}^{[c]} = \max_{j \in J^{[c]}} \mathcal{U}_{nj}^{[c]} \\ 0 & otherwise \end{cases}, \qquad \forall c$$

Appendix E: Ordered Choice Informational Conformity Model

E.1 Model Formulation

This formulation follows similarly to the binary choice formulation and will be written under an ordered choice decision with four choices. Begin by assuming a population *N* of decision makers where individuals are connected in a social network *G*. Each individual *n* is faced with a choice task where the individual must choose between four levels of response $y_n = \{0,1,2,3\}$. In this population, individuals may be influenced via informational conformity (class $\mathscr{b}$) or not influenced (class $\mathscr{a}$). This process is unobserved and will be modeled latently with discrete classes. Class membership is affected by the average response of an individual's connected social contacts $\sum_{q \in g(n)} \frac{y_q}{\|g(n)\|}$. This will be denoted by the information function $\mathcal{F}$, which will take the following linear-in-parameter form[62]:

$$\mathcal{F}_n = \alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|} + \varepsilon_n^{\mathcal{F}} \tag{126}$$

Assuming that the error term $\varepsilon_n^{\mathcal{F}}$ is IID logistic (with location 0 and scale 1), then the probability for an individual to be in the informed class takes the familiar binary logit (logistic regression) form as follows:

$$\pi_n^{[\mathscr{b}]} = \mathrm{Prob}(c_n = \mathscr{a}) = \frac{\exp(\mathcal{F}_n)}{1 + \exp(\mathcal{F}_n)} = \frac{\exp(\alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|})}{1 + \exp(\alpha z_n + \delta \sum_{q \in g(n)} \frac{y_q}{\|g(n)\|})} \tag{127}$$

[62] No time superscripts are used in this formulation as it is assumed that the modeler will choose an appropriate formulation. The model can be formulated for both the cross-sectional and dynamic cases.

Accordingly, the probability of being in class $\mathcal{a}$ follows:

$$\pi_n^{[\mathcal{b}]} = \mathrm{Prob}(c_n = \mathcal{b}) = 1 - \pi_n^{[\mathcal{a}]} \tag{128}$$

It is expected that individuals surrounded by others with larger responses may be able to reevaluate their preferences for the alternative under the new information they receive from being exposed to these stronger responses more often than other people. Thus, the preferences of these "more informed" individuals may vary compared to the "less informed" individuals. Assuming utility maximizing behavior for individuals, the utility an individual *n* for each class $\{\mathcal{a}, \mathcal{b}\}$ is given as follows:

$$\mathcal{U}_n^{[\mathcal{a}]} = \beta^{[\mathcal{a}]} x_n^{[\mathcal{a}]} + \beta^* x_n^* + \varepsilon_n^{[\mathcal{a}]} \tag{129}$$

$$\mathcal{U}_n^{[\mathcal{b}]} = \beta^{[\mathcal{b}]} x_n^{[\mathcal{b}]} + \beta^* x_n^* + \varepsilon_n^{[\mathcal{b}]} \tag{130}$$

where:

$x_n^{[\mathcal{a}]}, x_n^{[\mathcal{b}]}$ ≡ individual-level characteristics of individual *n* that specific to the choice models for class $\mathcal{a}$ and class $\mathcal{b}$ (may be the same as those in the class-membership model)

x_n^* ≡ individual-level characteristics of individual *n* that are shared between both class $\mathcal{a}$ and class $\mathcal{b}$ (may be the same as those in the class-membership model)

$\beta^{[\mathcal{a}]}, \beta^{[\mathcal{b}]}$ ≡ model parameters specific to class $\mathcal{a}$ and class $\mathcal{b}$

β^* ≡ model parameters shared by both class $\mathcal{a}$ and class $\mathcal{b}$

β ∈ $\{\beta^{[\mathcal{a}]}, \beta^{[\mathcal{b}]}, \beta^*\}$

$\varepsilon_n^{[\mathcal{a}]}, \varepsilon_n^{[\mathcal{b}]}$ ≡ unobserved effects on individual *n* for classes $\mathcal{a}$ and $\mathcal{b}$, distributed IID Logistic(0,1)

The systematic portion of the utility can be separated from the random portion as follows:

$$\mathcal{V}_n^{[a]} = \beta^{[a]} x_n^{[a]} + \beta^* x_n^* \tag{131}$$

$$\mathcal{V}_n^{[b]} = \beta^{[b]} x_n^{[b]} + \beta^* x_n^* \tag{132}$$

Additionally, since this is an ordered logit model, thresholds are used to denote the appropriate response depending on the individual's utility. For an individual in class a, the decision process is represented as follows:

$$\begin{aligned} y_n = 0 \text{ if } &\quad \psi_1 \geq U_n^{[a]} \\ y_n = 1 \text{ if } &\quad \psi_2 \geq U_n^{[a]} > \psi_1 \\ y_n = 2 \text{ if } &\quad \psi_3 \geq U_n^{[a]} > \psi_2 \\ y_n = 3 \text{ if } &\quad U_n^{[a]} > \psi_3 \end{aligned} \tag{133}$$

where:

$\psi_1, \psi_2, \psi_3 \quad \equiv$ response thresholds

Thus, the probability of observing a choice $y_n = \{0,1,2,3\}$ for individual *n* in class a is as follows:

$$\begin{aligned} P_{n0}^{[a]} &= P_n(y_n = 0 | c_n = a) = \frac{\exp(\psi_1 - \beta^{[a]} x_n + \beta^* x_n)}{1 + \exp(\psi_1 - \beta^{[a]} x_n + \beta^* x_n)} \\ P_{n1}^{[a]} &= \frac{\exp\left(\psi_2 - \beta^{[a]} x_n + \beta^* x_n\right)}{1 + \exp(\psi_2 - \beta^{[a]} x_n + \beta^* x_n)} - \frac{\exp\left(\psi_1 - \beta^{[a]} x_n + \beta^* x_n\right)}{1 + \exp(\psi_1 - \beta^{[a]} x_n + \beta^* x_n)} \\ P_{n2}^{[a]} &= \frac{\exp\left(\psi_3 - \beta^{[a]} x_n + \beta^* x_n\right)}{1 + \exp(\psi_3 - \beta^{[a]} x_n + \beta^* x_n)} - \frac{\exp\left(\psi_2 - \beta^{[a]} x_n + \beta^* x_n\right)}{1 + \exp(\psi_2 - \beta^{[a]} x_n + \beta^* x_n)} \\ P_{n3}^{[a]} &= 1 - \frac{\exp\left(\psi_3 - \beta^{[a]} x_n + \beta^* x_n\right)}{1 + \exp(\psi_3 - \beta^{[a]} x_n + \beta^* x_n)} \end{aligned} \tag{134}$$

The probabilities for class b follow similarly. Taken together, the probability of observing a choice y_n for individual *n* is as follows:

$$
\begin{aligned}
P_{n0}^{[a]} &= \pi_n^{[a]} P_{n0}^{[a]} + \pi_n^{[b]} P_{n0}^{[b]} \\
P_{n1}^{[a]} &= \pi_n^{[a]} P_{n1}^{[a]} + \pi_n^{[b]} P_{n1}^{[b]} \\
P_{n2}^{[a]} &= \pi_n^{[a]} P_{n2}^{[a]} + \pi_n^{[b]} P_{n2}^{[b]} \\
P_{n3}^{[a]} &= \pi_n^{[a]} P_{n3}^{[a]} + \pi_n^{[b]} P_{n3}^{[b]}
\end{aligned}
\tag{135}
$$

E.2 Likelihood Function

Using equation (135) and substituting values from equations (127), (131), (132), and (134), the likelihood of an observation for individual *n* can be written as follows:

$$
\begin{aligned}
\mathcal{L}_{ni} &= \mathcal{L}_{ni}(\alpha,\beta,\delta;y_n) = \prod_{i\in J}(P_{ni})^{y_{ni}} \\
\mathcal{L}_{ni} &= \pi_n^{[a]} \cdot \prod_{i\in J}\left(P_{ni}^{[a]}\right)^{y_{ni}} + \pi_n^{[b]} \cdot \prod_{i\in J}\left(P_{ni}^{[b]}\right)^{y_{ni}} \\
&= \frac{\exp(\mathcal{F}_n)}{1+\exp(\mathcal{F}_n)}\left[\left(\frac{\exp\left(\psi_1-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_1-\mathcal{V}_n^{[a]}\right)}\right)^{y_{n0}}\right. \\
&\qquad \cdot\left(\frac{\exp\left(\psi_2-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_2-\mathcal{V}_n^{[a]}\right)}-\frac{\exp\left(\psi_1-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_1-\mathcal{V}_n^{[a]}\right)}\right)^{y_{n1}} \\
&\qquad \cdot\left(\frac{\exp\left(\psi_3-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_3-\mathcal{V}_n^{[a]}\right)}-\frac{\exp\left(\psi_2-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_2-\mathcal{V}_n^{[a]}\right)}\right)^{y_{n2}} \\
&\qquad \left.\cdot\left(1-\frac{\exp\left(\psi_3-\mathcal{V}_n^{[a]}\right)}{1+\exp\left(\psi_3-\mathcal{V}_n^{[a]}\right)}\right)^{y_{n3}}\right] \\
&+\frac{1}{1+\exp(\mathcal{F}_n)}\left[\left(\frac{\exp\left(\psi_1-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_1-\mathcal{V}_n^{[b]}\right)}\right)^{y_{n0}}\right. \\
&\qquad \cdot\left(\frac{\exp\left(\psi_2-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_2-\mathcal{V}_n^{[b]}\right)}-\frac{\exp\left(\psi_1-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_1-\mathcal{V}_n^{[b]}\right)}\right)^{y_{n1}} \\
&\qquad \cdot\left(\frac{\exp\left(\psi_3-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_3-\mathcal{V}_n^{[b]}\right)}-\frac{\exp\left(\psi_2-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_2-\mathcal{V}_n^{[b]}\right)}\right)^{y_{n2}} \\
&\qquad \left.\cdot\left(1-\frac{\exp\left(\psi_3-\mathcal{V}_n^{[b]}\right)}{1+\exp\left(\psi_3-\mathcal{V}_n^{[b]}\right)}\right)^{y_{n3}}\right]
\end{aligned}
\tag{136}
$$

The likelihood and log-likelihood for a sample drawn randomly from a population under a simple random sample for a set of parameters are given as follows:

$$\mathcal{L} = \mathcal{L}(\alpha, \beta, \delta, \psi; y_n) = \prod_{n \in N} \mathcal{L}_{ni}$$
$$\mathcal{LL} = \mathcal{LL}(\alpha, \beta, \delta, \psi; y_n) = \sum_{n \in N} \log(\mathcal{L}_{ni}) \tag{137}$$

E.3 Case Study: Bicycle Ownership in the United States

Results from the class membership model are given in Table 35. Results from the ordered choice model are given in Table 36.

Table 35. Class Membership Model Estimation Results for Informational Conformity Latent Class Ordered Choice Model

Parameter	Informational Conformity LC
Class Membership Model:	
Class Constant	-2.74*
Mean MSA Bicycle Ownership~	1.97*
Less than HS Diploma or GED	-0.65*
Associate Degree	0.40*
Bachelor Degree or Higher	0.60*
African-American or Black	-0.72*
Asian-American or Asian	-1.22*
Native American/Pacific Islander	-0.20
Hispanic	-0.48*
Other Race, Non-Hispanic	0.05
Vehicles per Person in HH	0.77*
HH with No Vehicles	0.05
New England Census Division	-0.19
Middle Atlantic Census Division	0.19**
South Atlantic Census Division	-0.12
East North Central Division	0.15
West North Central Division	-0.22
East South Central Division	-0.40
West South Central Division	-0.37*
Mountain Census Division	0.03
HH Located in Hawaii	-0.02

Note: * denotes estimate p-value ≤ 0.05.
** denotes estimate p-value > 0.10 and < 0.05.
~ denotes estimate is the natural exponential function of the model estimated value

Table 36. Choice Model Estimation Results for Ordered Choice of Bike Ownership

Parameter	Informational Conformity LC	
	More Informed Class	Less Informed Class
Bike Ownership Choice Model: Threshold 0 | 1	0.00+	
Threshold 1 | 2	1.38*	
Threshold 2 | 3+	3.56*	
Constant	0.72*	-0.94*
Mean MSA Bicycle Ownership		
Mean MSA Bike Own Residual	0.66*	0.57**
Number of Adults (aged 18-54)	0.88*	0.29*
Number of Women (aged 18-54)	-0.59*	-0.24*
Number of Adults (aged 55+)	-0.26*	-0.54*
Number of Women (aged 55+)	-0.76*	-0.56*
Number of Children (aged 6-17)	0.96*	0.64*
Rent Home	-0.43*	-0.34*
Duplex	-0.11	-0.24**
Townhouse / Rowhouse	-0.60*	-0.27*
Apartment	-0.64*	-0.50*
Mobile Home	-0.40*	-0.88*
Single Person HH	-0.90*	0.19
Low Income HH (< $25k)	-0.53*	-0.53*
High Income HH (> $75k)	0.37*	0.51*
HH Income Unknown	0.02	-0.48*
Model Statistics: Log-Likelihood	Estimated w/ Bayesian Inference	
AIC	Estimated w/ Bayesian Inference	
BIC	Estimated w/ Bayesian Inference	
Number of Parameters	55 (34 choice + 21 class model)	

Note: * denotes estimate p-value ≤ 0.05.
** denotes estimate p-value > 0.10 and < 0.05.
+ denotes that this parameter was normalized to the given value
Blank cells denote parameters that were not included in that specific model.

Appendix F: A Choice Model with Heterogeneous Social Influence Processes

This appendix describes a formulation of a social influence choice model with heterogeneous social influence processes. The choice model uses a latent class structure to allow for to for varying social influence processes among different individuals in the population. A case study is planned for this section involving mode choice with the four conformity processes of automatic activation, majority influence, reactance, and deindividuation effects. As stated in Chapter 7, many different social influence mechanisms and motivations may be involved in an individual's decision process. The current state-of-the-art in travel behavior assumes homogeneous social influence among the population. Kuwano et al. (2012) develop a model where part of the population is affected by conformity while the other part does not. The example model developed in this chapter takes that a step forward by assuming multiple conformity motivations persist in the population which can add additional behavioral realism.

F.1 General Model Formulation

To model heterogeneous social influence processes, a latent class approach will be used. Each class will have a corresponding utility specification that is tailored to a different social influence motivation.

Begin by assuming a population *N* of decision makers where individuals are connected in social networks *G*. Each individual is faced with a choice task where the individual must choose between a set of mutually exclusive alternatives, i.e. the choice set *J*. In this population, individuals may be influenced by different social influence processes. These processes are modeled through a set of classes C and corresponding payoff specifications $\mathcal{P}^{[C]}$. For any individual *n* in the population with a social influence process *c*, the payoff $\mathcal{P}_{ni}^{[c]}$ he obtains from choosing alternative *i* takes the form:

$$\mathcal{P}_{ni}^{[c]} = \beta_i^{[c]} x_{ni} + \psi_i^{[c]} s_{ni}^{[c]} \left(G_n^{[c]}(w), m_{ni}(N), m_{ni}^*(N) \right) + \varepsilon_{ni}^{[c]} \tag{138}$$

where:

x_{ni} ≡ individual-level characteristics of individual *n* for alternative *i*

$s_{ni}^{[c]}(\cdot)$ ≡ social influence mechanism for individual *n* under social influence process *c* for alternative *i*

$G_n^{[c]}(w)$ ≡ individual *n*'s social contacts relevant for social influence mechanism *c* and the strength of these relationships with weighting function *w*

$m_{ni}(N)$ ≡ exogenous social influence sources of the population on individual *n* for alternative *i*

$m_{ni}^*(N)$ ≡ endogenous social influence sources of the population on individual *n* for alternative *i*

ε_{ni} ≡ unobserved effects on individual *n* for alternative *i* (includes correlated individual-level effects and alternative-specific unobservables)

$\beta_i^{[c]}, \psi_i^{[c]}$ ≡ model parameters (can be alternative-specific and class-specific)

Since there are varying social influence processes used by different segments of the population, the selection of the method of influence likely is unobserved by the modeler and can be modeled as a latent process. This specification will use a simplified structure where the probability of being influenced by a particular social influence process is homogeneous across the population. Thus, the probability of being in class *c* is given by:

$$P_n(C_n = c) = \pi^{[c]}, \qquad \forall n \in N \tag{139}$$

To satisfy the laws of probabilities, the following constraints must be imposed:

$$\sum_{c \in C} \pi^{[c]} = 1 \quad and \quad 0 \leq \pi^{[c]} \leq 1 \;\; \forall c \in C \tag{140}$$

Thus, the probability of observing a choice $y_{ni} = 1$ for individual *n* given a decision rule *d* is:

$$P_{ni} = \sum_{c \in C} P_n \left\{ d^{[c]} \left(\mathcal{P}_{ni}^{[c]} \right) \rightarrow (y_{ni} = 1) \right\} \tag{141}$$

where:

$d^{[c]} \left(\mathcal{P}_{ni}^{[c]} \right) \rightarrow (y_{ni} = 1) \equiv$ the decision rule for social influence process (class) *c* that transforms the payoff $\mathcal{P}_{ni}^{[c]}$ into a decision by individual *n* to choose alternative *i*

F.2 Example Model Formulation

In this model formulation, four different conformity mechanisms will be considered:

A. Automatic Activation.

B. Majority Influence

C. Reactance

D. Deindividuation Effects

In automatic activation, it is assumed that an individual just follows the direction of others by imitating behavior without outside influence. This is modeled through the formulation for class A given as follows:

$$\mathcal{U}_{ni}^{[A]} = \beta_i^{[A]} + \delta^{[A]}\bar{y}_{g(n)} + \varepsilon_{ni}^{[A]} \tag{142}$$

In majority influence, individuals' behaviors are influenced depending on how many other social contacts are exhibiting said behavior. This influence is proportional to the share of social contacts exhibiting the behavior. This mechanism is formulation in class B and is given as follows:

$$\mathcal{U}_{ni}^{[B]} = \beta^{[B]}X + \exp(\delta^{[B]})\bar{y}_{g(n)} + \varepsilon_{ni}^{[B]} \tag{143}$$

In reactance, some individuals in the population prefer an alternative when fewer people choose that alternative. This is similar to a counter-culture movement and parallels such reasoning as "I like it because no one else does" or "I don't like that because it is too popular." This mechanism is formulated in class C and is given as follows:

$$\mathcal{U}_{ni}^{[C]} = \beta^{[C]}X + \exp(\delta^{[C]})(1 - \bar{y}_{g(n)}) + \varepsilon_{ni}^{[C]} \tag{144}$$

In deindividuation effects, individuals may conform to local norms as defined by a group identity (Cialdini and Goldstein 2004). In this class, this is modeled through a social norm that is a function of other's behavior. Once a majority of the group's members perform an action, the norm begins to impact an individual's utility through a social influence term. This mechanism is formulated in class D and is given as follows:

$$\mathcal{U}_{ni}^{[D]} = \begin{cases} \beta^{[D]}X + \exp(\delta^{[D]}) + \varepsilon_{ni}^{[D]}, & if\ \bar{y}_{g(n)} \geq 0.50 \\ \beta^{[D]}X + \varepsilon_{ni}^{[D]}, & otherwise \end{cases} \tag{145}$$

Assuming utility maximization for all classes, the probability of observing a choice $y_{ni} = 1$ when an individual follows the social influence mechanism in class $\{A, B, C, D\}$ is as follows:

$$P_{ni}\left(\beta^{[A]}, \delta^{[A]}, \mathcal{U}_{max}\right) = Prob(\mathcal{U}_{ni}^{[A]} \geq \mathcal{U}_{nj}^{[A]}, \forall j \in J) \tag{146}$$

$$P_{ni}\left(\beta^{[B]}, \delta^{[B]}, \mathcal{U}_{max}\right) = Prob(\mathcal{U}_{ni}^{[B]} \geq \mathcal{U}_{nj}^{[B]}, \forall j \in J) \tag{147}$$

$$P_{ni}\left(\beta^{[C]}, \delta^{[C]}, \mathcal{U}_{max}\right) = Prob(\mathcal{U}_{ni}^{[C]} \geq \mathcal{U}_{nj}^{[C]}, \forall j \in J) \tag{148}$$

$$P_{ni}\left(\beta^{[D]}, \delta^{[D]}, \mathcal{U}_{max}\right) = Prob(\mathcal{U}_{ni}^{[D]} \geq \mathcal{U}_{nj}^{[D]}, \forall j \in J) \tag{149}$$

Thus, the probability of observing a choice $y_{ni} = 1$ for individual *n* is given as:

$$\begin{gathered} P_{ni} = \sum_{r \in \{A,B,C,D\}} \pi_n^{[r]} P_{ni}\left(\beta^{[r]}, \delta^{[r]}, \mathcal{U}_{max}\right), \\ where \sum_{r \in \{A,B,C,D\}} \pi_n^{[r]} = 1 \ \& \ 0 \leq \pi_n^{[r]} \leq 1 \end{gathered} \tag{150}$$

Appendix G: MLE Estimates from the Informational Conformity Case Study

The following tables show the results of maximum likelihood estimation of the informational conformity model from section 9.5. Latent class discrete choice models often have difficulties in maximum likelihood estimation due to the presence of multiple optima and areas along the likelihood surface which are nearly flat. Because of difficulty with inverting the Hessian matrix and multiple optima, a Bayesian inference procedure was used to obtain estimates of the standard errors. Then, using the Bayesian estimates as starting values, the model was estimated again using MLE in order to show that the estimates were similar and to provide a way to compare the models between the informational conformity, direct-benefit conformity, and non-social models shown in Table 30 and Table 31 by using log-likelihood, AIC, and BIC measures. The maximum likelihood estimation of the informational conformity model is shown below:

Table 37. MLE for Informational Conformity Model for Binary Choice of Bike Ownership

Parameter	Informational Conformity LC	
Class Membership Model: Class Constant	-2.75	
Mean MSA Bicycle Ownership˜	2.10	
Less than HS Diploma or GED	-0.56	
Associate Degree	0.40	
Bachelor Degree or Higher	0.53	
African-American or Black	-0.60	
Asian-American or Asian	-1.16	
Native American/Pacific Islander	-0.33	
Hispanic	-0.34	
Other Race, Non-Hispanic	-0.03	
Vehicles per Person in HH	0.60	
HH with No Vehicles	0.28	
New England Census Division	-0.31	
Middle Atlantic Census Division	0.09	
South Atlantic Census Division	-0.23	
East North Central Division	-0.02	
West North Central Division	-0.42	
East South Central Division	-0.45	
West South Central Division	-0.35	
Mountain Census Division	-0.01	
HH Located in Hawaii	0.00	
Bike Ownership Choice Model: Constant	1.25	-1.01
Mean MSA Bike Own Residual	0.26	
Number of Adults (aged 18-54)	2.39	0.16
Number of Women (aged 18-54)	-1.21	-0.12
Number of Adults (aged 55+)	0.06	-0.17
Number of Women (aged 55+)	-0.67	-0.48
Number of Children (aged 6-17)	0.74	
Rent Home	-0.47	-0.61
Duplex	0.05	-0.35
Townhouse / Rowhouse	-0.38	-0.44
Apartment	-0.57	-0.84
Mobile Home	0.35	-0.96
Single Person HH	-0.78	-0.93
Low Income HH (< $25k)	-0.67	-0.68
High Income HH (> $75k)	0.22	0.51
HH Income Unknown	-0.43	-0.27
Model Statistics: Log-Likelihood	-14709	
AIC	29519	
BIC	29935	
Number of Parameters	51 (30 choice + 21 class model)	

Bibliography

Abdelal, R., Herrera, Y. M., Johnston, A. I., & McDermott, R. (2009). Introduction. *Measuring Identity: A Guide for Social Scientists*, 1-13.

Abou-Zeid, M., & Ben-Akiva, M. (2011). The effect of social comparisons on commute well-being. *Transportation Research Part A: Policy and Practice*, 45(4), 345-361.

Adjemian, M. K., Lin, C. Y., & Williams, J. (2010). Estimating spatial interdependence in automobile type choice with survey data. *Transportation Research Part A: Policy and Practice*, 44(9), 661-675.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3), 715-753.

Akerlof, G. A., & Kranton, R. E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature*, 40(4), 1167-1201.

Akerlof, G. A., & Kranton, R. E. (2010). *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-being*. Princeton, NJ: Princeton University Press.

An, W. (2011). Models and methods to identify peer effects. In Scott, J., & Carrington, P. J. (Eds.) *The SAGE handbook of social network analysis (pp.* 515-532). London: SAGE.

Anderson, L. R., & Holt, C. A. (1997). Information cascades in the laboratory. *The American economic review*, 847-862.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6), 1086-1120.

Arentze, T. A., & Timmermans, H. J. P. (2008). Social networks, social interactions, and activity-travel behaviour: A framework for microsimulation. *Environment and Planning B: Planning and Design*, 35(6), 1012-1027.

Axhausen, K. W. (2008). Social networks, mobility biographies, and travel: survey challenges. *Environment and Planning B: Planning and Design*, 35(6), 981-996.

Axsen, J., & Kurani, K. S. (2012). Interpersonal influence within car buyers' social networks: Applying five perspectives to plug-in hybrid vehicle drivers. *Environment and Planning A*, 44(5), 1047-1065.

Axsen, J., Orlebar, C., & Skippon, S. (2013). Social influence and consumer preference formation for pro-environmental technology: The case of a UK workplace electric-vehicle study. *Ecological Economics*, 95, 96-107.

Baltas, G., & Saridakis, C. (2013). An empirical investigation of the impact of behavioural and psychographic consumer characteristics on car preferences: An integrated model of car type choice. *Transportation Research Part A: Policy and Practice*, 54, 92-110.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Bartle, C., Avineri, E., & Chatterjee, K. (2013). Online information-sharing: A qualitative analysis of community, trust and social influence amongst commuter cyclists in the UK. *Transportation Research Part F: Traffic Psychology and Behaviour*, 16, 60-72.

Bayer, P., & Timmins, C. (2007). Estimating Equilibrium Models of Sorting Across Locations. *The Economic Journal*. 117(518), 353–374.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 63(4), 841-890.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: MIT press.

Bhat, C. R. (1997). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science*, *31*(1), 34-48.

Bhat, C. R., & Gossen, R. (2004). A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B: Methodological*, *38*(9), 767-787.

Bhat, C., & Lockwood, A. (2004). On distinguishing between physically active and physically passive episodes and between travel and activity episodes: an analysis of weekend recreational participation in the San Francisco Bay area. *Transportation Research Part A: Policy and Practice*, *38*(8), 573-592.

Bhat, C. R., & Srinivasan, S. (2005). A multidimensional mixed ordered-response model for analyzing weekend activity participation. *Transportation Research Part B: Methodological*, *39*(3), 255-278.

Bhat, C. R., Srinivasan, S., & Sen, S. (2006). A joint model for the perfect and imperfect substitute goods case: application to activity time-use decisions. *Transportation Research Part B: Methodological*, *40*(10), 827-850.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 992-1026.

Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*. New York: Free Press.

Blume, L. E., Brock, W. A., Durlauf, S. N., & Ioannides, Y. M. (2011). Identification of social interactions. In J. Benhabib, M. O. Jackson & A. Bisin (Eds.), *Handbook of social economics* (Vol. 1B, pp. 853-964). North Holland, The Netherlands: Elsevier.

Brock, W. A., & Durlauf, S. N. (2001). Discrete choice with social interactions. *Review of Economic Studies*, 68(2), 235-260.

Brock, W. A., & Durlauf, S. N. (2002). A multinomial-choice model of neighborhood effects. *American Economic Review*, 92(2), 298-303.

Blume, L., & Durlauf, S. (2003). Equilibrium concepts for social interaction models. *International Game Theory Review*, *5*(3), 193-209.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. Los Angeles: SAGE.

Bourdeaudhuij, I. D., Teixeira, P. J., Cardon, G., & Deforche, B. (2005). Environmental and psychosocial correlates of physical activity in Portuguese and Belgian adults. *Public Health Nutrition*, *8*(07), 886-895.

Brock, W. A., & Durlauf, S. N. (2006). A multinomial choice model with social interactions. In Blume, L. E., Durlauf, S. N. (Eds.), *The economy as an evolving complex system III*. New York: Oxford University Press.

Brock, W. A., & Durlauf, S. N. (2007). Identification of binary choice models with social interactions. *Journal of Econometrics*, 140(1), 52-75.

Carrasco, J. A., & Cid-Aguayo, B. (2012). Network capital, social networks, and travel: An empirical illustration from Concepción, Chile. *Environment and Planning A*, 44(5), 1066-1084.

Carrasco, J. A., Hogan, B., Wellman, B., & Miller, E. J. (2008). Collecting social network data to study social activity-travel behavior: An egocentric approach. *Environment and Planning B: Planning and Design*, 35(6), 961-980.

Carrasco, J. A., & Miller, E. J. (2006). Exploring the propensity to perform social activities: a social network approach. *Transportation*, 33(5), 463-480.

Carrasco, J. A., & Miller, E. J. (2009). The social dimension in action: A multilevel, personal networks model of social activity frequency between individuals. *Transportation Research Part A: Policy and Practice*, 43(1), 90-104.

Castiglione, J., Bradley, M., & Gliebe, J. (2015). *Activity-Based Travel Demand Models: A Primer*. Washington: Transportation Research Board.

Çelen, B., & Kariv, S. (2004). Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review*, 484-498.

Chávez, O., Carrasco, J. A., & Tudela, A. (2015). Social Activity-Travel Dynamics with Core Contacts: Evidence from Two-Wave Data Collection Effort. *Transportation Research Board 94th Annual Meeting*. Washington: Transportation Research Board. Retrieved from http://amonline.trb.org/

Chen, Y. (2012). *Neighborhood design and the energy efficiency of urban lifestyle in China : Treating residence and Mobility as lifestyle bundle*, Ph.D. Dissertation. Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA.

Choi, S., Gale, D., & Kariv, S. (2005). Behavioral aspects of learning in social networks: an experimental study. *Advances in Applied Microeconomics*, *13*, 25-61.

Chorus, C., Arentze, T., & Timmermans, H. (2008). A random regret minimization model of travel choice. *Transportation Research Part B*, 42(1), 1-18.

Chung, H., Flaherty, B. P., & Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 723-743.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Reviews of Psychology*, *55*, 591-621.

Cipriani, M., & Guarino, A. (2005). Herd behavior in a laboratory financial market. *American Economic Review*, 95(5), 1427-1443.

Clifton, K., & Handy, S. (2003). Qualitative methods in travel behaviour research. *Transport Survey Quality and Innovation*, 283-302.

Davis, J. B. (2010). *Individuals and identity in economics*. Cambridge: Cambridge University Press.

De Palma, A., Picard, N., & Inoa, I. (2014). Discrete choice decision-making with multiple decision-makers within the household. In Hess, H., & Daly, A. (Eds.) *Handbook of choice modelling* (pp. 363-382). Glos, UK: Edward Elgar.

Dickinson, J. E., Kingham, S., Copsey, S., & Hougie, D. J. P. (2003). Employer travel plans, cycling and gender: Will travel plan measures improve the outlook for cycling to work in the UK?. *Transportation Research Part D: Transport and Environment*, 8(1), 53-67.

Dill, J., & Voros, K. (2007). Factors affecting bicycling demand: Initial survey findings from the Portland, Oregon region. *Transportation Research Record: Journal of the Transportation Research Board*, 2031, 9-17.

Deutsch, K., & Goulias, K. (2013). Decision makers and socializers, social networks and the role of individuals as participants. *Transportation*, 40(4), 755-771.

Dugundji, E. R. (2003). Discrete choice on networks: Nested logit. In *Graduate Workshop in Computational Economics*, Santa Fe, NM. Retrieved from http://zia.hss.cmu.edu/econ/projects03.htm

Dugundji, E. R. (2012). Socio-dynamic discrete choice: Equilibrium behavior of the nested logit model with social interactions. In *Interdisciplinary Workshop on Information and Decision in Social Networks*. Cambridge, MA: Massachusetts Institute of Technology.

Dugundji, E. R., & Gulyás, L. (2003). An exploration of the role of global versus local and social versus spatial networks in transportation mode choice behavior in the Netherlands. In *Challenges in Social Simulation*. University of Chicago, Argonne National Laboratory, Chicago, 437-454.

Dugundji, E. R., & Gulyás, L. (2008). Sociodynamic discrete choice on networks in space: Impacts of agent heterogeneity on emergent outcomes. *Environment and Planning B: Planning and Design*, 35(6), 1028-1054.

Dugundji, E. R., & Gulyás, L. (2012a). Sociodynamic discrete choice applied to intercity travel demand: Issues in estimation. In *Federated Conference on Computer Science and Information Systems* (FedCSIS), Wroclaw. Available in IEEE Xplore.

Dugundji, E. R., & Gulyás, L. (2012b). Sociodynamic discrete choice on networks in space: Role of utility parameters and connectivity in emergent outcomes. *Procedia Computer Science*, 10, 827-832.

Dugundji, E. R., & Gulyás, L. (2013). Structure and emergence in a nested logit model with social and spatial interactions. *Computational and Mathematical Organization Theory,* dx.doi.org/10.1007/s10588-013-9157-y.

Dugundji, E. R., Páez, A., & Arentze, T. A. (2008). Social networks, choices, mobility and travel. *Environment and Planning B: Planning and Design*, 35(6), 956-960.

Dugundji, E. R., Páez, A., Arentze, T. A., & Walker, J. L. (2011). Transportation and social interactions. *Transportation Research Part A: Policy and Practice*, 45 (4), 239-247.

Dugundji, E. R., Poorthuis, A., & Van Meeteren, M. (2011). Modeling User Behavior in Adoption and Diffusion of Twitter Clients. In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom)* (pp. 1372-1379). Boston: IEEE.

Dugundji, E. R., Poorthuis, A., & Van Meeteren, M. (2013). Capturing unobserved correlated effects in diffusion in large virtual networks. In Ghoshal, G., Poncela-Casasnovas, J., Tolksdorf, R. (Eds.), *Complex networks IV, studies in computational intelligence, Vol. 476* (pp 81-92), Berlin Heidelberg: Springer-Verlag.

Dugundji, E. R., Scott, D. M., Carrasco, J. A., & Páez, A. (2012). Urban mobility and social-spatial contact – introduction. *Environment and Planning A*, 44 (5), 1011-1015.

Dugundji, E. R., & Walker, J. L. (2005). Discrete choice with social and spatial network interdependencies: An empirical example using mixed generalized extreme value models with field and panel effects. *Transportation Research Record: Journal of the Transportation Research Board,* 1921, 70-78.

Durlauf, S. N., & Ioannides, Y. M. (2010). Social interactions. *Annual Review of Economics,* 2, 451-478.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, *5*, 17-61.

Federal Highway Administration. (2005). 2001 NHTS (4.0) [Data files and codebook]. Retrieved from http://nhts.ornl.gov/download.shtml

Federal Highway Administration (2010). The national bicycling and walking study: 15-year status report. Retrieved from *American Trails* website: http://www.americantrails.org/resources/trans/FHWA-National-Bicycling-Walking%20Study-2010.html

Fellows, I., & Handcock, M. S. (2012). Exponential-family random network models. Working paper. Retrieved from *Cornell University Library* website: http://arxiv.org/abs/1208.0121

Field, J. (2003). *Social capital*. London: Routledge.

Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In Anselin, L., Florax, R. J. G. M, & Rey S. J. (Eds.) *Advances in spatial econometrics* (pp. 145-168). Berlin Heidelberg: Springer.

Fowler-Brown, A. G., Ngo, L. H., Phillips, R. S., & Wee, C. C. (2009). Adolescent obesity and future college degree attainment. *Obesity,* 18(6), 1235-1241.

Freese, J. & Long J.S. (2006). *Regression Models for Categorical Dependent Variables Using Stata.* College Station, TX: Stata Press.

Frei, A. (2012). *Networks, geographies and travel: Travel between infrastructure and social structure*, Ph.D. Dissertation. Institute for Transport Planning and Systems, ETH Zürich, Zürich.

Frei A., & Axhausen K. W. (2007). Size and structure of social network geographies. Arbeitsberichte Verkehrs-und Raumplanung, 439, IVT, ETH Zürich, Zürich.

Fukuda, D., & Morichi, S. (2007). Incorporating aggregate behaviour in an individual's discrete choice: An application to analyzing illegal bicycle parking behaviour. *Transportation Research Part A: Policy and Practice*, 41(4), 313-325.

Gaker, D., Zheng, Y., & Walker, J. L. (2010). Experimental economics in transportation. *Transportation Research Record: Journal of the Transportation Research Board*, 2156, 47-55.

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, *56*(4), 1055-1067.

Garrett, E. S., Eaton, W. W., & Zeger, S. (2002). Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in medicine*, *21*(9), 1289-1307.

Goetzke, F. (2008). Network effects in public transit use: evidence from a spatially autoregressive mode choice model for New York. *Urban Studies*, 45(2), 407-417.

Goetzke, F., & Andrade, P. M. (2010). Walkability as a summary measure in a spatially autoregressive mode choice model: an instrumental variable approach. In Páez, A., Buliung, R. N., Le Gallo, J., Dall'erba, S. (Eds.), *Progress in spatial analysis: Methods and applications*. Springer-Verlag, Berlin Heidelberg, pp. 217-229.

Goetzke, F., & Rave, T. (2011). Bicycle use in Germany: explaining differences between municipalities with social network effects. *Urban Studies*, 48(2), 427-437.

Goetzke, F., & Weinberger, R. (2012). Separating contextual from endogenous effects in automobile ownership models. *Environment and Planning A*, 44(5), 1032.

Grannis, R. (1998). The Importance of Trivial Streets: Residential Streets and Residential Segregation. *American Journal of Sociology*, 103(6), 1530-1564.

Grinblatt, M., Keloharju, M., & Ikäheimo, S. (2008). Social influence and consumption: Evidence from the automobile purchases of neighbors. *The Review of Economics and Statistics*, *90*(4), 735-753.

Guevara-Cue, C. A. (2010). *Endogeneity and sampling of alternatives in spatial choice models*, Ph.D. Dissertation. Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA.

Gulyás, L. & Dugundji, E. R. (2006). Emergent opinion dynamics on endogenous networks. In Trajkovski, G.P., Collins, S.G. (Eds.), *Interaction and emergent phenomena in societies of agents*, Technical Report FS-06-05. AAAI Press, Menlo Park, CA.

Habib, K. M., Carrasco, J. A., & Miller, E. J. (2008). Social context of activity scheduling: discrete-continuous model of relationship between "with whom" and episode start time and duration. *Transportation Research Record: Journal of the Transportation Research Board*, 2076, 81-87.

Habib, K. M. & Carrasco, J. A. (2011). Investigating the role of social networks in start time and duration of activities. *Transportation Research Record: Journal of the Transportation Research Board*, 2230, 1-8.

Hällsten, M., Edling, C., & Rydgren, J. (2015). The effects of specific occupations in position generator measures of social capital. Social Networks, 40, 55-63.

Hampton, K. N. (2011). Comparing bonding and bridging ties for democratic engagement: Everyday use of communication technologies within social networks for civic and civil behaviors. *Information, Communication & Society*, *14*(4), 510-528.

Hampton, K. N., Sessions, L. F., Her, E. J., & Rainie, L. (2009). "Social Isolation and New Technology." Retrieved from *Pew Research Centers Internet American Life Project* website: http://www.pewinternet.org/2009/11/04/social-isolation-and-new-technology/.

Handcock, M. S. & Gile, K. J. (2007). Modeling social networks with sampled or missing data. Working paper 75, Center for Statistics and the Social Sciences. Retrieved from *University of Washington* website: http://www.csss.washington.edu/Papers/

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris M. (2003). statnet: Software tools for the Statistical Modeling of Network Data. Retrieved from http://statnetproject.org

Handy, S., Cao, X., & Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. *Transportation Research Part D: Transport and Environment*, *10*(6), 427-444.

Handy, S. L., Xing, Y., & Buehler, T. J. (2010). Factors associated with bicycle ownership and use: a study of six small US cities. *Transportation*, *37*(6), 967-985.

Hahn, J., & Hausman, J. (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review*, 93(2), 118-125.

Haustein, S., Klöckner, C. A., & Blöbaum, A. (2009). Car use of young adults: The role of travel socialization. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(2), 168-178.

Henningsen, A., & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), 443-458.

Hess, S. (2014). Latent class structures: Taste heterogeneity and beyond. In Hess, H., & Daly, A. (Eds.) *Handbook of choice modelling* (pp. 311-329). Glos, UK: Edward Elgar.

Hess, S., Ben-Akiva, M., Gopinath, D., & Walker, J. (2011). Advantages of Latent Class Over Continuous Mixture of Logit Models. Working Paper. Retrieved from http://www.stephanehess.me.uk/papers/Hess_Ben-Akiva_Gopinath_Walker_May_2011.pdf

Hess, S., Stathopoulos, A., & Daly, A. (2012). Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation*, 39(3), 565-591.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, *8*(3), 691-711.

Huck, S., & Oechssler, J. (2000). Informational cascades in the laboratory: Do they occur for the right reasons?. *Journal of Economic Psychology*, *21*(6), 661-671.

Humphries M. D., & Gurney K. (2008) Network 'small-world-ness': A quantitative method for determining canonical network equivalence. PLoS ONE 3(4): e0002051. doi:10.1371/journal.pone.0002051

Hunecke, M., Blöbaum, A., Matthies, E., & Höger, R. (2001). Responsibility and environment ecological norm orientation and external factors in the domain of travel mode choice behavior. *Environment and Behavior*, 33(6), 830-852.

Hung, A. A., & Plott, C. R. (2001). Information cascades: Replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review*, 1508-1520.

Hunter, D. R. (2007). Curved exponential family models for social networks. Social networks, 29(2), 216-230.

Ioannides, Y. M. (2006). Topologies of social interactions. *Economic Theory*, 28(3), 559-584.

Ioannides, Y. M., & Zabel, J. E. (2008). Interactions, neighborhood selection and housing demand. *Journal of Urban Economics*, 63(1), 229-252.

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings.* Oxford University Press, Oxford.

Kim, J. H., & Kim, K. S. (2012). The role of sensation seeking, perceived peer pressure, and harmful alcohol use in riding with an alcohol-impaired driver. *Accident Analysis & Prevention*, *48*, 326-334.

Kowald, M., & Axhausen, K. W. (2012). Focusing on connected personal leisure networks: Selected results from a snowball sample. *Environment and Planning A*, 44(5), 1085-1100.

Kowald, M., & Axhausen, K. W. (2014). Surveying data on connected personal networks. Travel Behaviour and Society, 1(2), 57-68.

Kowald, M., Frei, A., van den Berg, P., Carrasco, J. A., Arentze, T., Axhausen, K. W., & Wellman, B. (2013). Distance patterns of personal networks in four countries: a comparative study. *Journal of Transport Geography*, 31, 236-248.

Krackhardt, D., & Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, 51(2), 123-140.

Kübler, D., & Weizsäcker, G. (2004). Limited depth of reasoning and failure of cascade formation in the laboratory. *The Review of Economic Studies*, *71*(2), 425-441.

Kuwano, M., Chikaraishi, M., & Fujiwara, A. (2013). An Analysis of Personal Mobility Promoting Factors Considering the Social Network in Old Newtown. In *Proceedings of the Eastern Asia Society for Transportation Studies (Vol. 9).*

Kuwano, M., Tsukai, M., & Matsubara, T. (2012). Analysis on promoting factors of electric vehicles with social conformity. Paper presented at *13th International Conference on Travel Behaviour Research* (IATBR) [CD], Toronto.

Kuwano, M., Zhang, J., & Fujiwara, A. (2011). Dynamic discrete choice model for multiple social interactions. *Transportation Research Record: Journal of the Transportation Research Board*, 2231, 68-75.

Larsen, J., Urry, J., & Axhausen, K. W. (2006). *Mobilities, networks, geographies*. Aldershot, England: Ashgate.

Larsen, J., Urry, J., &Axhausen, K. W. (2008). Coordinating face-to-face meetings in mobile network societies. *Information, Communication and Society*, 11(5), 640-658.

Lenk, P. (2014). Bayesian estimation of random utility models. In Hess, H., & Daly, A. (Eds.) *Handbook of choice modelling* (pp. 311-329). Glos, UK: Edward Elgar.

Li, J., & Lee, L. F. (2009). Binary choice under social interactions: an empirical study with and without subjective data on expectations. *Journal of Applied Econometrics*, 24(2), 257-281.

Liggett, Thomas M. (1985). *Interacting Particle Systems*. New York: Springer Verlag.

Liggett, T. M. (1997). Stochastic models of interacting systems. *The Annals of Probability, 25*(1), 1-29.

Lin, N. (1982). Social resources and instrumental action. In P.V. Marsden and N. Lin (Eds.), *Social Structure and Network Analysis* (pp 131-46). Beverly Hills, CA: SAGE.

Lin, N. (2001). Building a network theory of social capital. In Lin, N., Cook, K., Burt, R.S. (Eds.), *Social capital: Theory and research* (pp. 3-29). New York: Aldine de Gruyter.

Lin, N., & Dumin, M. (1986). Access to occupations through social ties. *Social networks*, *8*(4), 365-385.

Lin, N., Fu, Y. C., & Hsung, R. M. (2001). The position generator: Measurement techniques for investigations of social capital. In Lin, N., Cook, K., Burt, R.S. (Eds.), *Social capital: Theory and research* (pp. 57-81). New York: Aldine de Gruyter.

Lin, T., & Wang, D. (2014). Social networks and joint/solo activity-travel behavior. *Transportation Research Part A: Policy and Practice*, *68*, 18-31.

Lovejoy, K., & Handy, S., (2011). Social networks as a source of private-vehicle transportation: The practice of getting rides and borrowing vehicles among Mexican immigrants in California. *Transportation Research Part A: Policy and Practice*, 45(4), 248-257.

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Matous, P., Todo, Y., & Mojo, D. (2013). Boots are made for walking: interactions across physical and social space in infrastructure-poor regions. *Journal of Transport Geography*, 31, 226-235.

Maness, M. (2012). Bicycle ownership in the United States: Empirical analysis of regional differences. *Transportation Research Board 91st Annual Meeting*. Washington: Transportation Research Board. Retrieved from http://amonline.trb.org/

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

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press, Cambridge, MA.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 72(5), 1329-1376.

Martínez-Sánchez, A., Pérez-Pérez, M., Vela-Jiménez, M. J., &de-Luis-Carnicer, P. (2008). Telework adoption, change management, and firm performance. *Journal of Organizational Change Management*, 21(1), 7-31.

Milgram, S. (1967). The small world problem. Psychology today, 2(1), 60-67.

Mote, J. E., & Whitestone, Y. (2011). The social context of informal commuting: Slugs, strangers and structuration. *Transportation Research Part A: Policy and Practice*, 45(4), 258-268.

Nöth, M., & Weber, M. (2003). Information Aggregation with Random Ordering: Cascades and Overconfidence. *The Economic Journal*, *113*(484), 166-189.

Ortuźar, J. de D. & Willumsen, L. (2011). *Modelling transport* (4th ed.). Oxford: Wiley-Blackwell. Páez, A., Scott, D.M., 2004. Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal*, 61(1), 53-67.

Owen, N., De De Bourdeaudhuij, I., Sugiyama, T., Leslie, E., Cerin, E., Van Van Dyck, D., & Bauman, A. (2010). Bicycle use for transport in an Australian and a Belgian city: associations with built-environment attributes. *Journal of Urban Health*, *87*(2), 189-198.

Páez, A., & Scott, D. M. (2007). Social influence on travel behaviour: A simulation example of the decision to telecommute. *Environment and Planning A*, 39(3), 647-665.

Páez, A., Scott, D. M., & Volz, E. (2008). A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design*, 35(6), 1055.

Parkin, J., Ryley, T., & Jones, T. (2007). Barriers to cycling: An exploration of quantitative analyses. *Cycling and Society*, 67-82.

Pike, S. (2014). Travel Mode Choice and Social and Spatial Reference Groups. *Transportation Research Record: Journal of the Transportation Research Board*, 2412, 75-81.

Pike, S. (2015). Endogeneity in Social Influence and Transportation Mode Choice Using Ego Networks. *Transportation Research Board 94th Annual Meeting*. Washington: Transportation Research Board. Retrieved from http://amonline.trb.org/

Pinjari, A. R., Bhat, C. R., & Hensher, D. A. (2009). Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B: Methodological*, *43*(7), 729-748.

Pinjari, A. R., Pendyala, R. M., Bhat, C. R., & Waddell, P. A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, *38*(6), 933-958.

Prell, C. (2012). Social network analysis: History, theory and methodology. Los Angeles: SAGE.

Primerano, F. (2006). Mobility considerations in restricting choice sets in modal choice models. *Transport Engineering in Australia*, *10*(1), 31.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In Liu, L., & Zsu, M. T. (Eds.) (2009). *Encyclopedia of database systems* (pp. 532-538). New York: Springer.

Reicher, S.D., Spears R., & Postmes T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6(1), 161-198

Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, *39*(3), 347-366.

Robins, G. & Daraganova, G. (2013). Social selection, dyadic covariates, and geospatial effects. In D. Lusher, J, Koskinen, & G. Robins (Eds.), *Exponential random graph models for social networks: Theory, methods and applications* (pp. 91-101). Cambridge: Cambridge University Press.

Robins, G., Elliott, P., & Pattison, P. (2001). Network models for social selection processes. *Social networks*, *23*(1), 1-30.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007). Recent developments in exponential random graph (*p**) models for social networks. *Social networks*, *29*(2), 192-215.

Roy, P., Martínez, A. J., Miscione, G., Zuidgeest, M. H. P., & van Maarseveen, M. F. A. M. (2012). Using Social Network Analysis to profile people based on their e-communication and travel balance. *Journal of Transport Geography*, *24*, 111-122.

Ryley, T. (2008). Encouraging Bicycle Use in Residential Neighborhoods: Insight from Edinburgh. Paper presented at the *5th Cycling and Society Symposium*, Bristol, UK.

Sadri, A.M., Lee, S., & Ukkusuri, S. (2015). Exploring the Effects of Social Ties on Joint Trip Frequency Ego-centric Social Network. *Transportation Research Board 94th Annual Meeting*. Washington: Transportation Research Board. Retrieved from http://amonline.trb.org/

Schmid, K., Hewstone, M., & Al Ramiah, A. (2011). Self-categorization and social identification: making sense of us and them. In *Theories in social psychology* (pp. 211-231). Malden, MA: Wiley-Blackwell.

Scott, D.M., Dam, I., Páez, A., Wilton, R.D., 2012. Investigating the effects of social influence on the choice to telework. *Environment and Planning A*, 44(5), 1016-1031.

Sener, I. N., Eluru, N., & Bhat, C. R. (2009). Who are bicyclists? Why and how much are they bicycling?. *Transportation Research Record: Journal of the Transportation Research Board*, *2134*, 63-72.

Sherwin, H., Chatterjee, K., & Jain, J. (2014). An exploration of the importance of social influence in the decision to start bicycling in England. *Transportation Research Part A: Policy and Practice*, *68*, 32-45.

Sidharthan, R., Bhat, C. R., Pendyala, R. M., Goulias, K. G., 2011. Model for children's school travel mode choice: Accounting for effects of spatial and social interaction. *Transportation Research Record: Journal of the Transportation Research Board*, 2213, 78-86.

Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional science and urban economics*, 40(5), 292-298.

Snijders, T. A., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. Sociological methodology, 36(1), 99-153.

Snijders, T. A. B., Steglich, C. E. G., & Schweinberger, M. (2007). Modeling the co-evolution of networks and behavior. In: Van Montfort, K., Oud, H., & Satorra, A. (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 41-71). Mahwah, NJ: Lawrence Erlbaum.

Snijders, T. A., Van de Bunt, G. G., Steglich, C. E., (2010). Introduction to stochastic actor-based models for network dynamics. *Social networks*, 32(1), 44-60.

Snyder, T. D., & Dillow, S. A. (2013). Digest of Education Statistics, 2012. NCES 2014-015. Retrieved from *National Center for Education Statistics* website: http://nces.ed.gov/pubs2014/2014015.pdf

Spiwoks, M., Bizer, K., & Hein, O. (2008). Informational cascades: A mirage?. *Journal of Economic Behavior & Organization*, *67*(1), 193-199.

Stadje, W. (1990). The collector's problem with group drawings. *Advances in Applied Probability*, 22(4) 866-882.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518-529.

Sun, L., Axhausen, K. W., Lee, D. H., & Huang, X. (2013). Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34), 13774-13779.

Tajfel, H. (1978). Interindividual Behaviour and Intergroup Behavior. In H. Tajfel (Ed.), *Differentiation between social groups: Studies in the social psychology of intergroup relations* (pp. 27-60). London: Academic Press.

Tajfel, H., & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations*. Monterey, CA: Brooks-Cole .

Tilahun, N., & Li, M. (2015). The Geography of Close Contacts and Face-to-Face Meetings. *Transportation Research Board 94th Annual Meeting*. Washington. Washington: Transportation Research Board. Retrieved from http://amonline.trb.org/

Train, K. E. (2009). *Discrete choice methods with simulation*. New York: Cambridge University Press.

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

Van den Berg, P., Arentze, T. A., & Timmermans, H. J. P. (2008). Social networks, ICT use and activity-travel patterns: Data collection and first analyses. In *9th International Conference on Design and Decision Support Systems in Architecture and Urban Planning*, Netherlands: DDSS.

Van den Berg, P., Arentze, T. A., & Timmermans, H. J. P. (2010). Factors influencing the planning of social activities. *Transportation Research Record: Journal of the Transportation Research Board*, 2157, 63-70.

Van den Berg, P., Arentze, T. A., & Timmermans, H. J. P. (2012). A latent class accelerated hazard model of social activity duration. *Transportation Research Part A: Policy and Practice*, 46(1), 12-21.

Verbrugge, L. M. (1977). The structure of adult friendship choices. *Social Choices*, 56, 576-597.

Verhaeghe, P. P., Van de Putte, B., & Roose, H. (2013). Reliability of Position Generator Measures across Different Occupational Lists A Parallel Test Experiment. Field Methods, 25(3), 238-261.

Vij, A., Carrel, A., & Walker, J. L. (2013). Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transportation Research Part A: Policy and Practice*, *54*, 164-178.

Walker, J. L., Ehlers, E., Banerjee, I., & Dugundji, E. R. (2011). Correcting for endogeneity in behavioural choice models with social influence variables. *Transportation Research Part A: Policy and Practice*, 45(4), 362-374.

Walker, J. L., & Li, J. (2007). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1), 77-101.

Waller, J. A. (1971). Bicycle ownership, use, and injury patterns among elementary school children. *Pediatrics*, *47*(6), 1042-1050.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440-442.

Weinberger, R., & Goetzke, F., 2010. Unpacking preference: How previous experience affects auto ownership. *Urban Studies*, 47(10), 2111-2128.

Whalen, K. E., Páez, A., Bhat, C. R., Moniruzzaman, M., & Paleti, R. (2012). T-communities and sense of community in a university town: Evidence from a student sample using a spatial ordered-response model. *Urban Studies*, 49(6), 1357-1376.

Wigan, M. R. (1984). Bicycle ownership use and exposure: Participation and activity patterns in Melbourne, Australia. *Transportation Research Part A: General*, *18*(5), 379-398.

Wilton, R. D., Páez, A., & Scott, D. M. (2011). Why do you care what other people think? A qualitative investigation of social influence and telecommuting. *Transportation Research Part A: Policy and Practice*, 45(4), 269-282.

Woittiez, I., & Kapteyn, A. (1998). Social interactions and habit formation in a model of female labour supply. *Journal of Public Economics*, 70(2), 185-205.

Wong, L. H., Pattison, P., & Robins, G. (2006). A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360(1), 99-120.

Wu, L., Zhang, J., & Chikaraishi, M. (2013). Representing the influence of multiple social interactions on monthly tourism participation behavior. *Tourism Management*, *36*, 480-489.

U.S. Census Bureau (n.d.). Census regions and divisions of the United States. Retrieved from *US Census Bureau* website: https://www.census.gov/geo/maps-data/maps/pdfs/reference/us_regdiv.pdf

Yamamoto, T. (2009). Comparative analysis of household car, motorcycle and bicycle ownership between Osaka metropolitan area, Japan and Kuala Lumpur, Malaysia. *Transportation*, *36*(3), 351-366.

Zanella, G. (2007). Discrete Choice with Social Interactions and Endogenous Memberships. *Journal of the European Economic Association*, 5(1), 122-153.

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer Psychology*, 17(1), 3-18.