Summer 2014

Three essays on social networks and the diffusion of innovation models

Tae-Hyung Pyo
University of Iowa

THREE ESSAYS ON SOCIAL NETWORKS AND THE DIFFUSION OF INNOVATION MODELS

by

Tae-Hyung Pyo

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

August 2014

Thesis Supervisors: Professor Thomas S. Gruca
Professor Gary J. Russell

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

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

Tae-Hyung Pyo

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Business Administration at the August 2014 graduation.

Thesis Committee:
Thomas S. Gruca, Thesis Supervisor

Gary J. Russell, Thesis Supervisor

Dhananjay Nayakankuppam

Sheila T Goins

Sanghak Lee

ABSTRACT

The Bass model has been used extensively and globally to forecast the first purchases of new products. It has been named by INFORMS as one of the top 10 most influential papers published in the 50-year history of *Management Science*. Most models for the diffusion of innovation are deeply rooted in the work of Bass (1969). His work provides a framework to model the underlying process of innovation adaption among first-time customers.

Potential customers may be connected to one another in some sort of network. Prior research has shown that the structure of a network affects adoption patterns (Dover et al. 2012; Hill et al. 2006; Katona and Sarvary 2008; Katona et al. 2011; Newman et al. 2006; Shaikh et al. 2010; Van den Bulte and Joshi 2007). One approach to addressing this issue is to incorporate network information into the original Bass model. The focus of this study is to explore how to incorporate network information and other micro-level data into the Bass model.

First, I prove that the Bass model assumes all potential customers are linked to all other customers. Through simulations of individual adoptions and connections among individuals using a Random Network, I show that the estimate of q in the Bass model is biased downward in the original Bass model. I find that biases in the Bass model depend on the structure of the network. I relax the assumption of the fully connected network by proposing a Network-Based Bass model (NBB), which incorporates the network structure into the traditional Bass model. Using the proposed model (NBB), I am able to recover the true parameters.

To test the generalizability and to enhance the applicability of my NBB model, I tested my NBB model on the various network types with sampled data from the population network. I showed that my NBB model is robust across different types of networks, and it is efficient in terms of sample size. With a small fraction of data from the population, it accurately recovered the true parameters. Therefore, the NBB model can be used when we do not have complete network information.

The last essay is the first attempt to incorporate heterogeneous peer influence into the NBB model, based on individuals' preference structures. Besides the significant extension of the NBB (Bass) Model, incorporating high-quality data on individual behavior into the model leads to new findings on individuals' adoption behaviors, and thus expands our knowledge of the diffusion process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The Bass Model has been used extensively and globally to forecast first purchases of new products. It has been named by INFORMS as one of the top 10 most influential papers published in the 50-year history of *Management Science*. Most models for the diffusion of innovation are deeply rooted in the work of Frank Bass. His work provides a framework to model the underlying process of innovation adoption among customers. The Bass model states that there are two processes driving new adoption with respect to the first time trial of a new product. The first one is called the innovation process, which involves external influence outside of the social system and does not interact with prior adopters. Another process is the imitation process, which originates from interactions between those who adopted the innovation (adopters) and those who have not yet adopted it (non-adopters). Bass saw word-of-mouth (social interaction) as a key driver for the diffusion of innovation. Successive studies on the diffusion model confirm that internal influence is a key driving force for innovation diffusion (Goldenberg et al. 2001a; Mahajan et al. 1990b).

Recently booming online social network services have delivered a powerful vehicle for firms to utilize the word-of-mouth effect using social network marketing campaigns, which are often called "viral" or "buzz" marketing. Social networks formed by customers provide firms with alternative marketing tools that may be more powerful than traditional marketing media.

Prior research on social networks has shown that the structure of the network affects adoption patterns (Dover et al. 2012; Goldenberg et al. 2009; Hill et al. 2006; Katona and Sarvary 2008; Katona et al. 2011; Newman et al. 2006; Shaikh et al. 2010). Gupta and

Mela (2008) found that having network effects at the early stage of a product is crucial for a firm's economic success.

However, there is little research on how network structure affects the parameter estimates of the Bass model, which is the most widely applied diffusion model in marketing and specifies that social interaction between adopters and non-adopters is the major driver for new adoptions. In this study, I explore the influence of network structure on the Bass model estimates, using both simulated and empirical data. I further look into how to incorporate complete network information into the original Bass model, while preserving its parsimonious form. For this purpose, I specify the individual-level adoption process that becomes the Bass model when we aggregate it. At the same time, I derive a more general Bass model that renders more accurate estimates of the imitation parameter (q), as well as that of the innovation parameter (p).

First, in essay 1 I show analytically that the Bass model assumes all potential customers are linked to all other customers. However, this assumption of a fully connected network needs to be integrated with more realistic settings (a partially connected network), so I relax the assumption of the fully connected network and show how the distribution of connections among people affects the estimates of the imitation parameter (q). I propose the Network-Based Bass (NBB) model, which can fully accommodate network structures into the Bass model. Moreover, I provide proof that the Bass model is nested within the NBB model.

I simulate random networks of connections between potential customers using the Random Network[1]. I then show that the estimated imitation coefficient *q* is biased downward in the traditional Bass model, and that the estimated innovation coefficient estimate of *p* is also biased due to the negative correlation between the *p* and *q* estimates. I found that these biases depend on the structure of the network. Using the NBB model, I am able to recover the true parameters from simulated data.

I further test the NBB model on empirical data from an online music social network, which is not necessarily a random network. From a network formed in the real world, I estimate the Bass model parameters using the NBB model and the traditional Bass model. The results confirm that the NBB model better captures the diffusion dynamics through social influence. More importantly, the estimated effect of social interaction (imitation), measured by *q* is much lower for the traditional Bass model than the NBB model. This suggests that by ignoring the effect of network structure, the original Bass model depreciates the impact of social interactions on new adoptions.

A subsequent study in essay 2 focuses on extending the NBB model's practical application by network sampling when the entire network information is not available, or the network is too big to analyze. I have tested various network types and network sampling methods and have identified the best network sampling method, the snowballing sampling (SBS) method, for the NBB model. I found that with a small sample size, data sampled by SBS contain the best information for the NBB model. A simulation study shows that small samples by SBS yield highly accurate estimates for the

[1] Random Network: The probability that two individuals are connected is determined randomly.

NBB model parameters; thus, it is a highly efficient sampling method for the NBB model.

I further tested via simulations whether the NBB model is robust against various network types and demonstrated its superiority in recovering true parameters in all of the network types tested. This is due to the fact that the NBB model utilizes all individual-level connection information. I believe that the NBB model can be applied to any type of network and requires no prior knowledge of network type.

The two major findings in essay 2—that SBS is the best network sampling method for the NBB model, and that the NBB model is robust to variation in network structure—provide a strong incentive for practitioners to use my model in order to measure the impact of social influence.

The last essay extends the Bass model further by modifying the NBB model. With a large volume of high-quality data involving individuals' past product consumption, I showed how to take advantage of these "big data" by simply adding one more parameter to the NBB model. I further refined the method in order to visualize these "big data," which is one of the urgent issues in the marketing literature. Moreover, the preference map that visualizes customers' preference structures can be used for many practical purposes, such as segmentation and identifying optimal target seeding points for new products.

The modified NBB model with preference information (NBB.Pref model) demonstrates that preference similarity among customers determines the speed of new product adoption. Thus, more managerial-relevant findings are provided in the NBB.Pref model.

This dissertation has followed a series of steps to enhance the most well-known diffusion model, both theoretically and practically. Three essays have successfully enhanced the original Bass model by incorporating social network information first and then micro-level preference information. Moreover, this dissertation provides remedies for the data requirement of the proposed model by network sampling. In sum, my dissertation will provide significant academic contributions to the existing marketing literature, as well as a more accurate and insightful decision marketing tool for industry practitioners.

CHAPTER 1: THE ROLE OF SOCIAL NETWORK STRUCTURE ON THE BASS MODEL PARMETER ESTIMATION

1.1 Literature Review

Substantial academic endeavors have been devoted to examining how network structure determines the shape of the diffusion curve, and to increasing our understanding of dynamics involving the diffusion of innovation through peer influence (Bell and Song 2007; Garber et al. 2004; Goldenberg et al. 2001b; Liu et al. 2005).

A specific segment named, *influentials*, has been recognized as the driver of the social process (Hinz et al. 2011; Katz and Lazarsfeld 1955; Van den Bulte and Joshi 2007). They tend to adopt innovation earlier (Katz 1957) and are more clustered (highly connected) in the network map (Aral and Walker 2012). Therefore, locating those influentials in network is key to understand the dissemination of information and behaviors (Aral and Walker 2012). It has been well also documented that early adopters have higher number of incoming connections (Iyengar and Bulte 2011; Katona et al. 2011). Katona et al. (2011) argue that the probability of adoption is positively correlated with the number of connections to the adopters. Thus, a high level of this asymmetric relationship of early adopters to the mass of non-adopters may shift the diffusion curve up sharply. Goldenberg et al. (2009) found that people with exceedingly high connection (called hub) are more likely to adopt earlier and these hubs play roles in determining the adoption speed and market size.

Consequently, a large body of work has explored the relation between the connection structure among network members and the diffusion dynamics of innovation (Abrahamson and Rosenkopf 1997; Delre et al. 2006; Dover et al. 2012; Shaikh et al.

2010; Watts 2002). Delre et al. (2006) and Watts (2002) found that the speed of adoptions varies with the network structures i.e., the distribution of the degree (connection) determines the time-path of innovation dissemination. In terms of a parameter estimation, Hill et al. (2006) found that incorporating network information improves model performance, and Shaikh et al. (2010) showed that incorporating network information improves penetration forecasts.

Studies on the impact of specific types of networks on penetration have also been popular among academics. Delre et al. (2006) and Shaikh et al. (2010) explored the diffusion dynamics in scale-free networks. Recently, Dover et al. (2012) compared diffusion patterns on four different types of network structures and provided empirical evidence that different types of networks have differently shaped diffusion curves.

However, current studies on social networks and the diffusion of innovation literature have not yet provided tools to incorporate full network information into the aggregate diffusion model (e.g. within the framework of the Bass model). Information on individuals' networks and behavioral data are aggregated to the summary measures, such as the average or variance of degree distribution (Delre et al. 2006; Dover et al. 2012). In short these studies ignore the individual network data that may be available.

Furthermore, literature on the micro-level diffusion process is rare. Chatterjee and Eliasberg (1990) modeled the effect of individual characteristics on the micro-level diffusion process, reflecting the heterogeneity among individuals. The micro-process is then aggregated to predict the diffusion curve. However, their model does not incorporate the social network information.

As shown in Figure C1, I propose the model that utilizes full network information within the basic form of Bass model as well as a micro-model of the adoption process that models individual-level heterogeneity of adoption probability. Heterogeneous network effect is determined by a person's number of friends and her friends' behaviors (network structure and social influence). I prove that in the limit condition, this proposed model is equivalent to the Bass model. Furthermore, when we aggregate the micro-level adoption process it is equivalent to the NBB model, which is the macro-level model. However, the proposed model does not add complexity nor change the interpretation parameters in the Bass model. It will preserve parsimonious form of the Bass model.

This study examines how network structure affects the Bass model parameter estimates (Figure C2). I expect that the imitation coefficient estimate from the traditional Bass model will be biased because it assumes the homogenous adoption probability for all non-adopters, regardless of their position in the network, i.e., different numbers of connections to the adopters. The extreme case will be that all adopters and non-adopters are all isolated (no connections thus no imitation process at all). Even in such a case the Bass model assumes that there is social interaction since no actual network information is built into the Bass model. When we model the heterogeneous peer to peer influence due to different connection structures into the Bass model, the estimates will calibrate unbiased parameter estimates, and therefore, we will observe better performance of the Bass model.

To best of my knowledge, the macro-level diffusion model, which incorporates explicit and full network information and micro-level behavior data together into the original Bass model, has never been proposed and empirically tested. Current studies on

the network-based diffusion model depart from the original forms and simple underlying behavioral assumptions of the Bass model. Because the original Bass model has been the most popular model to forecast or describe the adoption process on a macro-level, I believe that extending the Bass model by combining network information within the basics of the Bass model would provide a meaningful contribution to the diffusion literature.

The rest of paper is laid out as follows. First, I start with the analytic results. I will provide the proof that the Bass model is equivalent to the NBB model when the assumption of the fully connected network is met. Then, I relax this assumption and modify the Bass model to accommodate a more realistic assumption, a partially connected network. I will then suggest the micro-level adoption process that underlies the NBB model and that is also consistent with the Bass model. In the subsequent section, I simulate individual adoption data with various levels of innovation and imitation influence across different network structures. Then, I compare the NBB model's performance with the traditional Bass model in recovering the true parameters from the simulated data. I further confirm the findings from the simulation study using empirical adoption and network data that have been collected from a large number of online social network users on multiple products. This paper finalizes with conclusions and ample suggestions for future research.

1.2 Model Development

1.2.1 From the Original Bass Model

The Bass model (Bass 1969) specifies two processes that drive new adoptions, innovation and imitation. Diffusion of new adoptions is first initiated by innovators who

accept new ideas or who try new products without any influence from the prior adopters. Its impact on increasing adoption does not come from the social interaction with prior adopters, so it is often referred to an external influence.

The other process is imitation, which is described in the second term of the Bass model. The group of adopters, Y_t, interact with the group of non-adopters, $m - Y_t$. Their interactions within social system accelerate diffusion of innovation. As more people adopt, more social pressure, $\frac{Y_t}{m}$, will be passed on to the non-adopters. Because this imitation process due to social interaction takes place within social system it is often called internal influence, social influence, or social contagion.

$$S_t = p\left[m - Y_t\right] + q\left[\frac{Y_t}{m}\left(m - Y_t\right)\right]$$

Where m = total number of adopters during time t

p = coefficient of innovation

q = coefficient of imitation

S_t = number of new adopters at t

Y_t = cumulative adopters at t

To describe the Bass model in individual level, it is formed by the hazard function, $\frac{f_t}{1 - F_t} = p + q\frac{Y_t}{m}$, which defines that the probability of a non-adopter becomes an adopter

at time *t*, f_t. It assumes that f_t is a linear function of the proportion of previous adopters, $\frac{Y_t}{m}$. The probability of new adoption f_t is attributed to innovation and imitation processes.

Since innovation is not affected by number of adopters at current time (Y_t) its coefficient parameter, p, is static and measures influences that do not depend on social interaction. The imitation coefficient parameter, q, measures the impact of social influence (also called the word-of-mouth effect), which is proportional to the size of cumulative adopters, Y_t. In other words, the value of q indicates how much of an increase in the adoption probability is attributed to the social interaction with prior adopters.

Another way to see how the Bass model describes social interaction on micro-level is looking at $\frac{q}{m}$. The probability of a non-adopter becoming an adopter is increased by $\frac{q}{m}$ for every increase in the number of adopters. That is, there will be more social pressure to a non-adopter as proportion of adopter increases. However, the social pressure or social influence is constant across all non-adopters, implying that the Bass model assumes a *homogeneous* influence of adopters on non-adopters.

To calculate the aggregate number of adopters, the coefficients of two processes, innovation and imitation, are multiplied by the number of non-adopters, $m - Y_t$. For the innovation process, the impact is not related to the social interaction between adopters and non-adopters (external influence from social system); thus, there are no multipliers other than $m - Y_t$ for the innovation coefficient, p. The impact of interaction between peers is modeled by $Y_t(m - Y_t)$. Previous adopters Y_t interact with non-adopters $m - Y_t$ and all non-adopters are influenced by all adopters.

1.2.2 A Network-Based Bass Model

The first step to extend the Bass model is to look at the assumptions and to search alternative expression of social interaction term in the Bass model. Since there is no social interaction that plays a role in the innovation parameter, p, I follow all assumptions and maintain the same expression with the Bass model.

For the social interaction term $\left(Y_t\left(m-Y_t\right)\right)$, I further examine the assumptions of the Bass model in term of the network structure, i.e., underlying assumptions of connections between adopters and non-adopters. I express this term with two vectors and one matrix, which is commonly used in the social networks and spatial literatures (Bradlow et al. 2005; Jackson 2010; Leenders 2002).

First, I define two vectors which have a size of $1*m$, as follows.

$\mathbf{a_t}$ is $1*m$ row vector and $a_{it}\begin{cases}1 & \text{if } i \text{ is an adopter at time } t\\ 0 & \text{otherwise}\end{cases}$ and

$\mathbf{1_m}$ is $1*m$ row vector with all entries equal to 1

Then Y_t and $m-Y_t$, representing the number of adopters and non-adopters, are sum of all elements in $\mathbf{a_t}$ and $\mathbf{1_m}-\mathbf{a_t}$ respectively, i.e., $Y_t=\sum_{i=1}^{m}a_{it}$ and $m-Y_t=\sum_{i=1}^{m}\left(1-a_{it}\right)$. Next, I borrow the matrix representation of connections between individuals from social network and spatial literature, which is called an adjacency matrix (also called connection matrix). An adjacency matrix, $\mathbf{C}$, is a symmetric square matrix in which each element

represents connections between all possible pair individuals in the network. All individuals are indexed by both rows and columns, and the values of the element in the corresponding rows and columns are either 0 if they are not connected or 1 if they are connected. The general adjacency matrix can be generated as follows:

$$\text{Adjacency matrix } \mathbf{C} \text{ is a } m * m \text{ matrix,}$$

$$\text{where } c_{ij} \begin{cases} 1 \text{ if } i \text{ and } j \text{ are connected and } i \neq j \\ 0 \text{ otherwise or } i = j \end{cases}$$

Following the definition for a general adjacency matrix, I define the size of the $m * m$ square matrix, $\mathbf{C_F}$, where all off-diagonal elements are 1 and the diagonal elements are 0. This is an adjacency matrix for a fully connected network.

Suppose all ultimate adopters in the market size of m, are connected, i.e., a fully connected network. Then for any Y_t, the social interaction term in the Bass model, $Y_t(m - Y_t)$, can be expressed as follows.

$$Y_t(m - Y_t) = \mathbf{a_t C_F} \left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}} = \begin{bmatrix} a_{1t} & \dots & a_{mt} \end{bmatrix} \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & 0 & 1 & \vdots & 1 \\ \vdots & 1 & \ddots & 1 & \vdots \\ 1 & \vdots & 1 & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 - a_{1t} \\ \vdots \\ 1 - a_{mt} \end{bmatrix}$$

Where $\mathbf{C_F}$ is the $m * m$ adjacency matrix for a fully connected network

Next, I column standardize[2] the adjacency matrix $\mathbf{C_F}$ and denote it as $\mathbf{C_F^*}$. Then $\mathbf{C_F^*} = \frac{1}{m-1}\mathbf{C_F}$, since the sum of elements in each column is same, $m-1$. Finally, the matrix expression of the Bass model is as follows.

For a fully connected adjacency matrix $\mathbf{C_F}$ and large *m*

$$S_t = p\left[m - Y_t)\right] + q\left[\mathbf{a_t}\mathbf{C_F^*}\left(\mathbf{1}_m - \mathbf{a_t}\right)^{\mathbf{T}}\right]$$

$$= p\left[m - Y_t\right] + q\left[\frac{Y_t}{m}\left(m - Y_t\right)\right]$$

Where $\mathbf{C}^*$ is column standardized version of $\mathbf{C}$

(See Appendix A1 for proof)

To see this with simple general example, suppose we have $m = 5$ and $Y_t = 3$ then,

[2] $\mathbf{C}^*$=Column-standardized matrix for $\mathbf{C}$, *where* $c_{ij}^* = \frac{c_{ij}}{\sum_j c_{ij}}$

$$\mathbf{a_t C_F}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}}=\begin{pmatrix}1&0&1&1&0\end{pmatrix}\begin{pmatrix}0&1&1&1&1\\1&0&1&1&1\\1&1&0&1&1\\1&1&1&0&1\\1&1&1&1&0\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\\1\end{pmatrix}$$

Column-standardized $\mathbf{C_F}=\mathbf{C_F^*}$,

$$\mathbf{a_t C_F^*}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}}=\begin{pmatrix}1&0&1&1&0\end{pmatrix}\begin{pmatrix}0&1/(5-1)&1/(5-1)&1/(5-1)&1/(5-1)\\1/(5-1)&0&1/(5-1)&1/(5-1)&1/(5-1)\\1/(5-1)&1/(5-1)&0&1/(5-1)&1/(5-1)\\1/(5-1)&1/(5-1)&1/(5-1)&0&1/(5-1)\\1/(5-1)&1/(5-1)&1/(5-1)&1/(5-1)&0\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\\1\end{pmatrix}$$

$$=\frac{6}{4}=\left(\frac{1}{5-1}\right)\times 3\times 2=\frac{1}{m-1}Y_t\left(m-Y_t\right)\approx\frac{1}{m}Y_t\left(m-Y_t\right)$$

I show here thought a universal example that the Bass model assumes a fully connected network and a devised matrix expression of the Bass model, which incorporates the network structure. I call this model the *Network-based Bass Model* (NBB). Note that I chose all diagonal elements to be 0 since an adopter cannot influence herself to adopt because she already adopted (also c_{ii} indicates the connection to herself), and this definition on diagonal elements is consistent with the sociology and spatial literatures.

Since it is natural to assume that some people are not connected to each other in most of social networks, I assume that $\mathbf{a_t C_F^*}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}}$ is not equal to $\mathbf{a_t C^*}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}}$, where $\mathbf{C}$ and $\mathbf{C_F}$ are an adjacency matrix that we may observe in the real world (a partially connected adjacency matrix), and that the Bass model assumes (a fully

connected adjacency matrix), respectably. Note that *density*[3] measures the ratio of observed connections to the all possible connections among all members in a social system. Thus when density equals 1 all member in the social system are fully connected. When *density*=1, i.e., a fully connected network, the Bass model and the NBB model is equivalent.

It is easy to see that the estimate of the imitation coefficient from the Bass model (which assumes a fully connected network) will be biased downward, since the estimation of q in the Bass model is confounded with the level of connections in network. By assuming a fully connected network, the number of potential social interaction in the Bass model is always higher than the actual. Therefore, the model always underestimates q.

The NBB model measures, however, internal influence consistently regardless of the different network density because it accurately measures the connections between the network members by $\mathbf{C}$. Without actual link information among individuals, the measure of social interaction in the Bass model is only determined by cumulative number of adopters, Y_t.

Suppose we have some unconnected individuals as shown in adjacency matrix $\mathbf{C}$ below and $m = 5$ and $Y(t) = 3$, the same with the previous example.

---

[3] For undirected graph, $density = \dfrac{\mathbf{1^T C\ 1}}{m(m-1)}$ where $c_{ij} = (i, j)^{th}$ element of matrix $\mathbf{C}$ and m = number of row (or column) of $\mathbf{C}$.

$$\text{Suppose } \mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \text{ then } \textit{density} = \frac{14}{5(5-1)} = 0.7 < 1 \text{ and}$$

$$\mathbf{a_t C^*}\left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 1/2 & 1/4 & 1/2 \\ 0 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 1/4 & 1/2 & 0 & 1/2 \\ 0 & 1/4 & 0 & 1/4 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$= \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right) + \left(0 + 0 + \frac{1}{2}\right) = \frac{5}{4} < \frac{6}{4} = \frac{1}{5-1} \times 3 \times 2 = \frac{1}{m-1} Y_t \left(m - Y_t\right)$$

Given the same number of current adopters ($Y(t)$), the measure of social interaction from the NBB model will be different because the adjacency matrix will be adjust it based on the actual connections between adopters and non-adopters while the Bass model always computes the maximum of social interaction.

The first finding in this study is that above matrix expression for the social interaction term in the NBB model allows us to incorporate explicit network information and individual adoption data over time into the estimation process of the imitation parameter. Also the Bass model is a special case of the NBB model when every adopter is linked to every non-adopter. At the same time I show by previous example that the Bass model gives us inconsistent estimates for the imitation coefficient, and its estimates depend on the density of the network map, while the NBB model does not. In the next

section, I will propose individual-level adoption process which is consistent with the NBB model as well as the Bass model and then describe the statistical properties of the NBB model.

1.2.3 Individual Adoption Process and Heterogeneous Social Influence

Because the NBB model reflect heterogeneous social influence due to different connections across individuals, it is necessary to specify the individual level adoption process that underlies the NBB model and also the Bass model, the special case of the NBB model. The proposed individual level adoption process will illustrate how heterogeneous social influences across non-adopters are incorporated in aggregated level through social interaction term (matrix expression of social interaction) in the NBB model.

We assume that the probability of adopting in time is given by

$$s_i \mid t \sim Bernoulli\left(p + q \times r_{it}\right) \text{ and } p > 0,\ q > 0,\ p + q < 1$$

where i = non-adopter

$s_i \mid t$ = binary random variable indicating non-adopter i becomes an adaptor at time t

$$p = P\left(s_i = innovator \mid t\right) \text{ and } q \times r_{it} = P(s_i = imitator \mid t)$$

$$r_{it} = \frac{\text{\# of } i\text{'s friends who adopted as of } t}{\text{\# of all of } i\text{'s friends in the set of the ultimate adopters}}$$

i and j are friends when i and j are connected

I assume that the binary random variable, $s_i|t$, follows the Bernoulli distribution. I assume further that an event of a non-adopter becomes an adopter at time *t* consists of two processes of innovation (*p*) and imitation (*q*), but happens at the same time.

Adoption probability of non-adopter *i* due to external influence, p, is constant across time and individuals. The probability of adoption due to the social influence or pressure, $q \times r_{it}$, is different across time and across individuals.

I defined the heterogeneous social influence (r_{it}) as a linear function of the proportion of adopter friends to all friends of non-adopter *i*. This individual-level model differentiates itself from the Bass model by this term. Social influence at each time period is measured by the ratio of the adopter friends to all of her friends whom she knows (r_{it}) rather than by that of all adopters to all ultimate adopters whom she might not know or observe $\left(\frac{Y_t}{m}\right)$. This model isolates the social influence within the personal network only, not the entire network. It is also assumes that the non-adopter adopts faster as she is exposed to more adopter friends relative to all of her friends. r_{it} measures the *proportion* of adopter friends because the influence of one adopter friend will be different for an individual with a few friends and one with many friends (Granovetter 1973; Katona et al. 2011).

Now r_{it} equals $\mathbf{a_t C^*_{\cdot i}}$ in terms of matrix expression of the NBB model, where $\mathbf{a_t}$ is a row vector that indicates adopters at *t* and $\mathbf{C^*_{\cdot i}}$ is the i^{th} column of the column standardized adjacency matrix, $\mathbf{C}^*$ and r_{it} equals 0 if *i* has no friend.

Because of the assumptions of independence between the two adoption processes and of the concurrent event of two processes, it turns out that $E(s_i \mid t) = p + q \times r_{it}$ because being an adopter involves two independent Bernoulli trials. To achieve the aggregate level of adoption, S_t, I further assume that the event of being an adopter at the current time is independent among non-adopters. Then, I sum up the expected values over all non-adopters.

Because of the assumption of independence among adopters at the current time *t*, S_t follows the Poisson-binomial distribution (Wang 1993). It refers to the probability distribution of the number of successes in *n* binary outcome trials with different success probabilities ($p_1, p_2 \ldots p_n$). The expected number of success from the Poisson-binomial distribution is the sum over all different probabilities ($\mu = \sum_{i=1}^{n} p_i$). Therefore, in the NBB model, the expected number of new adopters from the two adoption processes over all non-adopters at *t* then becomes:

$$E(S_t) = \sum_{i \notin A} E(s_i = 1 | t) \quad \text{where } A = \{ i \mid a_{it-1} = 1 \}$$

$$= \sum_{i \notin A} (p + q \times r_{it}) = \sum_{i \notin A} p + \sum_{i \notin A} q \times r_{it} = p \sum_{i \notin A} 1 + q \sum_{i \notin A} r_{it}$$

$$= p(m - Y_t) + q \sum_{i \notin A} r_{it}$$

(See Appendix A2.1 for proof)

Finally, the above expression of the expected value from the Poisson-Binomial process is equivalent to the NBB model since $\sum_{i \notin A} r_{it} = \mathbf{a_t C^*} \left(\mathbf{1_m} - \mathbf{a_t} \right)^{\mathbf{T}}$ (See Appendix A2.2 for proof)

All assumptions in the NBB model are the same as those in the Bass model except for the assumption of the heterogeneous social influence, r_{it} which was assumed to be homogeneous in the Bass model, $\frac{Y_t}{m}$.

1.2.4 Proposed Model

Based on the NBB model I derived previously, I propose the following model and method to estimate the imitation parameter, q, as well as the innovation parameter, p from aggregated adoption data.

$$S_t = p\left[m - Y_t\right] + q\left[\mathbf{a_t C^*} \left(\mathbf{1_m} - \mathbf{a_t} \right)^{\mathbf{T}}\right]$$

The NBB model states that if we have full network information of all ultimate adopters we can simply use OLS (Ordinary Least Squares) to estimate the two parameters. Even though the NBB model opts for the simple and parsimonious form to estimate parameters, the aggregated (macro-level) measures in the model are accurate if my assumption of micro-level adoption process is satisfied. The NBB model can be estimated with the simple ordinary least square OLS method without an intercept. However, the estimates of the NBB model retain the interpretation as the parameters from the poison-binomial distribution.

Along with the other benchmark Bass models, I will test how accurately the NBB model can recover the parameters of the Poisson-binomial process from data simulated with different network structures and heterogeneous adoption probabilities in the next section.

1.3 Simulations and Benchmarking

The simulation study serves two purposes. One is to test how well the NBB model and the Bass model uncover the true parameters when varying the levels of network of connections and the model parameters. Because the NBB model incorporates network information to estimates the imitation parameter, it should not be affected by the network density, while the Bass model would be. Another purpose is whether the estimates from the NBB model are comparable to those from the Bass model when network structure becomes closer to being a fully connected network, which the Bass model assumes.

Based on assumptions about the NBB model, I generate micro-level adoption data by the following steps. First, I set true parameters for *density*, m, p, and q. Then I generate a random network map for m individuals with the given density and compute the adjacency matrix $\mathbf{C}$. For individual random adoption data, two adoption processes are applied to each non-adopter simultaneously for each time period *t*. For the given p and $q \times r_{it-1}$ [4], the probability of the non-adopter adopting innovation at the current time t

[4] Note that r_{it-1} corresponds to $\frac{Y_{t-1}}{m}$ in the discrete analog of the Bass Model to fit the empirical data, $S_t = p\left[m - Y_{t-1}\right] + q\left[\frac{Y_{t-1}}{m}\left(m - Y_{t-1}\right)\right]$ (Bass, 1969)

is $p + q \times r_{it-1}$, i.e., $P(a_{it} = 1 | t) = Bernoulli\left(p + q \times r_{it-1} \right)$. This study does not examine two adoption processes separately, even though Van den Bulte and Joshi (2007) suggest that there are two distinct segments of influentials and imitators in the market. After both processes are applied simultaneously to all non-adopters, I obtain $\mathbf{a_t}$ and compute S_t. Every combination of the parameters has been iterated five times.

Three models are used to recover the true parameters p, and q from the aggregated simulation data. The first model is a non-linear square model from Srinivasan and Mason (1986), which is call here the Bass.NLS Model here:

$F_t = \frac{1 - \exp[-(p+q)t]}{1 + q/p \exp[-(p+q)t]}$ (see Schmittlein and Mahajan (1982) for derivation of this formula from the Bass model).

The second model is the original Bass model to fit the discrete data. I call it the Bass.OLS model here: $S_t = p\left[m - Y_{t-1}\right] + q\left[\frac{Y_{t-1}}{m}\left(m - Y_{t-1}\right)\right]$. The last model is the NBB model: $S_t = p\left[m - Y_{t-1}\right] + q\left[\mathbf{a_{t-1}C^*}\left(\mathbf{1_m - a_{t-1}}\right)^{\mathbf{T}}\right]$. Estimates from the three models are calibrated from the simulated data with the same parameters. Since there are five iterations for each combination of the true parameters, the final model estimates are calculated by averaging five estimates from all five iterations.

Figure C3 shows the aggregated penetration data from individual simulations. The pattern of diffusion shows the typical shape of diffusion, gradually increasing to the peak and decreasing adoptions after the peak. The dissemination of innovation is symmetric, since the connections between adopters follow a random process, indicating that individual Bernoulli trial assumptions replicate the macro-level diffusion pattern.

For the networks with high density (*density*=0.017 and *m*=10000), the three models fit the data equally well. Estimates for social contagion from the NBB model is exact with the true values (0.233), but there are no significant differences across the three models (0.221 for Bass.OLS and 0.236 for Bass.NLS).

However, when the connection probability of two individuals are low (*density*=0.00021 and 0.00005) in Figure C4, the Bass model shows poor fits, especially in the early stage of diffusion, while fits from the NBB model closely follow the actual adoption curves. Inaccurate estimate for the coefficient of imitation contributes to poor fits of the Bass model. The true value of *q* is 0.233 the same as the high density in Figure C4, but the estimates from the Bass models are 0.104 and 0.111 for *density*=0.00021, and they are even lower for *density*=0.00005 (0.020 and 0.022).

The Bass model fits for the lowest density (0.000005) network in Figure C4 indicate that the peak of adoption is the first time period, but the actual peak is seventh time period. This is caused by the large underestimate of imitation parameter in the Bass model since the estimates from both of the Bass model (0.02 and 0.022) is much less than the actual (0.233). The fact that *q* estimate is less than *p* estimate (0.027 and 0.028) induces non-mirror shape of fits (S-shaped curve for the cumulative adopters). On the contrary, the NBB model consistently yields superior fits irrespective of the network density level and the Bass model performs poorly in a sparse network.

Figure C3 and Figure C4 show that the distribution of connection among network members can introduce bias in the Bass model estimates. Regardless of network density, the NBB model consistently recovers the true parameters of the social process**.**

I further test the estimates for the three models for the wide ranges of *density* and *q*. Because of the high dimensionality of parameter space (four parameters to determine for the simulation), I test on the single values for *m* and *p*. The most frequently found values of innovation parameter from the empirical studies range from 0.01 to 0.03. So I choose true *p*=0.022 whereas for *q*, the true values cover most of the possible range, 0 to 0.7.

In Figure C5, the true *p* values are recovered all density levels from the NBB model. For the Bass.OLS model, the *p* estimates converge to the true values as density increases, but Bass.OLS estimates for *p* deviate from the true values of 0.022 for low densities.

Figure C6 suggests that this overestimation of *p* is caused by an underestimation of *q*. It is quite more noticeable that the *p* estimates from the Bass.NLS model largely depend on the value of *q*, regardless of the network density. As one might expect, because the NBB model correctly measures the magnitude of social influence by incorporating social network information, the estimates of *q* from the NBB model show no relationship with the network structure. On the contrary, for Bass.NLS and Bass.OLS, both model estimates for the imitation coefficient considerably depend on the network density.

To see how many friends a network member has, on average, to have the good estimates from the Bass model parameter, I mapped two random networks for two density levels, 0.00171 (High Density) and 0.00005 (Low Density) in Figure C7 since the Bass model gives us good approximation of *q* when density is above 0.00171.

The average number of connections given density equals $density * m$ where *density* = mean number of link [5]. Since the graphs are generated from a

[5] $density = \frac{\mathbf{1}^{\mathbf{T}}\mathbf{C}\,\mathbf{1}}{m(m-1)} \rightarrow$ mean number of link $= \frac{\mathbf{1}^{\mathbf{T}}\mathbf{C}\,\mathbf{1}}{m-1} = m \times density$

random network, each individual in a high density network has approximately 17 friends. For a low density (0.00005) graph, a randomly selected network member has a half chance (0.5) of having one friend, for example we have *m*=10000 and *density*=0.00005 and $0.5 = 10000 \times 0.00005$.

Considering that connections to measure social interaction in the Bass model are among those who have ultimately adopted an innovation, for most of innovations (unless innovation is very successful), it would be rare for an adopter to have 17 or more adopter friends on average. In other words, average connections for the most of innovation will be lower than 17, and social influence in the Bass model is more likely to be underestimated. The network maps among the final adopters in the real world will be more likely to be ones between a low density and high density graph (if it is random network) in Figure C7.

Correlations between the network structures and estimates are summarized in Table B1. Estimates of *q* from Bass.NLS (q.Bass.NLS) and Bass.OLS (q.Bass.OLS) are positively correlated with density (0.24 and 0.21), which indicates that estimates of imitation from both models increase with density. Since estimates of peer influence from the NBB model (q.NBB) are already modified by network information in the estimation process, they show no relation with density (the correlation is 0). It is interesting that innovation coefficient estimates from Bass.NLS (p.Bass.NLS) are negatively correlated with imitation coefficient estimates, whereas those from Bass.OLS and NBB have a positive relationship with imitation coefficients. The correlation between two coefficients is more prominent at a low density, as shown in Figure C5 and Figure C6, where innovation drives the most of increases in new adoptions.

I look closer in Figure C8 at the effect of network connectivity on the estimates for extremely sparse networks. Within the density range from 0.00001 to 0.001, the Bass model (Bass.NLS and Bass.OLS) estimates are biased downward more as the network becomes sparser. The plane shape of q.NBB suggests that the NBB model yields accurate and consistent estimates even for extremely sparse networks.

I examine why the Bass model underestimates social influence by comparing the difference in social interaction measures between the Bass model and the NBB model. First, I computed the *difference*, where $difference = \frac{Y_t}{m}(m - Y_t) - \mathbf{a_t C^*}\left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}}$. Since the social interaction term is the only difference between the two models, examining the differences in the social interaction measures between the two models will explain why the Bass model underestimates *q*. Figure C9 shows that even when there is almost no social interaction between adopters and non-adopters (*density* equals 0.00005 . i.e., almost all adopters are innovators) , the Bass model assumes that all adopters influence non-adopters (see the red lines for the Bass.OLS Model). The difference in the social interaction measures between the two models become wider as the density of network decreases, and the two models agree more on the magnitude of social influence as the network gets denser, as shown in Figure C10.

1.4 Empirical Analysis

This study seeks to validate the arguments beyond a simulation. Empirical data on a large number of network members and their explicit network information have been used to cross-validate the findings in the simulation study. Furthermore, in order to examine the consistency of my findings on various scenarios (different network density,

market size, *p*, and *q*), three models have been applied for multiple products generated from identical networks. Unlike the networks in the simulation study, these networks are not necessarily random networks. There are diverse levels of connections, i.e. some individuals with large number of connections and other with few or no connections.

1.4.1 Data Description

Utilizing a web-crawling agent, I collected individual adoption data (listening behavior of songs) from an online music social network along with all individuals' friendship information. This online music social network, Last.fm, publicizes all users' friendship information on a user's personal webpage. Most of users' track histories are also publically accessible unless a user sets the option to hide them. Approximate 97% users' music libraries are available to the public. The software developed by Last.fm, "Scrobbler," extracts user's music library histories from her portable devices, such as cellular phone, iPod, and tablet PC, as well as web-streaming records, and saves library histories from these difference sources on her personal website. I collected data for approximately 360,000 users' complete track history lists from their electronic devices (offline behavior) and web streaming (online behavior) from Jan. 2011 to Mar. 2012.

Each track history uploaded to a user's personal webpage includes the title of a song, the singer, and the time it was played so I was able to locate the date of the users' first trial of the new music. The singer's name and the title of song are used to locate a specific song from a user's library, and then the earliest time from the selected song histories is identified as the first trial. If there was no single track record of a song on a user's personal library during the 15 months of data collection period, such users were excluded from the ultimate adopters of the song.

New singles that were on Billboard's Top 100 and released during Feb. to May 2011 were selected to find the adopters and their first trial date. Songs that yields the negative imitation coefficient estimates from both the Bass model and the NBB model or total adopters totaled less than 1000 were excluded from the analysis. Most of excluded songs have the exponentially decaying adoption curve so there is no tipping-point on adoption curve and peak time is the first period. Finally, a total of 22 songs have been analyzed in this study.

1.4.2 Empirical Results

The three diffusion models, NBB, Bass.OLS, and Bass.NLS, are compared. Descriptive statistics for the parameters estimates are provided in

Table B2. The q.Ratio is defined by q.NBB/q. Bass.OLS. A q.Ratio greater than 1 implies that the NBB model estimate for the imitation coefficient is greater than the estimate from the Bass model. Adoption data are analyzed by week. The number of weeks traced for the diffusion of new songs range from 52 to 66 weeks, which is long enough for all songs' adoption rate to reach far over their peak. Densities of the network cover a wide range (minimum=0.0003 and maximum=0.0069), which allows me to examine the impact of density on parameter estimates across the three models. The smallest market size is 1242 and the most successful song has a total number of 48,613 adopters: "Someone Like You," from Adele.

Figure C11 shows fits for two songs, "Roll Up" and "Just Can't Get Enough". The empirical adoption pattern is not symmetric. There is a mass of new adoptions at the early stage of diffusion and a long lingering tail. The sudden upheavals during the early time period might indicate other external effects such as the start of radio airplay. Most of

the songs' diffusion curves show a low adoption rate at a later time of the data collection period, implying that the data collection periods are long enough to include most adopters.

Contrary to the Bass model (Bass.OLS and Bass.NLS), the NBB model fits are not smooth. Fits from the NBB model mimic spikes in the actual penetration curve followed by an immediate increase in adoptions, which are probably caused by social interaction between adopters and non-adopters. In Table B3, I list the estimates of external and internal influence from the three models as well as network density for 22 songs. The adjusted R^2 from the NBB model (NBB.R^2) show a higher fit than the Bass model (Bass.OLS. R^2) for 13 songs out of 21 and a lower fit can be found for only one song, but the difference in R^2 is 1%.

The NBB model show better fit than the Bass model. However, more important value as a model for a decision-supporting is accuracy of the parameter estimates. Consistent with the results from the simulation studies, the NBB model gives higher values of imitation parameters in all songs, i.e., q.Ratios are greater than 1 for all songs while the estimates for innovation coefficients are similar across all models.

From the Boxplot in Figure C12, we can easily notice that the ratios of *p* between p.NBB and p.Bass.OLS (p.Ratio) are centered right below 1. This suggests that external influence estimates are not different much between the NBB model and the Bass model and that the network density has little impact on the estimate of innovation. On the contrary, q.Ratios are all above 1 and there are large variations among songs, supporting my argument that the Bass model underestimates the magnitude of social influence and that network structure affects the estimates of parameter. Figure C12 shows another same

finding from Figure C5 and Figure C6 in the simulation study. Even though p.Ratios are not differ much they are always below 1, which implies p estimates from the Bass model is always greater than those from the NBB model; Underestimation of q leads to overestimates of p due to the effect of the network structure.

It is worthwhile to examine how network density affects the two penetration processes. Table B4 provides correlation coefficients among density and the model parameter estimates. Similar with previous simulation results, we observe higher correlations between *density* and the traditional Bass model estimates for q (0.35 for q.Bass.NLS and 0.30 for q.Bass.OLS) and lower correlations for the NBB model (0.18).

1.5 Conclusion

1.5.1 Summary and Contribution

In this study, I demonstrated analytically and empirically that the traditional Bass model is biased downward when a network is characterized by sparse connections. Then I proposed the extension of the Bass model and named it Network-Based Bass (NBB) model. The NBB model relaxes the assumption of the fully connected network, and the Bass model is the special case of the NBB model. It directly incorporates explicit network information into the original Bass model, and thus eliminates bias due to the network structure while preserving its parsimonious form.

In theoretical perspective, this study proposed micro-model of adoption processes that underlies the Bass model. The original Bass model recovers (converges to) the true parameters when data have been generated by my assumption of two adoption processes

(independent Bernoulli trials for innovation and imitation but heterogeneous social influence).

I further empirically tested my assumptions on both adoption mechanisms, and the NBB model's performance with micro-level data in the real world. Complete network information of 360,000 individuals and product usage data have been used for the empirical analysis. The NBB model shows superior fits over the Bass model. It successfully captures fluctuations in the adoption curve by the social process.

In addition to an enhanced fit, the NBB model provides us with more useful and accurate information from the model parameters than the Bass model. Because the NBB model assumes the same Bernoulli trials for innovation and imitation processes, the two estimates, both p and q, are the "probability" of the binary event, not the "influence" (Mahajan et al. 1990a). This allows us to directly compare the impacts of the two processes.

This study showed that social influence is underestimated when there is a network effect. This finding holds significant importance for practical purpose in marketing. The impact of social (internal) influence is much greater than external influence. Most of adoption patterns are characterized S-shaped curve, i.e., new product adoptions are mostly driven by social interaction. Underestimation of social influence leads to the overestimation of external influence such as a firm's marketing mix. A firm might overly evaluates the effect of its marketing efforts (external influence) by using the Bass model and thus ignoring the effect of social network.

From a manager's point of view, thanks to the development in world-wide-web and information technology, individual-level adoption and network data become readily

available these days. The NBB model delivers a tool to utilize these latest big data source using the Bass model, which a manager is familiar with.

1.5.2 Limitations and Future Research

Simulation results show that the Bernoulli trial assumption on diffusion process is well supported by the Bass model. It will be interesting to find analytical description of how the Bass model estimates converge to the parameters for aggregated Bernoulli trials (Poisson-binomial), especially for non-linear square model (NLS Model) which is the special case of Gamma/Shifted Gompertz distribution (Bemmaor 1994).

This study limits its attention on random network. However, it is well stated in social network literatures that many social networks are known to be a scale-free network (Barabási and Albert 1999). Scale-free network is characterized by a power law distribution of connections among agents. Small numbers of nodes that have a large number of links called "hub" are short cuts to the mass of members in the network map. Rapid penetration by social contagion will start or pass through these hubs. Therefore given same density, the estimates for imitation parameter in the Bass model might be more biased downward for scale-free network than that for random network.

I found strong negative correlation between p and q estimates from empirical data in Table B4, but this study has not come up with a clear explanation of this finding. It might indicate that the NBB model (also the Bass model) needs more specification to address this strong correlation or it can be also explained by the network structure. Closer examination on this finding would be intriguing.

One limitation of the Bass model is that it requires two points (take-off and slowdown) on penetration to calibrate good parameter estimates (Heeler and Hustad 1980;

Srinivasan and Mason 1986). Such requirement raised questions on effectiveness of diffusion models as forecasting tool (Kohli et al. 1999; Mahajan et al. 1990a). By the time we observe these two points a firm already made most of risky investment decisions. The model's value as normative tool has been significantly diminished (Hauser et al. 2006). Thus, it is critical for the diffusion model to provide good estimates before major decisions have been made. Since the NBB model shows better fits over the Bass model at the early stage of diffusion in Figure C4, it will have better forecasting power. Empirical study on forecasting performance of the NBB model using only early stage of penetration data seems promising.

Because the NBB model presents individuals network information adjacency matrix form it allows us to compute transitive relations among individuals. This study only measure the impact of direct link between adopters (density) but the social contagion might more preeminent if her adopter friends are also friends to each other, i.e., more social pressure. Therefore the clustering of connections will modify diffusion pattern and there is little study on the effect of clustering on diffusion (Peres et al. 2010).

CHAPTER 2: THE NBB MODEL AND NETWORK SAMPLING ON LARGE SOCIAL NETWORK

In essay 1, I proposed an extension of the Bass model, namely the NBB model, to utilize explicit network information. The NBB model is free from bias caused by network structure because it fully utilizes all network information and demonstrates superior performance in recovering the true parameters as shown in the simulation study. In addition, it fits my empirical data set better. Essay 2 will examine the performance of NBB model on other network types such as the scale-free and test the effectiveness of different network sampling methods for various network types.

The simulations in the previous study have been conducted on the simplest network structure, the random network, where the distribution of connections has a Poisson distribution. Although the NBB model estimates may be used with various types of network structure, it is useful to examine how network sampling methods and network structures may influence the parameter estimates. This study is especially important in cases when we do not have full network information since we would have to rely on network sampling. Complete network data is typically hard to obtain for researchers. In other cases, utilizing whole network information might be computationally too expensive. Previous studies on network sampling suggests that performance of different sampling methods may depend on the structure of the network.

Regarding the choice of sampling methods for the diffusion model, the questions I will answer in this essays are:

1. Is the NBB model robust across different types of networks?
2. Is there a better sampling methods for the Bass or the NBB model?

3. What sampling method and which model is more efficient in terms of the sample size?
4. Are they any interactions among models, sampling methods, and network types that affect the parameter estimates?

2.1 Degree Distribution and Network Types

2.1.1 Random Network and Poisson distribution

A random graph (or network) describes the probability distribution (random process) used to generate a network graph and it refers exclusively to the Erdős–Rényi random graph (Erdos and Rényi 1960). In this graph, the probability distribution of any member i has k connections follows binomial distribution. Thus, for large size networks, the distribution of connections (i.e., degree for each node) follows Poisson distribution[6].

$$P\left(\#\text{ of connection for } i = k\right) = \binom{m-1}{k} p^k \left(1-p\right)^{m-1-k}$$

Where i = network member

k = # of connection (degree)

m = a network size

p = connection probability between two members

[6] As $m \to \infty$, $p \to 0$, and $mp \to \mu$, Binominal$(p) \to$ Poisson(μ)

Most relevant characteristic of this network graph to this study is that the network is constructed by connecting each member randomly. The probability of any pair of members to be connected is independent with their connections with others, i.e., there is no hub or clustering in the network.

2.1.2 Scale-Free Network and Power-law Distribution

While Erdos-Renyi random graph is the simplest method to generate social networks, most social network in the real world, especially online social networks are characterized as a scale-free network (Barabási and Albert 1999). In a scale-free network, the number of connections follow a power law distribution. That is,

$$P(\#\text{ of connection for } i = k) = k^{-\gamma}$$

Where γ is a parameter for power-law distribution

For a scale-free network, the typical value of γ falls between 2 and 3. Scale-free networks have large big clusters centered on a few key members. In this essay I am especially interested in this type of structure which could introduce high degree of bias in the estimates of social influence on diffusion.

As shown in Figure C13 random network do not have a hub or clustering while scale-free network has a few members with large number of connections. Both networks have the same network density of 0.002 and network size of 20,000.

2.1.3 Watts and Strogatz Network

The Watts and Strogatz Network (Watts and Strogatz 1998) has the property of a small-world network where most members are not near neighbors, but most of members

can be reached by a small number of steps. This small-world-network property is often called "six degrees of separation." A Watts and Strogatz (hereafter WS) network has short average path lengths and a high degree of clustering, but does not have a hub. This network can be generated by first creating lattice and then rewinding the links of the lattice with the uniform random probability *p*. Since this type of network has been used in the diffusion of innovation literature (Shaikh et al. 2010). I also included WS network to examine the impact of sampling on the parameter estimates of the diffusion model.

2.2 Literature Reviews

Conceptually, we can choose three sampling methods: randomly select network members (nodes) or connections (edges) across whole network or select exclusively a part of whole network (Non-random sampling). However, random sampling on connections (Random Edge Sampling) requires prior knowledge on the entire network structure. Therefore, I selected Random Node Sampling (RNS) for random selections method. For non-probabilistic sampling, we can either use Snow-Balling Sampling (SBS) or Forest-Fire Sampling (FFS.) In SBS, we first select one seed sample and we recruit this seed node's acquaintances (first layer of sampling), then further recruit seed node's acquaintances' acquaintances (2^{nd} layer of sampling) and so on until we reach the predetermined sample size. Thus the sampled group appears to grow like a rolling snowball. Since SBS is not including all the connected members for the last layer of recruits, i.e., draw boundary on the last layer, this method is a potential candidate for the edge effect (Griffith 1985), which results from ignoring interference outside of boundary in spatial literature.

FFS is a mixture of RNS and SBS. First we pick randomly one node and then choose z of its outgoing neighbors where z is geometrically distributed with $p/(1-p)$ and p is a parameter. However, SBS is the only network sampling method feasible for web-crawling (Ahn et al. 2007) therefore I choose RNS and SBS for this study.

Because the choice of the sampling method will have significant impact on inference on network topology, a handful of researchers have tried to find better network sampling methods to capture characteristics of network topology or to obtain more accurate the impact of social interaction across different networks (Ahn et al. 2007; Chen and Chen 2008; Ebbes et al. 2008; Gjoka et al. 2009; Lee et al. 2006; Leskovec and Faloutsos 2006).

Direct comparison of network sampling in identifying graph property (e.g., clustering coefficient and density) and has been studied by Leskovec and Faloutsos (2006). Lee et al. (2006) showed why a bias in estimates of graph parameters is introduced by the sampling procedure. Gjoka et al. (2009) developed new sampling method in an online social network by crawling social network. Ahn et al. (2007) examined the validity of snowball sampling across large online social network sites such as Cyworld, MySpace, and orkut. With complete data on social graph of Cyworld, they evaluated SBS and found that SBS is better sampling technique to discover graph parameters. Their study argues the superiority of Random Node Sampling over Random Edge sampling and in general Forest-Fire performs the best with 15% of whole network sample. Regarding which sampling method is better for a specific type of network, Ebbes et al. (2008) found that SBS work the best for the networks with Poisson degree

distribution and Random Walk (Markov Process) is the best for the power-law degree distribution.

The most relevant previous study to this essay would be from Chen and Chen (2008). This study focuses on how the estimate of social influence is influenced by the choice of sampling technique and showed that samples from network lead to underestimate of social influence. Using spatial regression model, they found that SBS has a disadvantage in measuring the effects of social interaction in a scale-free network.

However, Chen and Chen (2008) study focuses on a spatial model, not a diffusion of innovation model. Furthermore, different models might have different impact from the selection of sampling as well as network structure because the model specification for social influence or interactions is different. I will examine how the social influence (social pressure) defined in Bass and my NBB model is affected by network sampling methods. I will identify the best sampling method across different network structures. The remaining of Essay 2 is organized as follows.

First, I will attempt to describe how I simulated the different types of network graphs and the range of graph parameters as well as how network structure and graph parameter alter the shape of diffusion curve through social interaction. Next, I will compare the estimates of the Bass and the NBB model using two sampling methods, RNS and SBS. Final analysis includes sampling of real world network and estimation of models to revalidate the findings in my simulation. This essay will conclude with summary of findings and discussion.

2.3 Simulation on Network Structure and Sampling

2.3.1 Network Simulations, Degree Distributions, and Diffusion Curves

Using the procedure suggested by Erdos and Rényi (1960), random graphs with Poisson degree distribution were generated. Because random networks with various level of density have been examined in Essay1 and I need to compare the networks types of the same density, all types of networks have the same density of 0.002 which is in the range of the most frequently found network densities in empirical study in Essay 1. Figure C13 shows an example of random network. It is evident that there is no hub or locally heavy cluster, i.e., average number of linkages varies little across members.

A set of scale-free networks (with the same density of 0.002) were generated with different power-law distribution parameters, ranging from 0.1 to 3. Figure C14 is example of network graphs with the same density but different power-law distribution parameters. Scale-free networks with low power-law parameter (0.01) yields graph similar with Poisson degree distribution, but for high power-law parameter links are highly centered on a single network member. It is easy to speculate that the diffusion patterns for these networks would be very dissimilar.

From Figure C15, the degree distribution from low power-law distribution parameter shows similar curve with Poisson distribution. There is no member with a high degree. Also, the diffusion pattern has smooth curve and gradual diffusion of innovation over long period time. In contrast, the power-law parameter in Figure C16 shows that a small fraction of members (in this network only one member) dominate most of the connections. The diffusion curve shape in Figure C16 is marked by a sudden jump in new adoptions when a hub adopted and massive new adoptions follows. The diffusion process reaches a full marketing potential immediately after the adoption by the hub member.

2.3.1.1 Scale-free Network

I utilized two sampling methods, RNS and SBS. The purpose of this essay is to examine how the selection of network sampling method affects the estimates of diffusion parameters and to provide guidance for data collection procedures.

In addition, I attempted to find how sample sizes contribute to recovering the true model parameters. I set the population size as 20,000 and sample sizes from 1% to 30% of population size. The true parameter for innovation coefficient (p) is fixed at 0.025, which is within the rage of most commonly founded value in empirical studies. Since the one of main foci of this essay is to examine the interaction between sampling methods and the estimates of social influence (q). I adopted two different degrees of social influence, 0.2 and 0.4 which also fall within the range of estimates from the previous empirical studies. More importantly, to see how clusters in a scale-free networks affect the parameter estimates, I simulated the networks with different level of clustering by varying scale-free network parameter (i.e., power law distribution parameter) from 0.1 to 3. For each combination of the true diffusion parameters, scale-free network parameters, and different sample sizes, the estimation of the Bass model and the NBB model has been repeated 10 times and average of estimates these 10 repeats are reported.

Figure C17 shows the estimates of p from different sample sizes with the true diffusion parameters, $p = 0.025$ and $q = 0.4$. As one might expect, recovered parameters in the Bass model do not depend on sampling methods since it does not pertain any information on the network. Estimates of the innovation parameter from both RNS and SBS show consistent bias, especially network is characterized by a high level of global clustering (high value of power-law distribution parameter on x-axis).

However, there is a significant interaction between sampling methods and size and the NBB model in term of innovation parameter. *p*.NBB.ests (*p* estimates of the NBB model) from RNS are highly varying across samples sizes, and they tend to be overestimated for the scale-free networks. Estimated innovation parameters on scale-free network with power-law distribution parameter over 2 considerably overestimate the impact of external influence on new product adoptions. Figure C17 clearly indicates the superiority of SBS for NBB model over RNS. All lines for different sample sizes stay close on the true parameter line (dotted line). With minimal influence of sample size, SBS correctly recovered the true parameter on the innovation coefficient using the NBB model.

Figure C18 demonstrates the same finding for social influence estimates (*q*.NBB.est). The Bass model underestimates social influence for the less clustered network while it overestimates the influence for the scale-free network. For the NBB model, RNS yields highly unstable estimates for scale-free networks. Therefore SBS is the best sampling method for the NBB model and the NBB model performs extremely well in recovering true parameter of population from sampled network. Its estimates are not much affected by the sample sizes.

Furthermore, I found that if there is stronger social influence in the diffusion process, the biasness in the model estimates will be aggravated. We can see in Figure C19 and Figure C20 where the true social influence coefficient is very high, q=0.6, that the Bass model suffers from more in case of the scale-free networks. Even though the NBB model's performance on low power-law parameter has been compromised a bit, it

proves a strong robustness against network structures and different levels of social influence.

2.3.1.2 Sampling Methods and Density of the Scale-free Network

Boxplots in Figure C21 compare the estimates of the Bass model and the NBB model from RNS across different average number of connections (density) of network members on x-axis. Average number of connections are defined by network density * network size. At each density level, the estimates of both models from different sample sizes are shown in each boxplots. The location of the boxplot indicates where the estimates from the model lie and the height of boxplot shows how the estimates varies across different sample sizes.

With RNS, the NBB model overestimates both the innovation coefficient and the social influence coefficient. When scale-free network has low density, the NBB model returns q estimates that are highly unrealistic (over 1). Although the Bass model yields more stable estimates on different densities, they still depend on network density and are distant from the true parameters.

In contrast, the snowball sampling results show the superiority of the NBB model over the Bass model. Evidence from Figure C22 are that the NBB model with SBS performs well on all network densities for both p and q parameters. Its estimates are consistently close to true parameters (p=0.025, q=0.4) from low density to high density. Variations of estimates due to different sample size are much lower than those of Bass model. Therefore, I conclude that the NBB model and SBS is the best candidate for scale-free network and it is the most efficient combination in terms of sample size.

2.3.1.3 Random Network

To simulate a random network is very straightforward because there is no cluster or hub and the number of connections follows a Poisson distribution. Generating a random network only requires network size for the given network density because there are equal probabilities that two nodes are connected. For a scale-free network, if a member has a higher connections with others there is more probability that a new member will connect to her/him in the graph generating process. We do not need a parameter for network structure (clustering) to generate a random network. This allows me to plot the estimates across different network density level in Figure C23 and Figure C24.

Apparent from both figures is that the NBB model estimates from SBS are better recovering the true parameters of innovation coefficients from a random network than the estimate from the scale-free network, and the variations in estimates are much smaller than those from the scale-free networks.

Furthermore, I confirm my finding in Essay 1 on sampled data from Figure C23 and Figure C24, when the network is sparse, the Bass model underestimates the impact of social influence and overestimate external influence. As the network structure becomes denser, the Bass model estimates converges to the true parameters. This also revalidates my finding in Essay 1 that there is the strong negative correlation between these two parameter estimates. However, the NBB model estimates are consistently close to true parameters regardless of the network density levels when SBS is used to collect data.

2.3.1.4 Watts and Strogatz Network

WS (Watts and Strogatz) network is similar with random network in that there is no hub, but it shares the characteristic of the scale-free network, i.e., the small-world

phenomenon. Since the network structure in the WS network is similar with the random networks (no hub) the diffusion curves are closer to ones from the random networks; symmetric and steady increase and decrease of new adoption over time.

For WS networks, both estimates of the NBB model performs better on data from SBS, which is the same finding from the previous two network types. We also observe the negative correlation between two Bass model estimates from the sparse networks and these bias disappear as the network density increases.

2.3.2 Interaction Between Network Types and Models

To match the best sampling method with the diffusion models, I tested the Bass and my NBB model performances across different network types and network sampling methods through simulations. Table B5 summarizes findings of my simulation study.

For the Bass model, there is no appropriate sampling to recover the true parameters when network is sparse. Inaccuracy in the Bass model estimates suffers more in case of the scale-free networks since diffusion pattern on this network is more dependent on whether a hub adopts new product or not. When a hub becomes an adopter, a sudden peak of sales follows, which is properly captured by my NBB model. The NBB model does recover the true parameters when we use RNS. The estimates of the NBB model from RNS depends on sample size. Therefore, I conclude that RNS is not desirable sampling method for the NBB model. However, the NBB model consistently shows better performance across different sample sizes and network types when SBS is selected. Therefore the finding in my simulation study strongly suggests that the NBB model and SBS is the best match for accurate parameter estimation. In the next section, I will test whether these simulation findings holds in the real world.

2.4 Empirical Study

Last.fm claims that 30 million active users (Jones 2009). Due to large size of network size of Last.fm approximately 400,000 users' data have been sampled from this website. First I randomly selected 300 sub communities on Last.fm among 5,000 groups which have group member size below 10,000 but greater than 500. Then I collected all members' data including connected users and music library data, within these 300 sub communities. I assume that there will be more probability for members in the same sub-communities to be connected. This sampling method will better captures the clusters in the real population network. I consider this set of 400,000 users as the entire population for this empirical study.

2.4.1 RNS Vs. SBS and Empirical Model Fits

Among the 400,000 users, I randomly sampled 20,000 members using RNS to fit the Bass model and the NBB model. Identical estimation procedures have been used to estimate the innovation and social influence parameters. Among songs that I have been traced since their first release after January 1st 2012, I selected top 50 songs that have the most adaptors. Once I fit the Bass model on these 50 songs, I excluded the songs data that had negative social influence coefficient from the Bass model (q.OLS <0). This leaves 44 songs to compare the performance of various sampling methods.

For the sample size of RNS, I initially set sample size as 10,000, but the NBB model do not return estimates so I doubled the sample size to 20,000. SBS was applied for the sample size 10,000, half size of RNS. First I selected a seed node and collected all of his friends (first layer of SBS), and then I further recruited a seed node's friends' friends (second layer). In the middle of the third layer of snowballing I reached the

predetermined sample size of 10,000. The members included in SBS data set are different from the 400,000 data set available.

Assuming the estimates from the bigger sample size are close to the true parameters of the population, I compared the Bass model and the NBB model estimates from RNS and SBS with those from all 400,000 data set across 44 songs. Figure C27 shows the two estimates (*p* and *q*). X-axis represents the 44 songs in this empirical analysis and Y-axis shows the parameter estimates from RNS, SBS, and all 400,000 data set. Songs on x-axis have been sort by increasing order based on the *q* estimates using all 400,000 user data set.

Since we do not know the true parameters unless we have population data, the efficiency of sampling method can be guessed by comparing the estimates from the bigger sample size with those from smaller sample sizes. In Figure C27, the Bass model estimates (p.OLS and q.OLS) indicates minimal differences from two sampling methods. However, as I showed in essay 1 the Bass model gives us biased estimates if we assume Poisson-binomial adoption process.

Figure C28 is a key finding of this essay. The *p* estimates of the NBB model (p.NBB) from both sampling methods do not show significant improvement over those of the Bass model and *q* estimates of the NBB model (q.NBB) been highly overestimates using RNS. Where *q* estimates for all data set are high, RNS data shows *q* estimates over upper boundary of 1. However, considering much smaller sample size (400,000 Vs. 10,000), SBS sampling yields the NBB model *q* estimates (red line on bottom left in Figure C28) close to those using 40 times higher sample size.

2.4.2 Sampling Methods and Captured Network Structure

The reason why RNS is not a good choice for the NBB model is that RNS cannot capture enough links to properly estimate social influence coefficient in the NBB model. Figure C29 shows network graphs among adaptors of a song drawn from the samples by RNS and SBS. Graph from RNS shows less links and also there is no distinct hub (which is red color). SBS allows us capture more links and clusters presented in the graph, which is essential to fit the NBB model.

Furthermore, the graph generated by SBS is more similar with the graph by using all data in Figure C30. This similarity of graphs between larger data set and small data from SBS here leads to parallel estimates of the NBB model, thus SBS is more economical sampling method from the large social network.

2.5 Conclusion

2.5.1 Summary and Conclusion

The main foci of Essay 2 are 1) the most efficient sampling method for the NBB model and 2) testing the NBB model on other network structures. The former focus is more geared toward the practical applicability of the NBB model. Since the NBB model requires all adopters' network information, it is very expensive model and often these whole network information is not readily accessible. I examined whether from sampled data the NBB model can recover the true parameters. This essay suggests that snowball sampling provides us the most efficient solution for the large social network. Because SBS detect more connections in the network than RNS the estimates from the NBB model with SBS approximate the true parameter well. Moreover, the NBB model is free from the edge effect (Griffith 1985) when using SBS. It recovers the true parameters

from the sub-region of whole network. Similarly in empirical study I found that with small sample size from SBS the NBB model provides us very close estimates to those from much large sample.

The latter focus is intended to test the idea that by incorporating all the network information into to the model we can apply such model to any network types. The NBB model was proven to be robust to different network types. In sum this study shows the my NBB model still can be a useful (and also accurate) tool when there is less information about the social network and limited resource for data collection.

2.5.2 Limitations and Future Directions

Although my study suggests that SBS is very good idea for estimating, caution should be exercised when generalizing this finding since SBS can be swamped in a typical group of cluster (Chen and Chen 2008). For example, if all samples from SBS are collected from the group of rock music fans who are connected tightly, the sampled data do not represent population at large. One possible alternative for this issue will be multiple SBS starting from different seed nodes that have various degrees or locations in the network. SBS with multiple seed nodes would be interesting future study especially population network is composed of diverse members.

Empirical analysis in this study used large size sampled data as a proxy for a population network. Test of sampling methods was performed on an assumption that larger sample will provide more accurate estimates on population parameter. Although this is true in general, obtaining whole real world network data thus comparison with the true parameters would provide a firm conclusion on the findings in this study.

CHAPTER 3: HETEROGENEOUS SOCIAL INFLUENCE MODERATED BY PREFERENCE COMPATIBILITY

3.1 Literature Review

Extensions of diffusion models have been popular with recent advances in social network studies. This parallel research boom deepens our understanding of the social contagion of behaviors or ideas. There have been studies on heterogeneous susceptibility to adoption, due to varying degrees of individual ties to prior adopters (Brown and Reingen 1987). However, this heterogeneous adoption probability is explained only by external variation across individuals, such as the number of friends or the level of clustering. Disregarding individual specific variation in adoption probability may oversimplify both internal and external influences in the diffusion model, but there is little research addressing this issue.

I will examine my idea that an individual's susceptibility to adoption of new product by her/his adaptor friends (social influence) would depend on her/his preference compatibility with friends, i.e., the higher preference compatibility with a friend, the higher social influence from that friend is.

In his study on new product adoption, Rogers (1983) defined compatibility as "degree to which an innovation is perceived as consistent with existing values, past experience, and needs of potential adopters." He argues that compatibility affects the rate of adoption of innovations. Rogers (1983) also suggests that homophily (love of the same), the similarity between people in term of certain attributes, forms stronger ties, and Brown and Reingen (1987) found that homophily facilitates word-of-mouth.

It is a reasonable assumption that people with similar preference have strong ties. Therefore, they are more susceptible to influence from a peer with similar tastes and vice versa. For example, a person might not listen to a new song in the rock music genre, even if all of her friends are listening to it if she has no preference for that type of music. On the other hand, if she is a fan of rock music (strong preference), observing even one friend listening to new rock song may be sufficient to turn her into an adopter. This study will test this conjecture by modeling heterogeneous susceptibility to internal influence due to *preference compatibility* among network members.

To my best knowledge, this is the first attempt to model tie strength that is an individual trait and to incorporate this heterogeneous word-of-mouth effect into a new product diffusion model.

3.2 Model Development

3.2.1 Distance as a Measure of Preference Compatibility

The first step is to create continuous measures for preference compatibility between connected individuals is to position each individual on a preference map. Once I create the preference map, I will calculate distances among all connected individuals. These distances indicate how much two connected persons' preferences are similar or dissimilar (preference compatibility.) If the two friends are listening to different genres of music (low preference compatibility), the distance between them on the map will be greater and vice versa. For example, individual A in Figure C31 has two friends, B and C, but C is located far away from A, which implies their preferences are very different (she likes rock music exclusively) thus I assume social influence of A on C will not as strong as on B who is closer to A.

3.2.2 Decaying Social Influence by Distance

Tobler's "First Law of Geography" (Tobler 1970) states that interaction between two locations decreases as distance between them increases. Therefore, I assume that the degree of social influence between two connected people diminishes as their location on the preference map becomes farther away. In addition, I also assume that social influence has always as positive impact on new adoption (greater than or equal to zero.) To measure the degree of decrease I add "Distance Decaying Parameter", β for distance measure between two connected people, d_{ij}.

$$\text{Social Influence between } i \text{ and } j = \begin{cases} 1-\beta d_{ij} & \text{if } 1-\beta d_{ij} > 0 \\ 0 & \text{if } 1-\beta d_{ij} < 0 \end{cases}$$

where β = distance decaying parameter

d_{ij} = distance between connected individual i and j on the preference map

Note that when $\beta = 0$ there is no difference among connected people in terms of magnitude of social influence (homogeneous social influence). The relationship between the level of social influence and distance decay parameter is shown in Figure C32. The high value of β indicates peer influence works only between two friends who shares very similar tastes. The value of $\beta \neq 0$ implies the peer influence in fact takes a continuous form, which mean there is heterogeneous contagion effect. When $\beta = 0$ there is no interaction between the magnitude of social influence and preference compatibility. The model reduces to the NBB model, which assumes homogeneous social influence (all social influence among connected people equals 1).

3.2.3 NBB model Incorporating Preference Compatibility

I introduce this modification to accommodate the following assumption: the higher preference compatibility (short distance), the stronger the social influence is. This heterogeneous probability of adoption can be modeled through the imitation term in the NBB model by redefining adjacency matrix as follows:

$$S_t = p\left[m - Y_{t-1}\right] + q\left[\mathbf{a_{t-1}}\mathbf{C_g^*}\left(\mathbf{1_m} - \mathbf{a_{t-1}}\right)^{\mathbf{T}}\right]$$

where $\mathbf{C_g}[i,j] = \begin{cases} 1-\beta d_{ij} & \text{if } 1-\beta d_{ij} \geq 0 \\ 0 & \text{if } 1-\beta d_{ij} < 0 \text{ or } i \text{ and } j \text{ are not connected} \end{cases}$, $\mathbf{C_g}[i,j] \in [0,1]$

β = distance decaying parameter

d_{ij} = distance between individual i and j on the preference map

$\mathbf{C_g^*}$ is a colmun standardized matrix of $\mathbf{C_g}$

$\mathbf{a_t}$ is an adopter vector and $\mathbf{a_t}[i]$=1 if individula i is an adopter by time t

S_t = number of new adoptor at time t

m = market size parameter

p = innovation parameter

q = social influence (imitation) parameter

Y_t = cumulative number of adopters at time t

The new adjacency matrix will compute the tie strength by measuring the preference similarity among network members. I will call this model "NBB.Pref model"

hereafter. The NBB.Pref model is a special case of the NBB model because when $\beta = 0$ the NBB.Pref model is equivalent to the NBB model.

3.2.4 Social Pressure and Decaying Social Influence by Distance

In the NBB model, the denominator for social pressure term, (r_{it}) is calculated by the total number of friends, i.e., sum of binary measures indicating connections. Since I created continuous measure for each tie in the adjacent matrix, new social pressure term in the NBB.Pref model provides us new interpretation on this term.

Using the column-standardized adjacency matrix of $\mathbf{C_g}$,

$$r_{ij}^{*} = \mathbf{C}_{\mathbf{g}}^{*}[i, j] = \frac{\sum_{j \in A_{t-1}} \left(1 - \beta d_{ij}\right)}{\sum_{k \in U} \left(1 - \beta d_{ik}\right)}$$

Where U = a set of ultimate adopter friends of individual i

A_t = a set of adopter friends of individual i as of time t

Now, the social pressure measure, r_{ij}^{*}, implies that when a friend (even one) with more similar tastes adopted the innovation, there will be more social pressure. On the other hands, this term also means that she might not feel much social pressure when most (not all) of her adopter friends are located far away in the preference map (highly dissimilar preference.)

3.3 Data and Measure of Preference

3.3.1 Mapping Individuals' Preferences Using Social Tags Data

Since its launch in 1996, social tagging (also called "social bookmarking" or "online tagging") has become a major characteristic of online content. Social tagging allows users to categorize the contents and share collaborative vocabulary on the contents (new products) with other users. Last.fm summarizes social tags from users for each song as shown in Figure C33. Users in Last.fm can put tags on the songs they listened to and organize their music libraries. The tags of a song provide us with aggregated opinions on the product characteristics. For example, in Figure C33 the most frequent two tags regarding the genre of "Party Rock Anthem" are "dance" and "electronic" (with the largest font).

In a user's music library, Last.fm provides 20 tags that are most popular among the tracks on each webpages (Figure C34). During one year of data collection period, a user has a several hundred track history webpages (Figure C34) on average, which means I have a several hundred sets of top 20 tags for that user.

This study suggests a novel way to utilize this social tag data to measure individual's preferences. Moreover, this is the first study to utilize social tag data in the marketing literature. From the sets of 20 tags collected from approximately 400,000 users, I selected 34 music genre (product characteristics) related tags such as "pop", "hip-hop", and "rock" etc. Next I calculated the proportion of frequency of each 34 genre tags across all sets of top 20 tags for each users during the year 2013. This proportional measure (x_{ik}) indicates how much time an individual spends on listening to a certain genre of a music compared to all of her time listening music. I use this consumer's past

usage behavior data to measure her/his music preference. The degree of preference for each genre are defined as follows:

$$x_{ik} = \frac{\sum_{v} x_{ikv}}{\sum_{k}\sum_{v}\left(x_{ikv}\right)}$$

$$= \frac{\text{Sum of freqency of genre } k \text{ in individual } i\text{'s top 20 tags}}{\text{Sum of the frequencies of all genre tags in individual } i\text{'s top 20 tags}}$$

Where i = individual,

v = page number of a user's music library

k = genre

x_{ik} = coordinate for genre k axis for individual i

After calculating preference on each genre, I use these data as coordinates to plot the individual preferences on a map with 34 genres on axes. Figure C31 shows the 3 dimensional preference map simplified with only 3 genres: hip-hop, pop, and country.

Since preference measures are proportions which always sums to 1, they are compositional data that span a 34 dimensional simplex. Therefore I use to Aitchsion Distance (Aitchison 1982), which is the best measurement for distances on compositional data (Martín-Fernández et al. 1998; Otero et al. 2005).

$$\text{Aitchison Distance} = d_{ij} = \left[\sum_{k=1}^{K}\left(\ln\left(\frac{x_{ik}}{g\left(x_i\right)}\right) - \ln\left(\frac{x_{jk}}{g\left(x_j\right)}\right)\right)^2\right]^{\frac{1}{2}}$$

where x_{ik} are user i's coordinates for genre k

i, j = inidvidual

$g\left(x_i\right)$ = geometric mean of x_i

3.3.2 Song Data

Individual level data has been collected for 400,000 users from Jan 1st 2011 to December 31st 2012. Among songs released after Jan 1st 2012, the top 50 songs with the most adopters from collected data set are included this analysis. However after fitting the Bass model, I excluded six songs that yielded negative estimates for imitation (social influence) coefficient. For 44 songs included, the daily new adoptions and adjacency matrixes with preference compatibility have been computed to fit the NBB.Pref model.

3.4 Empirical Analysis

3.4.1 Dynamics of Distance among Interacting Users

I believe that it is meaningful to examine how distances between adopters and non-adopters change over the product life cycle. For this quick descriptive study, I calculated average distances between connected adopters and non-adopters.

$\tilde{d}_t$ =mean distance between interacting adoptors and non-adoptors at time t

$$= \frac{\mathbf{a_{t-1}C_d}\left(\mathbf{1_m - a_{t-1}}\right)^{\mathbf{T}}}{\mathbf{a_{t-1}C}\left(\mathbf{1_m - a_{t-1}}\right)^{\mathbf{T}}} \text{ where } \mathbf{C_d}[i,j] = d_{ij} \text{ if } i \text{ and } j \text{ areconnected o.w. } 0$$

The denominator is the social interaction term in the NBB model and it standardizes the distances in term of matrix size (number of ultimate adopters). Figure C35 shows typical pattern I observed across most of songs. At the early stage of product diffusion, interactions happen among closely located members in the preference map and distances are increasing overtime. 5

To statistically test this finding, I applied simple OLS, where $\tilde{d}_t = \alpha + \delta \times t + e_t$ where t=week. In Table B6, 40 out of 44 songs' $\hat{\beta}$s are positive and significant, supporting my assumption that preference compatibility decreases among interacting members as time increases. In other world, early adopter friends have more similar preferences than laggards do not.

I also found that $\hat{\alpha}$ are negatively correlated with imitation coefficients estimates of the Bass model ($r = -0.28$) and correlation between $\hat{\delta}$ and imitation coefficients is also negative ($r = -0.11$). These correlations shows that strong preference compatibility in the network accelerates the diffusion process.

3.4.2 Estimation of the NBB.Pref Model

3.4.2.1 Model Modification and Estimation Procedure

NLS (non-linear least squares) has been chosen to estimate parameters for the NBB.Pref model because the social influence parameter (q) and decaying distance parameter (β) are not linear. One obstacle in applying NLS is that NLS uses numerical approximation to estimate parameters and the NBB.Pref model has a large adjacency matrix, which requires extensive computation time. Therefore I devised two remedies. First I used a smaller sample collected by SBS (snowball sampling). Essay 2 shows that SBS is efficient and precise sampling method. Secondly, to reduce calculation time in a statistical software (in my case R), I revised the previous NBB.Pref model to the equivalent form that expressed by matrix form to reduce computing time.

$$S_t = p\left[m - Y_{t-1}\right] + q\left[\mathbf{a_{t-1}C_g^*}\left(\mathbf{1_m} - \mathbf{a_{t-1}}\right)^{\mathbf{T}}\right]$$
$$= p\left[m - Y_{t-1}\right] + q\mathbf{D}[t,t]$$

where $\mathbf{D} = diag\left(\mathbf{A_{t-1}C_g^*}\left(\mathbf{1_{tm}} - \mathbf{A_{t-1}}\right)^{\mathbf{T}}\right)$

$\mathbf{A_t}$ is $t \times m$ matrix and $\mathbf{A}[t.] = \mathbf{a_t}$ and $\mathbf{A}[1.] = \mathbf{0}$

$\mathbf{1_{tm}} = t \times m$ matrix with all entries equal to 1

The t^{th} element in a diagonal matrix, $\mathbf{D}[t,t]$, matrix is equivalent to $\mathbf{a_{t-1}C_g^*}\left(\mathbf{1_m} - \mathbf{a_{t-1}}\right)^{\mathbf{T}}$ in the NBB.Pref model.

Since I assume that the individual adoption process follows Bernoulli trials with different success probabilities (Poisson-binomial process) which are bounded below 1, I impose restriction on the upper bound of each social pressure to be 1 for all user, i.e., $q\left(1 - \beta d_{ij}\right) = 1$ if $q\left(1 - \beta d_{ij}\right) > 1$.

Starting values of NLS method are given by the grid created by each combination of three parameters. For the external influence estimate, grid has only one value that is the estimate from the NBB model to save computation time. The range of values for social influence was from 0 to 10, and the values for decaying distance parameter range from 0 to 0.5 which cover most of possible values for them based on Figure C32. Social Influence and Distance Decaying Parameter

3.4.2.2 Summary of Analysis

I selected 46 songs with a market size above 200 for an initial fit. I excluded 14 songs with negative estimates for the imitation coefficient from the Bass model and also removed 6 songs whose NLS estimations had not been converged. Four of these 6 songs

have the lowest imitation coefficient from the Bass model. Finally estimates of 26 songs by the NBB.Pref model are reported in Table B7. Innovation coefficient estimates from the Bass model range from 0.001 to 0.019, and imitation coefficient estimates range from 0 to 0.06. These ranges are much lower than the typical ranges of both parameters. Mahajan et al. (1995) found that the average $\hat{p} = 0.03$ and $0.3 < \hat{q} < 0.5$. I suspect that this is due to atypical diffusion patterns for new songs, e.g., they are highly asymmetric and have multiple peaks.

The distribution of $\hat{\beta}$ is spread out within the range of the values in the grid that I tried in NLS, which indicates that NLS found the optimal values of $\hat{\beta}$ within the given range of the grid (Otherwise, NLS will find the solution at the maximum or minimum value, 0 or 0.5). The imitation parameter estimates ($\hat{q}_{NBB.Pref}$) are skewed to the right; most of their values lie below 1. Two of $\hat{q}_{NBB.Pref}$ are at the upper bound of the grid of NLS.

To show how the level of preference compatibility influences market size (m), I computed the average distance among connected adopters ($\tilde{d}$) and then regress m on $\tilde{d}$. Regressing the 26 songs from the data shows that there is a significant (p-value=0.008) negative relation (-278) among them. This suggests that less successful songs have a lower degree of preference similarity. In other words, if the song finds a highly homogeneous group of people, the dissemination of innovation could reach farther into the network. However, this conjecture needs more exhaustive examination.

The estimates of $\hat{\beta}$ for songs are significant at a p-level=0.05 for 16 out of 26 songs. The $\hat{q}_{NBB.Pref}$ are significant, except for 8 songs. For 12 songs, the parameter

estimates are statistically significant. Based on this result, I argue that new preference compatibility contains information to explain adoption behaviors via peer influence.

Next, I computed the MSE[7] (Mean Squared Error) to compare the fit of three models, NBB.Pref, NBB, and Bass. After calculating MSE, I computed the difference in MSE between the two models divided by its own MSE. For example, the fit comparison between the Bass model and the NBB model is given by the following formula:

$$MSE_{BASS-NBB} = \frac{MSE_{BASS} - MSE_{NBB}}{MSE_{BASS}} \times 100$$

Where $MSE_{BASS} = MSE$ from the Bass Model

A positive value indicates that the NBB model fits better than the Bass model.

It is interesting to observe that model fits (*MSE*) are strongly correlated with distance ($\tilde{d}$). Figure C36 shows the fit comparisons, the NBB model vs. the NBB.Pref model ($MSE_{NBB-NBB.Pref}$) and the Bass model vs. the NBB.Pref model ($MSE_{BASS-NBB.Pref}$). Note that negative values in Table B8 imply that the NBB.Pref model fits are not better than the corresponding model ($MSE_{BASS} < MSE_{NBB.Pref}$). As the distances among adopters increase, the performance of the NBB.Pref model deteriorates.

I further investigated this finding by a correlation matrix. Table B9 clearly shows that mean distances have a strong significant relationship with other estimates and

---

[7] $MSE = \left(S_t - \hat{S}_t\right)^2 \Big/ T$

measures. In fact, mean distance is negatively related with $MSE_{NBB-NBB.Pref}$ (ρ = -0.42) and $MSE_{BASS-NBB.Pref}$ (ρ = -0.27), which implies that when a network has a higher degree of heterogeneity, we should opt for the NBB model or the Bass model. I conjecture that when there are too much differences in preferences among adopters and sample size is small (small market size) the NBB.Pref model cannot identify β, resulting poor performance. This is somewhat evident from Table B7. $\hat{\beta}$s in the upper portion of Table B7 are not statistically significant where the average distances ($\tilde{d}$) are high thus the market sizes (*m*)are smaller. Future investigation with bigger sample size can verify my conjecture.

3.5 Conclusion

3.5.1 Summary of Findings

By incorporating an individual's preference information via the adjacency matrix, I enhanced the NBB model, the NBB.Pref model. The NBB.Pref model, in general, fits for the empirical data better. More importantly, it utilizes rich information available these days and provides us more insight into customer adoption behavior. For example, utilizing preference information, I showed that early adopters have similar preferences so word-of-mouth works very quickly and more widely leading to a larger averaged market size.

The distance decaying parameter contains important managerial information. A higher value of distance decay implies that social influence is limited to friends whose preferences are highly compatible. The estimates of the distance decaying parameter are correlated with imitation estimates in the Bass model, indicating that this preference

compatibility influences the word-of-mouth effect. This provides insight with why and when buzz-marketing did or did not work. The distance decaying parameter may answer for this question. Furthermore, if the peer influence cannot reach out because of low level of preference similarity in the network, a manger might go for other marketing approaches other than buzz-marketing.

One challenge of big data is how to visualize and summarize big data. By creating the preference map, I showed how we can plot these large scale data on past usage behavior and summarize them with a single measure (distance). With the preference map, we can see how customer preferences are distributed in term of each product characteristics.

3.5.2 Limitation and Future Directions

I found that there is some degree of correlation between innovation parameter estimates and distance decaying parameter in the covariance matrix of parameter estimates. This indicates that the external influence also is moderated by an individual's preferences. Incorporating preference compatibility between product characteristics and a customer via external influence term in the NBB model would be interesting.

From the average distance data over time in Figure C35, we see that early adopters are located nearby in the preference map. It maybe be fruitful to examine where are the optimal seeding points in the network are, using with preference information. One promising research direction would be to examine how different seeding in term of network structure and preference compatibility affects the diffusion pattern.

There might be a model identification issue between social influence and the distance decaying parameter. Follower research is need to better separate these two

parameters during estimation process. Future modeling should also look into the limitation of the current NBB.Pref model that requires long calculation times and high computing power.

Lastly, we gain more insight into the underlying adoption process through preference compatibility if we depart from the Bass model frame work and develop individual level dependent variable. Current NBB.Pref model aggregates new adoptions as a dependent variable. We might lose some information on individuals by using this aggregated measure. A micro-model such as non-linear discrete-time hazard model may provide more insights into the micro-level adoption process.

CHAPTER 4: GENERAL CONCLUSION AND DISCUSSION

Recent unprecedented developments in information technology have brought us into a data-rich environment. This data-driven world, along with enhanced computing power, excites academics, as it allows us to test theories and models that were previously impossible. For example, I can collect whole online network information, including individual-level behavior data by web-crawling if the data are in the public domain. Most online network managers can easily obtain such big data. However, there is little literature on how we can take advantage of these big data.

My dissertation explores the underlying micro-level underlying adoption behaviors and models them into the Bass model, which is a macro model. This disaggregated approach to the Bass model has never been studied. The proposed individual adoption process is fully consistent with the Bass model. Another theoretical finding is that the Bass model assumes a fully connected social network; thus, it over-counts the number of social interactions and underestimates word-of-mouth impact. I showed that this issue is more evident when the social network is sparse in a simulation study. Online social networks further confirm that real-world social networks are sparse, and the Bass model always returns lower estimates for social influence than my NBB model does. I argue that by using the Bass model, fewer resources for buzz marketing might be allocated than what is appropriate.

In the study of network sampling, I demonstrated that the NBB model estimates are robust against different types of social networks. More importantly, it provides us with accurate parameter estimates on populations with sampled data. Snowballing sampling is the most efficient sampling method for the NBB model, which recovers

population parameters precisely with small sample sizes. Therefore, the NBB model can be used when we do not have all of the network information.

The last study searched for simple methods to utilize big data and social tagging, which is a new data source. I web-crawled over 2 billion data points on 400,000 online users' music- listening behaviors over one year. Using social tagging data, I measured each individual's relative preferences on product characteristics (music genre) and drew a map to visualize these big data. I believe that my preference map drawn from the big data can provide managers with an easy-to-understand summary and useful information about the market. I further demonstrated how to combine information from the preference map with the NBB model. By modifying the adjacency matrix, I successfully enhanced the NBB model, which tells us how different levels of preference structures in the network influence new product success.

Although the three essays add theoretical contributions to the best-known diffusion model and introduce a novel approach to understand individual-level adoption behavior, they still need to be supplemented by further investigation. First, the last study regarding the impact of preference compatibility on social influence can use other dependent variables than aggregated numbers of new adoption. Two new models in this study were drawn from the Bass model, which is an aggregate model. More insights into individuals' behavior can be drawn if we look at each individual's adoption behavior separately.

Secondly, the NBB.Pref model, using the nonlinear least squares method, did not converge for some songs. It is not clear whether no convergence is due to the small market size or whether the model needs to be redefined. Future research with more

samples should follow. Also for the network sampling study, I did not use population data to compare the performance of network sampling methods; rather, a population parameter should be used as a reference.

Thirdly, I showed one way to use social tag data that has never been previously tried in the marketing literature. Since social tags change as more adopters put tags on new products, a more important question might be whether social tag data have predictive power for new product success, especially in the early stages.

Finally, I believe that a preference map has much potential as a market-analyzing or data- summarizing tool. For example, a study on whether this map is useful in identifying the optimal cluster for seeding in a new product campaign is interesting. It is one of the most key questions among new product managers. Subsequent studies on the applications of preference maps will be both promising and fruitful.

APPENDIX A: PROOFS

A1. Proof of the equivalence between the Bass model and the NBB Model

Theorem: For a fully connected adjacency matrix $\mathbf{C^F}$ and a large m,

$$S_t = pm - Y_t + q\left[\mathbf{a_t C_F^*}\left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}}\right] = p\left[m - Y_t\right] + q\left[\frac{Y_t}{m}\left(m - Y_t\right)\right]$$

Proof:

$$S_t = mp + \left[q - p\right]Y_t - \frac{q}{m}Y_t^2 \quad \text{(Frank M. Bass, 1969)}$$

$$= p\left[m - Y_t\right] + q\left[\frac{Y_t}{m}\left(m - Y_t\right)\right]$$

Let $\mathbf{a_t}$ be a $1*m$ row vector and $a_{it}\begin{cases}1 & \text{if } i \text{ is an adopter at time } t\\ 0 & \text{otherwise}\end{cases}$ and

$\mathbf{1}_m$ be a $1*m$ row vector with all the entries equal to 1. Then:

$$Y_t = \sum_i^m a_{it} \text{ and } m - Y_t = \sum_i^m \left(1 - a_{it}\right)$$

Let $\mathbf{C}$ denotes a $m \times m$ adjacency matrix, i.e. $c_{ij}\begin{cases}1 & \text{if } i \text{ and } j \text{ are connected and } i \neq j\\ 0 & \text{otherwise}\end{cases}$

Suppose all i and j are connected to each other, i.e. $c_{ij} = 1$ for all $i \neq j$

i.e. $\mathbf{C_F}$ is the adjacency matrix for the fully connected network then:

$$\mathbf{a_t C_F}\left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}} = \begin{bmatrix}a_{1t} & \cdots & a_{mt}\end{bmatrix}\begin{bmatrix}0 & 1 & \cdots & 1 & 1\\ 1 & 0 & 1 & \vdots & 1\\ \vdots & 1 & \ddots & 1 & \vdots\\ 1 & \vdots & 1 & 0 & 1\\ 1 & 1 & \cdots & 1 & 0\end{bmatrix}\begin{bmatrix}1 - a_{1t}\\ \vdots\\ 1 - a_{mt}\end{bmatrix}$$

$$= \sum_i^m \sum_j^m a_{it} c_{ij} \left(1 - a_{it}\right) = \sum_i^m a_{it} \sum_j^m c_{ij} \left(1 - a_{jt}\right)$$

$$= \sum_i^m a_{it} \sum_j^m \left(c_{ij} - c_{ij} a_{jt}\right)) = \sum_i^m a_{it} \left[\sum_j^m c_{ij} - \sum_j^m c_{ij} a_{jt} \right]$$

$$= \sum_i^m a_{it} \sum_j^m c_{ij} - \sum_i^m a_{it} \sum_j^m c_{ij} a_{jt} = \sum_i^m a_{it} \left(m - 1\right) - \sum_i^m a_{it} \sum_j^m c_{ij} a_{jt}$$

$$= Y_t (m-1) - \left[\sum_i^m a_{it} \sum_j^m a_{jt} - \sum_i^m a_{it}^2 \right] \text{ since } c_{ij} = 1 \text{ for all } i \neq j \text{ and } c_{ij} = 0 \text{ for all } i = j$$

$$= Y_t (m-1) - \left[\sum_i^m a_{it} \sum_j^m a_{jt} - \sum_i^m a_{it} \right] \text{ since } a_{it}^2 = a_{it} \text{ for all } i \text{ by definition}$$

$$= Y_t (m-1) - \sum_i^m a_{it} \left[\sum_j^m a_{jt} - 1 \right] = Y_t (m-1) - Y_t \left(Y_t - 1\right)$$

$$= Y_t (m - Y_t) \qquad (1)$$

Let the column-standardized matrix of $\mathbf{C}$ be $\mathbf{C}^*$, *where* $c_{ij}^* = \dfrac{c_{ij}}{\sum_j c_{ij}}$ then:

$$\mathbf{C}_\mathbf{F}^* = \frac{1}{m-1} \mathbf{C}_\mathbf{F} \qquad (2)$$

Given any arbitrary $\varepsilon > 0$, let us choose N s.t. $\dfrac{1}{N} < \varepsilon$ and,

we can find a $m(m-1)$ s.t. $m(m-1) > N$

$$\left| \left(\frac{1}{m} - \frac{1}{m-1} \right) - 0 \right| = \left| \left(\frac{m-1-m}{m(m-1)} \right) - 0 \right| = \left| \frac{1}{m(m-1)} \right| = \frac{1}{m(m-1)} < \frac{1}{N} < \varepsilon$$

$$\Rightarrow \lim_{m \to \infty} \left(\frac{1}{m} - \frac{1}{m-1} \right) = \lim_{m \to \infty} \left(\frac{1}{m(m-1)} \right) \to 0$$

$$\Rightarrow \frac{1}{m} = \frac{1}{m-1} \text{ for a large } m \qquad (3)$$

Finally for a large *m*,

$$\frac{1}{m}Y_t\left(m-Y_t\right)=\frac{1}{m-1}Y_t\left(m-Y_t\right) \quad \text{from (3)}$$

$$=\frac{1}{m-1}\mathbf{a_t C_F}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}} \quad \text{from (1)}$$

$$=\mathbf{a_t C_F^*}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}} \quad \text{from (2)}$$

$$\therefore\ p\left[m-Y_t\right]+q\left[\mathbf{a_t C_F^*}\left(\mathbf{1_m}-\mathbf{a_t}\right)^{\mathbf{T}}\right]=p\left[m-Y_t\right]+q\left[\frac{Y_t}{m}\left(m-Y_t\right)\right]$$

for a fully connected adjacency matrix **C** and a large *m*.

A2. Proof of the NBB Model

A2.1

Proposed Model:

$$S_t = p\left[m - Y_t\right] + q\left[\mathbf{a_t C_F^*}\left(\mathbf{1_m} - \mathbf{a_t}\right)^{\mathbf{T}}\right]$$

If $s_i \mid t \sim Bernoulli\left(p + q \times r_{it}\right)$

where $s_i \mid t$ = Binary random variable indicating a non-adopter i adopts at t

$$p = P\left(s_i = innovator \mid t\right)$$

$$q \times r_{it} = P(s_i = imitator \mid t)$$

Let $r_{it} = \begin{cases} \mathbf{a_t C_{.i}^*} & \text{if } i \text{ has at least one friend (connection) in the set of the ultimate adopters} \\ 0 & \text{otherwise} \end{cases}$

where $\mathbf{C_{.i}^*} = i$-th column of $\mathbf{C}^*$ and

$$r_{it} = \frac{\text{\# of } i\text{'s friends who adopted as of } t}{\text{\# of all of } i\text{'s friends in the set of the ultimate adopters}}$$

Then the expected value for the each potential adopter at the time t is

$$E(s_i \mid t) = p + q \times r_{it}$$

Assuming independence among the adopters at the current time t,

$S(t)$ follows Poisson-binomial distribution (Y.H. Chang, 1993)

$$E\left(S_t\right) = \sum_{i \notin A} E\left(s_i = 1 \middle| t\right) \text{ where } A = \left\{i \mid a_{it} = 1\right\}$$

$$= \sum_{i \notin A}\left(p + q \times r_{it}\right) = \sum_{i \notin A} p + \sum_{i \notin A} q * r_{it} = p\sum_{i \notin A} 1 + q\sum_{i \notin A} r_{it}$$

$$= p\left[m - Y_t\right] + q\sum_{i \notin A} r_{it}$$

A2.2

$$\sum_{i\notin A} r_{it} = \sum_{i\notin A} \frac{\sum_{j}^{m} a_{jt} c_{ji}}{\sum_{j}^{m} c_{ji}} = \sum_{i\notin A} \sum_{j}^{m} \left(a_{jt} \frac{c_{ji}}{\sum_{j}^{m} c_{ji}} \right) = \sum_{i\notin A} \sum_{j}^{m} a_{jt} c_{ji}^{*}$$

$$= \sum_{i}^{m} \sum_{j}^{m} a_{jt} c_{ji}^{*} - \sum_{i\in A} \sum_{j}^{m} a_{jt} c_{ji}^{*} = \sum_{i}^{m} \sum_{j}^{m} a_{jt} c_{ji}^{*} - \sum_{i}^{m} \sum_{j}^{m} a_{jt} c_{ji}^{*} a_{it}$$

$$= \sum_{i}^{m} \sum_{j}^{m} a_{jt} \left[c_{ji}^{*} - c_{ji}^{*} a_{it} \right] = \sum_{i}^{m} \sum_{j}^{m} a_{jt} c_{ji}^{*} \left(1 - a_{it} \right)$$

$$= \mathbf{a_t C^{*} \left(1_m - a_t \right)^{T}}$$

Note: For the fully connected network,

$$\sum_{i}^{m} \sum_{j}^{m} a_{jt} c_{ji}^{*} \left(1 - a_{it} \right)$$

$$= \sum_{i}^{m} \sum_{j\neq i}^{m} a_{jt} c_{ji}^{*} \left(1 - a_{it} \right) - \sum_{i}^{m} a_{it} c_{ii}^{*} \left(1 - a_{it} \right)$$

$$= \sum_{i}^{m} \sum_{j\neq i}^{m} a_{jt} \frac{1}{m-1} \left(1 - a_{it} \right) - \sum_{i}^{m} a_{it} c_{ii}^{*} \left(1 - a_{it} \right)$$

$$= \frac{1}{m-1} \sum_{i}^{m} \sum_{j\neq i}^{m} a_{jt} \left(1 - a_{it} \right) - 0 \text{ since } a_{it} \left(1 - a_{it} \right) = 0 \text{ for all } i \text{ by defintion}$$

$$= \frac{1}{m-1} \sum_{i}^{m} \sum_{j}^{m} a_{jt} \left(1 - a_{it} \right)$$

$$= \frac{1}{m-1} \sum_{j}^{m} a_{jt} \sum_{i}^{m} \left(1 - a_{it} \right) = \frac{1}{m-1} Y_t \left(m - \sum_{i}^{m} a_{it} \right)$$

$$= \frac{1}{m-1} Y_t \left(m - Y_t \right)$$

APPENDIX B: TABLES

Table B1. Correlation between Network and Estimates from Models

N=580

	density	p.Bass.NLS	p.Bass.OLS	p. NBB	q. Bass.NLS	q. Bass.OLS	q.NBB
density	1.00						
p.Bass.NLS	-0.12	1.00					
p.Bass.OLS	-0.02	0.92	1.00				
p.NBB	0.00	0.92	1.00	1.00			
q.Bass.NLS	0.24	-0.18	0.19	0.19	1.00		
q.Bass.OLS	0.21	-0.28	0.26	0.25	0.99	1.00	
q.NBB	0.00	-0.22	0.26	0.23	0.80	0.86	1.00

Table B2. Descriptive Statistics on Selected Songs

	N	Mean	sd	Min	Max	Range
density	22	0.001	0.001	0.0003	0.007	0.007
p.NBB	22	0.014	0.010	0.0005	0.040	0.039
p.OLS	22	0.017	0.009	0.003	0.041	0.037
p.NLS	22	0.015	0.010	0.001	0.042	0.041
q.NBB	22	0.152	0.081	0.025	0.387	0.362
q.OLS	22	0.075	0.031	0.017	0.138	0.120
q.NLS	22	0.081	0.036	0.016	0.161	0.145
m	22	14971	13105	1242	48613	47371
NBB.R2	22	0.798	0.154	0.392	0.937	0.545
OLS.R2	22	0.785	0.164	0.344	0.936	0.592
t	22	62	7	41	66	25

Table B3. Estimates for Songs

song	density	p.NBB	p.Bass .OLS	p. Bass .NLS	q.NBB	q.Bass .OLS	q.Bass .NLS	m	NBB.R^2	Bass.OLS.R^2	t	q.Ratio
Someone Like You	0.00029	0.009	0.010	0.009	0.075	0.057	0.060	48613	0.86	0.86	66	**1.33**
Party Rock Anthem	0.00029	0.006	0.006	0.003	0.142	0.090	0.097	28611	0.89	0.89	66	**1.58**
Give Me Everything	0.00050	0.008	0.011	0.008	0.172	0.097	0.102	16953	**0.92**	0.90	58	**1.77**
Born This Way	0.00051	0.024	0.027	0.026	0.092	0.060	0.060	32910	**0.51**	0.49	66	**1.53**
Roll Up	0.00052	0.010	0.014	0.013	0.162	0.065	0.065	6425	**0.74**	0.70	66	**2.47**
E.T.	0.00059	0.020	0.023	0.023	0.071	0.048	0.048	31275	0.92	0.92	66	**1.47**
The Lazy Song	0.00062	0.010	0.012	0.012	0.111	0.070	0.073	17973	0.93	0.93	66	**1.58**
Look At Me Now	0.00063	0.010	0.014	0.013	0.136	0.057	0.057	5921	**0.90**	0.88	66	**2.41**
Where Them Girls At	0.00063	0.008	0.010	0.008	0.171	0.095	0.101	13860	**0.83**	0.82	55	**1.79**
S&M	0.00072	0.040	0.041	0.042	0.028	0.017	0.016	27657	0.94	0.94	66	**1.60**
The Edge of Glory	0.00085	0.017	0.019	0.007	0.172	0.108	0.140	23017	**0.39**	0.34	60	**1.60**
Dirt Road Anthem	0.00098	0.001	0.005	0.003	0.387	0.093	0.098	1258	**0.84**	0.81	66	**4.16**
Tonight Tonight	0.00107	0.000	0.003	0.001	0.266	0.118	0.125	4228	**0.92**	0.90	62	**2.24**
Just Can't Get Enough	0.00110	0.013	0.018	0.018	0.134	0.074	0.071	11734	**0.91**	0.90	66	**1.81**
On The Floor	0.00123	0.010	0.016	0.015	0.236	0.115	0.113	6267	**0.80**	0.78	64	**2.05**
Backseat	0.00137	0.009	0.015	0.014	0.213	0.073	0.070	1580	**0.85**	0.80	65	**2.93**
Blow	0.00172	0.036	0.036	0.035	0.025	0.020	0.027	13701	0.89	0.89	66	**1.25**
Honey Bee	0.00195	0.020	0.025	0.025	0.147	0.045	0.045	1242	0.77	0.77	52	**3.25**
Motivation	0.00242	0.014	0.018	0.017	0.113	0.059	0.064	5305	0.80	**0.81**	57	**1.90**
She Ain't You	0.00326	0.012	0.015	0.012	0.128	0.062	0.068	2921	**0.63**	0.59	65	**2.06**
In the Dark	0.00688	0.008	0.008	0.001	0.227	0.138	0.161	1571	**0.50**	0.48	53	**1.65**

NBB.R^2= Adj. R^2 for the NBB model; Bass.OLS.R^2= Adj. R^2 for the Bass.OLS model

t:Week; q.Ratio= q.NBB/q.Bass.OLS

Table B4. Correlation between Parameters and Estimates from Songs

N=22

	density	q.Ratio	p.Bass .NLS	p.Bass .OLS	p.NBB	q.Bass .NLS	q.Bass .OLS	q.NBB
density	1.00							
q.Ratio	0.02	1.00						
p. Bass.NLS	-0.17	-0.24	1.00					
p. Bass.OLS	-0.09	-0.31	0.97	1.00				
p.NBB	-0.07	-0.40	0.94	0.99	1.00			
q. Bass.NLS	0.35	0.01	-0.79	-0.66	-0.64	1.00		
q. Bass.OLS	0.30	0.09	-0.81	-0.72	-0.72	0.98	1.00	
q.NBB	0.18	0.71	-0.70	-0.70	-0.75	0.67	0.73	1.00

q.Ratio= q.NBB/q.Bass.OLS

Table B5. Sampling Methods Vs. Model Estimates

	NTW Type	Bass model	NBB model
Innovation Coefficient (*p*)	Scale-free	X	SBS
	Random	X	SBS
	WS	X	SBS
Social Influence Coefficient (*q*)	Scale-free	X	SBS
	Random	X	SBS
	WS	X	SBS

X: Not proper Sampling

Table B6. Regression on Mean Distance

song	m	$\hat{\delta}$	$\hat{\alpha}$	R^2
All Your Gold	12687	0.002***	8.707***	0.31
Anna Sun	8635	0.002***	9.038***	0.38
Anything Could Happen	17120	0.005***	9.120***	0.71
As Long as You Love Me	5168	0.010***	8.925***	0.67
Beauty and a Beat	5028	0.004***	8.965***	0.26
Cola	14212	0.000	9.537***	-0.02
Come & Get It	6697	0.021***	8.919***	0.81
Die Young	15061	0.010***	9.369***	0.59
Don't Wake Me Up	7476	0.007***	8.952***	0.51
Don't You Worry Child	4449	0.029***	8.837***	0.61
Feel This Moment	5505	0.004***	9.073***	0.15
Gentleman	8108	0.026***	9.225***	0.41
Girl on Fire	8271	0.013***	8.908***	0.69
Goldie	9249	0.004***	8.851***	0.57
Heart Attack	7964	0.026***	9.063***	0.89
Heaven	13123	0.008***	9.125***	0.47
Home	6713	0.002***	9.034***	0.18
How We Do	113	-0.032***	9.994***	0.40
I'll Be Alright	11963	0.002**	8.791***	0.07
I Cry	8356	0.012***	8.797***	0.71
Kiss You	7024	0.015***	9.025***	0.36
Late Night	12009	0.006***	8.640***	0.34
Laura	14278	0.002***	8.696***	0.28
Let Me Love You	922	0.011***	9.053***	0.16
Live While We're Young	7966	0.009***	9.033***	0.65
Locked Out of Heaven	19471	0.006***	9.236***	0.21
Lover of the Light	16819	0.001***	9.030***	0.26
Madness	28232	-0.006***	9.611***	0.78
No Church in the Wild	4600	0.004***	8.796***	0.19
Octopus	11536	0.001	8.787***	0.06
One More Night	18780	0.009***	9.163***	0.50
Ride	20687	-0.002	9.536***	0.03
Runaways	17812	-0.001	9.504***	0.02
She Wolf	609	0.035***	9.316***	0.74
Sleep Alone	15013	0.009***	9.046***	0.88
Something Good	20078	0.009***	8.355***	0.94
Spin Spin Sugar	3493	0.002**	9.327***	0.04
Sun	14288	0.005***	9.055***	0.75
Take A Walk	20764	0.003***	8.909***	0.34
This Is Love	5149	0.003***	9.159***	0.13
We Are Never Ever Getting Back	17359	0.013***	9.318***	0.59
Whistle	13597	0.008***	9.017***	0.62
Wildest Moments	12383	0.003***	8.706***	0.18
Your Body	9834	0.008***	9.292***	0.12

***: Sig. at p=.01; **: Sig. at p=.05; *: Sig. at p=.1

Table B7. Summary of the NBB.Pref Model

song	m	$\tilde{d}$	$\hat{\beta}$	$\hat{q}_{NBB.Pref}$	$\hat{p}_{NBB.Pref}$
Feel This Moment	218	10.182	0.043	0.203***	0.000
Home	456	9.769	0.103***	0.102***	0.003***
Don't You Worry Child	211	9.603	0.129	0.305***	0.004***
Beauty and a Beat	343	9.511	0.121***	0.102***	0.006***
Cola	926	9.450	0.181	0.814***	0.015***
This Is Love	163	9.344	0.078	0.203***	0.005**
I Cry	335	9.315	0.112***	0.203***	-0.001
One More Night	922	9.308	0.198**	0.407	0.004***
Live While We're Young	596	9.240	0.155	0.407***	0.013***
Come & Get It	531	9.233	0.121	0.102	0.018***
Locked Out of Heaven	922	9.210	0.112***	0.102***	0.01***
Don't Wake Me Up	400	9.151	0.112***	0.102***	0.001
Whistle	559	9.142	0.172***	0.61***	0.007***
Anything Could Happen	1020	9.130	0.207	1.729	0.004***
Anna Sun	607	9.086	0.138***	0.203***	0.004***
Goldie	599	8.826	0.147***	0.305***	0.002
Girl on Fire	386	8.759	0.172*	0.203	0.011***
Lover of the Light	869	8.679	0.103	0.102***	0.002
I'll Be Alright	830	8.611	0.19***	10.000	0.017***
Sleep Alone	855	8.594	0.181**	0.814***	0.003***
Heaven	383	8.513	0.129**	0.407***	0.004**
All Your Gold	679	8.509	0.164***	10.000	0.012***
Laura	767	8.429	0.138**	0.102***	0.011***
Take A Walk	969	8.418	0.181***	4.881	0.012***
Something Good	1110	8.281	0.19***	0.814***	0.004***
Late Night	669	8.111	0.181	6.734	0.012***

*: Sig. at p=0.1; **: Sig. at p=0.05; *** Sig. at p=0.01

Table B8. Model Fit Comparison

song	$\tilde{d}$	$MSE_{NBB.Pref}$	MSE_{NBB}	MSE_{BASS}	$MSE_{NBB-NBB.Pref}$	$MSE_{BASS-NBB.Pref}$	$MSE_{BASS-NBB}$
Feel This Moment	10.182	1.54	1.47	1.48	-5.05	-3.91	1.08
Home	9.769	2.87	2.91	2.91	1.61	1.38	-0.24
Don't You Worry Child	9.603	1.14	1.19	1.05	3.62	-8.64	-12.73
Beauty and a Beat	9.511	4.03	4.14	3.75	2.54	-7.57	-10.37
Cola	9.450	122.85	129.18	124.30	4.90	1.17	-3.93
This Is Love	9.344	1.24	1.22	1.15	-1.65	-7.86	-6.11
I Cry	9.315	2.43	2.43	2.10	0.16	-15.51	-15.70
One More Night	9.308	25.55	25.76	25.79	0.83	0.93	0.10
Live While We're Young	9.240	14.04	15.06	15.45	6.79	9.12	2.50
Come & Get It	9.233	19.51	19.34	18.16	-0.87	-7.42	-6.50
Locked Out of Heaven	9.210	12.50	12.05	11.70	-3.80	-6.85	-2.94
Don't Wake Me Up	9.151	2.30	2.29	2.62	-0.39	12.10	12.45
Whistle	9.142	3.29	3.46	3.48	5.06	5.60	0.57
Anything Could Happen	9.130	34.20	35.49	34.83	3.62	1.80	-1.89
Anna Sun	9.086	3.35	3.54	3.37	5.26	0.42	-5.11
Goldie	8.826	8.92	9.15	9.50	2.58	6.11	3.62
Girl on Fire	8.759	5.55	5.33	4.93	-4.15	-12.65	-8.16
Lover of the Light	8.679	87.96	87.83	79.32	-0.15	-10.89	-10.72
I'll Be Alright	8.611	57.03	57.99	57.26	1.66	0.39	-1.28
Sleep Alone	8.594	31.87	33.47	32.29	4.78	1.27	-3.68
Heaven	8.513	20.10	20.35	19.57	1.24	-2.69	-3.98
All Your Gold	8.509	23.08	26.53	28.60	12.99	19.30	7.25
Laura	8.429	25.10	24.70	24.64	-1.64	-1.86	-0.22
Take A Walk	8.418	86.91	89.05	84.51	2.40	-2.84	-5.38
Something Good	8.281	6.82	7.41	7.14	7.97	4.59	-3.67
Late Night	8.111	58.80	65.18	62.59	9.78	6.05	-4.13

*: Sig. at p=0.05; ** Sig. at p=0.01; $MSE_{A-B} > 0$ means model B fits better

Table B9. Correlations between Estimates

	$\tilde{d}$	m	$\hat{\beta}$	$MSE_{NBB,NBB.Pref}$	$MSE_{BASS,NBB.Pref}$
$\tilde{d}$	1.00				
m	-0.51	1.00			
$\hat{\beta}$	-0.56	0.69	1.00		
$MSE_{NBB,NBB.Pref}$	-0.41	0.32	0.56	1.00	
$MSE_{BASS,NBB.Pref}$	-0.27	0.28	0.38	0.68	1.00

APPENDIX C: FIGURES

Figure C1. Aggregate Diffusion Model and Network

Individual Level Data	Social Network Information	
	No	Yes
No	Bass Model (1969) Extensions of the Bass Model	Dover et al. (2012) Shaikh et al. (2010)
Yes	Chatterjee and Eliashberg (1990)	Proposed (NBB) Model

Figure C2. Bass Model with Network

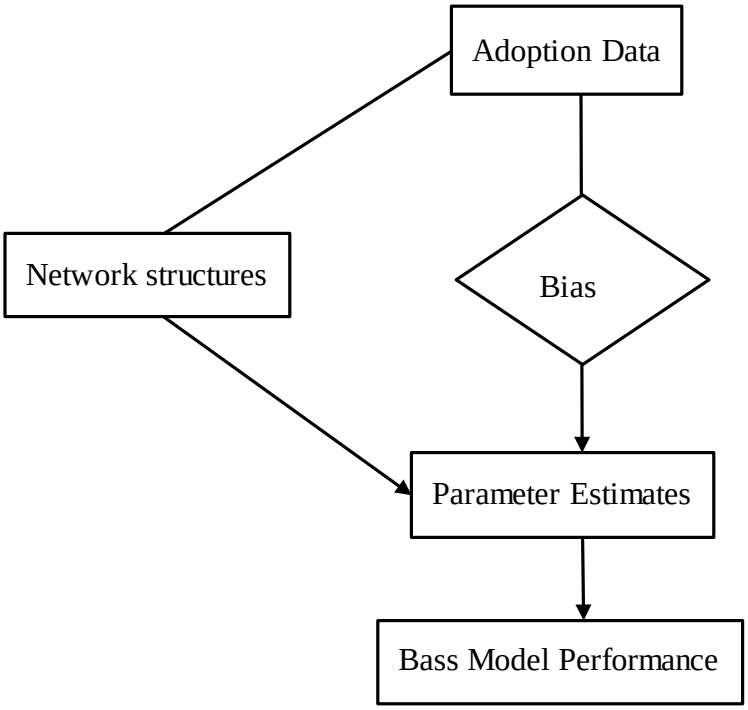


Figure C3. Fits from the High Density Network

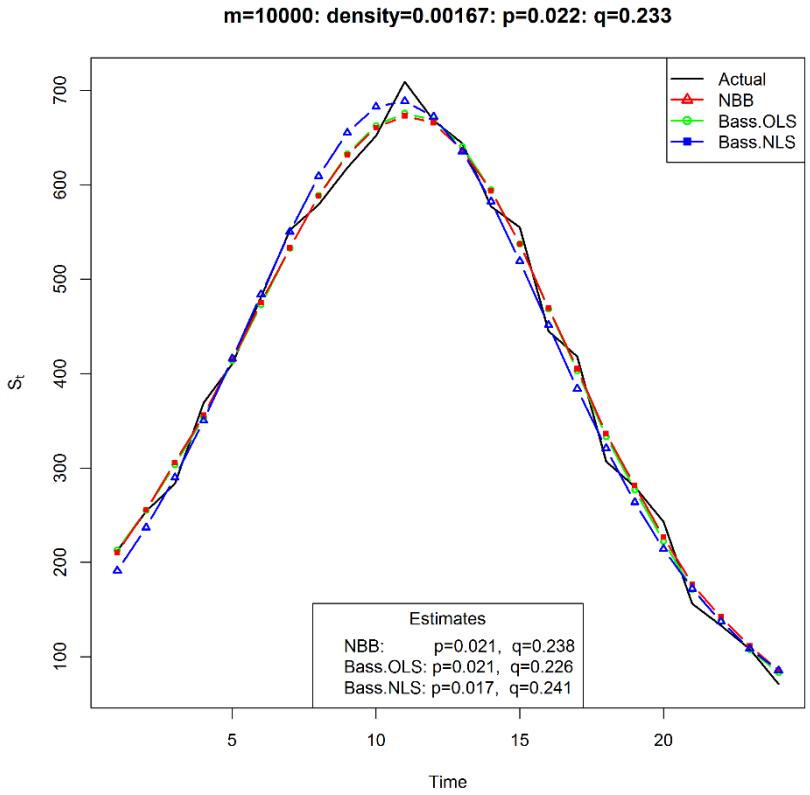


Figure C4. Fits from the Low Density Network

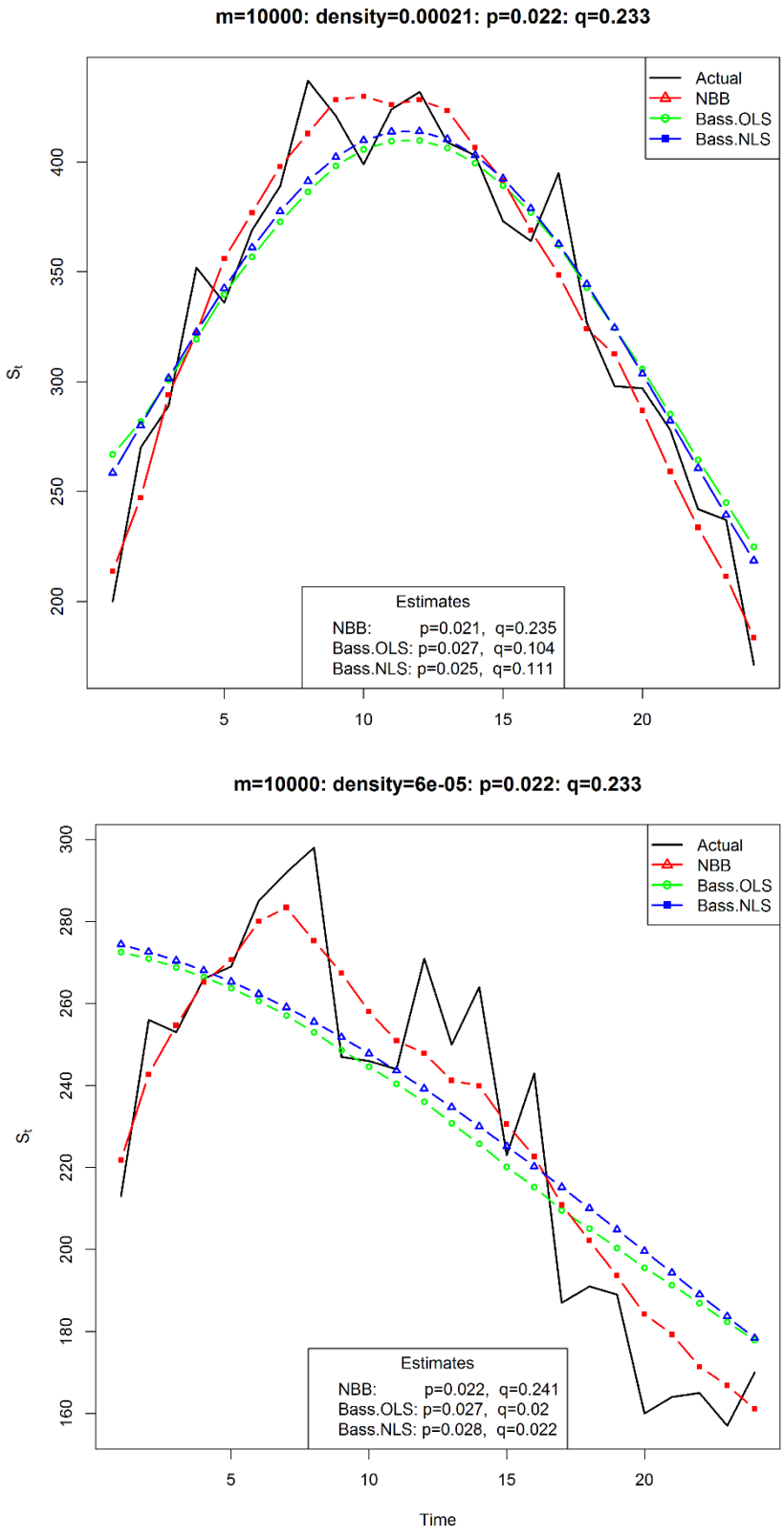


Figure C5. p Estimates from the Models

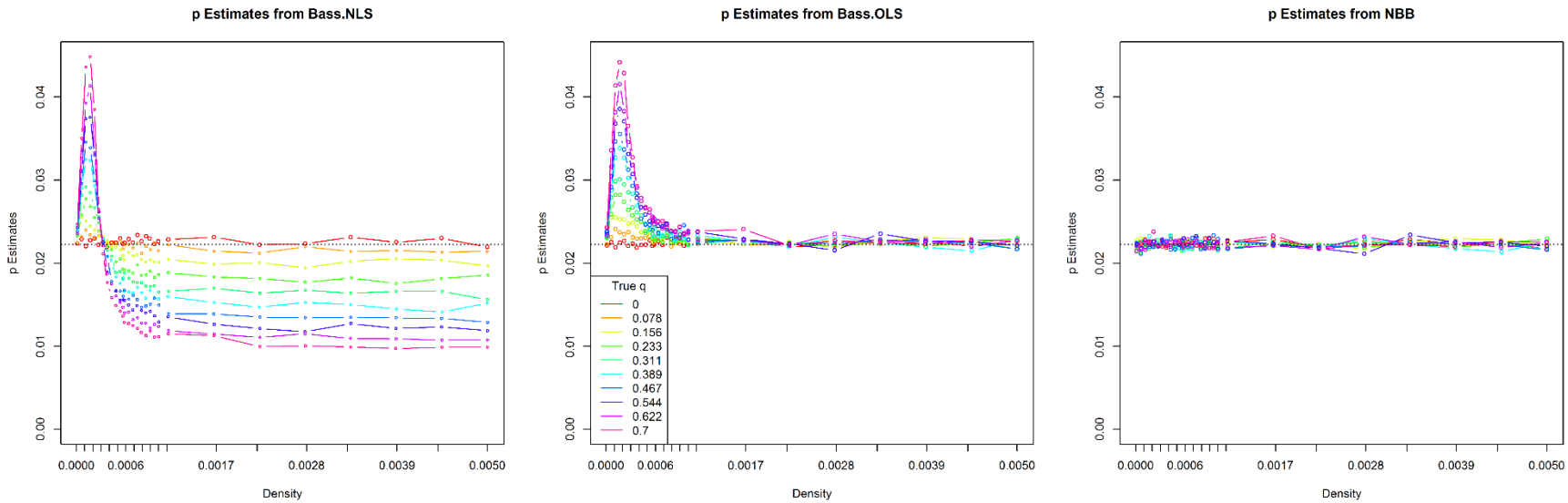


Figure C6. q Estimates from the Models

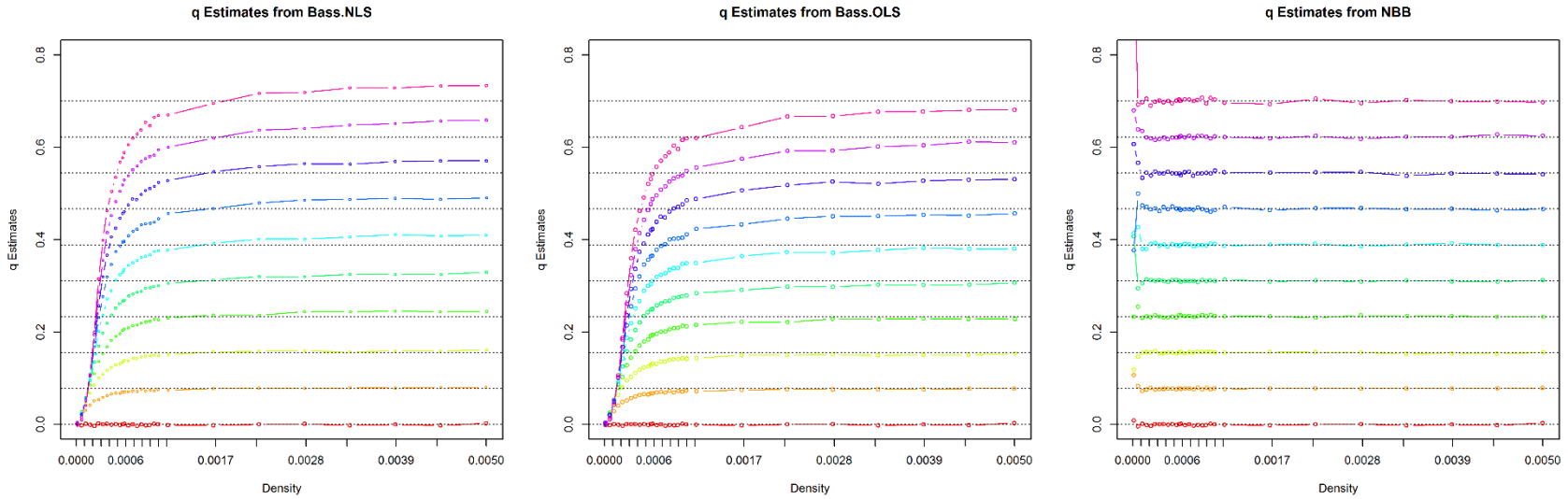


Figure C7. High vs. Low Density Random Network Map

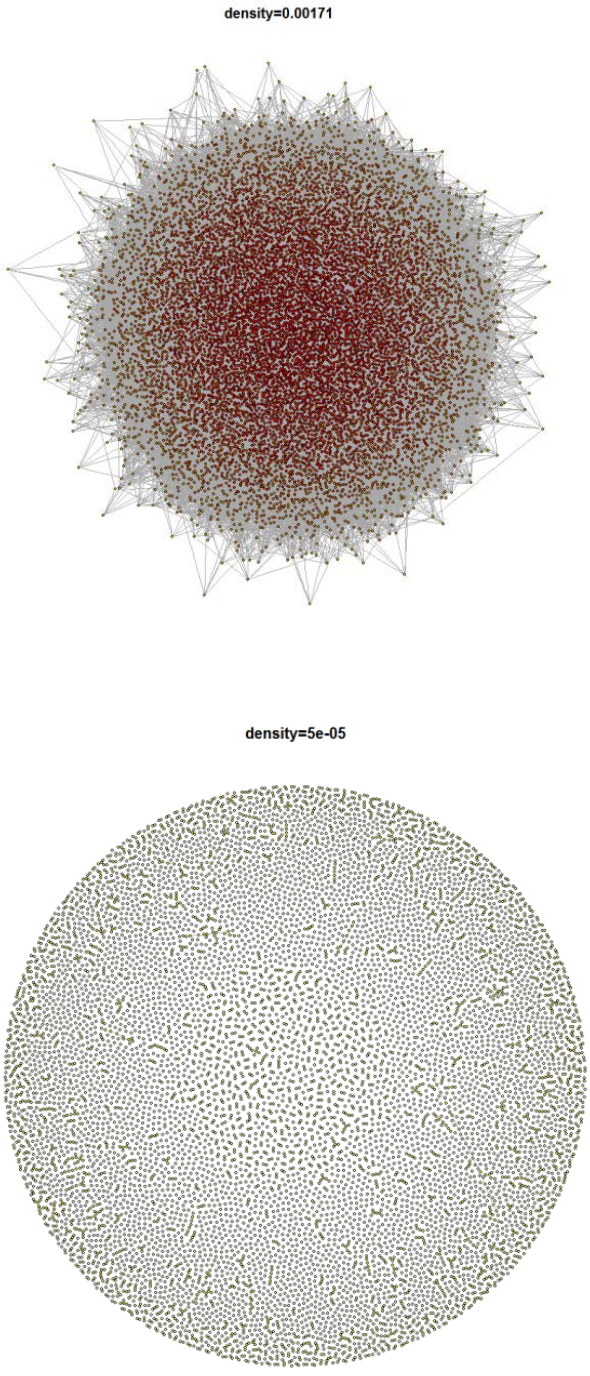


Figure C8. Low Density and Imitation Coefficients

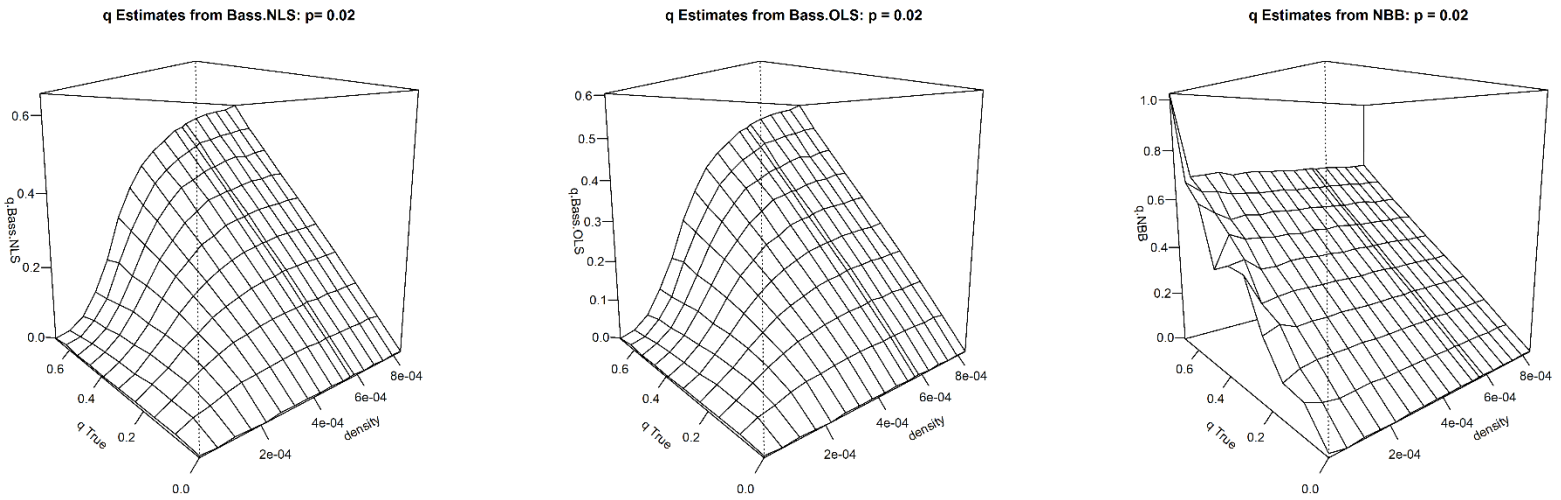


Figure C9. Comparison of Social Interaction Measures

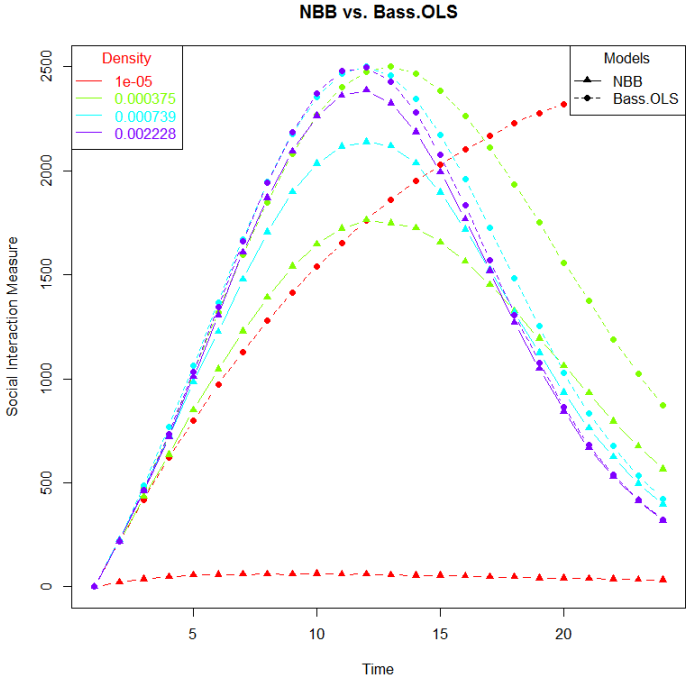


Figure C10. Difference in Social Interaction Measures

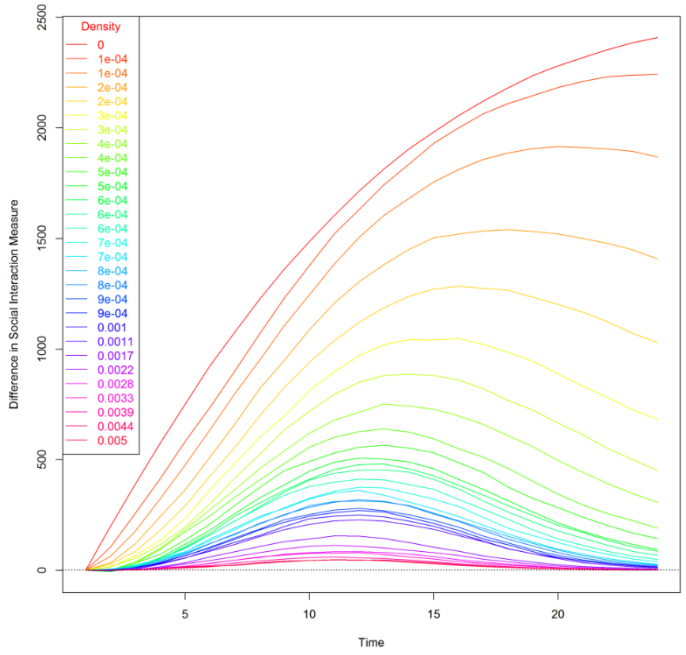


Figure C11. Fits from Songs

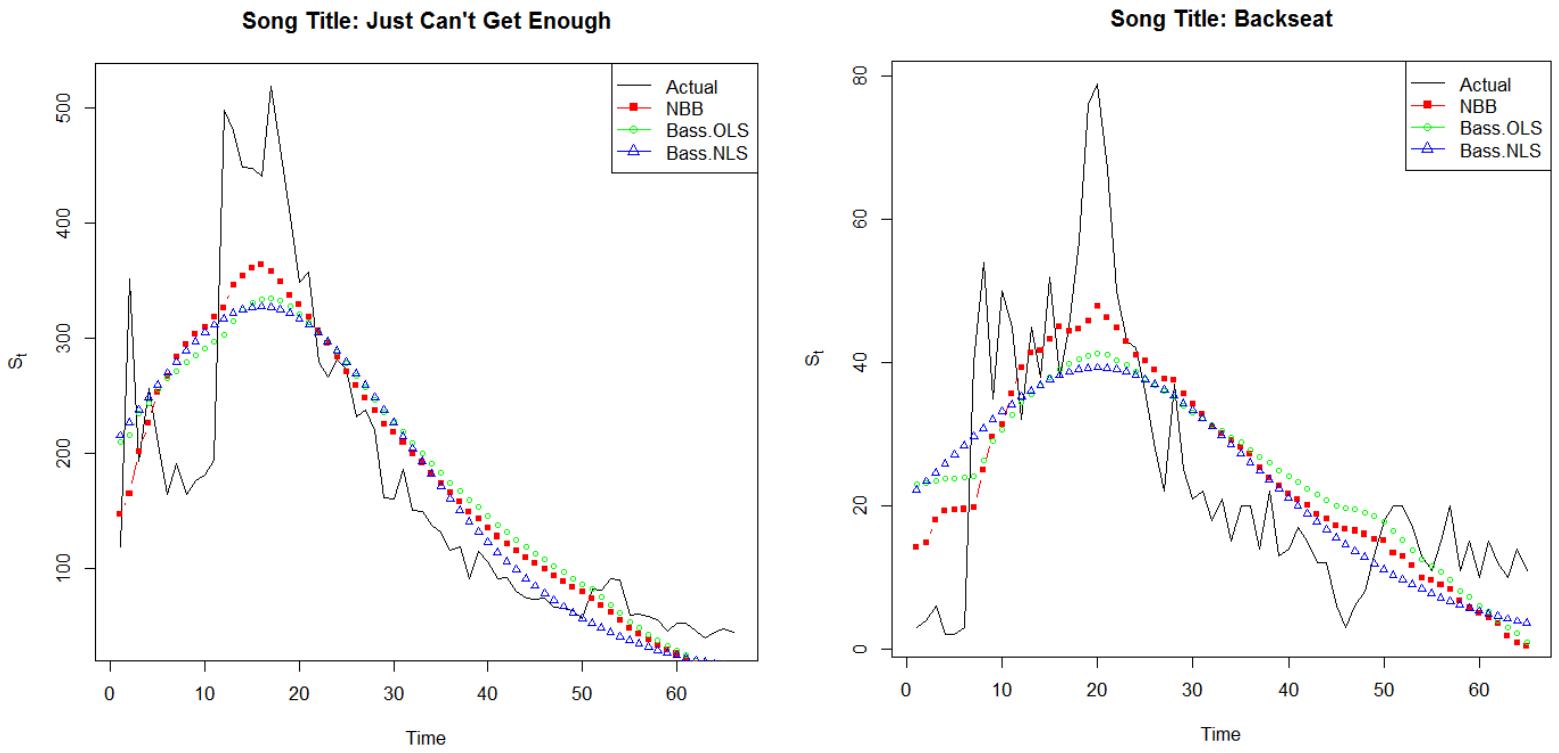


Figure C12. Boxplot of the Ratios of the Estimates

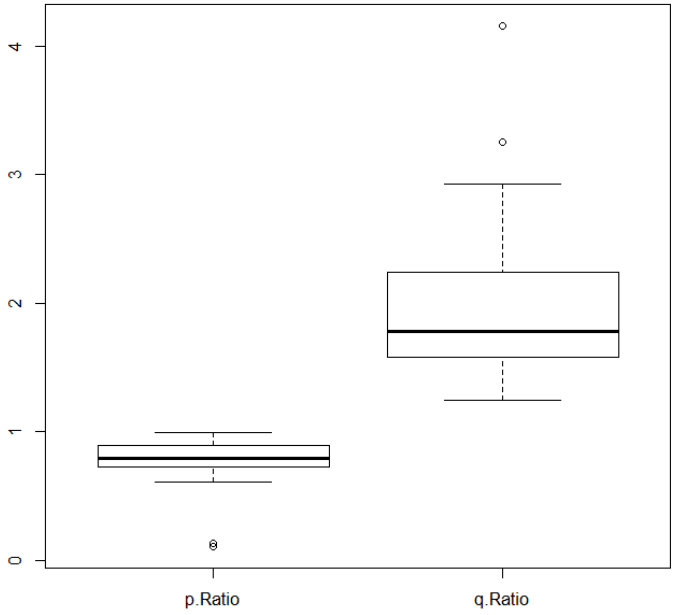


Figure C13. Random vs. Scale-Free Network

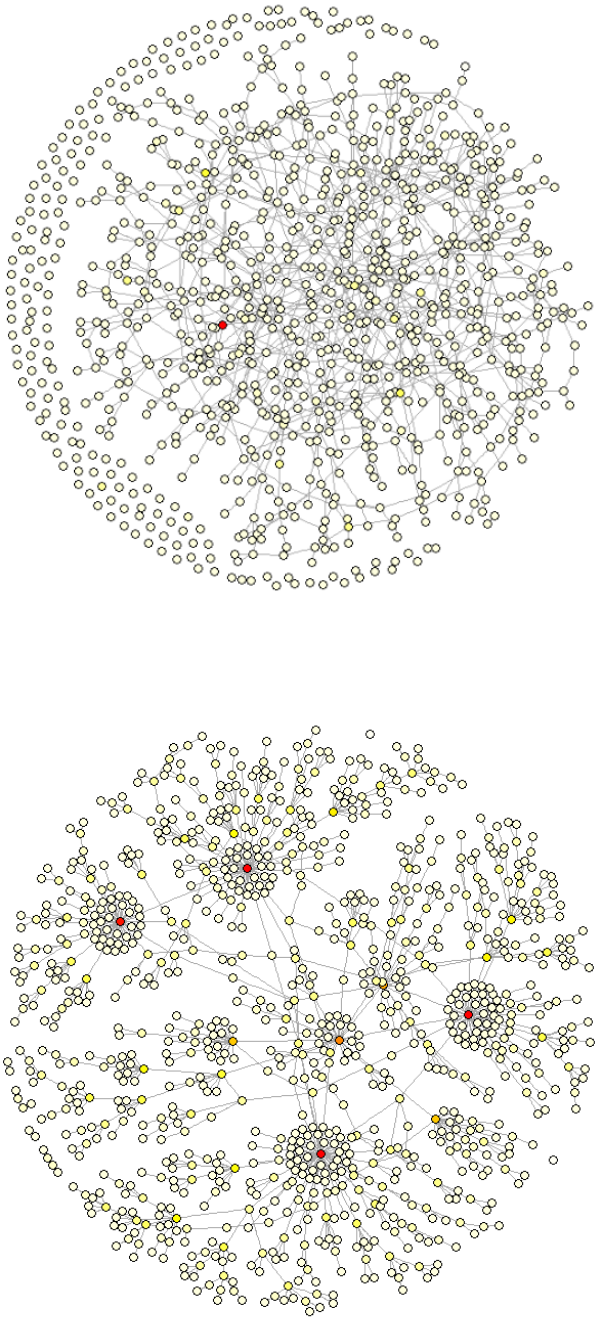


Figure C14. Scale-free Network (pwr=.1 Vs. 2.38)

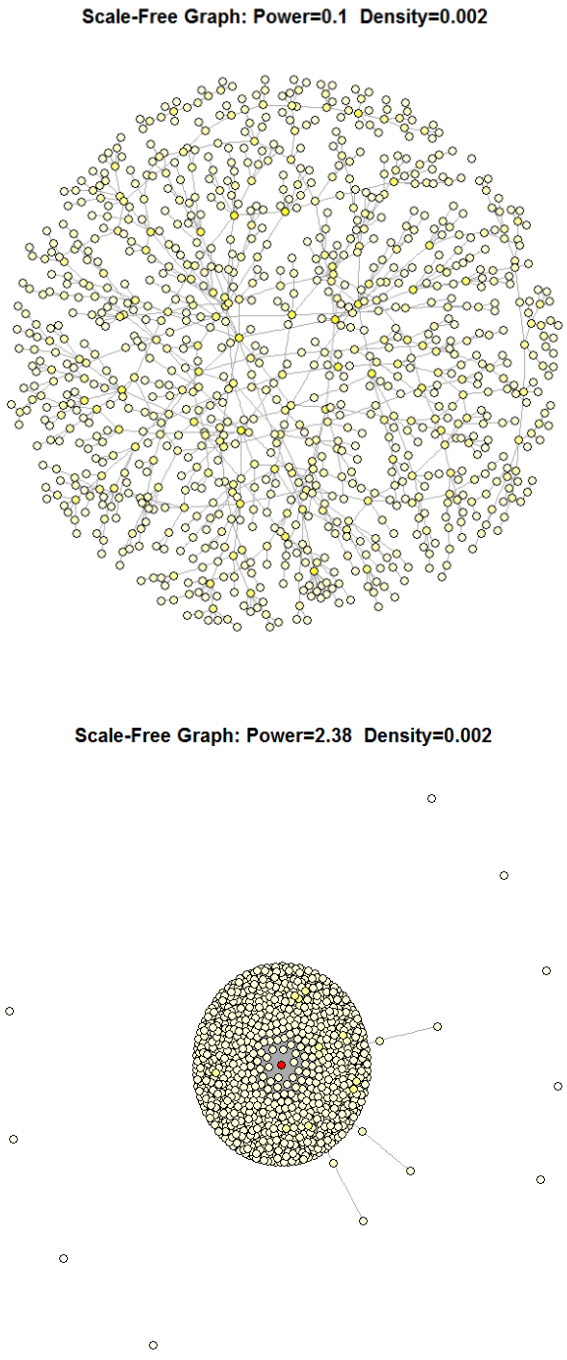


Figure C15: Degree Distribution from Pwr=0.01

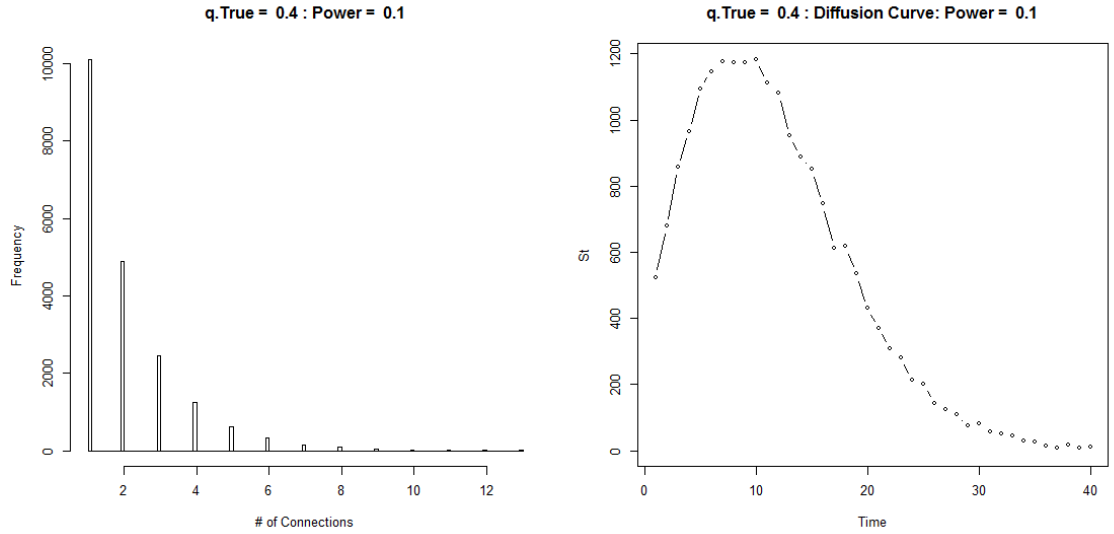


Figure C16: Degree Distribution from Pwr=2.38

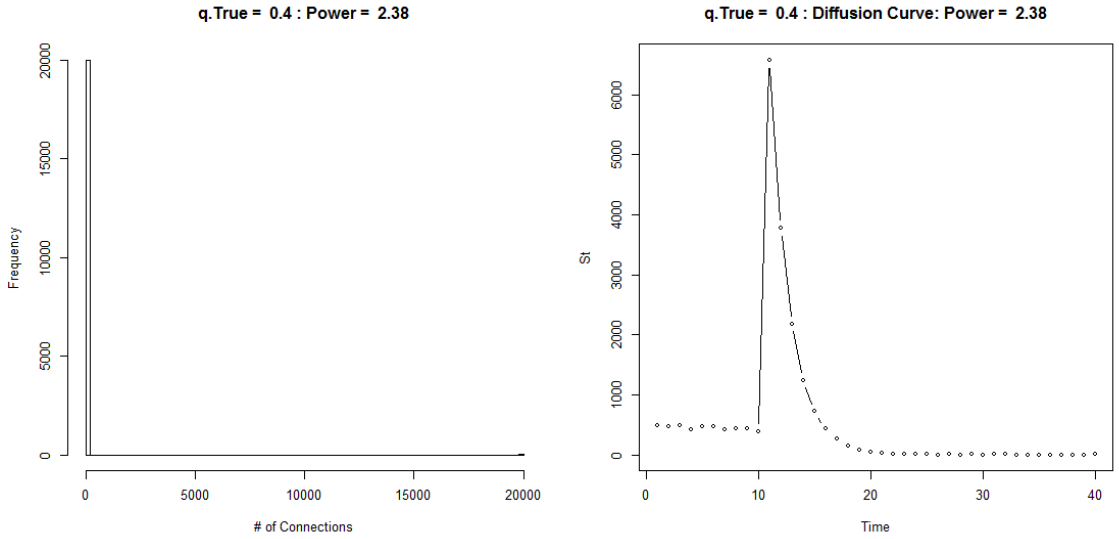


Figure C17. *p* Estimates from RNS Vs. SBS

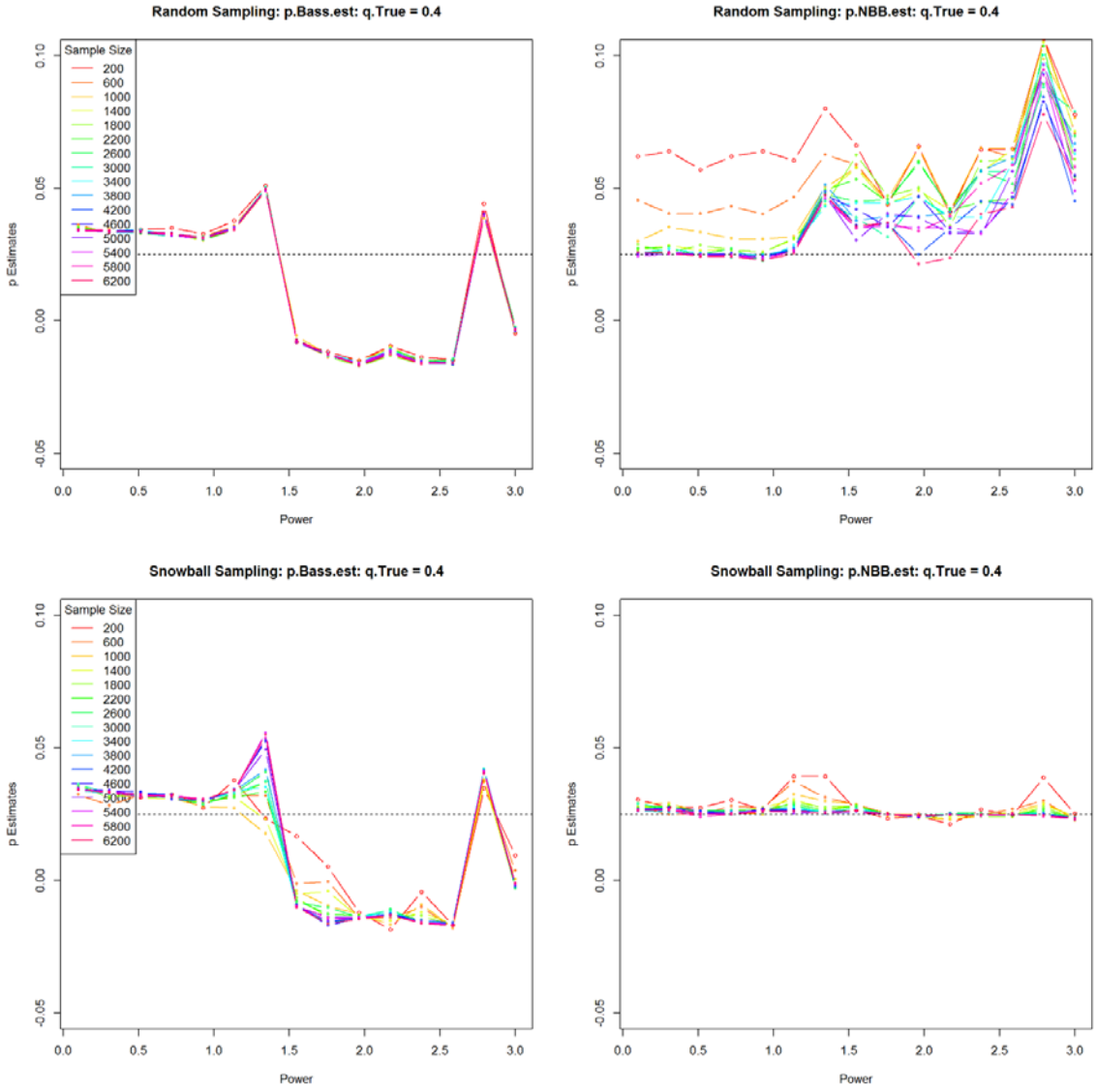


Figure C18. *q* Estimates from RNS Vs. SBS

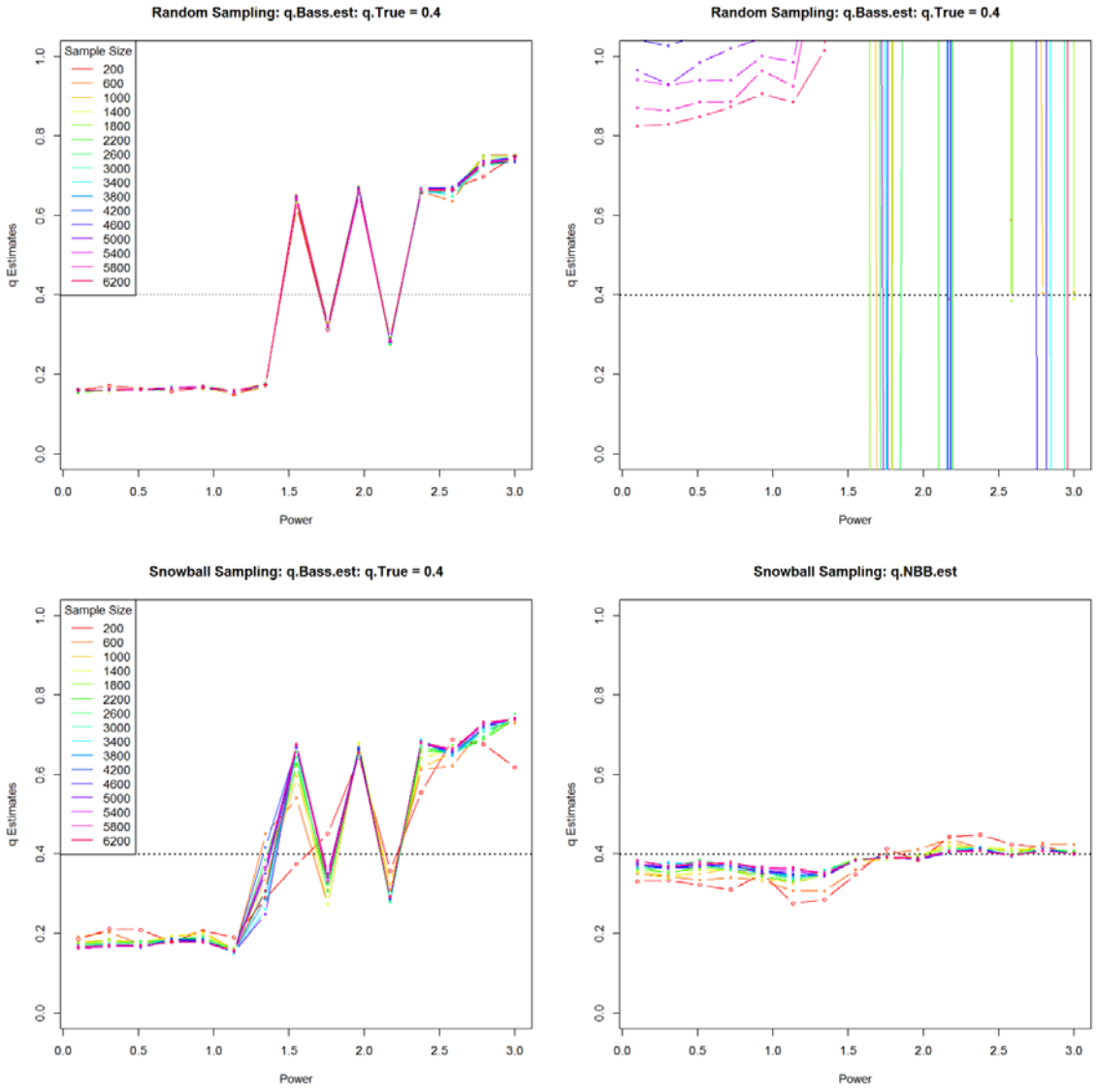


Figure C19. *p* Estimates from RNS Vs. SBS from High True *q*

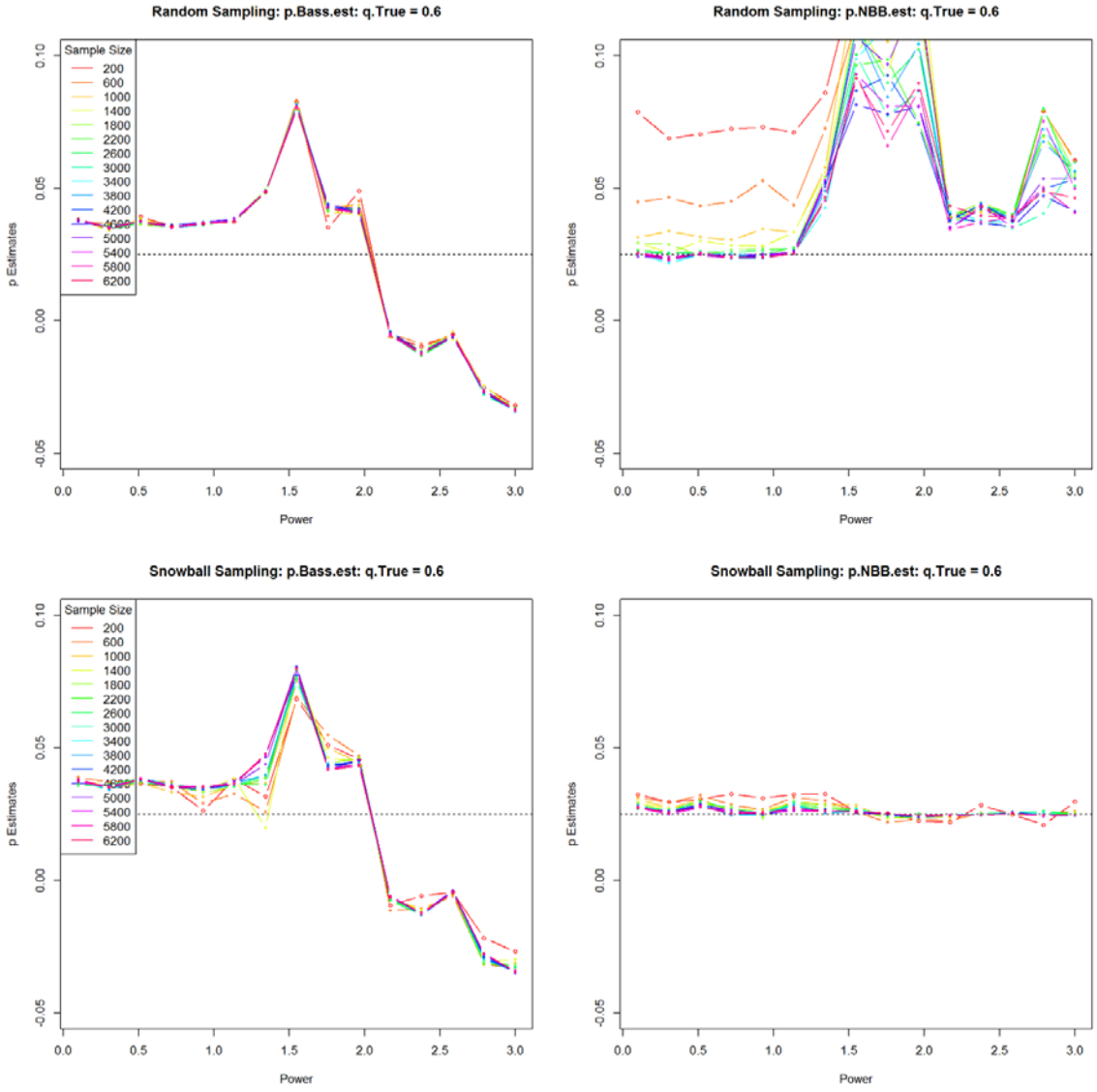


Figure C20. *q* Estimates from RNS Vs. SBS from High True *q*

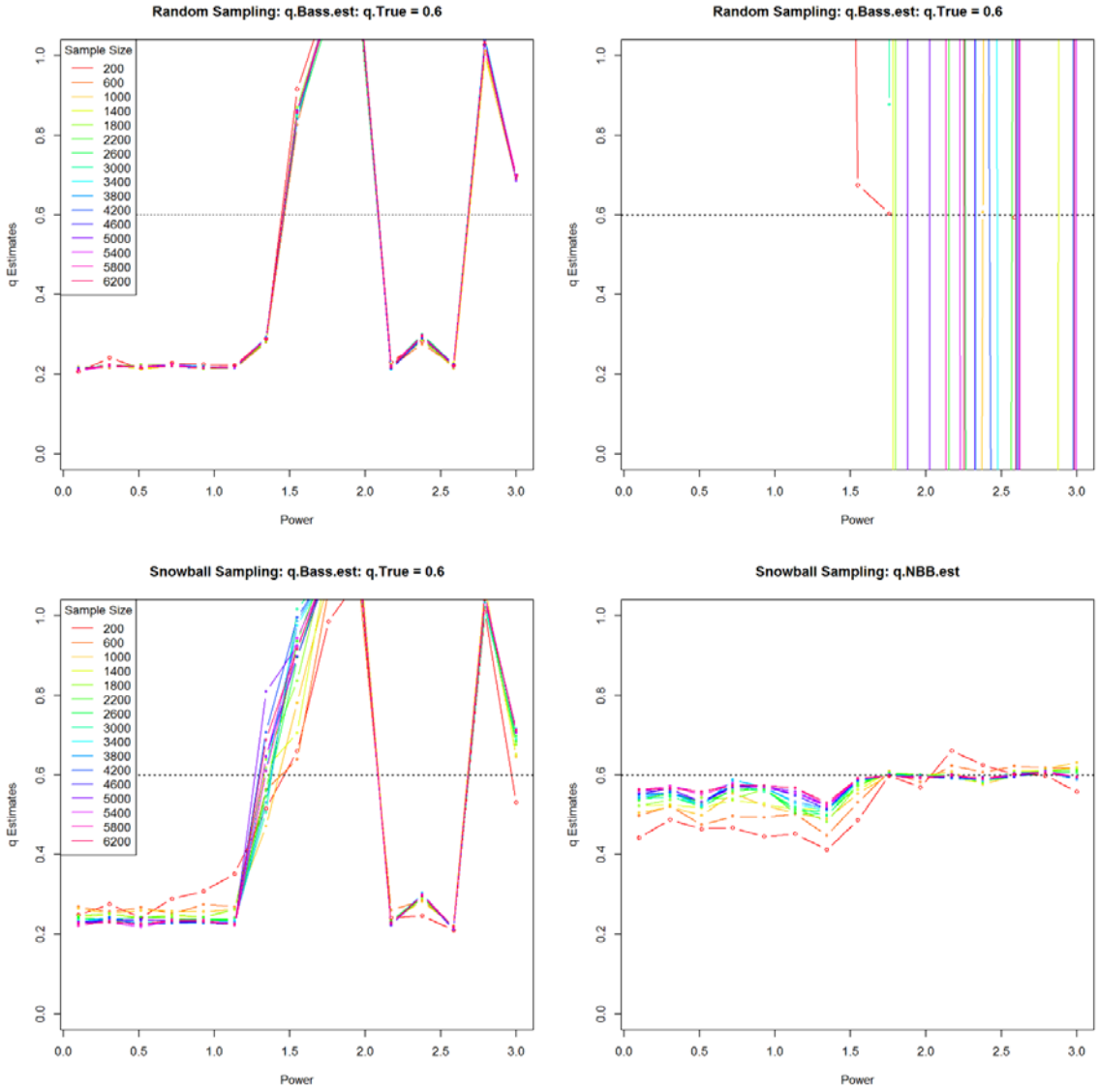


Figure C21. Random Sampling and Density of Scale-free Network

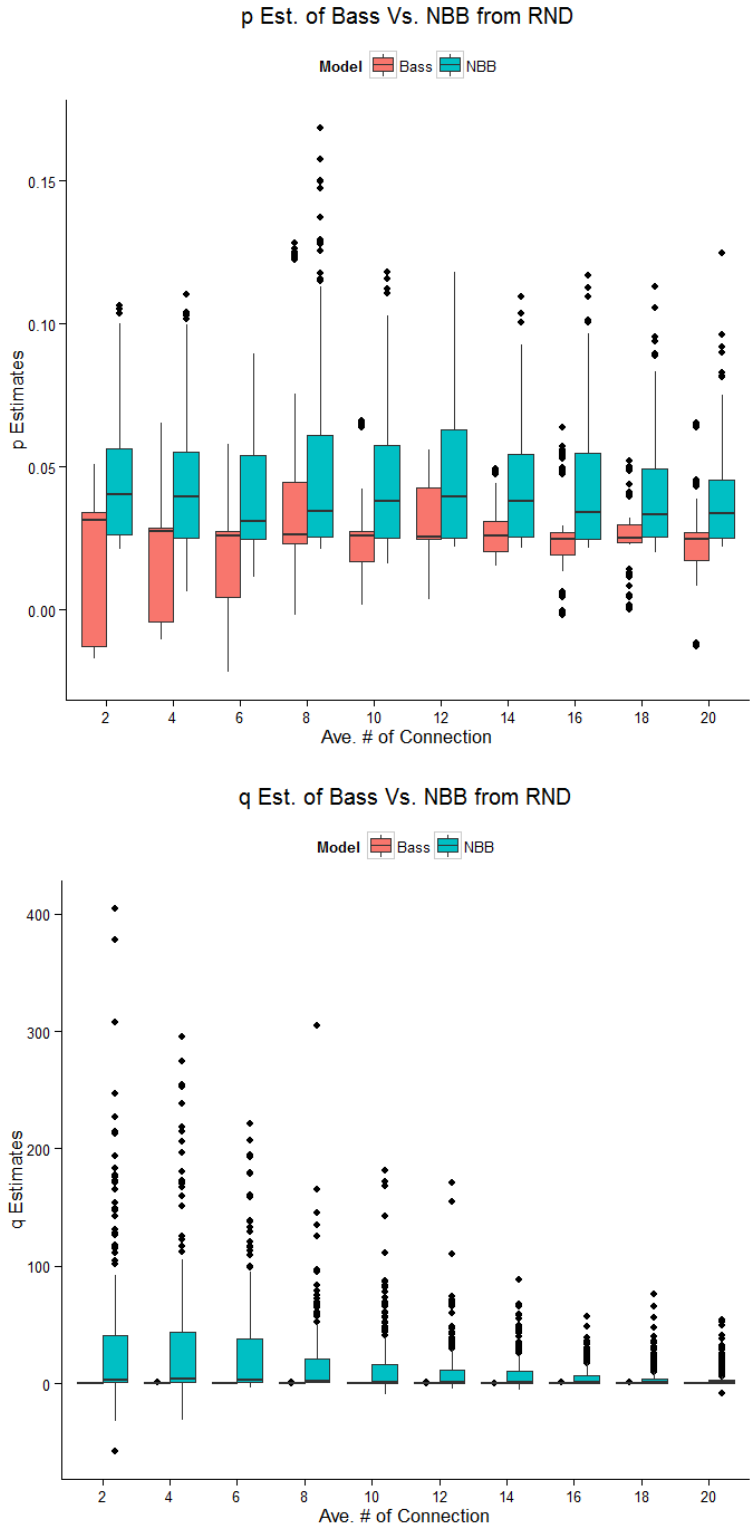


Figure C22. Snowball Sampling and Density of Scale-free Network

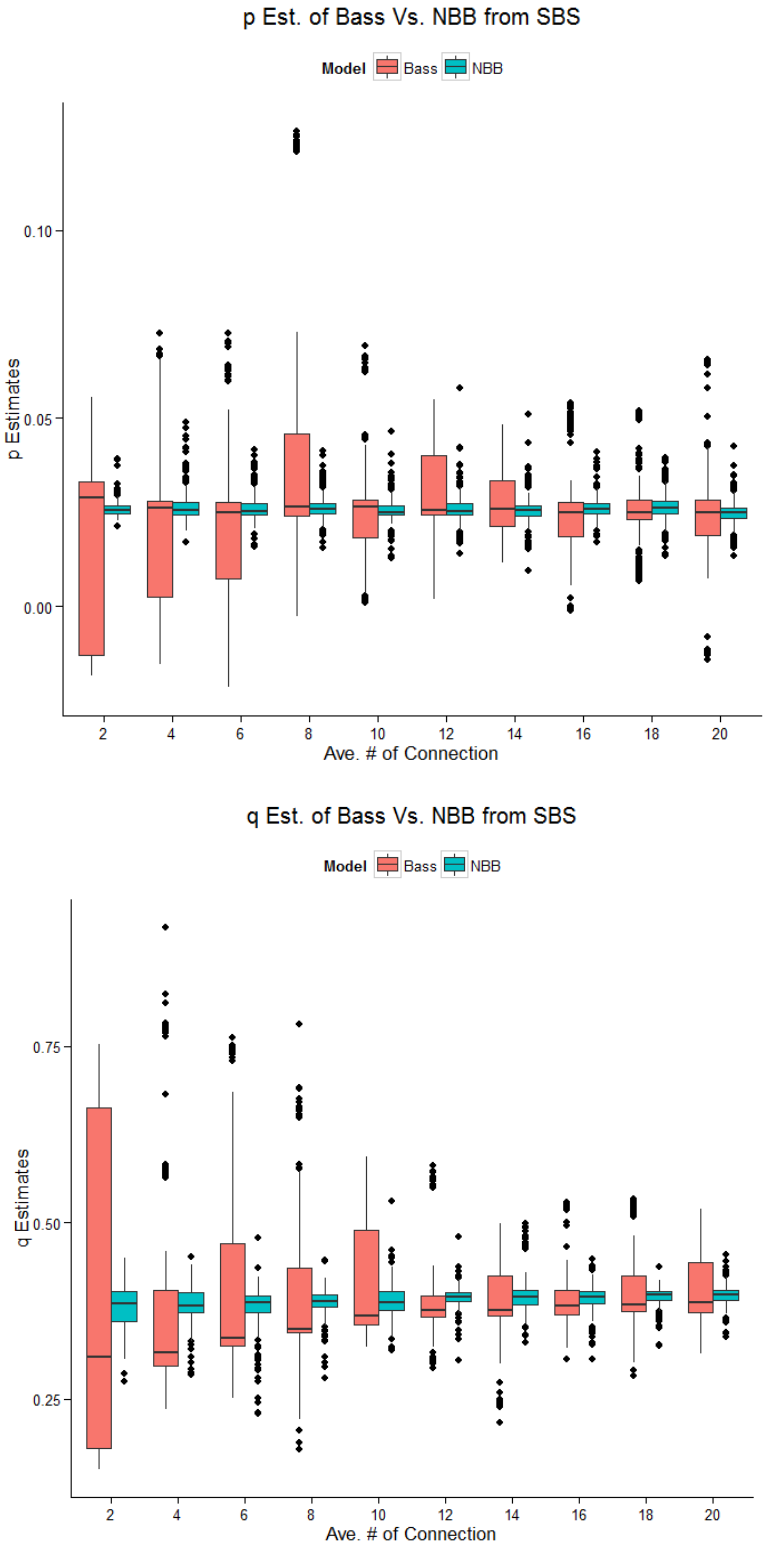


Figure C23. *p* Estimates from RNS Vs. SBS from Random Network

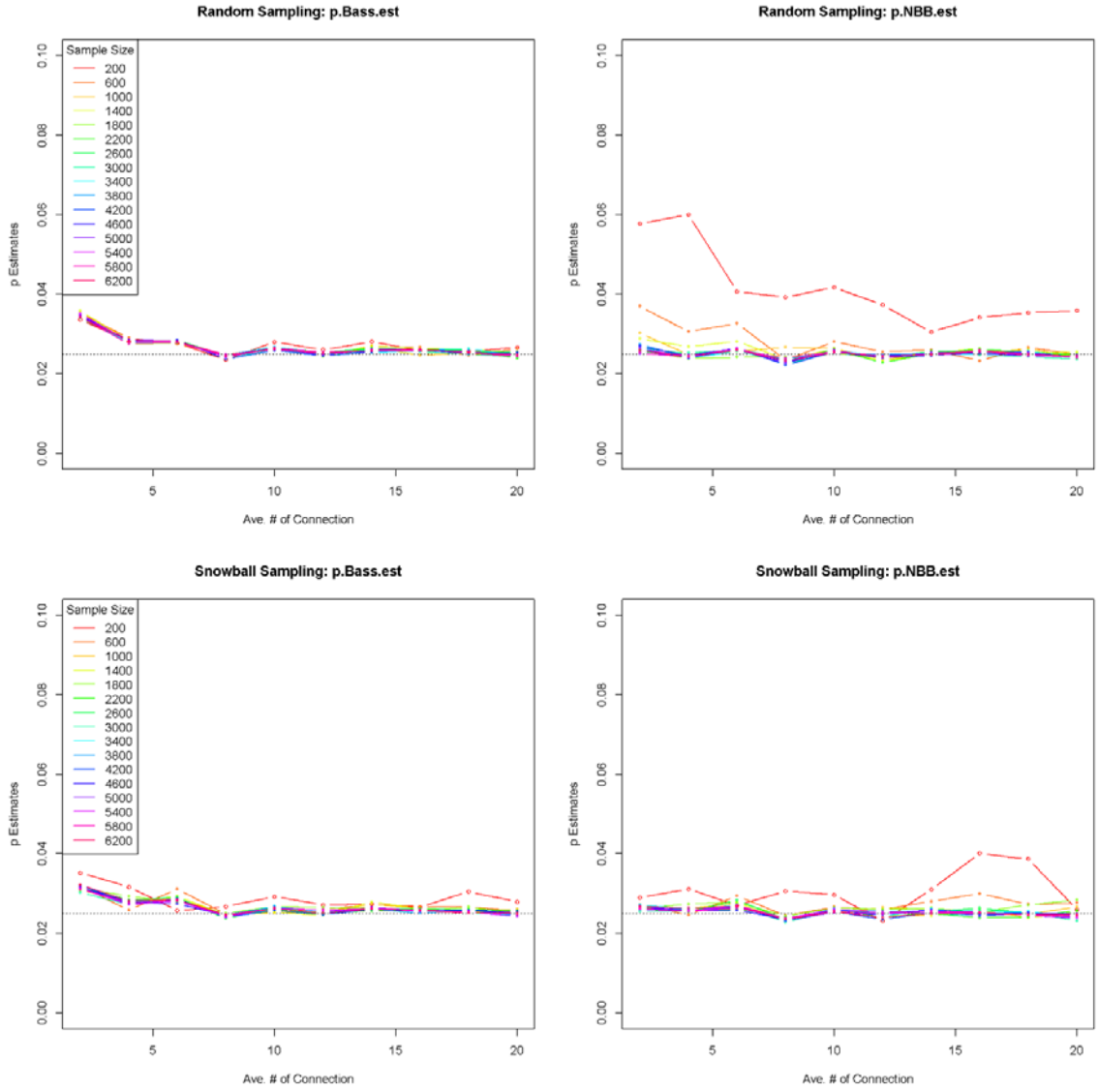


Figure C24. *q* Estimates from RNS Vs. SBS from Random Network

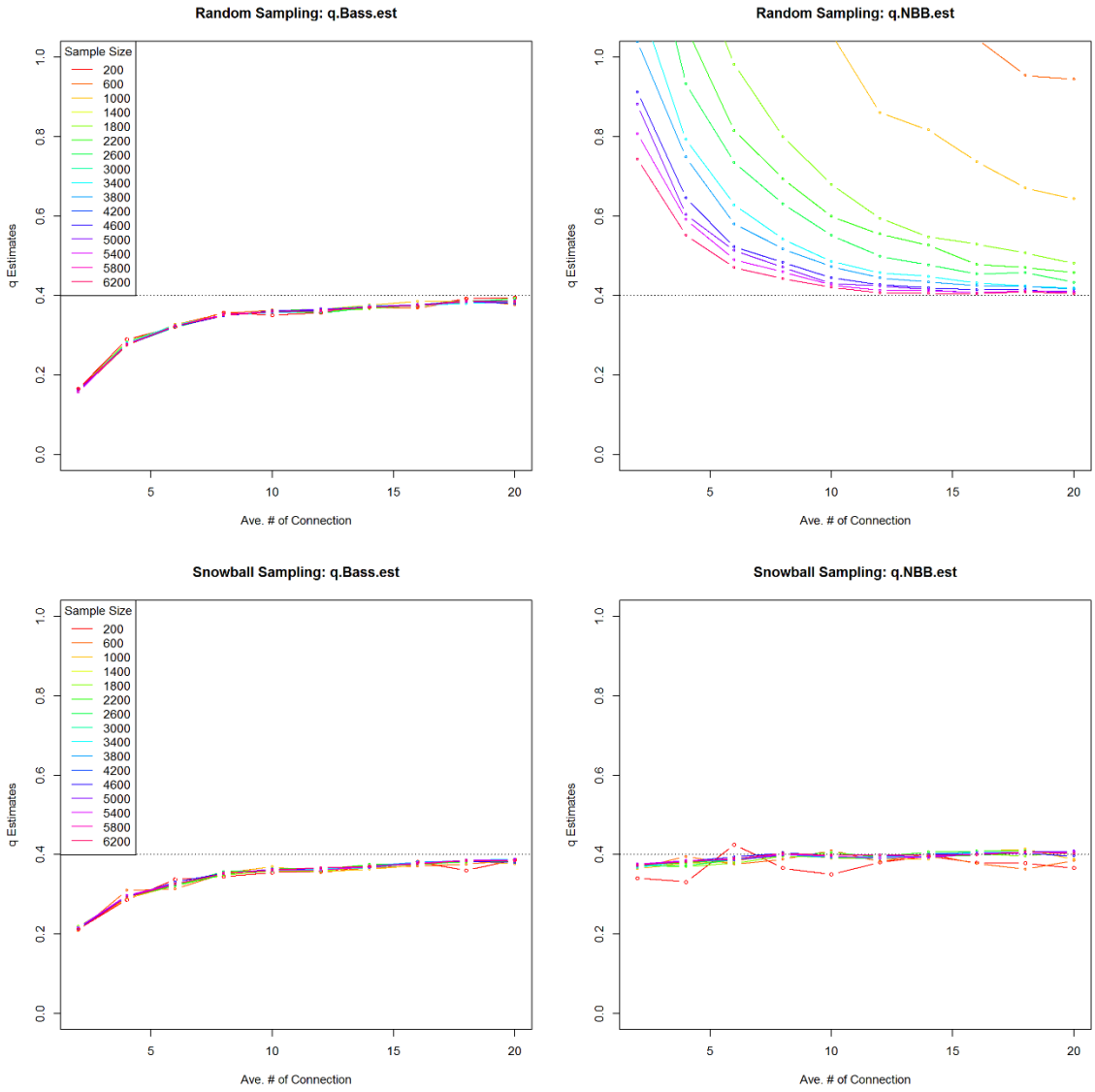


Figure C25. *p* Estimates from RNS Vs. SBS from WS Network

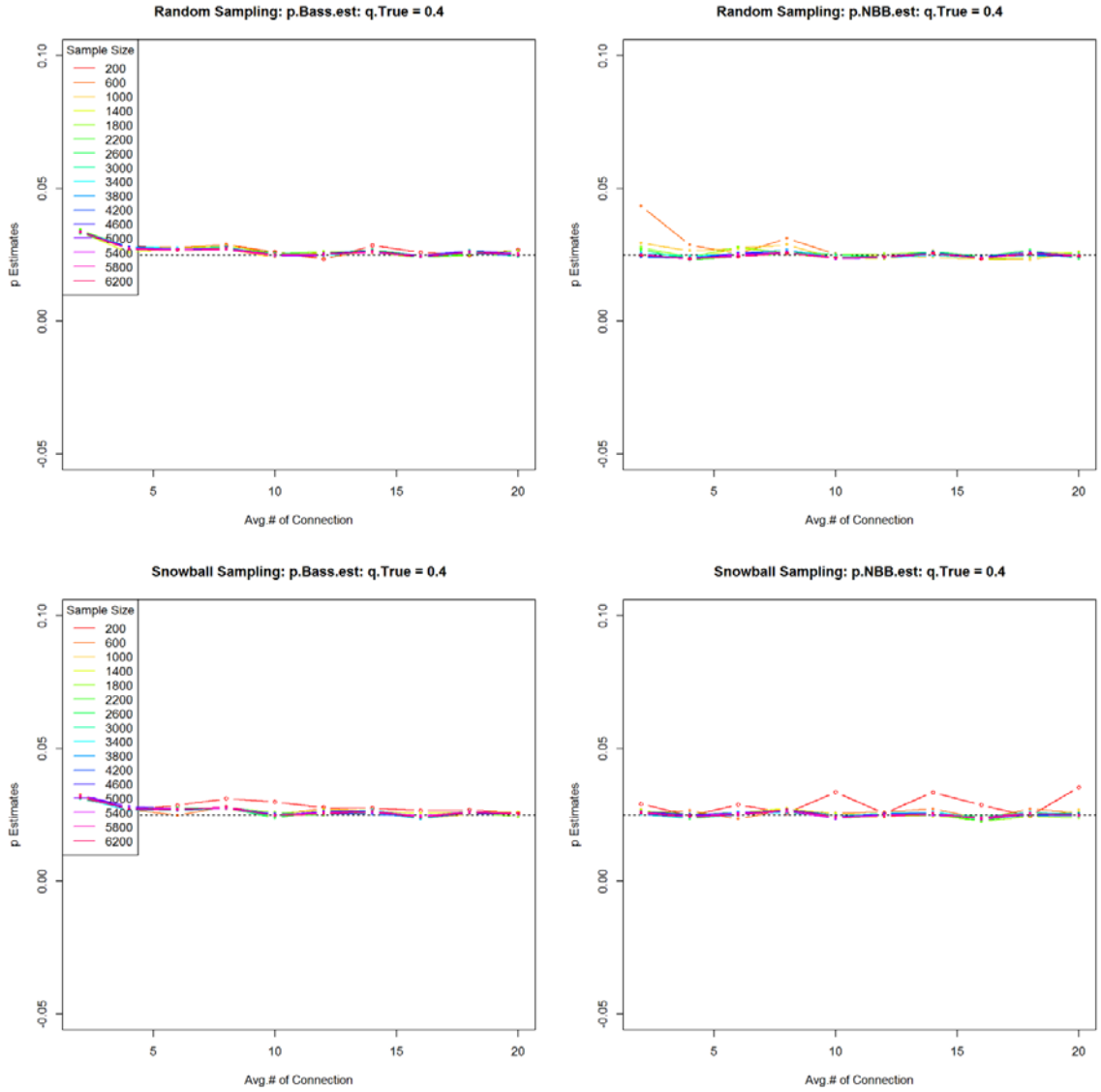


Figure C26. *q* Estimates from RNS Vs. SBS from WS Network

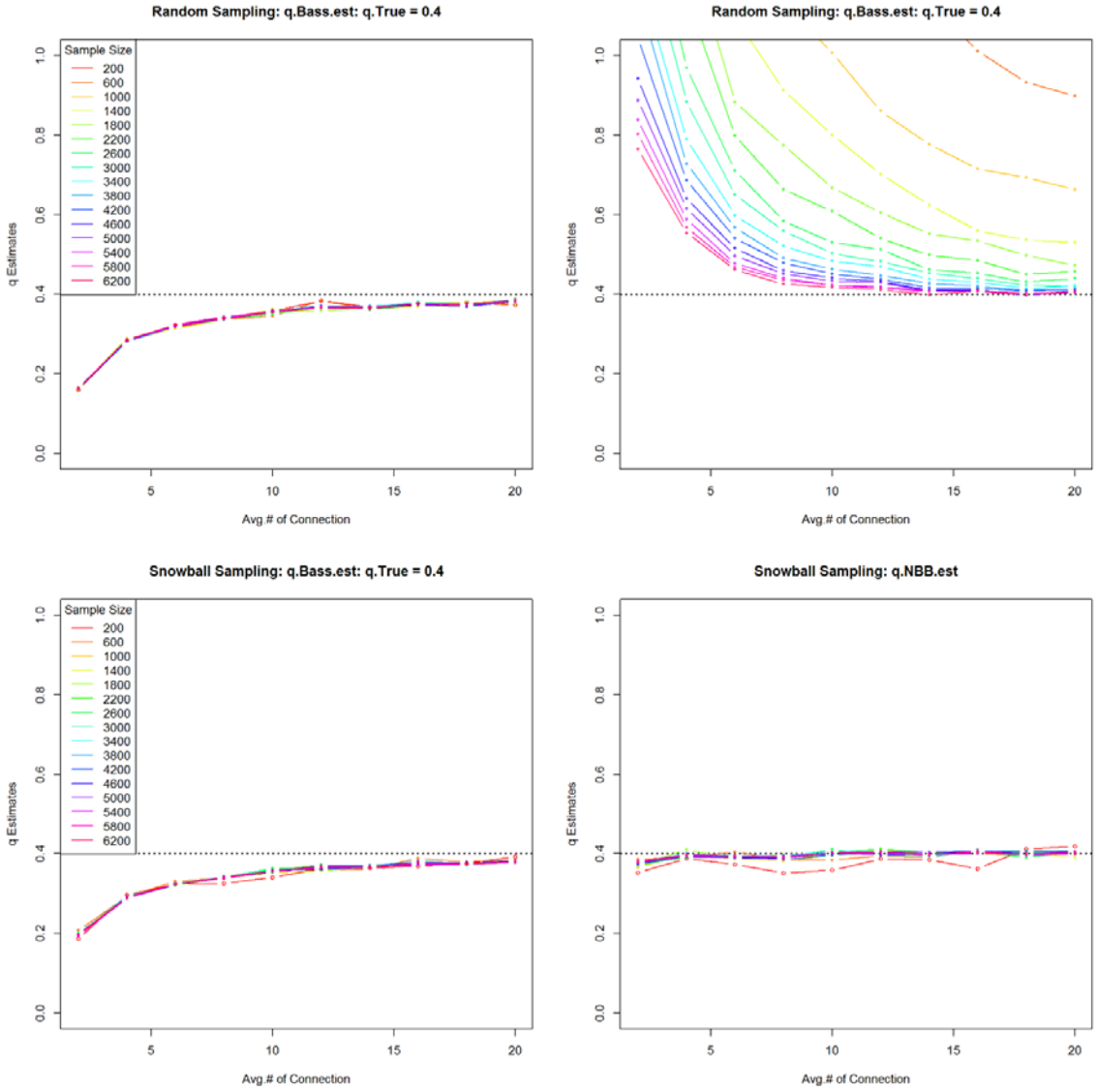


Figure C27. Bass model Estimates from Samples on Empirical Data

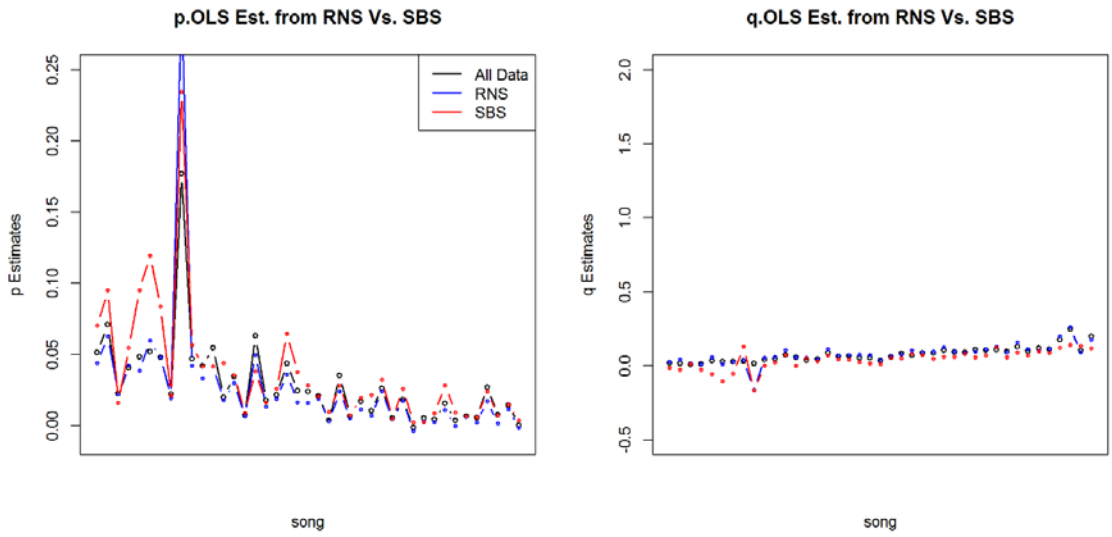


Figure C28. NBB Model Estimates from Samples on Empirical Data

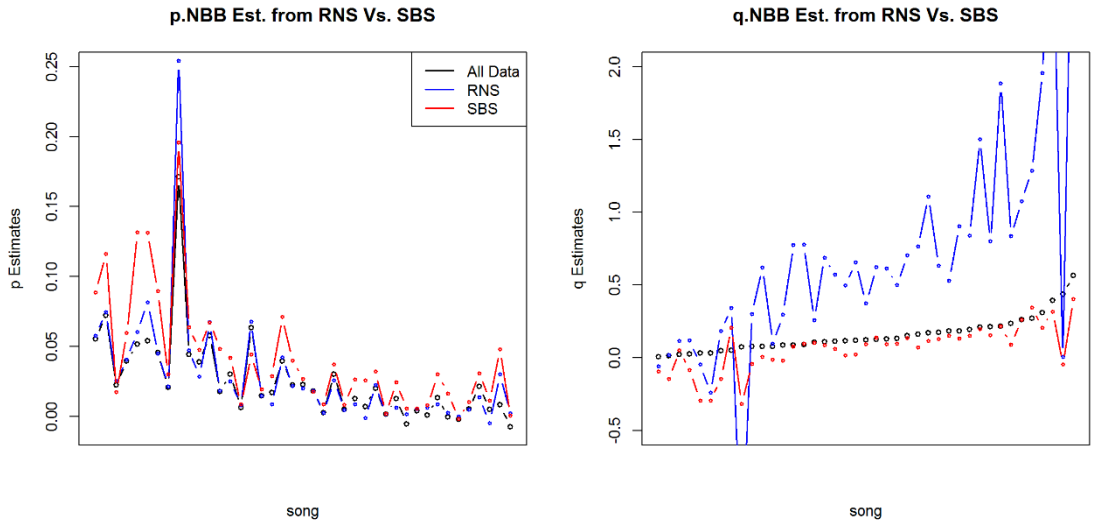


Figure C29. Adopter Network Graphs from RNS Vs. SBS

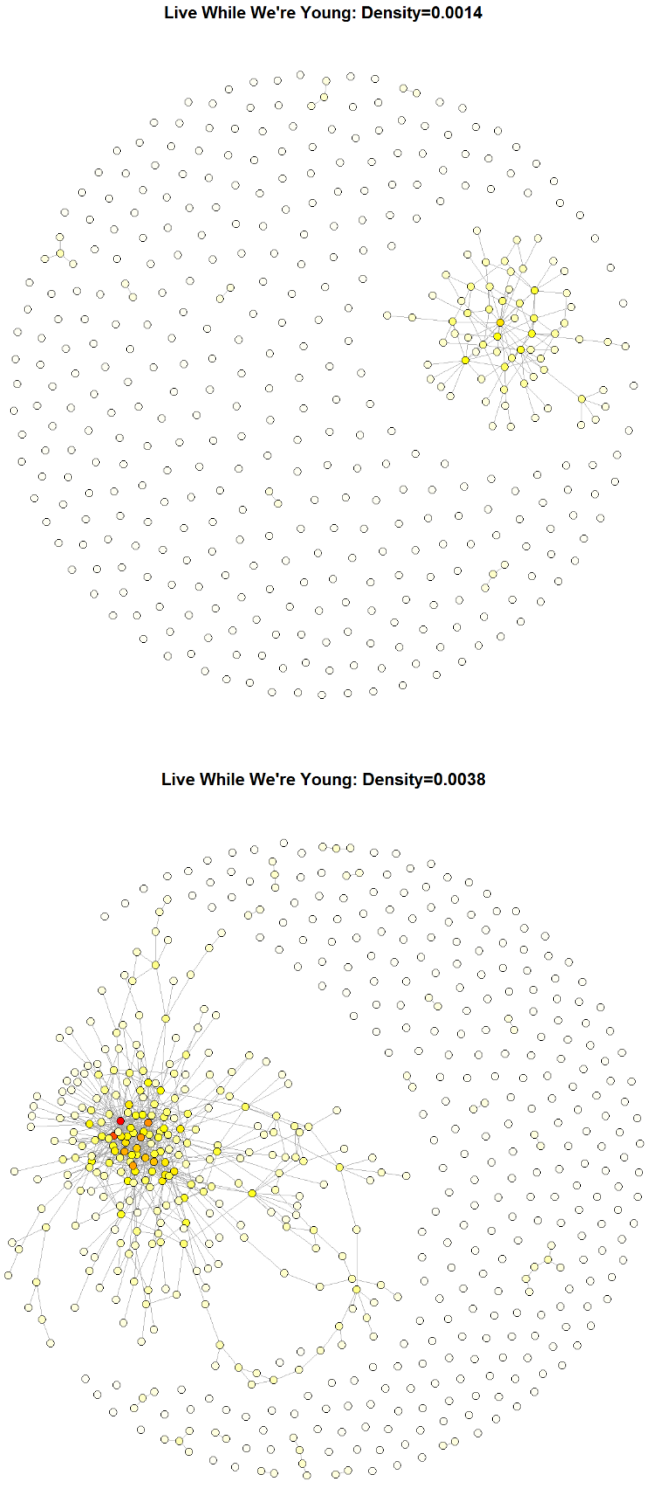


Figure C30. Adopter Network Graph from All Data

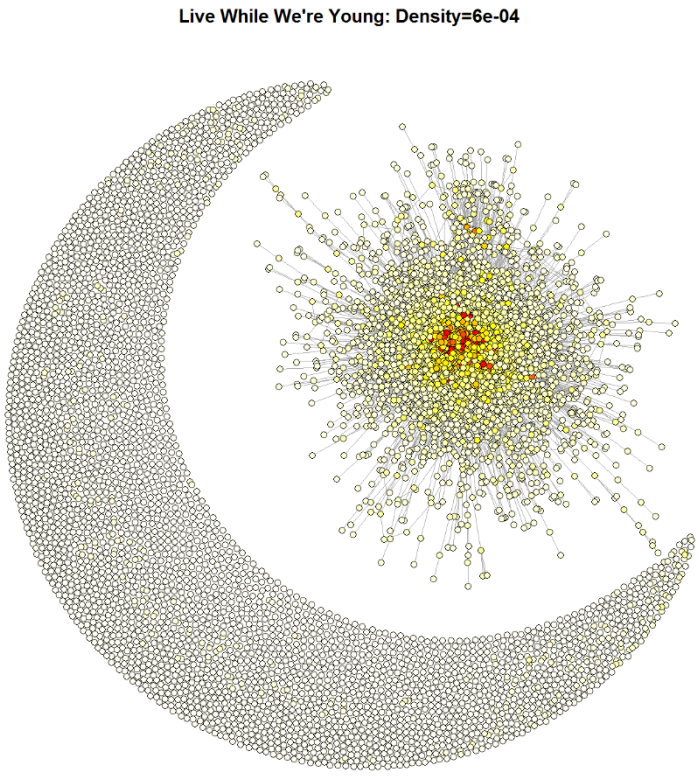


Figure C31. Preference Map

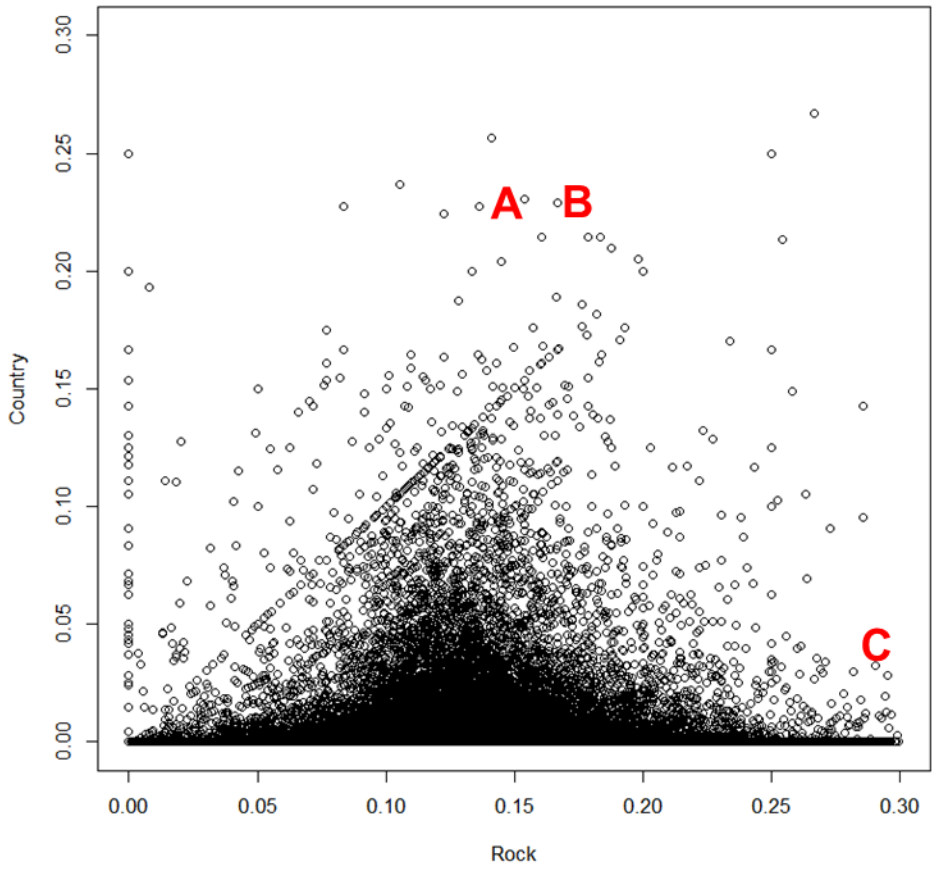


Figure C32. Social Influence and Distance Decaying Parameter

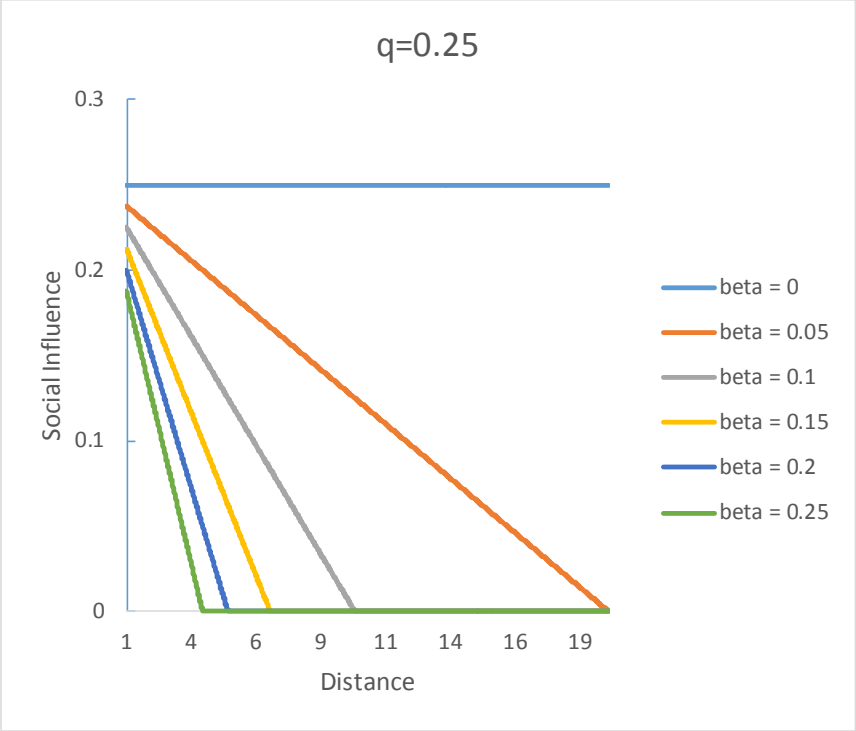


Figure C33. Social Tags for a Song

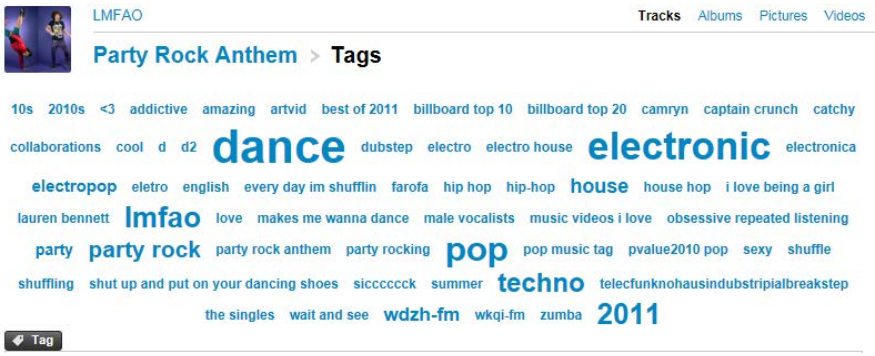


Figure C34. Social Tags Data from a User's Music Library

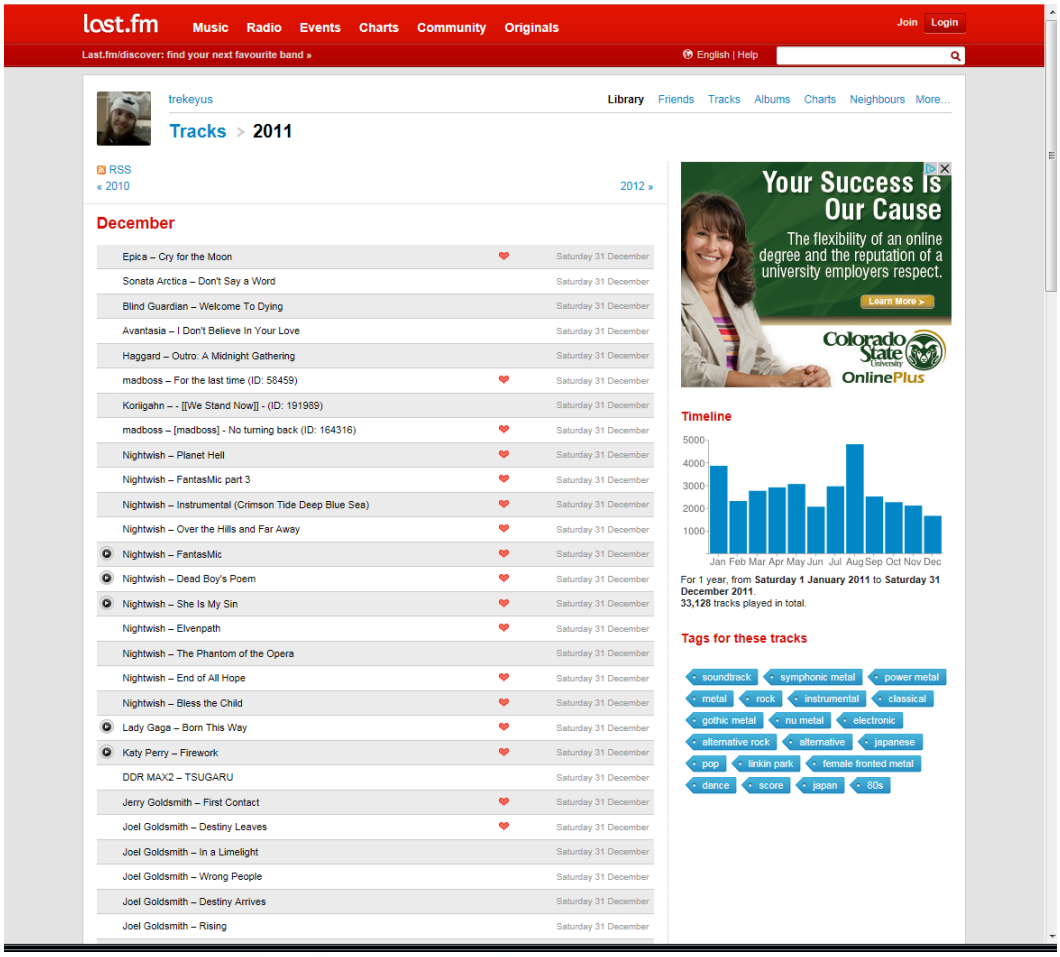


Figure C35. Mean of Distance Overtime

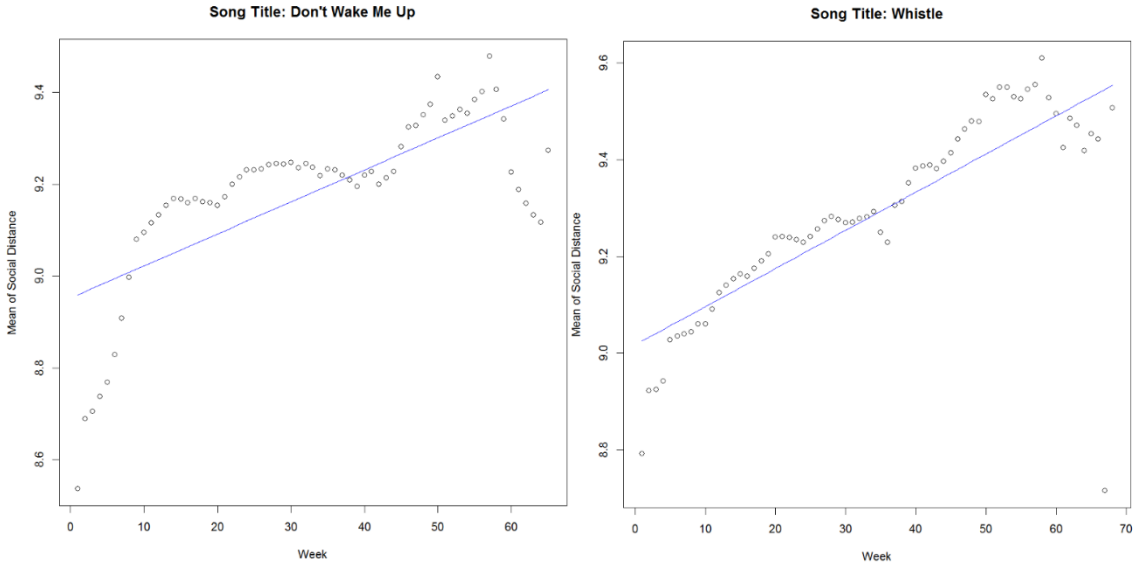
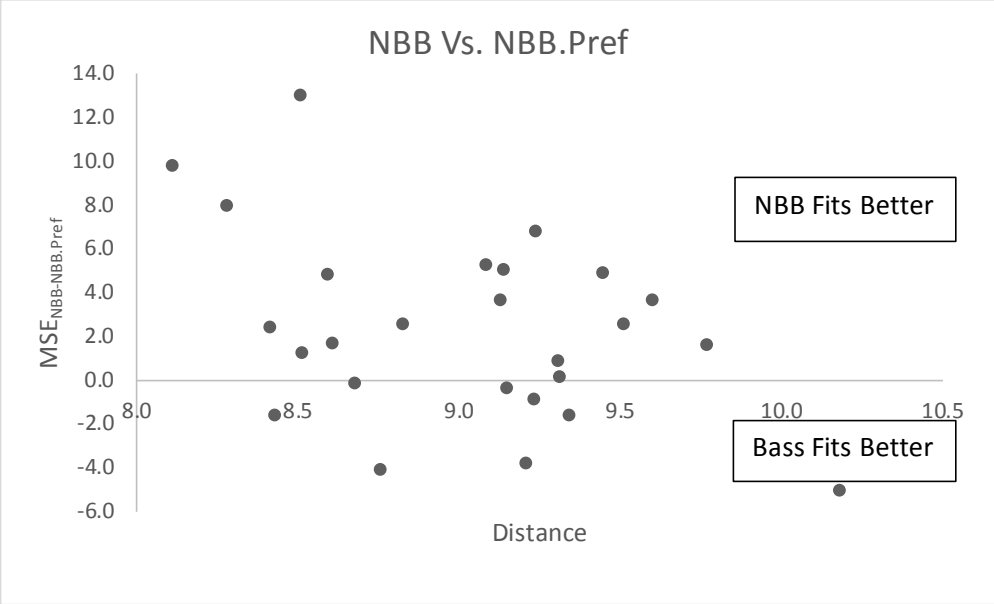
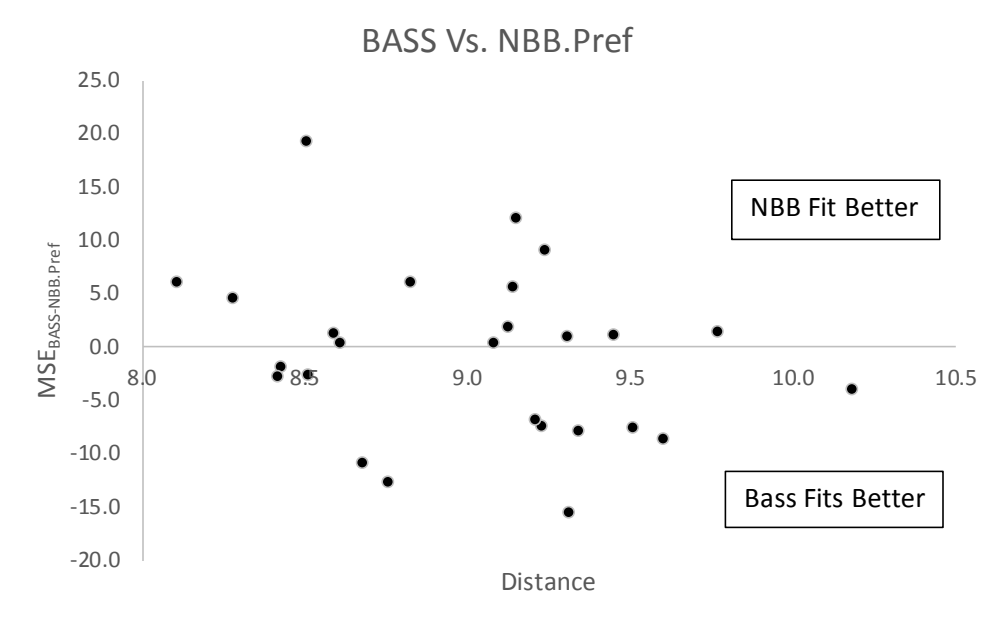


Figure C36. Distance and Model Fits





REFERENCES

Abrahamson, Eric and Lori Rosenkopf (1997), "Social Network Effects on the Extent of Innovation Diffusion: A Computer Simulation," *Organization Science*, 8 (3), 289-309.

Ahn, Yong-Yeol, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong (2007), "Analysis of topological characteristics of huge online social networking services," in Proceedings of the 16th international conference on World Wide Web: ACM.

Aitchison, John (1982), "The statistical analysis of compositional data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 44 (2), 139-77.

Aral, Sinan and Dylan Walker (2012), "Identifying Influential and Susceptible Members of Social Networks," *Science*, 337 (6092), 337-41.

Barabási, Albert-László and Réka Albert (1999), "Emergence of scaling in random networks," *Science*, 286 (5439), 509-12.

Bass, Frank M (1969), "A New Product Growth Model for Consumer Durables," *Management Science*, 15 (January), 215-27.

Bell, David R. and Sangyoung Song (2007), "Neighborhood effects and trial on the Internet: Evidence from online grocery retailing," *Quantitative Marketing and Economics*, 5 (4), 361-400.

Bemmaor, Albert C. (1994), "Modeling the Diffusion of New Durable Goods: Word-of-Mouth Effect Versus Consumer Heterogeneity," in *Research Traditions in Marketing*: Springer Netherlands.

Bradlow, Eric T, Bart Bronnenberg, Gary J Russell, Neeraj Arora, David R Bell, Sri Devi Duvvuri, Frankel Ter Hofstede, Catarina Sismeiro, Raphael Thomadsen, and Sha Yang (2005), "Spatial models in marketing," *Marketing Letters*, 16 (3-4), 267-78.

Brown, Jacqueline Johnson and Peter H Reingen (1987), "Social ties and word-of-mouth referral behavior," *Journal of Consumer Research*, 350-62.

Chatterjee, Rabikar and Jehoshua Eliasberg (1990), "The innovation diffusion process in a heterogeneous population: A micromodeling approach," *Management Science*, 36 (9), 1057-79.

Chen, Yuxin and Xinlei Chen (2008), "The Impact of Sampling and Network Topology on the Estimation of Social Inter-Correlations," *Available at SSRN 1319699*.

Delre, Sebastiano A., Wander Jager, and Marco A. Janssen (2006), "Diffusion dynamics in small-world networks with heterogeneous consumers," *Computational and Mathematical Organization Theory*, 13 (2), 185-202.

Dover, Y., J. Goldenberg, and D. Shapira (2012), "Network Traces on Penetration: Uncovering Degree Distribution from Adoption Data," *Marketing Science*, 31 (4), 689-712.

Ebbes, Peter, Zan Huang, and Arvind Rangaswamy (2008), "Sampling Large-scale Social Networks: Insights from Simulated Networks," in 18th Annual Workshop on Information Technologies and Systems. Paris, France.

Erdos, Paul and Alfréd Rényi (1960), "{On the evolution of random graphs}," *Publ. Math. Inst. Hung. Acad. Sci*, 5, 17-61.

Garber, Tal, Jacob Goldenberg, Barak Libai, and Eitan Muller (2004), "From Density to Destiny: Using Spatial Dimension of Sales Data for Early Prediction of New Product Success," *Marketing Science*, 23 (3), 419-28.

Gjoka, Minas, Maciej Kurant, Carter T Butts, and Athina Markopoulou (2009), "A walk in facebook: Uniform sampling of users in online social networks," *arXiv preprint arXiv:0906.0060*.

Goldenberg, Jacob, Sangman Han, Donald Lehmann, and Jae Hong (2009), "The role of hubs in the adoption processes," *Journal of Marketing*, 73 (2).

Goldenberg, Jacob, Barak Libai, and Eitan Muller (2001a), "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letter* (12:3), 211-23.

---- (2001b), "Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata," *Academy of Marketing Science Review*, 2001 (9), 1-20.

Granovetter, Mark (1973), "The Strength of Weak Ties," *American Journal of Sociology*, 78 (6), 1360-80.

Griffith, Daniel A (1985), "An evaluation of correction techniques for boundary effects in spatial statistical analysis: contemporary methods," *Geographical analysis*, 17 (1), 81-88.

Gupta, Sunil and Carl F Mela (2008), "What Is a free Customer Worth?," *Harvard Business Review*, 86 (11), 102-09.

Hauser, John, Gerard J. Tellis, and Abbie Griffin (2006), "Research on Innovation: A Review and Agenda for "Marketing Science"," *Marketing Science*, 25 (6), 687-717.

Heeler, Roger M and Thomas P Hustad (1980), "Problems in predicting new product growth for consumer durables," *Management Science*, 26 (10), 1007-20.

Hill, Shawndra, Foster Provost, and Chris Volinsky (2006), "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science*, 21 (2), 256-76.

Hinz, Oliver, Bernd Skiera, Christian Barrot, and Jan U. Becker (2011), "Seeding Strategies for Viral Marketing: An Empirical Comparison," *Journal of Marketing*, 75 (6), 55-71.

Iyengar, Raghuram and Christophe Van den Bulte (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30 (2), 195-212.

Jackson, Matthew O (2010), Social and economic networks: Princeton University Press.

Jones, Richard (2009), "Last.fm Radio Announcement."

Katona, Z. and M. Sarvary (2008), "Network Formation and the Structure of the Commercial World Wide Web," *Marketing Science*, 27 (5), 764-78.

Katona, Zsolt, Peter Pal Zubcsek, and Miklos Sarvary (2011), "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research*, 48 (3), 425-43.

Katz, Elihu and Paul Felix Lazarsfeld (1955), Personal Influence: The Part Played by People in the Flow of Mass Communications: Transaction Pub.

Kohli, Rajeev, Donald R Lehmann, and Jae Pae (1999), "Extent and impact of incubation time in new product diffusion," *Journal of Product Innovation Management*, 16 (2), 134-44.

Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong (2006), "Statistical properties of sampled networks," *Physical Review E*, 73 (1), 016102.

Leenders, Roger Th AJ (2002), "Modeling social influence through network autocorrelation: constructing the weight matrix," *Social Networks*, 24 (1), 21-47.

Leskovec, Jure and Christos Faloutsos (2006), "Sampling from large graphs," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining: ACM.

Liu, Ben Shaw-Ching, Ravindranath Madhavan, and D. Sudharshan (2005), "DiffuNET: The impact of network structure on diffusion of innovation," *European Journal of Innovation Management*, 8 (2), 240-62.

Mahajan, Vijay, Eitan Muller, and Frank M. Bass (1995), "Diffusion of New Products: Empirical Generalizations and Managerial Uses," *Marketing Science*, 14 (3_supplement), G79-G88.

---- (1990a), "New Product Diffusion Models in Marketing: A Review and Directions for Research," *Journal of Marketing*, 54 (1), 1-26.

Mahajan, Vijay, Eitan Muller, and Rajendra K Srivastava (1990b), "Determination of adopter categories by using innovation diffusion models," *Journal of Marketing Research*, 37-50.

Martín-Fernández, JA, C Barceló-Vidal, V Pawlowsky-Glahn, A Buccianti, G Nardi, and R Potenza (1998), "Measures of difference for compositional data and hierarchical clustering methods," in Proceedings of IAMG Vol. 98.

Newman, Mark, Albert-László Barabási, and Duncan J. Watts (2006), The Structure and Dynamics of Networks: Princeton University Press, Princeton, NJ.

Otero, N, R Tolosana-Delgado, A Soler, V Pawlowsky-Glahn, and A Canals (2005), "Relative vs. absolute statistical analysis of compositions: A comparative study of surface waters of a Mediterranean river," *Water Research*, 39 (7), 1404-14.

Peres, Renana, Eitan Muller, and Vijay Mahajan (2010), "Innovation diffusion and new product growth models: A critical review and research directions," *International Journal of Research in Marketing*, 27 (2), 91-106.

Rogers, Everett M (1983), "Diffusion of Innovation," New York NY The Free Press.

Schmittlein, David C and Vijay Mahajan (1982), "Maximum likelihood estimation for an innovation diffusion model of new product acceptance," *Marketing science*, 1 (1), 57-78.

Shaikh, Nazrul I., Arvind Rangaswamy, and Anant Balakrishnan (2010), "Modeling the Diffusion of Innovations Through Small-World Networks," *Working paper, Pennsylvania State University, University Park*.

Srinivasan, V. and Charlotte H. Mason (1986), "Nonlinear Least Squares Estimation of New Product Diffusion Models," *Marketing Science*, 5 (2), 169-78.

Tobler, Waldo R (1970), "A computer movie simulating urban growth in the Detroit region," *Economic geography*, 234-40.

Van den Bulte, C. and Y. V. Joshi (2007), "New Product Diffusion with Influentials and Imitators," *Marketing Science*, 26 (3), 400-21.

Wang, Y. H. (1993), "On the number of successes in independent trials," *Statistica Sinica*, 3 (2), 295-312.

Watts, Duncan J and Steven H Strogatz (1998), "Collective dynamics of ‘small-world’networks," *nature*, 393 (6684), 440-42.

Watts, Duncan J. (2002), "A simple model of global cascades on random networks," *PNAS*, 99 (9), 5766–71.