Measuring Information Diffusion in an Online Community

Rajiv Garg (rg@cmu.edu) *
Ph.D. Candidate
School of Information Systems and Management, Heinz College
Carnegie Mellon University

Michael D. Smith (mds@cmu.edu)
Professor of Information Technology and Marketing
School of Information Systems and Management, Heinz College
and Tepper School of Business
Carnegie Mellon University

Rahul Telang (rtelang@andrew.cmu.edu)
Professor of Information Systems and Management
School of Information Systems and Management, Heinz College
Carnegie Mellon University

* contact author

The authors thank David Krackhardt and Ramayya Krishnan for their valuable support and advice in completing this study. We also like to thank seminar participants at the 2009 Statistical Challenges in Electronic Commerce Research, the 2009 Conference on Information Systems and Technology, the 2010 University of Utah Winter Conference on Business Intelligence, and the 2011 Hawaii International Conference on Systems Science for valuable comments on this research. Finally, we are grateful to Rob Kauffman and three anonymous reviewers for very helpful feedback and suggestions.

Measuring Information Diffusion in an Online Community

Abstract

Measuring peer influence in social networks is an important business and policy question that has become increasingly salient with the development of globally interconnected ICT networks. However, in spite of the new data sources available today, researchers still face many of the same measurement challenges that have been present in the literature for over four decades: homophily, reflection and selection problems, identifying the source of influence, and determining pre-existing knowledge. The goal of this paper is to develop an empirical approach for measuring information diffusion and discovery in online social networks that have these measurement challenges. We develop such an approach and apply it to data collected from 4,000 users of an online music community. We show that peers on such network significantly increase music discovery. Moreover, we demonstrate how future research can use this method to measure information discovery and diffusion using data from other online social networks.

Keywords: information diffusion, peer influence, new content discovery, online music community, social influence, empirical research, data mining

1. Introduction

Empirical studies of information diffusion date back to the mid-twentieth century and focus on the diffusion of innovations; for example, new drugs in medical physician's networks [13], or new process and technique imitation by corporations [39]. Researchers also evaluated how product related word-of-mouth triggered the diffusion of information [3]. This created interest in evaluating the process of diffusion, especially for product marketing [4]. Over the next three decades, interest in information diffusion continued to develop among researchers in the social sciences, marketing [8, 37], and computer sciences [31] disciplines.

Online social communities provide a new channel for diffusing information, but at the same time estimating diffusion is now more challenging because of the large amount of information being exchanged on the Internet and the added uncertainty in identifying the information source [35]. The Internet has contributed to this uncertainty because personal communication has become more diversified: users now communicate in person, over analog channels (e.g. phones), and over new digital channels (e.g., email, social networks, discussion forums, instant messages). Thus our research focuses on these digital channels, which present challenges for estimating diffusion because of the large volume of untraceable information flows between individuals.

Measuring information diffusion online has become more important in the last decade in part because of significant growth in the use of social networks. A study by eMarketer [52] found that 41% of Internet users in the US visited a social network website at least once a month in 2008, an increase of 11% from 2007. Based on statistics from Alexa (www.alexa.com), the combined daily reach of Facebook (www.facebook.com) and Twitter (www.twitter.com) was 50% of daily Internet consumption in February 2011. While the growth of online social networks suggests a significant impact on online community members, empirical research is only beginning to analyze how online social communities help users discover and diffuse new content [2, 23, 24, 42].

At the same time, there are many empirical challenges to measuring diffusion in online social networks. For example, researchers have found that, in some contexts, online peers may not significantly influence diffusion because of the presence of a large numbers of peers causing limited interactions between those peers [28]. This is understandable when one considers the large number of peers one might interact with online. For example, users have an average of 130 “friends” on Facebook [18]. However, even with this large number it seems likely that some of these 130 connected friends are more valuable than others for diffusing new information to users. Likewise, peers and friends in online social networks tend to be self selected, leading to a significant selection problem. There are also empirical challenges from homophily [43] and contamination due to outside sources influencing diffusion [1].

Our goal in this paper is to develop an empirical method to measure information discovery and information diffusion that addresses these challenges and that can be used in the context of the data available in online social networks. After outlining our empirical method, we apply it to data on the music listening behavior of over 4,000 users of Last.fm, an online social network that allows users to consume, discover, and discuss music. Last.fm also allows us to isolate users for whom the platform provided by Last.fm is the only mode of communicating with each other. Using our empirical method, we identify a statistically significant causal influence of peers on music discovery in the network. Specifically we show that, on average, peers are six times more likely to diffuse a new song to other network users than they would be in the absence of those peers.

2. Literature Review

Information diffusion in social networks can be broadly classified into two categories: influence (by a system or a peer) and discovery (by active search or observational learning). Influence from peers occurs when individuals influence other individuals directly. Prior work has shown that peer influence has a positive effect in a variety of contexts [11, 16]. Influence through systems commonly occurs through recommender systems, which are used to influence and inform potential customers [34, 46].

We can further separate the discovery literature into discovery by active search and by observational learning. This lets us differentiate between two scenarios: one where a user makes an effort to find content and another where a user comes across content serendipitously, without significant additional effort. Active search on the Internet is accomplished through search engines or by seeking help on discussion forums. In this case the user knows what to look for, but her behavior towards the new content is unobservable. Observational learning has long been studied in the psychology literature, but has more recently attracted interest in business and economics literatures, and is classified as learning by either observable action or observable signal [7].

Peer influence may be described as a type of observable signal where actions from peers influence the decision of a consumer. For example, user generated content available on online forums can provide signals to influence other consumers [17]. The literature has shown that online forums such as blogs [29] and message boards [6] can be more effective in influencing consumers than direct marketing channels are. Our research extends this prior work to the context of online social networks and focuses on identifying the extent of additional learning from the presence of peers in an online platform.

Specifically, our research analyzes the role of peers in influencing others and the diffusion of new information. The literature has observed positive effects from peer based online marketing approaches such as online word-of-mouth [23] and viral marketing [32] when used as a means of influencing potential consumers. This influence happens not only because of the presence of the peers but also because of online word-of-mouth [22], which can build trust [45] and foster cooperation in online marketplaces [16]. Research has also shown that word-of-mouth helps consumers to make better and quicker decisions [26]. But we have also seen that word-of-mouth diffuses not only positive information but also negative information, which dominates in many cases [36]. Finally, the literature has analyzed social influence in a variety of online settings such as computer mediated communication [50], email [48], and instant messaging [44].

The literature has also shown that informed consumers prefer differentiated products [12], and since music is a highly differentiated product, information shared in online social networks may make consumers more aware of music available in the market. This strengthens the need for measuring the extent of peer influence and information diffusion in an online social network.

Thus, in this paper we attempt to identify peer influence and to quantify the extent of diffusion in an online community for music, while attempting to use the unique characteristics of our data to address the estimation challenges commonly faced in existing studies: selection, homophily (tendency of individuals to associate with similar others), identification of the diffusion source, a user's pre-existing knowledge, and the size of a user's personal online network. Thus, one key contribution of our study is to provide an empirical approach whereby traditional estimation challenges could be reduced when analyzing large datasets available on the Internet.

In addition to contributing to the peer influence literature outlined above, our research also contributes to the growing literature in Information Systems analyzing the impact of ICT systems on online networks,[1] and the growing literature in marketing and information systems analyzing word-of-mouth in online markets.[2]

3. Methodology

In this section we discuss empirical challenges in studying information diffusion in online social networks, an ideal experimental scenario for detecting diffusion, and a feasible approach for analyzing an available archival dataset to cleanly identify diffusion.

3.1. Estimation Challenges

Owing to the openness of information, the Internet, at a macro level, has simplified the measurement of diffusion of a new product in a social community: once a new product is launched, one can study how quickly a product can diffuse in a community [4]. However, it becomes harder to identify whether the information was diffused due a particular online platform, especially for existing non-novel content. Because of this, we examine diffusion at a micro level: between two individuals in an online community and because of the communication and interaction medium provided by that community. In other words, we seek to identify whether members of an

[1] See for example several recent papers in the Journal of Management Information Systems such as [51] in the context of positive influence on technology use, [30] in the context of diffusion of software, [14] in the context of review creation, [19] in context of co-creation and cooperation between consumers, and [5] in the context of music sales in presence of piracy.

[2] Representative papers in this literature include [33] which suggests positive and negative effect of self-selection on consumer reviews, [27] which analyzes role of network structure on information diffusion when bidding for secret reserve price auctions, [10,49] which analyze the role of influencers, imitators, and opponents on diffusion of innovation, [15] which compares various diffusion models to estimate the effect of consumer reviews on box office sales, and [21] which discusses the consumer's valuation of products in the presence of alternate secondary markets.

online social platform discover something new from their peers on that platform and because of the existence of that platform.

Although information diffusion and peer influence has been studied by social scientists for many years, estimation of micro-level diffusion on online channels [25] is still attracting innovative identification strategies. There are several notable challenges that exist with online diffusion studies: the reflection problem, homophily, the confounding effect of diffusion source, media influence, noise in data, and the availability of an actual dataset. We outline these challenges in more detail in the following sections.

3.1.1. Reflection Problem and Homophily

Most social influence studies face the reflection problem (e.g., [40] and [47]), which suggests that the behavior of individuals could be a reflection of the peers they associate with or other environmental factors. This adds complexity to cleanly estimate diffusion in an online social network because of the presence of endogenous effects, which can be defined as an environment *"wherein the propensity of an individual to behave in some way varies with the prevalence of that behavior in the group."* [41, p.1]

This endogenous effect implies that the behavior of individuals may be similar because of shared characteristics that could be interpreted incorrectly as influence. These shared characteristics can arise from homophily [43], which is often expressed with the adage "birds of a feather flock together." In online social networks any two users may have an inherent propensity to discover the same piece of information because of homophily (shared behaviors, beliefs, interests, or characteristics). Thus diffusion from one user to another may not be cleanly identified. In the case of

Last.fm (the platform used in this study), two individuals might share the same interests in genre, artist, band, broadcast station, and fan-base and thus any discovery could arise because of that intersection of interests and not diffusion of information from one individual to another.

A correlated effect, which can be defined as a situation *"wherein individuals in the same group tend to behave similarly because they face similar institutional environments or have similar individual characteristics"* [41, p.1] is a significant obstacle in empirical studies of diffusion. In the case of Last.fm, correlated effects suggest questioning whether diffusion between two individuals is a result of shared information or of shared environments (e.g., residential neighborhood, school/college, or workplace). The correlated effect also suggests analyzing if two individuals share the same characteristics. For example two saxophone playing women in their mid-thirties might tend to discover the same new song by "Jazzmasters" because of their (shared) interests, as opposed to from information diffusion on the social network.

3.1.2. Confounding Effect of Diffusion Source

The reflection problem is a significant challenge in estimating diffusion in traditional environments, and online platforms exacerbate this issue by introducing challenges in identifying the source of diffusion. For example, if we observe diffusion of the song "Touch and Go" by "Jazzmasters" from one individual to her peer, we still have significant uncertainty around the source of the diffusion. One cannot confidently identify the source of diffusion as the observed peer because of the possibility of influence by other peers and outside media.

The size of a social network on Facebook, MySpace, or LinkedIn is frequently more than 150 peers or friends. Thus the probability that a single piece of information was diffused from a specific individual becomes much lower. It is possible that an individual discovers new information from any of the other peers in the network, and this can cause an overestimation of diffusion. Similarly a large number of social connections could simultaneously exist on other online or offline platforms, creating multiple channels for diffusion. Thus, the true value of diffusion on a social platform could actually be much lower than observed diffusion without controlling for the true source of diffusion.

*****Insert Figure 1 Here *****

To illustrate this with an example consider Figure 1, which shows a hypothetical case of the diffusion of four songs in an online community from four peers of an individual (in the middle). If we do not control for the true source of diffusion, we will see that each peer causes an individual to discover one new song. However, after controlling for the sources of diffusion (Figure 2), we see that in reality there was only one song discovered from one peer (an average of 0.25/peer) on the online social network. This happened because the other three songs were not diffused by the directly connected peers, but rather (1) through the media, (2) indirectly through other peers, and (3) through an offline channel. This highlights the importance of accounting for the source of diffusion.

*****Insert Figure 2 Here *****

3.1.3. Noise and Media

Assuming that individuals are only discovering information from their peers in an online community is an overly ambitious assumption for diffusion studies. Instead, we need to allow for the possibility that users discover information from media or through sampling of content. This issue is more prevalent in a study of music diffusion because of the availability of a vast number of technologies allowing users to sample music in many locations and at many different times. To address this issue, most studies use a control user to account for “chance” discovery and we follow suit in this paper. Additionally, we use the strength of a large group of homophilic peers of a user to further control for any false positives in music diffusion.

3.1.4. Data

Another challenge faced when observing diffusion in online social networks is obtaining rich data on a user’s behavior. This is especially challenging in online social networks because of concerns of privacy. For this reason, we selected music as the target of our observation for diffusion because of the reduced chance that observing music listening behavior will reveal evidence leading to identification of an individual’s identity. This allows users to track and share much of their listening behavior online.

Music is also a useful setting for diffusion studies because of its large consumption volume and because it features two distinct dimensions of measurement: bands and songs. These two metrics are important because songs represent a single unit of information and bands represent an aggregated information category. To better identify this distinction, throughout the remainder of the paper we refer to “music” when the statement is independent of a particular band or song and we refer to “bands” and “songs” when there is a dependence on the granularity of information.

3.2. The Ideal Experiment

How can a researcher address these empirical estimation challenges? To answer this question, first consider an ideal experiment to measure peer influence in an online social network. To cleanly measure peer influence in an online social network we would need to observe all interaction between two random users in a closed environment, while preventing any flow of information into the network from external sources. We would also need control users who are not interacting with other users, to control for diffusion that might occur because of other uncontrolled sources or inherent propensity to discover music. Additionally, we would need to observe diffusion of completely novel or niche information to account for any pre-existing knowledge of a user. Thus, observation of a controlled exchange of niche content in a closed online environment can allow us to estimate the peer influence in the experimental network.

Unfortunately, conducting this experiment in a real world environment is not only difficult but also poses challenges in the selection of participating candidates. Therefore in this study we use an alternative approach that mimics this ideal scenario by utilizing a large volume of archival data from Last.fm to estimate peer influence.

3.3. Alternate Approach

Because of the difficulty in conducting an "ideal experiment," in this study we pick "neighbors" as the potential source of diffusion. These neighbors (recommended peers) are typically strangers to the user and are recommended by Last.fm based on an observed matched interest in music. Thus these peers have no other mode of communication with the users except for modes offered by the Last.fm network. From this, we gain access to most of the content exchanged between users and their neighbors on Last.fm.

We also know a user's playlist before she connects to a new neighbor, and from this we can readily identify if songs from the peers diffused to the user. Together, the use of neighbors as peers and control over the diffused music emulates, to some extent, the “ideal” closed environment for diffusion estimation discussed above.

To account for any pre-existing knowledge of a user, we remove all songs or bands listened to by the user and her peers (non-new neighbors) from the list of songs or bands available for diffusion. This reduced playlist represents content that could be diffused to a user by the new neighbors, and includes only those songs that have not been played by the user or anyone else in the user’s neighbor network.

Finally, we need a control user to account for any by-chance discovery of the pool of songs available for diffusion. This control user population needs to be similar to the target users — users that are connected to and discovering content from neighbors. Therefore we pick control users that share a similar interest in music as the target users, but who don’t directly influence the target user’s behavior in the observed time period. To ensure similarity and the absence of a current connection we identify potential control users by observing network dynamics and new ties formed in future time period. This allows us to select control users who are similar to the music diffusing peers, but who are not connected to them at the time of observation. This set of control users then allows us to estimate the discovery of new content from sources other than the diffusing neighbor, and to adjust our estimate of peer influence accordingly.

With this setup, we are able to account for common challenges in the measurement of peer influence or information diffusion. Selection issues are addressed by using system recommended neighbors (who are not friends). Endogenous and correlated effects are reduced by removing any homophily between the music discovering users and her neighbors during the selected timeframe. Pre-existing knowledge is accounted for by screening out music already played by the user. Finally diffusion, “by chance” or from external sources is controlled by using a control group of users who are similar to, but are not currently connected to the target user.

3.3.1. Empirical Model

In this section, we explain our empirical approach mathematically. For simplicity we summarize our notations in Table 1. Within this notation we express the total number of music discovering users i as n_i, and the total number of music diffusing (new) neighbors j to a user i as $n_{i,j}$, and any other connected neighbors k as $n_{i,k}$. We also express the total number of distinct songs played by all users i and neighbors j and k as n_s and total number of distinct music bands as n_b. Then a binary row vector S indicating all songs (and B indicating all bands) listened to by an individual i between time periods t_1 and t_2 is given as follows:

$$S_{i,(t_1,t_2)} = [s_1 \quad \cdots \quad s_{n_s}] \tag{1}$$

$$B_{i,(t_1,t_2)} = [b_1 \quad \cdots \quad b_{n_b}] \tag{2}$$

*****Insert Table 1 Here *****

For the sake of simplicity, we use M to denote music that represents either songs or bands. This allows us to create one equation with M, where M is used in lieu of S or B. Thus:

$$M_{i,(t_1,t_2)} = [m_1 \quad \cdots \quad m_{n_m}] \tag{3}$$

Now assume that there are three non-overlapping time periods of interest: pre-connection, connection, and post-connection. Our goal then is to estimate diffusion from a new peer j who is connected to our music-discovering peer i during the “connection” period. We then detect the music that was played by this peer j during the “pre-connection” period and discovered by user i during “post-connection” period. Thus diffusion from user j to user i can be presented as an intersection (or dot-product) of their respective M vectors across two distinct time periods:

$$D_{i,j}(T,T+\Delta t) = M_{i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)} \tag{4}$$

Here the time interval *(0, T - t_c)* represents the "pre-connection" time period, *(T, T + Δt)* represents the "post-connection period, and *(T – t_c, T)* represents the "connection" period. We use the "connection" period to (1) account for uncertainty around the actual time of the connection of the users and (2) dilute the correlated effect or diffusion because of other environmental elements (for example, radio stations).

To account for a user's pre-existing knowledge we remove all music previously listened to by user *i* (in both the pre-connection and connection time periods). Thus, Equation 4 becomes:

$$D_{i,j}(T,T+\Delta t) = \left[M_{i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)}\right] \cdot \left[1 - M_{i,(0,T)}\right] \tag{5}$$

Still, there is a possibility that the diffused music really came from other peers of user *i* and not peer *j*. To address this issue, we remove all content that could possibly be diffused from other peers.

$$D_{i,j}(T,T+\Delta t) = \left[M_{i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)}\right] \cdot \left[1 - M_{i,(0,T)}\right] \cdot \prod\nolimits_{k \neq (i \text{ or } j)} \left[1 - M_{k,(0,T)}\right] \tag{6}$$

$$= \left[M_{i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)}\right] \cdot \prod\nolimits_{k \neq (j)} \left[1 - M_{k,(0,T)}\right] \tag{7}$$

Thus, Equation 7 represents the diffusion from neighbor *j* to user *i* after reducing the effect of homophily and the uncertainty of other peers as a source of diffusion. Since we selected neighbors as peers who are diffusing music, our issue of a diffusion source from an alternate platform is also minimized. Still we need to address the issues of correlated effect and noise and media.

To further minimize the effect of media and noise we use a control user who is homophilic to the neighborhood, but who is not connected to the music diffusing neighbors *J*. First, we find all potential control users from the list of neighbors *K* that are not connected to music-diffusing neighbors *J*. Then from that list we find one control user c_i that is most similar (in terms of music listening behavior) to the target user *i*. This allows us to have a strong control for media and noise using homophily because control user is very similar to target user *i* and all of the music-diffusing neighbors *J* but is not connected to those neighbors.

Using Equation 7, we can define the vector of all music diffused by all neighbors *J* as:

$$D_{i,J}(T,T+\Delta t) = 1 - \prod_{j\in J}\left[1 - \left[\left[M_{i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)}\right] \cdot \prod_{k\neq(j)}\left[1 - M_{k,(0,T)}\right]\right]\right] \tag{8}$$

$$D_{c_i,J}(T,T+\Delta t) = 1 - \prod_{j\in J}\left[1 - \left[\left[M_{c_i,(T,T+\Delta t)} \cdot M_{j,(0,T-t_c)}\right] \cdot \prod_{k\neq(j)}\left[1 - M_{k,(0,T)}\right]\right]\right] \tag{9}$$

Equation 8 estimates diffusion from the newly connected music diffusing neighbors *J* to connected target user *i* and Equation 9 estimates diffusion from the newly connected music diffusing neighbors *J* to a non-connected control user c_i. Taking the dot product of these binary vectors for multiple users results in a vector that has 1s for music that is played by every individual. Similarly a dot product of a complementary vector *(1-M)* gives a list of songs that are not played by any individual in the community. Thus equations 8 and 9 strategically account for most of the empirical challenges discussed above. Notice that after accounting for various controls, the music available for diffusion will be mostly niche music. Thus our estimates should be viewed as a lower bound on the effect of peers in diffusing music.

Binary vector $D_{i,J}$ represents all music that was diffused to user *i*. Given this, we define two other variables: (1) binary variable $Y_{1,i}$ indicating the existence of diffusion and (2) an integer (count) variable $Y_{2,i}$ indicating the total music diffused to user *i*. These two variables are given as:

$$Y_{1,i} = \begin{Bmatrix} 0 \; if \; d = 0 \; \forall \; d \in D_{i,J}(T, T + \; \Delta t) \\ 1 \qquad\qquad\qquad\qquad\qquad otherwise \end{Bmatrix} \tag{10}$$

$$Y_{2,i} = \; \sum_{d\epsilon \, D_{i,J}(T,T+\,\Delta t)} d \tag{11}$$

Using the above strategy to dissect the archival data, we can conservatively estimate the extent of peer influence on the online community. Our last challenge is obtaining data for the analysis.

3.4. Data

We use the online community created by Last.fm (www.last.fm) as our empirical setting. Last.fm is a platform to share, listen and discuss music. The site is part-owned by CBS Interactive and has an estimated user-base of 40 million active users in over 200 countries. Based on statistics from Alexa in December 2009, Last.fm reaches 0.33% of daily Internet users (globally) and is the second most popular music community online, behind Pandora. An interesting feature of Last.fm is scrobbling, which allows Last.fm's users to track the music they have listened to online or off-line (from personal computers or portable music players). From a research perspective, scrobbling provides us with access to an extremely rich dataset, enabling us to observe both the current and past listening behavior of users. Additionally, Last.fm allows users to socially connect with friends and other registered users on the website. Finally, Last.fm recommends a list of 60 neighbors who share a similar taste in music to the user, allowing the user to socially connect and interact with both existing friends and new similar users. Figure 3 provides a snapshot of the neighborhood provided by Last.fm. These recommended neighbors change as users' music listening behaviors change over time.

*****Insert Figure 3 Here *****

To collect data from this network, we needed a random set of target users. To achieve randomness we captured a list of 500 active users from the 40 million registered users on Last.fm during five different time periods in April 2008. Of these 500 users we used a simple random number generation algorithm to pick 50 target users. We selected our random users from all registered users because music listening data was available only for members of the community. We then found the neighbors of these target users and identified 50 control users who were similar to these target users, but who were not connected with the music diffusing neighbors. We then collected data on the neighbors of these 100 users (50 target users and 50 control users). This resulted in data for about 4,017 neighbors leading to 21 million data points over nine months of "historical usage." Our final dataset contains network information for the 50 target users during three non-overlapping time periods (January to April 2008, April to July 2008, and July to September 2008), and the playlist (songs and the time the user played the song) for each user. Figure 4 provides a snapshot of a representative user's playlist.

*****Insert Figure 4 Here *****

During the analysis we found that some users had missing playlist data, possibly because of a change in their privacy settings. Dropping these users from the study, we ended up with 35 target users and 40 control users who had data available for the entire nine-month period. Table 2 lists summary statistics for this data.

*****Insert Table 2 Here *****

The selection of these time periods was especially important in our research methodology. We consider three different time periods: pre-connection (or creation) from January to April 2008, connection from May to July 2008, and post-connection (or discovery) from August to September 2008.

During the connection period we observe changes in the network, and specifically the entry of music diffusing new neighbors who played songs that were new to the entire network. Increasing the duration of connection period (t_c) would allow us to better control for any environmental effect but would reduce the probability of diffusion from a peer because of additional delays. To balance the two considerations, and observing an average of one new music-diffusing neighbor replaced per week, we selected t_c to be about 10 weeks to have about 10 new music diffusing neighbors.

The pre-connection duration (T-t_c) was selected to observe the formation of networks and listening behavior of all users and to control for any pre-existing knowledge of a user. A large pre-connection duration could cause selection issues associated with the user's length of membership on the platform, and a shorter duration could cause underestimation of pre-existing knowledge. To balance the two considerations, and since Last.fm launched the free music initiative in January 2008, we selected a pre-connection period (T-t_c) of 16 weeks from January to April 2008.

During the post-connection (July and September 2008) time period (Δt) we observe the discovery of new songs that were introduced by the new neighbors. A small duration may provide no observations and a large period could increase complexities from network dynamics. To balance these two considerations we selected Δt to be 10 weeks. Summary statistics for each phase are given in Table 3.

*****Insert Table 3 Here *****

Figure 5 summarizes our empirical process. To explain our data more clearly, consider a target user “Sue” (who is connected to say, 10 new neighbors) and a control user “May” (whose taste in music is similar to Sue but who is not connected to any of Sue’s 10 new neighbors). Suppose Sue and May have 60 other neighbors that they are already connected to. Let’s assume 10 of those new neighbors played about 500 songs, of which 300 were played by the other neighbors as well. Eliminating all songs played by other neighbors (other users in the network) and by Sue herself, we find that new neighbors expose Sue to 87 new songs. Similarly, “May” is exposed to 98 new songs. Of these potential songs that can possibly diffuse, we observed diffusion for Sue and May to be 10 and 3 respectively. Controlling for other characteristics, the difference in Sue’s and May’s diffusion rate is the effect of the peers. Put another way, Sue is discovering additional new content as compared to May because she is connected to the new neighbors.

*****Insert Figure 5 Here *****

4. Analysis

We assume diffusion has happened when a song that was played by a music diffusing new neighbor (*J)* in the pre-connection period shows up in the music discovering user’s (i) playlist in the post-connection period. As discussed previously, we pick only songs that are new to the entire network of a user: that is, only songs or bands that are not played by the user or any of her neighbors (*K*) in any of the time periods prior to diffusion. To ensure that a song is indeed diffused, we consider diffusion only when the user played the song at least two times. A simple regression model could be defined as follows:

$$Diffusion \ = \alpha * (Target/Control\ Dummy) + \textstyle\sum_u \beta_u \ * (user\ characteristics) \qquad (12)$$

Here the dependent variable, *diffusion*, takes the form of a binary occurrence of diffusion ($Y_{1,i}$) or, a count of music (bands/songs) diffused to a user ($Y_{2,i}$). The two variations of the dependent variable allow us to not only estimate the existence of music diffusion between peers but also to quantify the extent of diffusion because of online peers. The count of diffusion of bands and songs for both target and control users is given in Figure 6 below.

*****Insert Figure 6 Here *****

The independent variables are the users' music listening characteristics: the number of unique bands or songs listened during the post-connection period, the number of new bands or songs the new-neighbors made available for diffusion, and the listening heterogeneity of a user (described below). The parameter of interest is the coefficient on the target/control indicator variable.

4.1. User Characteristics

When evaluating diffusion, we need to control for user characteristics that may influence users' music listening behavior and hence diffusion. We consider the following characteristics:

Quantity of music played is the number of unique bands or songs in a user's playlist. Two music listeners could be very different in terms of their exploratory nature. A user listening to a larger diversity of music may be more interested in discovering new music. Since the average quantity of music played is large and has a large variance, we use the log value of this characteristic in our regressions.

Quantity of new music exposed reflects the amount of music exposed to a user. Since each user gets exposed to a different set of music diffusing new neighbors who may bring in a different quantity of new content, we would expect that more exposure will lead to higher diffusion. Since the average quantity of music exposure is large and has a large variance, we use the log value of this characteristic in our regressions.

Heterogeneity in listening behavior captures a user's propensity to listen to more diverse music. We capture this heterogeneity by the Gini coefficient [20], which measures the inequality or statistical dispersion in the data. Since diversity of music in a user's playlist follows approximately a Lorenz curve with unique bands/songs on the x-axis and the number of repetitions on the y-axis, we define the Gini coefficient as follows:

$$G = \frac{A_{Equality} - A_{Lorenz}}{A_{Equality}} \qquad (13)$$

$$\text{or, } G_i = 1 - 2 * \left(\frac{\sum_{p=1}^{n_m} (f_{i,p})}{max_p(f_{i,p}) * (n_m + 1)} \right) \qquad s.t. \{ f_{i,p} \in F_{i,(T,T+\Delta t)} \} \qquad (14)$$

Here $F_{i,(T,T+\Delta t)}$ is a vector representing the frequency of each piece of music (band or song) played by the user i during post-connection period $(T,T+\Delta t)$. The summation in numerator computes the total entries in a user's playlist and $max_p()$ in the denominator represents the maximum frequency value of a piece of music (band or song) in the playlist. A smaller Gini coefficient represents more diversity in the music listened to by a user. The representation of four different listening behaviors is illustrated in Figure 7 below.

*****Insert Figure 7 Here *****

Over the entire nine months we collected comprehensive data on 35 target users i and 40 control users c_i. On average, the number of new neighbors ($n_{i,j}$) was 15. The music listening statistics for these users are presented in Table 4.

*****Insert Table 4 Here *****

4.2. Control Users

Since we picked control users based on homophily and absence of connections with music diffusing neighbors, it is important to compare the relative similarity of both target and control users with the music diffusing neighbors. To avoid any bias in the measurement of influence, we test the extent of similarity in music listening behavior for both the target and control users we use various distance measurements to compare their behaviors.

One potential metric is the Euclidean (ordinary) distance between two users. We define this distance measures as the difference in the music listening patterns of two users and is computed by taking the distance between the two vectors representing the frequency of each intersecting song played by each user. Let the Euclidian distance between a target user *i* and her neighbor *j* is presented as *E(i, j)* and the distance between a control user c_i and *j* be represented as *E*(c_i*, j)* as follows.

$$E(i,j) = \sqrt{\sum_{p=1}^{n_m} \left| f_{i,p} - f_{j,p} \right|^2} \qquad s.t. \begin{Bmatrix} f_{i,p} \epsilon\ F_{i,(T,T+\Delta t)} \\ f_{j,p} \epsilon\ F_{j,(T,T+\Delta t)} \end{Bmatrix} \tag{15}$$

$$E(c_i,j) = \sqrt{\sum_{p=1}^{n_m} \left| f_{c_i,p} - f_{j,p} \right|^2} \qquad s.t. \begin{Bmatrix} f_{c_i,p} \epsilon\ F_{c_i,(T,T+\Delta t)} \\ f_{j,p} \epsilon\ F_{j,(T,T+\Delta t)} \end{Bmatrix} \tag{16}$$

Here $f_{i,p}$ is the p^{th} element of the frequency vector F_i. Thus, for each user *i* we need to test if the Euclidean distances between *i* and *j*, and c_i and *j* are similar. The paired *t-test* for both set of Euclidian distances has a *p-value* of 0.0761. Thus we can say, with 90% confidence, that both the target and control users are similar to music diffusing new neighbors based on Euclidean distances. Euclidean distance measures of both target and control users are presented in Figure 8.

*****Insert Figure 8 Here *****

We also tested the similarity of the target and control users with the music diffusing new neighbors using a Gini coefficient that measures statistical dispersion in the music listening behavior of two sets of users. We find here the *p-value* from a paired *t-test* is 0.0937, which is also within the 90% confidence interval for both kinds of users being similar to the music diffusing new neighbors. This dispersion measure doesn't really compute the distance between the users, but gives us a better understanding of listening behavior as a combination of diversity and repetition of songs in a user's playlist. The model to compute the Gini coefficients for both target and control users is presented in equations 17 and 18 below and actual measures are shown in Figure 9.

$$G(i,j) = 1 - 2 * \left[\frac{\sum_{p=1}^{n_m}(|f_{i,p} - f_{j,p}|)}{(n_m + 1) * max_p(|f_{i,p} - f_{j,p}|)}\right] \quad s.t. \begin{Bmatrix} f_{i,p} \in F_{i,(T,T+\Delta t)} \\ f_{j,p} \in F_{j,(T,T+\Delta t)} \end{Bmatrix} \tag{17}$$

$$G(c_i,j) = 1 - 2 * \left[\frac{\sum_{p=1}^{n_m}(|f_{c_i,p} - f_{j,p}|)}{(n_m + 1) * max_p(|f_{c_i,p} - f_{j,p}|)}\right] \quad s.t. \begin{Bmatrix} f_{c_i,p} \in F_{c_i,(T,T+\Delta t)} \\ f_{j,p} \in F_{j,(T,T+\Delta t)} \end{Bmatrix} \tag{18}$$

*****Insert Figure 9 Here *****

Thus from the above two measures — Euclidean distance and the Gini coefficient — we can say, with 90% confidence, that both target and control users are similar to the music diffusing neighbors. This strengthens our selection of control users in measuring peer influence.

5. Results

We estimate Equation 12 with diffusion as the dependent variable and report the results in Tables 5 and 6. There are two interesting observations here: (1) evidence of discovery because of online peers in the presence of control users and (2) quantifying the extent of discovery.

First we evaluate *diffusion* as a binary variable (specifically, 1 if diffusion occurs and 0 otherwise) using a logit model and report the findings in Table 5. We find that the coefficient for the target/control dummy variable is positive (3.4 for bands and 6.1 for songs) and significant at a 10% level. This suggests that diffusion of new bands is 3.4 times more likely (6.1 times more likely for new songs) to occur in the target group than in the control group.

*****Insert Table 5 Here *****

Additionally we see that users who listen to more songs are more likely to see diffusion. A 1% increase in the average number of distinct bands listened to (104) increases the odds-ratio of diffusion of a new band by 0.07. Similarly a 1% increase in the average number of distinct songs listened to (439) increases the odds-ratio of diffusion of a new song by 0.13. This is intuitive because the more music a user listens to, the more she is prone to discovering.

Finally, users who are exposed to a larger volume of new content are also likely to see more diffusion — a 1% increase in the average number of distinct bands played by peers (485) increases the odds-ratio of diffusion by 0.04. This change is approximately the same for the 1% increase in the average number of distinct songs played by peers (3,166). This follows a similar intuition as before, except the results are driven by the behavior of the neighbors whereas previously they were driven by the behavior of the user. In other words, users who are close to peers who are listening to more new songs, tend to get a spillover effect in new music discovery.[3]

Next we evaluate diffusion as a count of the number of unique bands/songs diffused to a user. Because of over-dispersion in the count data (seen from non-zero values of α in Table 6) we use

[3] We note that we did not observe is a significant role of the Gini coefficient in this model. This implies that the statistical dispersion in music listening behavior does not play a role in suggesting the presence of diffusion.

a negative binomial regression [9, 53]. We believe that the music listening behavior and heterogeneity among random users is the cause of this over-dispersion. A chi-squared test for dispersion in data provided a *p-value* equal to zero rejecting the null hypothesis (*$\alpha = 0$)*. Thus we use a negative binomial regression for this analysis and report the resulting estimates in Table 6.

*****Insert Table 6 Here *****

In the case of individual songs, the marginal effect for the target/control dummy is positive (2.7) and significant at the 5% level, suggesting that peer influence leads to diffusion of 2.7 additional unique songs to a target user.

Additionally a 1% increase in the number of songs played by a user suggests diffusion of an additional 2.3 unique songs, and a 1% increase in exposure to new songs increases the diffusion by 1.9 songs. We also see that a 1 standard deviation (0.137) increase in the Gini coefficient leads to a diffusion of 0.5 additional songs. In the case of diffusion of bands, the coefficient of the target/control indicator variable is positive (0.4738) but insignificant (p-value: 0.16).

Since the Gini coefficient is a function of the two other independent variables (unique music listened to and unique music exposure), a possible concern could be the correlation between the variables. But from Tables 7 and 8 we see that the correlation between the variables is very low, especially for songs.

*****Insert Table 7 Here *****

*****Insert Table 8 Here *****

Testing for multicollinearity, we found that the variance inflation factor (VIF) is 1.0 for the bands and songs regression, suggesting that multicollinearity is not a problem in our data [38].

6. Discussion

6.1. Findings and Contribution

In this paper, we find that online peers have a positive influence on the diffusion of new music. Users are 6.1 times more likely to discover a new song and 3.4 times more likely to discover a new band as a result of peer influence. There are two key contributions of our work:

First, from a methodological perspective we provide an empirical approach to test for diffusion in online networks and to overcome many key challenges in estimating peer effects. Moreover, we do this in a field setting as opposed to the more commonly used survey or laboratory setting. Thus our paper provides a roadmap for using a large yet noisy dataset for estimation of peer effects in online social networks.

Second, from a managerial and research perspective, we provide empirical evidence that even a network with extremely weak ties and where peers don't know one another can aid information discovery among users. We observed that new songs seem to diffuse in such a network, suggesting a significant power of online networks in content discovery. We believe this is a notable finding as marketers seek to both measure and harness the power of online networks to diffuse information about their products.

Indeed, we believe that recommending peers, as modeled by Last.fm, could be a new trend in marketing that could benefit from high consumer involvement, increased online trust between peers, and "pull" marketing strategies. While peer recommendation may not guarantee diffusion of a product, we think that the methodology outlined here will be effective in measuring influence and possibly in matching products to the customers who value those products.

In terms of managerial implications, our results suggest that adding a social platform to an existing online forum could accelerate the diffusion and discovery of relevant information. Our results could also be used by marketing managers in evaluating a conservative estimate of the return on investment for marketing a new product using social media. Specifically, our results indicate that the diffusion of songs is six times more likely when using peers than otherwise; and managers could use our proposed methodology to evaluate whether this result generalizes to other product categories. With this information, marketing professionals could plan and justify investment in social media when compared to non-social platforms.

6.2. Limitations and Future Work

One notable limitation of our study arises from our conservative approach to identifying diffusion, which causes us to ignore a potentially large volume of data that may include diffusion of other popular songs. This means that we may be underestimating the actual influence of online peers. Thus a next logical step is to analyze a larger volume of non-dissected data, which will allow researchers to test the diffusion of more popular music.

Another limitation pertains to the use of control users. Although control users do provide a baseline for diffusion in the absence of peers, there is still a possibility that target users discover a song or a band outside of the network and in a way that is not accounted for by our controls. Since control users cannot perfectly account for this issue, we have tried to further minimize the extent of diffusion from unobserved sources by screening the music played by all homophilic neighbors.

A related limitation arises from the need to use the neighbors of the target user to identify the music diffused to the control users. Ideally, we would have been able to use the control user's actual neighbors at the time the diffusion occurred. Since control users were identified after the post-connection time period, and historical network information is unavailable in the data, we were not able to screen out all songs played in the control user's neighborhood. This results in an overestimation of diffusion to the control user because some songs that could have been eliminated from the control user's playlist end up contributing to diffusion to the control user. This makes our estimates conservative and strengthens our finding of music diffusion in the online community of music listeners.

We also note that the only information available about the recommendation models used by Last.fm reveals that recommendations are based on the similarity and frequency of the music played by users (bands, artists, and genre). However, while we were unable to obtain detailed information about the specific recommendation system that Last.fm uses, our model is somewhat independent of the recommendation system. There are two reasons for this: (1) our model requires that users have Last.fm as the only platform for communication with their peers and any recommendation system suffices this for requirement; and (2) our model requires that recommendation system is consistent in matching users, thus if the performance of the model is higher or lower the diffusion estimates for both target users and control users will shift synchronously causing relatively much smaller change in the net diffusion estimates.

Future work could also extend our results by incorporating music genre in the diffusion estimates. For example, it is possible that a user listens to "pop" music, and discovers new music in the similar genre. Unfortunately, there are two challenges with using genre: First, there is no single recognized tagging system for music genre. Second, there is a possibility of high correlations between different music genres (e.g. “pop” may not be meaningfully different from “rock” in the same way that “pop” and “jazz” are different).

Further, although our approach using longer time periods allows us to dilute the instantaneous effect of media and other environmental factors, it would still be useful to explore applying our approach to shorter time periods for measuring information diffusion. Using shorter time periods would allow managers to evaluate the instantaneous effect of social media advertising, and researchers to estimate the role of micro-level network dynamics in the diffusion process.

We have modeled user’s behavior based on observable music listening characteristics. Future research could consider a consumer’s behavior like curiosity and willingness to discover new music. This will allow researchers to model diffusion as a function of market and social signals.

In conclusion, our results measuring the extent of diffusion are statistically significant yet conservative because our approach (of necessity) only considers the diffusion of niche music that was repeated by individuals after diffusion had occurred. In reality any single instance of use should be considered as potential diffusion, and popular content may be more likely to be diffused than niche content. We also note that our approach is just a starting methodology to analyze large datasets available on the Internet to statistically estimate the extent of information diffusion. We believe this and subsequent methodologies will create new perspectives to address the non-trivial challenges of measuring information diffusion in online ICT networks.

References

1. Adar, E. and Adamic, L.A. Tracking information epidemics in blogspace. In J. Liu, C. Liu, M. Klusch, N. Zhong, and N. Cercone (eds.), *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne Cedex, France: IEEE Computer Society, 2005, pp. 207-214.
2. Aral, S.; Muchnik, L.; and Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences,* 106, 51 (December 2009), 21544-21549.
3. Arndt, J. Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research*, 4, 3 (August 1967), 291-295.
4. Bass, F.M. A new product growth for model consumer durables. *Management Science,* 15, 5 (January 1969), 215-227.
5. Bhattacharjee, S.; Gopal, R.; Lertwachara, K.; and Marsden, J. Consumer search and retailer strategies in the presence of online music sharing. *Journal of Management Information Systems,* 23, 1 (Summer 2006), 129–159.
6. Bickart, B. and Schindler, R.M. Internet forums as influential sources of consumer information. *Journal of Interactive Marketing,* 15, 3 (July 2001), 31-40.
7. Bikhchandani, S.; Hirshleifer, D.; and Welch, I. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives,* 12, 3 (Summer 1998), 151-170.
8. Brown, J.J. and Reingen, P.H. Social ties and word-of-mouth referral behavior. *The Journal of Consumer Research,* 14, 3 (December 1987), 350-362.
9. Cameron, A.C. and Trivedi, P.K. *Regression Analysis of Count Data*. New York, NY: Cambridge University Press, 1998.
10. Cavusoglu, H.; Hu, N.; Li, Y.; and Ma, D. Information technology diffusion with influentials, imitators, and opponents. *Journal of Management Information Systems,* 27, 2 (Fall 2010), 305 - 334.
11. Chevalier, J.A. and Mayzlin, D. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research,* 43, 3 (August 2006), 345-354.
12. Clemons, E.; Gao, G.; and Hitt, L. When online reviews meet hyperdifferentiation: A study of the Craft beer industry. *Journal of Management Information Systems*, 23, 2 (October 2006), 149–171.
13. Coleman, J.; Katz, E.; and Menzel, H. The diffusion of an innovation among physicians. *Sociometry,* 20, 4 (December 1957), 253-270.
14. Dellarocas, C.; Gao, G.; and Narayan, R. Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems*, 27, 2 (Fall 2010), 127-158.
15. Dellarocas, C.; Zhang, X.; and Awad, N.F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21, 4 (2007), 23-45.
16. Dellarocas, C. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science,* 49, 10 (October 2003), 1407-1424.
17. Dhar, V. and Chang, E.A. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing,* 23, 4 (November 2009), 300-307.

18. *Facebook.com*. Facebook Statistics. Palo Alto, CA, February 1, 2011 (available at http://www.facebook.com/press/info.php?statistics).
19. Füller, J.; Mühlbacher, H.; Matzler, K.; and Jawecki, G. Consumer empowerment through Internet-based co-creation. *Journal of Management Information Systems*, 26, 3 (Winter 2009), 71–102.
20. Gastwirth, J.L. The estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics,* 54, 3 (August 1972), 306-316.
21. Ghose, A.; Telang, R.; and Krishnan, R. Effect of electronic secondary markets on the supply chain. *Journal of Management Information Systems*, 22, 2 (Fall 2005), 91–120.
22. Godes, D.; Mayzlin, D.; Chen, Y.; Das, S.; Dellarocas, C.; Pfeiffer, B.; Libai, B.; Sen, S.; Shi, M.; and Verlegh, P. The firm's management of social interactions. *Marketing Letters*, 16, 3 (2005), 415-428.
23. Godes, D. and Mayzlin, D. Using online conversations to study word-of-mouth communication. *Marketing Science,* 23, 4 (Fall 2004), 545-560.
24. Godes, D. and Mayzlin, D. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science,* 28, 4 (January 2009), 721-739.
25. Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. Information diffusion through blogspace. In S. Feldman and M. Uretsky (eds.), *Proceedings of the 13th International Conference on World Wide Web*, New York, NY: Association for Computing Machinery (ACM), 2004, pp. 491 - 501.
26. Hennig-Thurau, T. and Walsh, G. Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the Internet. *International Journal of Electronic Commerce*, 8, 2 (Winter 2003), 51–74.
27. Hinz, O. and Spann, M. The impact of information diffusion on bidding behavior in secret reserve price auctions. *Information Systems Research,* 19, 3 (September 2008), 351-368.
28. Huberman, B.A.; Romero, D.M.; and Wu, F. Social networks that matter: Twitter under the microscope. *First Monday [online]*, 14, 1 (December 2008), (available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063).
29. Jackson, A.; Yates, J.; and Orlikowski, W. Corporate blogging: Building community through persistent digital talk. In R. Sprague (ed.), *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society, 2007, pp. 80.
30. Jiang, Z. and Sarkar, S. Speed Matters: The role of free software offer in software diffusion. *Journal of Management Information Systems*, 26, 3 (Winter 2009), 207–240.
31. Kumar, R.; Novak, J.; and Tomkins, A. Structure and evolution of online social networks. In T. Eliassi-Rad (ed.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA: Association for Computing Machinery (ACM), 2006, pp. 611-617.
32. Leskovec, J.; Adamic, L.A.; and Huberman, B.A. The dynamics of viral marketing. In J. Feigenbaum (ed.), *Proceedings of the 7th ACM conference on Electronic commerce*, New York, NY: Association for Computing Machinery (ACM), 2006, pp. 228-237.
33. Li, X. and Hitt, L.M. Self-selection and information role of online product reviews. *Information Systems Research,* 19, 4 (December 2008), 456-474.
34. Linden, G.; Smith, B.; and York, J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE *Internet Computing,* 7, 1 (January/February 2003), 76-80.

35. Loh, L. and Venkatraman, N. Diffusion of information technology outsourcing: Influence sources and the Kodak effect. *Information Systems Research,* 3, 4 (December 1992), 334-358.
36. Mahajan, V.; Muller, E.; and Kerin, R.A. Introduction strategy for new products with positive and negative word-of-mouth. *Management Science*, 30, 12 (December 1984), 1389-1404.
37. Mahajan, V. and Muller, E. Innovation diffusion and new product growth models in marketing. *The Journal of Marketing,* 43, 4 (Autumn 1979), 55-68.
38. Mansfield, E.R. and Helms, B.P. Detecting multicollinearity. *The American Statistician,* 36, 3 (August 1982), 158-160.
39. Mansfield, E. Technical change and the rate of imitation. *Econometrica,* 29, 4 (October 1961), 741-766.
40. Manski, C. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies,* 60, 3 (July 1993), 531-542.
41. Manski, C. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press, 1999.
42. Mayzlin, D. Promotional chat on the Internet. *Marketing Science,* 25, 2 (March 2006), 155-163.
43. McPherson, M.; Smith-Lovin, L.; and Cook, J.M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology,* 27, (2001), 415-444.
44. Nardi, B.A.; Whittaker, S.; and Bradner, E. Interaction and outeraction: Instant messaging in action. In W. Kellogg and S. Whittaker (eds.), *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, Philadelphia, PA: Association for Computing Machinery (ACM), 2000, pp. 79-88.
45. Pavlou, P.A. and Dimoka, A. The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, 17, 4 (December 2006), 392-414.
46. Senecal, S. and Nantel, J. The influence of online product recommendations on consumers' online choices. *Journal of Retailing,* 80, 2 (June 2004), 159-169.
47. Soetevent, A.R. Empirics of the identification of social interactions; An evaluation of the approaches and their results. *Journal of Economic Surveys,* 20, 2 (March 2006), 193-228.
48. Sproull, L. and Kiesler, S. Reducing social context cues: Electronic mail in organizational communication. *Management Science,* 32, 11 (November 1986), 1492-1512.
49. Van den Bulte, C. and Joshi, Y.V. New product diffusion with influentials and imitators. *Marketing Science*, 26, 3 (May-June 2007), 400-421.
50. Walther, J.B. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research,* 23, 1 (February 1996), 3-43.
51. Wattal, S.; Racherla, P.; and Mandviwalla, M. Network externalities and technology use: A quantitative analysis of intraorganizational blogs. *Journal of Management Information Systems*, 27, 1 (Summer 2010), 145-174.
52. Williamson, D.A. The rush to social networks. *eMarketer*, (February 9, 2009), (available at http://www.emarketer.com/Article.aspx?R=1006910).
53. Wooldridge, J.M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2002.

Data Tables

Table 1: Guide to Notations

Notation	Description
i	Target or music discovering target user
j	Music diffusing neighbor of i (vector of all j is J)
k	Non-music diffusing neighbor of i (vector of all k is K)
c_i	Control user that is similar to i but not connected to j
n_i	Number of music discovering target users i
$n_{i,j}$	Number of music diffusing new neighbors J of a user i
$n_{i,k}$	Number of non-music diffusing neighbors K of a user i
$S_{i,\ (t1,\ t2)}$	Binary vector of songs played by a user i between time period t_1 and t_2
$B_{i,\ (t1,\ t2)}$	Binary vector of bands played by a user i between time period t_1 and t_2
$M_{i,\ (t1,\ t2)}$	Binary vector of music (songs or bands) played by a user i between time period t_1 and t_2
$F_{i,\ (t1,\ t2)}$	Vector representing frequency of music (songs or bands) played by a user i between time period t_1 and t_2
$f_{i,p}$	Represents the element p of the frequency vector $F_{i,(t1,\ t2)}$
n_b	Total number of distinct bands played by all users and neighbors
n_s	Total number of distinct songs played by all users and neighbors
n_m	Total number of distinct music (bands or songs) played by all users and neighbors
$D_{i,j}\,(t_1,\ t_2)$	Binary vector of music diffused from neighbor j or user i during period t_1 and t_2
d	Represents elements of the diffusion vector $D_{i,j}(t_1,\ t_2)$
$(0,\ T\text{-}t_c)$	Pre-connection period when user i is not connected with music diffusing neighbors j
$(T\text{-}t_c,\ T)$	Connection period during which music diffusing neighbor is connected with user i
$(T,\ T + \Delta T)$	Post-connection period when user i discovers new music (songs or bands) from neighbor j
$Y_{1,i}$	binary variable indicating the existence of diffusion to user i
$Y_{2,i}$	an integer (count) variable indicating the total music diffused to user i
G_i	Gini coefficient measuring the inequality and statistical dispersion in music consumed by user i
$G(i,\ j)$	Gini coefficient representing the statistical dispersion in differences in music consumed by users i and j
$E(i,\ j)$	Euclidian distance based on differences in music consumed by users i and j

Table 2: Data Summary

Data Description	Value
Observed treated users	35
Observed control users	40
Unique neighbors (observed in April 2008 & July 2008)	4017
Entries in playlist for all observed users	21,104,040

Table 3: Music Listening Statistics during Various Time Periods

Variable	Time Period	Target User	Control User
Average Number of Bands Played per User	Pre-Connection	256.91	283.75
	Connection	232.25	204.82
	Post-Connection	164.17	212.52
Average Number of Songs Played per User	Pre-Connection	1319.91	1124.32
	Connection	1139.71	960.6
	Post-Connection	722.37	842.07

Table 4: Music Listening Behavior (standard deviation in parentheses)

Music Listening Statistics	Target User Value	Control User Value
Bands played between July 2008 and Sep 2008	164 (181)	213 (235)
New bands exposed to user	691 (660)	737 (649)
Gini coefficient for band listening heterogeneity	0.893 (0.073)	0.870 (0.119)
Number of bands diffused to each user	1.4 (2.8)	1.1 (2.1)
Songs played between July 2008 and Sep 2008	722 (602)	842 (798)
New songs exposed to user	4577 (4342)	4737 (4241)
Gini coefficient for song listening heterogeneity	0.846 (0.096)	0.759 (0.155)
Number of songs diffused to each user	9.7 (15.0)	3.8 (5.9)

Table 5: Logit Regression with Diffusion (binary) as Dependent Variable (standard deviation in parentheses)

Variable	Coefficient (Bands)	Odds Ratio (Bands)	Coefficient (Songs)	Odds Ratio (Songs)
Target/Control Dummy	1.220 (0.733)*	3.386 (2.482)*	1.808 (0.944)*	6.096 (5.757)*
Ln (Music Played)	1.949 (0.595)**	7.022 (4.177)**	2.583 (0.693)**	13.233 (9.172)**
Ln (Music Exposed)	1.496 (0.492)**	4.463 (2.196)**	1.389 (0.563)**	4.0121 (2.260)**
Gini Coefficient	6.142 (5.353)	464.774 (2487)	-2.569 (3.219)	0.077 (0.247)
Constant	-25.30 (6.624)**		-24.91 (6.886)**	
R^2	0.492		0.593	
* 10% significance, ** 5% significance				

Table 6: Negative Binomial Regression with Diffusion (count) as Dependent Variable (standard deviation in parentheses)

Variable	Coefficient (Bands)	Marginal Effect (Bands)	Coefficient (Songs)	Marginal Effect (Songs)
Target/Control Dummy	0.474 (0.338)	0.159 (0.124)	1.041 (0.274)**	2.682 (0.846)**
Ln (Music Played)	1.050 (0.221)**	0.343 (0.086)**	0.954 (0.164)**	2.272 (0.392)**
Ln (Music Exposed)	1.140 (0.238)**	0.372 (0.094)**	0.786 (0.155)**	1.870 (0.432)**
Gini Coefficient	5.396 (3.120)*	1.762 (.947)*	1.481 (1.302)*	3.524 (3.122)*
Constant	-18.02 (3.819)**		-12.94 (1.788)*	
R^2	0.279		0.171	
α	0.518		0.817	
* 10% significance, ** 5% significance				

Table 7: Correlation in Independent Variables (Bands)

Variable Name	Ln (Bands Played)	Ln (Bands Exposed)	Gini Coefficient (Bands)
Ln (Bands Played)	1		
Ln (Bands Exposed)	0.2263	1	
Gini Coefficient (Bands)	0.2029	0.1194	1

Table 8: Correlation in Independent Variables (Songs)

Variable Name	Ln (Songs Played)	Ln (Songs Exposed)	Gini Coefficient (Songs)
Ln (Songs Played)	1		
Ln (Songs Exposed)	0.2434	1	
Gini Coefficient (Songs)	0.0815	0.0334	1

Figures and Charts

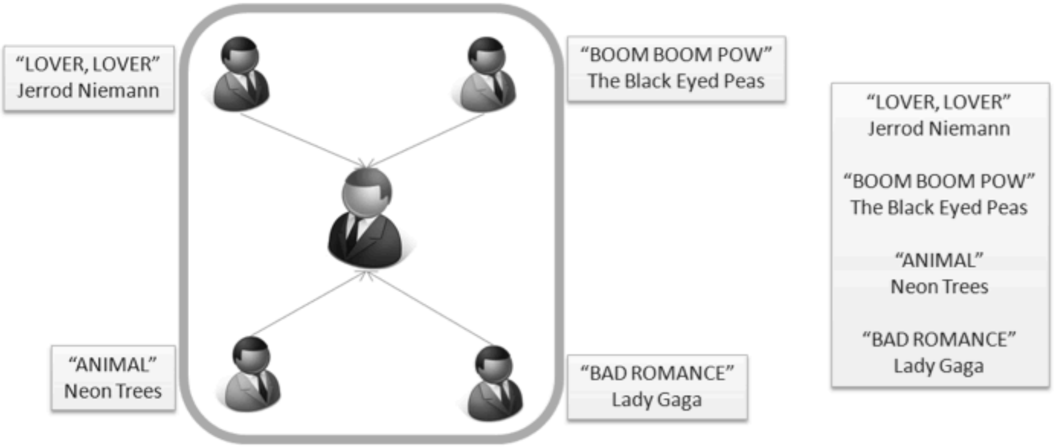


Figure 1: Observed Diffusion Sources

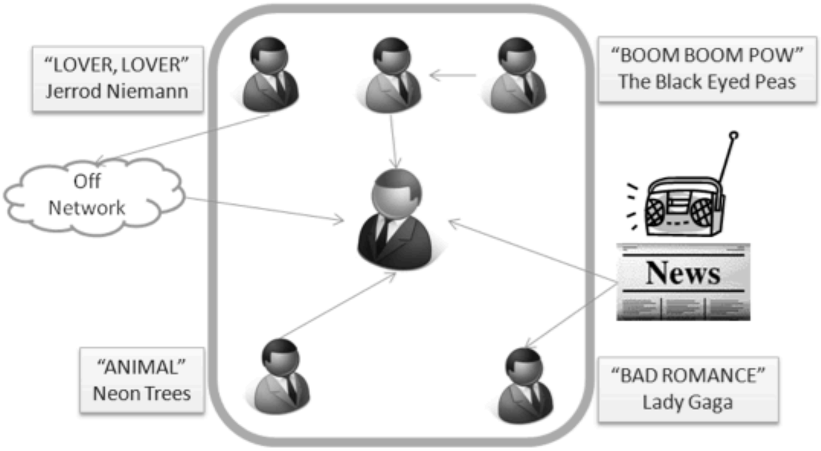
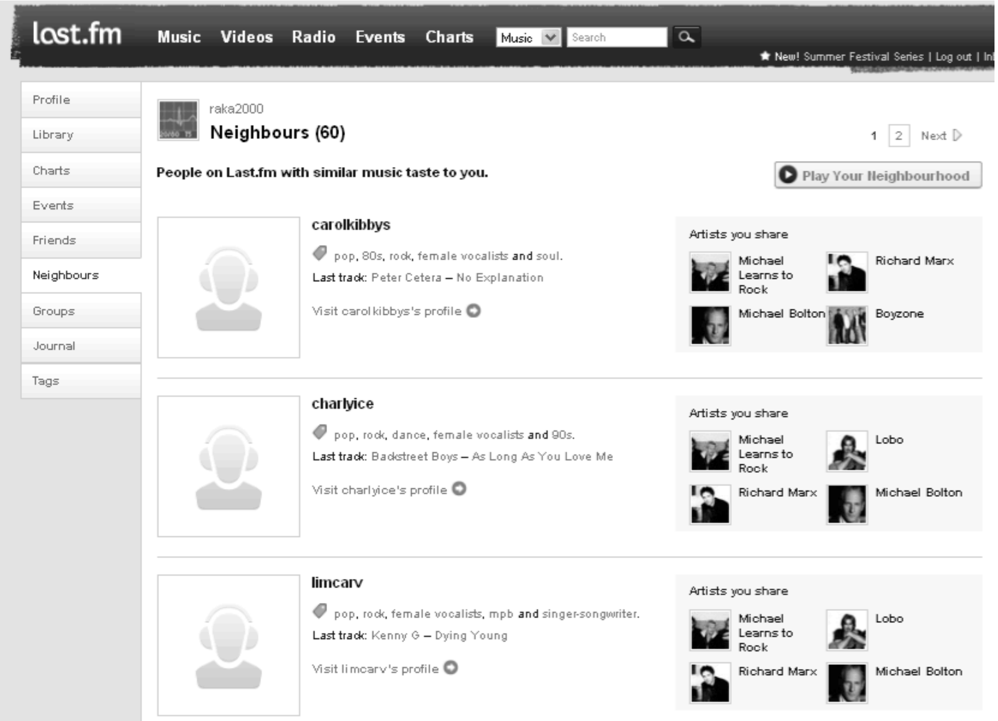


Figure 2: Actual Diffusion Sources



raka2000

Neighbours (60)

People on Last.fm with similar music taste to you.

carolkibbys

pop, 80s, rock, female vocalists **and** soul.

Last track: Peter Cetera – No Explanation

Visit carolkibbys's profile

Artists you share: Michael Learns to Rock, Richard Marx, Michael Bolton, Boyzone

charlyice

pop, rock, dance, female vocalists **and** 90s.

Last track: Backstreet Boys – As Long As You Love Me

Visit charlyice's profile

Artists you share: Michael Learns to Rock, Lobo, Richard Marx, Michael Bolton

limcarv

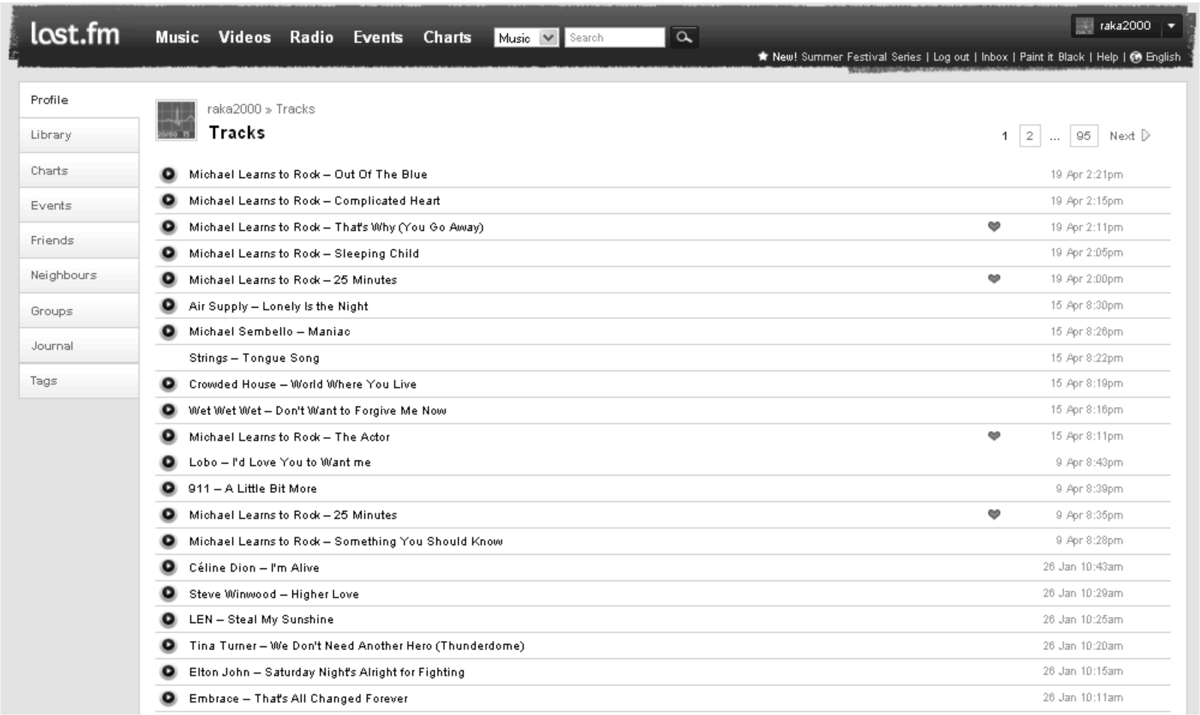
pop, rock, female vocalists, mpb **and** singer-songwriter.

Last track: Kenny G – Dying Young

Visit limcarv's profile

Artists you share: Michael Learns to Rock, Lobo, Richard Marx, Michael Bolton

Figure 3: Snapshot of Neighbors on Last.fm



raka2000 » Tracks

Tracks

Michael Learns to Rock – Out Of The Blue		19 Apr 2:21pm
Michael Learns to Rock – Complicated Heart		19 Apr 2:15pm
Michael Learns to Rock – That's Why (You Go Away)	♥	19 Apr 2:11pm
Michael Learns to Rock – Sleeping Child		19 Apr 2:05pm
Michael Learns to Rock – 25 Minutes	♥	19 Apr 2:00pm
Air Supply – Lonely Is the Night		15 Apr 8:30pm
Michael Sembello – Maniac		15 Apr 8:26pm
Strings – Tongue Song		15 Apr 8:22pm
Crowded House – World Where You Live		15 Apr 8:19pm
Wet Wet Wet – Don't Want to Forgive Me Now		15 Apr 8:16pm
Michael Learns to Rock – The Actor	♥	15 Apr 8:11pm
Lobo – I'd Love You to Want me		9 Apr 8:43pm
911 – A Little Bit More		9 Apr 8:39pm
Michael Learns to Rock – 25 Minutes	♥	9 Apr 8:35pm
Michael Learns to Rock – Something You Should Know		9 Apr 8:28pm
Céline Dion – I'm Alive		26 Jan 10:43am
Steve Winwood – Higher Love		26 Jan 10:29am
LEN – Steal My Sunshine		26 Jan 10:25am
Tina Turner – We Don't Need Another Hero (Thunderdome)		26 Jan 10:20am
Elton John – Saturday Night's Alright for Fighting		26 Jan 10:15am
Embrace – That's All Changed Forever		26 Jan 10:11am

Figure 4: Snapshot of Playlist on Last.fm

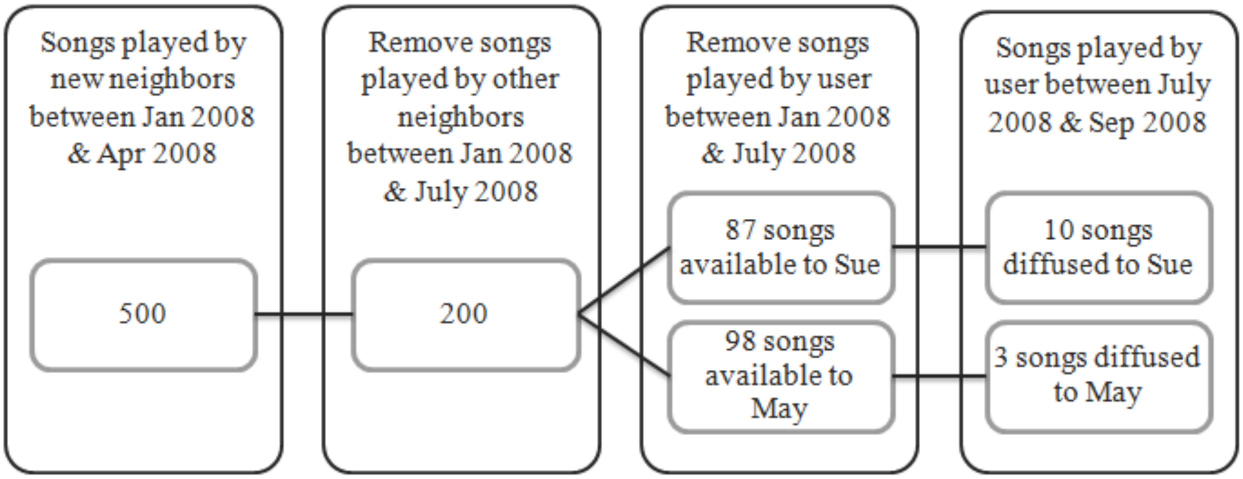

Songs played by new neighbors between Jan 2008 & Apr 2008
500
Remove songs played by other neighbors between Jan 2008 & July 2008
200
Remove songs played by user between Jan 2008 & July 2008
87 songs available to Sue
98 songs available to May
Songs played by user between July 2008 & Sep 2008
10 songs diffused to Sue
3 songs diffused to May

Figure 5: Estimation Process

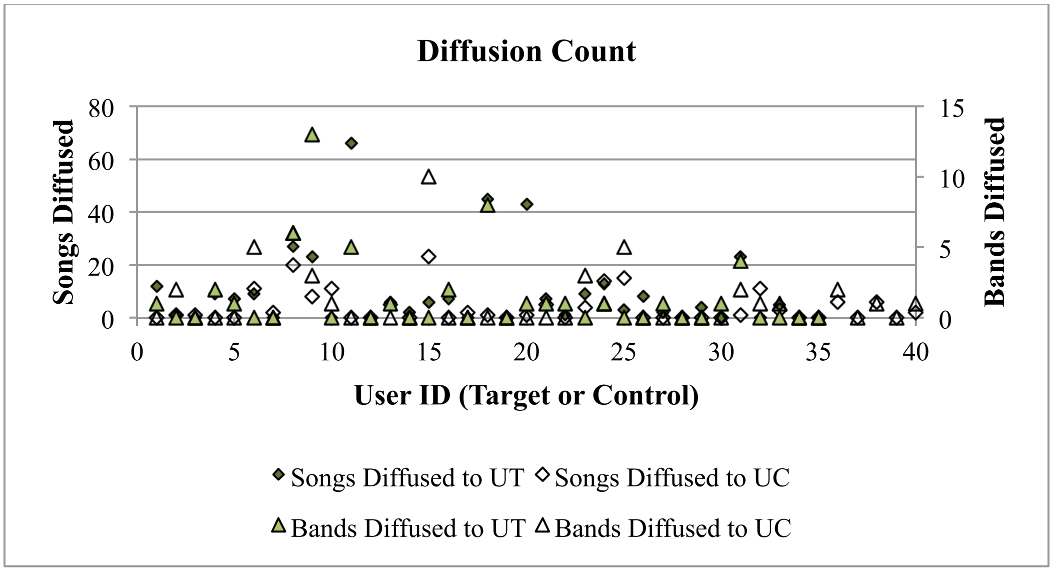


Figure 6: Number of Bands and Songs Diffused to Target (UT) and Control (UC) Users

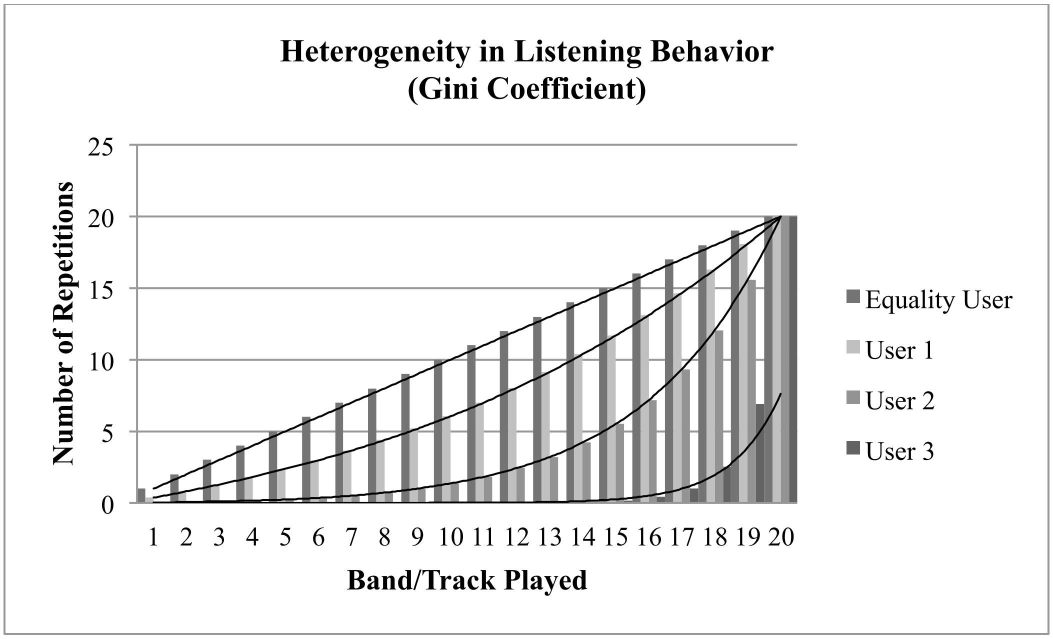


Figure 7: Hypothetical Scenarios of Listening Behaviors of 4 Different Users

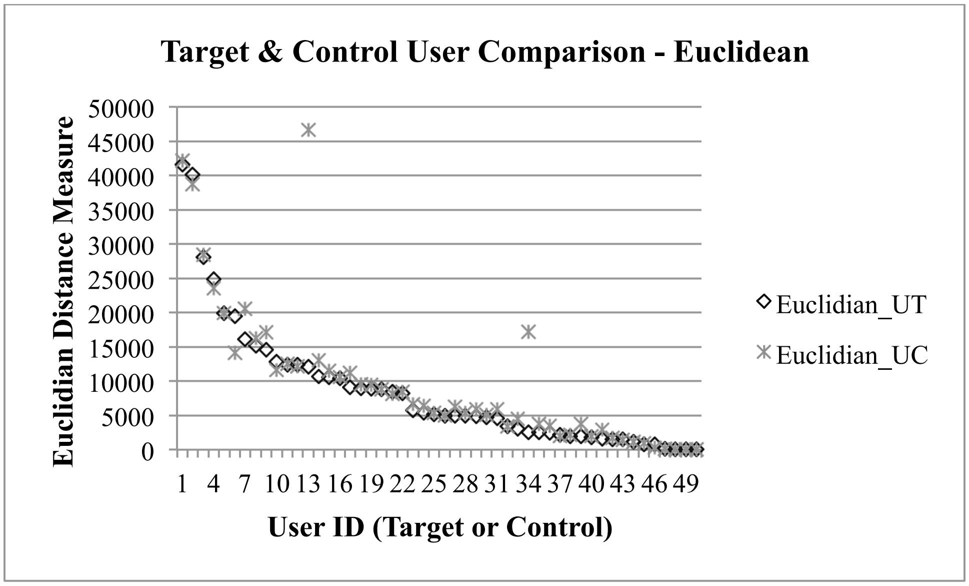


Figure 8: Euclidean Distance Comparison for Target (UT) and Control (UC) Users

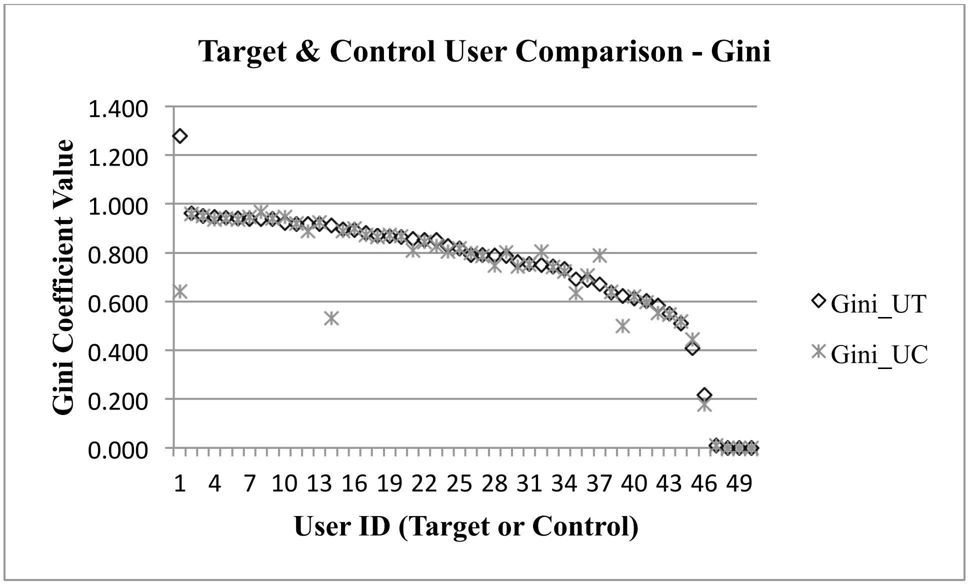


Figure 9: Gini Coefficient Comparison for Target (UT) and Control (UC) Users

Authors Bios

RAJIV GARG is a fourth year doctoral student at School of Information Systems and Management at Heinz College of Carnegie Mellon University in Pittsburgh, PA. He received graduate degrees in Computer Science and Electrical Engineering, both from University of Southern California in Los Angeles, CA and an undergraduate degree in Electrical Engineering from Indian Institute of Technology, Banaras Hindu University in Varanasi, India. His research interests are in the areas of social media, online marketing, Internet technologies, economics of information systems and artificial intelligence. Rajiv is a senior member of IEEE and has been serving on the board of various IEEE sections and multiple small corporations during past decade. Rajiv has also served as reviewer for Information Systems and Research (ISR), the Journal of Management Information Systems (JMIS), and various peer reviewed conferences.

MICHAEL D. SMITH is a Professor of Information Systems and Marketing at Carnegie Mellon University. He holds academic appointments at Carnegie Mellon University's School of Information Systems and Management and the Tepper School of Business. He received a Bachelors of Science in Electrical Engineering (*summa cum laude*) and a Masters of Science in Telecommunications Science from the University of Maryland, and received a Ph.D. in Management Science from the Sloan School of Management at MIT. Professor Smith's research relates to the impact of digital technologies on consumers, firms, and markets. Professor Smith has received several notable awards including the National Science Foundation's prestigious CAREER Research Award, and he was recently selected as one of the top 100 "emerging engineering leaders in the United States" by the National Academy of Engineering.

RAHUL TELANG is a Professor of Information Systems and Management at the Heinz College, Carnegie Mellon University. He received his Ph.D. in Information Systems from the Tepper School of Business, Carnegie Mellon University in 2002. His research interests include digital media and information security and privacy. He is the recipient of Sloan Foundation Industry Study fellowship and National Science Foundation's CAREER Research Award. His recent work includes the role of Internet and piracy on music and movie industry. His research has appeared in many top journals including Management Science, Marketing Science, Information systems research etc. He has been on the editorial board of Management Science and Information systems research.

Contact Information

Name	**Rajiv Garg***	**Michael Smith**	**Rahul Telang**
Affiliation	School of Information Systems and Management, Heinz College, Carnegie Mellon University		
Address	4800 Forbes Ave, Suite 3030 Pittsburgh, PA 15213	4800 Forbes Ave, Suite 3028 Pittsburgh, PA 15213	4800 Forbes Ave, Suite 3040 Pittsburgh, PA 15213
Phone	(412) 268-8717	(412) 268-5978	(412) 268-1155
Fax		(412) 268-5338	(412) 268-5337
Email	rg@cmu.edu	mds@cmu.edu	rtelang@andrew.cmu.edu
*** contact author**			