Discovery of Music through Peers in an Online Community

Rajiv Garg
Heinz College
Carnegie Mellon University
rg@cmu.edu

Michael D. Smith
Heinz College
Carnegie Mellon University
mds@cmu.edu

Rahul Telang
Heinz College
Carnegie Mellon University
rtelang@andrew.cmu.edu

Abstract

Peer influence in social networks has been studied for over four decades by social scientists and marketing researchers. Due to recent growth of internet technologies and online social networks, research on peer influence has gained even more attention over the last decade. But, most extant empirical work measuring peer effect faces challenges due to selection problem, difficulty in separating source of influence, and accounting for a user's pre-existing knowledge. In this paper we overcome these challenges by strategically dissecting the archival data and find that the online users have a small but statistically significant influence on other peers in discovery of music. We cleanly and empirically estimate that online users are 6 times more likely to discover new music because of their peers in an online community when compared to discovery in the absence of those peers.

1. Introduction

Empirical studies on information diffusion date back to the mid-twentieth century with findings in diffusion of innovation – new drugs in medical physician's network [1], or new process and technique imitation by corporations [2]. Researchers also evaluated how product related conversations or word-of-mouth (WoM) triggers the diffusion of information [3]. This created a wave of interest in evaluating the process of diffusion especially for product marketing [4]. Over the following three decades, interest in information diffusion continued to develop amongst the research communities of social sciences [5], marketing [6], and computer sciences [7].

Online social communities have provided a new channel for diffusing information, but at the same time, the estimation of diffusion has become more challenging because of the large amount of information being exchanged on the internet. This identification poses further challenges due to uncertainty in identifying the source of information, although some work has been done to estimate the true source of influence [8]. Traditional information diffusion studies triggered diffusion by introducing new information or by observing niche content from a defined source. Internet has changed this model for information diffusion; personal communication has become much more diversified creating barriers for diffusion estimation. Users now communicate in person, over analog channels (phones, etc.), or digital channels (emails, social networks, discussion forums, instant messengers, and more). In this research we focus on online social networks because they provide one of the fastest growing modes of communication with monotonically increasing number of users. Online social networks allow users to diffuse information by pushing content to their peers, and to discover information by pulling content from their peers.

Online social networks have seen a significant growth over the last decade. A study by eMarketer [9] found that 41% of Internet users in the US (about 80 million people) visited a social network website at least once a month in 2008, an increase of 11% from 2007. Based on statistics from Alexa.com (in December 2009), the combined daily reach of 3 popular social network websites (facebook.com, myspace.com and twitter.com) was 39% of daily Internet consumption. While the growth in online social networks suggests a significant impact on

1530-1605/11 $26.00 © 2011 IEEE

online community members, empirical research is only beginning to emerge ([10], [11], [12], [13]) on how online social communities help users discover and diffuse new content.

Diffusion can be broadly classified under two categories: influence (by a system or a peer) and discovery (by active search or observational learning). In more common terminology the process of influencing by a system is known as "recommendation system," which has been studied extensively over the last decade by computer as well as information scientists. Recommendation systems are mainly based on collaborative filtering model to influence potential consumers ([14], [15]).

Discovery by active search is different from discovery through observational learning. In the former the users makes an effort to find content while in latter the user comes across the content while going through his daily routine (without significant additional effort). Active search on the internet is accomplished by using search engines or by seeking help on discussion forums. In both cases the user knows what to look for but her behavior towards the new content is unobservable. Observational learning is traditionally studied heavily in psychology literature, which is beyond the scope of this research manuscript. But more recently, observation learning has been studied in relation to business and economics [16] and is classified as learning by either observable action or observable signal. One can argue that peer influence is a type of observable signal where actions and signals from predecessors influence the decision of a consumer. Observable actions or signals influence a decision because information or content is made available to user(s). One channel to enable the observable signals and actions on the internet is online forums. It has been found that these online forums, like blogs [17] or online message boards [18], can be more effective in influencing consumers when compared to direct marketing channels. This brings us to our motivation: how effective are these online peers, when compared to, ceteris paribus, a scenario where these peers are absent, in influencing other online users?

Peer influence for diffusing information is the most widely exploited area for product marketing. Viral marketing campaigns or positive word-of-mouth are repeatedly used for quicker growth and adoption in the market. Researchers have observed the positive effects of online word-of-mouth [10] and viral marketing [19] when used as means of influencing potential consumers. The influence happens not only because of the presence of the peers but also because of online word of mouth that builds trust and fosters cooperation in online marketplaces [15]. Most extant literature has also analyzed social influence in a variety of online settings such as computer mediated communication [20], email [21], and instant messaging [22]. The consensus is that peers play an important role in diffusion of information or content to a larger population.

Simultaneously, researchers have found that many online peers may not play a role in diffusion because of the presence of large number of peers and limited interactions [23]. This is understandable looking at the network statistics of an average of 130 and 150 friends on facebook.com and myspace.com respectively. But even with these numbers, could it be that some of these 130 connected friends are more valuable than any other source for diffusing new content to a user? This is exactly what we are interested in investigating and this paper takes the first step in that direction. We try to address this problem, using Last.fm - a website to listen, discover, and discuss music. Since Last.fm also provides a social platform where users can connect and communicate with their friends and other online users, we use this channel to examine how (if at all) peers diffuse music to each other. In particular we examine how big a role social networks play in users' discovering new content.

However, empirical work measuring peer effect is particularly challenging. Peers and friends tend to be self selected leading to a significant selection problem. There is also the issue of homophily [24] and contamination due to other sources influencing diffusion [25]. In this paper, we overcome these challenges by collecting and analyzing an extremely large dataset from Last.fm. In this dataset, we observe only those users for whom the social network on Last.fm is the only mode of communicating with each other. Using this archival dataset, we empirically estimate statistically significant peer influence on Last.fm. We find that, on average, peers are 6 times more likely to diffuse a new song and

lead to a diffusion of 3 additional songs to a user on the social network.

Thus, through this paper we cleanly and empirically identify peer influence and quantify the extent of diffusion in an online community while addressing the estimation challenges faced by extant research. The estimation issues that we have controlled in this research include selection problem, homophily, identification of the diffusion source, user's pre-existing knowledge, and size of a user's personal online network.

We present the research methodology in Section 2, followed by our findings in Section 3. Section 4 presents conclusions and directions for future.

2. Methodology

To better explain our research strategy we first discuss an ideal scenario to measure peer influence on an online social network. To cleanly measure peer influence on an online social network we would need to observe interaction between two users in a closed environment. Only a closed environment will allow us to observe all communications within the environment and prevent any flow of information into the network from external sources. Additionally, we would need to observe diffusion of completely novel or niche information to account for any pre-existing knowledge of a user. Finally, a control user would be required to adjust for any discovery due to factors like consumer behavior or by-chance discovery. Thus observation of controlled exchange of niche content in a closed online environment can allow us to cleanly estimate the online peer influence.

Running this experiment in real world is not only difficult but also poses challenges in selection of participating candidates. Therefore we use an alternative approach that mimics the ideal scenario by utilizing a large volume of archival data from an online social network for music (created on Last.fm) to estimate the peer influence.

Last.fm is a part of CBS Interactive with an estimated user-base of 40 million active users in over 200 countries. Based on statistics from Alexa.com in December 2009, Last.fm reaches 0.33% of daily internet users (globally) and is just behind Pandora.com, which has a daily reach of 0.39%. An interesting feature of Last.fm is scrobbling, which implies observing and tracking the music listened by users online or off-line (personal computers or portable music players). Scrobbling allows us to access a much richer dataset that enables us to observe both current and past listening behavior of users. Additionally, Last.fm allows users to socially connect with friends and other registered users on the website. Last.fm has taken another step in social networking by recommending a list of 60 neighbors that share the most similar taste in music.

In this study we pick these "neighbors" as peers for source of diffusion. These neighbors (recommended peers) are typically strangers to the user and are recommended based on their 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 itself. Additionally we can control (by screening out) all the music made available for diffusion by these neighbors. Thus, we gain control over the content exchanged between users and their neighbors on Last.fm. Together, the use of neighbors as peers and control over the diffused music emulates, to some extent, the closed environment for diffusion estimation.

To account for the pre-existing knowledge of a user we remove all of the songs/bands listened either by the user or her peers (neighbors) from the list of songs/bands available for diffusion. This reduced playlist that could be diffused to a user by the neighbors includes only those songs that have neither been played by the user nor 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 were matched, in future timeframe, with the target users for having the most similar taste in music. Additionally, we carefully selected these control users to be similar to the music diffusing peers but not connected with them in order to prevent any occurrence of peer influence to control group. This set of control users then allows us to estimate the discovery of new content from sources other than the diffusing neighbor and thus adjusts our estimate of peer

influence.

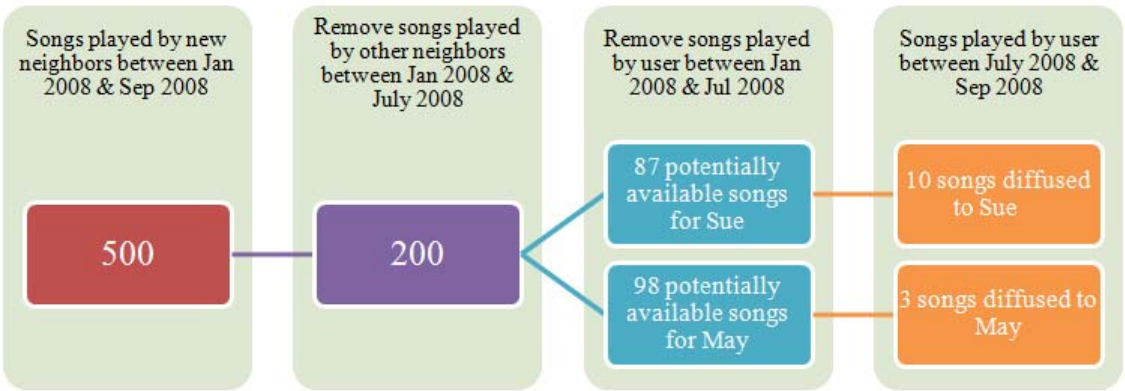
Thus with this alternate setup, we are able to account for common challenges in the measurement of peer influence or information diffusion. Selection issues are addressed by using system recommended neighbors (which are not friends). Endogeneity and pre-existing knowledge are accounted for by screening out music already played by the user. Finally, diffusion by-chance or from external sources is controlled by using a control group of users that are not connected to the new neighbors.

Although we started with about 500 random users and their network, the amount of data available for processing was beyond the computational power available to us, thus we picked 50 random target users, 50 control users, and had a network of about 4017 neighbors. For these 4117 people, we collected 21 million data points spanning 9 months of history and about 5 months of data collection period. The data contains the network information of these 50 target users during 3 non-overlapping time periods (ending on April 2008, July 2008, and September 2008), and the playlist (song listened with timestamp) for each user. During analysis we realized that some users had missing playlist data possibly because of the change in their privacy settings or non-tracked music listening. Thus for this study we used 35 target users and 40 control users that had data available for the entire 9 month period. Summary statistics of data is given in the Table 1.

Selection of time periods was especially important in our research methodology. We considered three different time periods: pre-connection (or creation), connection, and post-connection (or discovery). Pre-connection time period (of creation phase) is the period from Jan 2008 to April 2008 when we observed the formation of networks and listening behavior of all users. During the connection time period – from April 2008 to July 2008 – we observed the changes in the networks and entry of new neighbors (peers), who played some songs that were new to the entire network. During the post-connection time period (or discovery phase) we observed discovery of new songs that were introduced by these new neighbors. Some statistics of music listening during each phase is given in Table 2.

Our empirical process to account for all estimation issues and estimate diffusion is highlighted in Figure 1. 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 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 say those 10 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, let's assume we find that new neighbors expose Sue to 87 new songs. Analogously the number for "May" is 98. 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 peers. Put another way, Sue is discovering additional new content as compared to May because she is connected to those new neighbors.

Figure 1: Estimation Process



3. Analysis

We assume diffusion has happened when a song/band that was played by a music diffusing new neighbor (NN) in a pre-connection period (Jan-Apr 08) shows up in the users' playlist in the post-connection period (July-Sep 08). As discussed previously, we pick only the songs that are new to the entire network of a user i.e. the songs or bands that are not played by the user or any of her neighbors in any of the time periods prior to diffusion. To ensure that a song or band indeed diffused, we consider diffusion only if the user played the song at least two times. A simple regression model could be defined as follows:

$$Diffusion_i = \alpha * (Target/Control\ Dummy) + \sum_i \beta_i * (user\ characteristics) \quad \ldots (1)$$

Here the dependent variable, diffusion, takes the form of (i) a binary occurrence of diffusion or, (ii) a count of the number of bands/tracks diffused to a user. The independent variables are the music listening characteristics of users that include number of unique bands/tracks listened during the post-connection (or discovery) period (July 2008 to September 2008), number of new bands/tracks the new-neighbors has made available for diffusion, and the listening heterogeneity of a user (i.e., Gini coefficient described previously). The variable of interest then is the target/control dummy.

3.1. 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/tracks in a user's playlist. Two music listeners could be very different in terms of their exploratory nature. A user listening to a larger number of bands/tracks could be more interested in discovering new music.

Quantity of new music exposed: reflects the amount of music exposed to a user. Since each user gets exposed to a different set of new neighbors (NN) who may bring in different quantity of new content we would expect that a larger volume of exposure will lead to higher diffusion.

Heterogeneity in listening behavior: captures a user's propensity to listen to more diverse music. We capture this heterogeneity by the Gini coefficient [26]. Since diversity of music in a user's playlist follows approximately a Lorenz curve with unique bands/tracks on the x-axis and the number of repetitions on the y-axis, the Gini coefficient is then expressed as:

$$G = \frac{A_{Equality} - A_{Lorenz}}{A_{Equality}}$$

$$\text{or,} \quad G = 1 - 2 * \frac{N_{pl}}{N_r * (N_m + 1)}$$

Here N_m is the total number of unique bands/tracks played by the user, N_r is the maximum number of times a band/track was played by a user in his/her playlist and N_{pl} represents the total size of playlist for any user. Smaller Gini coefficient represents more diversity in music listened by a user.

Over the entire 9 months we collected comprehensive data on 35 target users (UT) and 40 control users (UC). On average the number of new neighbors (NN) was 15. The music listening statistics for these users is presented in Table 3.

3.2. Control group users

Before we dive deeper into the comparison of diffusion observed for either target or control group of users, it is important to evaluate the similarity between both kinds of users. To avoid any bias in the measurement of influence, we need to test the extent of similarity in the music listening behavior for both target (UT) and control (UC) group of users when compared to the behavior of music diffusing peers (NN). Thus to better evaluate the similarities we use various distance measurements to compare their behaviors. One common method is to estimate the Euclidean (ordinary) distance between two users. This distance measures the difference in the music listening patterns of the two users and is computed by taking the distance between the two vectors representing the frequency of each intersecting song

played by each user. The Euclidian distance between a target user UT_i and NN is presented as $E(UT_i)$ and the distance between a control user UC_i and NN is presented as $E(UC_i)$ below.

$$E(UT_i) = \sqrt{\sum_{j=0}^{n} |x_{UT_{ij}} - x_{NN_j}|^2}$$

$$E(UC_i) = \sqrt{\sum_{j=0}^{n} |x_{UC_{ij}} - x_{NN_j}|^2}$$

Here x is the frequency of repetition of a song by target user (UT_i), control user (UC_i), or music diffusing neighbors (NN). Thus for each user "i" we need to test if the Euclidean distances between UT and NN, and UC and NN are similar to each other. The paired t-test for both set of Euclidian distances has a p-value = 0.0761. Thus we can say with 90% confidence that the both target users and control users are similar to music diffusing new neighbors based on Euclidean distances.

We also tested the similarity of the two users (target and control) with the music diffusing new neighbors using a Gini coefficient that measures the statistical dispersion in music listening behavior of two sets of users. We find here the p-value from paired t-test to be 0.0937, which is also within the 90% confidence in 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 on the listening behavior as a combination of diversity and repetition of songs in a user's playlist.

$$G(UT_i) = \frac{1}{n}[1 - 2 * \sum_{j=0}^{n} \frac{(|x_{UT_{ij}} - x_{NN_j}|)}{(n * \text{Max}(|x_{UT_{ij}} - x_{NN_j}|)}$$

$$G(UC_i) = \frac{1}{n}[1 - 2 * \sum_{j=0}^{n} \frac{(|x_{UC_{ij}} - x_{NN_j}|)}{(n * \text{Max}(|x_{UC_{ij}} - x_{NN_j}|)}$$

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 the peer influence.

3.3. Results

We estimate equation (1) with diffusion as the dependent variable and report the results in Tables 4 and 5 in the appendix. The variable of interest in this model is the target/control dummy.

First we evaluate the diffusion as a binary variable (1 if diffusion occurs and 0 if not) using a simple logit model and report the findings in Table 4.

We find that the odds ratio for target/control dummy is positive (3.39 for bands and 6.1 for tracks) and significant at 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 target group than in control group.

Additionally we see that users who listen to more songs are more likely to see diffusion – a 1% increase in the playlist increases the likelihood of diffusion of a new band by 7 times and likelihood of diffusion of a new song by 13 times. This could be considered intuitive because the more music a user listens to, the more she is prone to discovering.

Thus there are two interesting observations here (i) quantifying the extent of discovery and (ii) evidence of discovery because of online peers in the presence of control users. Although these observations may not directly classify as influence but suggests a strong role of peers in discovery of new music.

Finally, users who get exposed to larger volume of new content are also likely to see more diffusion – a 1% increase in the availability of new music increases the likelihood of diffusion by 4 times. This follows similar intuition as before, except that the results are more directed towards the behavior of the neighbors whereas previously it was driven by the behavior of a user. In other words, users who are close to peers that are listening to a larger set of new songs, tend to get a spillover effect seen in new music discovery.

One thing we missed observing in this model is a significant role of Gini coefficient. This implies that the statistical dispersion in music listening behavior does not play a role in suggesting the presence of

diffusion, but we do observe that Gini coefficient does play a role in quantifying the diffusion.

Next we evaluate the diffusion as a count of number of unique bands/track diffused to a user by using negative binomial regression and report the estimate in Table 5. In case of individual songs the marginal effect for target/control dummy is positive (2.7) and significant at the 5% level suggesting that peer influence leads to a 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 additional 2.3 unique songs and 1% increase in exposure to new songs increases the diffusion by 1.9 songs. We also see that a 1 unit increase in Gini coefficient leads to a diffusion of 3.5 additional songs.

In case of diffusion of bands, the coefficient is positive (0.4738) but insignificant (p-value: 0.16).

Due to data limitations, we were not able to account for true neighbors of control users. Therefore, the neighbors for control users were assumed to be same as those of target users. But fortunately this underestimates the impact of diffusion because some songs that we could have eliminated from control user's playlist end up contributing to the diffusion to control user. This makes our estimates more conservative and strengthens our finding of music diffusion in the online community of music listeners.

4. Conclusion

In this paper, we find that there is a positive influence of online peers on 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. The key contributions of our paper are: (i) we provide a clean test for diffusion on online networks and are able to overcome many key challenges in estimating peer effects. Moreover, we do this in a field setting without resorting to survey or laboratory setting. Thus our paper provides a roadmap for using large yet noisy data and for estimation of peer effects with reasonable confidence, (ii) we provide empirical evidence that even a network with extremely weak-ties, where the peers don't know one another, can help in information discovery. We observed that new songs seem to diffuse on such a network thus suggesting a significant power of online networks in content discovery.

We believe that recommending peers, as modeled by Last.fm, could be the next wave of marketing that could possibly benefit from consumer curiosity, increased online trust between peers and "pull" marketing strategy. While peer recommendation may not guarantee diffusion of a product, we think that this methodology will be effective in matching products and customers that will value those products.

Since we strategically dissect the data to identify diffusion, we lose a large amount of data that may possibly include diffusion of other popular songs. Because of our very conservative approach in identification we may have overlooked the potentially larger influence of online peers. Thus as a next logical step, we would like to further analyze a much larger chunk of data and develop a methodology to test the diffusion of more popular music.

Another potential area to enhance our study is to account for consumer curiosity, and thus the willingness to discover a new piece of music. We have currently assumed user's behavior based on their macro-level music listening behavior. We are thus currently investigating approaches that will allow us to better understand online user's behavior at a micro-level. To better understand the online user's dynamics, we are investigating how individuals behave towards a music track when it is played by a certain group of peers and how or when it is adopted by her network. We believe future work should also consider domains that are different from music.

5. References

[1] J. Coleman, E. Katz, and H. Menzel, "The Diffusion of an Innovation Among Physicians," *Sociometry*, vol. 20, Dec. 1957, pp. 253-270.

[2] E. Mansfield, "Technical Change and the Rate of Imitation," *Econometrica*, vol. 29, Oct. 1961, pp. 741-766.

[3] J. Arndt, "Role of Product-Related Conversations in the Diffusion of a New Product," *Journal of Marketing Research*, vol. 4, Aug. 1967, pp. 291-295.

[4] F.M. Bass, "A New Product Growth for Model Consumer Durables," *Management Science*, vol. 15, 1969, pp. 215-227.

[5] J.J. Brown and P.H. Reingen, "Social Ties and Word-of-Mouth Referral Behavior," *The Journal of Consumer Research*, vol. 14, Dec. 1987, pp. 350-362.

[6] V. Mahajan and E. Muller, "Innovation Diffusion and New Product Growth Models in Marketing," *The Journal of Marketing*, vol. 43, Autumn. 1979, pp. 55-68.

[7] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA: ACM, 2006, pp. 611-617.

[8] L. Loh and N. Venkatraman, "Diffusion of Information Technology Outsourcing: Influence Sources and the Kodak Effect," *Information Systems Research*, vol. 3, 1992, pp. 334-358.

[9] D.A. Williamson, "The Rush to Social Networks - eMarketer," *emarketer.com*, Feb. 2009.

[10] D. Godes and D. Mayzlin, "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, vol. 23, Fall. 2004, pp. 545-560.

[11] D. Mayzlin, "Promotional Chat on the Internet," *Marketing Science*, vol. 25, 2006, pp. 155-163.

[12] D. Godes and D. Mayzlin, "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science*, vol. 28, 2009, pp. 721-739.

[13] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, 2009, pp. 21544-21549.

[14] J.A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, vol. 43, Aug. 2006, pp. 345-354.

[15] C. Dellarocas, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, vol. 49, Oct. 2003, pp. 1407-1424.

[16] S. Bikhchandani, D. Hirshleifer, and I. Welch, "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades," *The Journal of Economic Perspectives*, vol. 12, Summer. 1998, pp. 151-170.

[17] A. Jackson, J. Yates, and W. Orlikowski, "Corporate Blogging: Building community through persistent digital talk," *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, 2007, p. 80.

[18] B. Bickart and R.M. Schindler, "Internet forums as influential sources of consumer information," *Journal of Interactive Marketing*, vol. 15, 2001, pp. 31-40.

[19] J. Leskovec, L.A. Adamic, and B.A. Huberman, "The dynamics of viral marketing," *Proceedings of the 7th ACM conference on Electronic commerce*, Ann Arbor, Michigan, USA: ACM, 2006, pp. 228-237.

[20] J.B. Walther, "Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction," *Communication Research*, vol. 23, 1996, pp. 3-43.

[21] L. Sproull and S. Kiesler, "Reducing social context cues: electronic mail in organizational communication," *Manage. Sci.*, vol. 32, 1986, pp. 1492-1512.

[22] B.A. Nardi, S. Whittaker, and E. Bradner, "Interaction and outeraction: instant messaging in action," *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, Philadelphia, Pennsylvania, United States: ACM, 2000, pp. 79-88.

[23] B.A. Huberman, D.M. Romero, and F. Wu, "Social Networks that Matter: Twitter Under the Microscope," *SSRN eLibrary*, SSRN, 2008.

[24] M. McPherson, L. Smith-Lovin, and J.M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, 2001, pp. 415-444.

[25] E. Adar and L.A. Adamic, "Tracking Information Epidemics in Blogspace," *Proceedings of the 2005 IEEE/WIC/ACM*

International Conference on Web Intelligence, IEEE Computer Society, 2005, pp. 207-214.

[26] J.L. Gastwirth, "The Estimation of the Lorenz Curve and Gini Index," *The Review of Economics and Statistics*, vol. 54, Aug. 1972, pp. 306-316.

6. Appendix – Data Tables

Table 1: 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 2: 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 3: Music Listening Behavior (standard deviation in parentheses)

Music Listening Statistic for Bands	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)
Tracks played between July 2008 and Sep 2008	722 (602)	842 (798)
New tracks exposed to user	4577 (4342)	4737 (4241)
Gini coefficient for track listening heterogeneity	0.846 (0.096)	0.759 (0.155)

Table 4: Logit regression with diffusion (binary) as dependent variable

Variable	Coefficient (Bands)	Odds Ratio (Bands)	Coefficient (Tracks)	Odds Ratio (Tracks)
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-Squared	0.492		0.593	
* 10% significance, ** 5% significance				

Table 5: Negative binomial regression with diffusion (count) as dependent variable

Variable	Coefficient (Bands)	Marginal Effect (Bands)	Coefficient (Tracks)	Marginal Effect (Tracks)
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-Squared	0.279		0.171	
* 10% significance, ** 5% significance				