Will the Global Village Fracture into Tribes: Recommender Systems and their Effects on Consumers

Abstract

Personalization is becoming ubiquitous on the World Wide Web. Such systems use statistical techniques to infer a customer's preferences and recommend content best suited to him (e.g., "Customers who liked this also liked…"). A debate has emerged as to whether personalization has drawbacks. By making the web hyper-specific to our interests, does it fragment internet users, reducing shared experiences? We study whether personalization is in fact fragmenting the online population. Surprisingly, it does not appear to do so in our study. Personalization appears to be a tool for helping users widen their interests, which in turn creates commonality with others. This increase in commonality occurs for two reasons, which we term volume and taste effects. The volume effect is that consumers simply consume more after personalized recommendations, increasing the chance of having more items in common. The taste effect is that, conditional on volume, consumers buy a more similar mix of products after recommendations.

"Will the global village fracture into tribes?" – P. Resnick

1. INTRODUCTION

Recommender systems are becoming integral to how consumers discover media. They are used for all major types of media, such as books, movies, music, news, and television. They are commonplace at major online firms, such as Amazon, Netflix, and Apple's iTunes store. And they have a strong influence on what consumers buy and view. With music, Gartner and Harvard's Berkman Center predict that in 2010, over 25% of music sales will come from taste-sharing applications such as recommenders. With movies, Netflix reports that over 60% of their rentals originate from recommendations (Thompson 2008). With online news, Google News reports that recommendations increase articles viewed by 38% (Das et al. 2007). At Amazon, which sells music, books, and movies, 35% of sales are reported to originate from recommendations (Lamere & Green 2008).

The value that recommenders offer is personalization: the consumption experience is personalized to each user's taste. A personalized radio station plays music not for the general public but for each particular user. A personalized newspaper does not show the same front page to everyone but customizes it for each reader. A retailer arranges its online shelves and displays based on who is browsing at that moment. Such personalization is valuable in modern media markets, which can have millions of products to choose from. Indeed, personalization has become a major theme of research in Information Systems (e.g., Murthi and Sarkar 2003; Dellarocas 2003; Brynjolfsson et al. 2006; Clemons et al. 2006) and Marketing (e.g., Ansari et al. 2000; Manchanda et al. 2006; Shaffer and Zhang 1995; Rossi et al. 1996), with its origins in targeted and customized marketing.

The following examples show how recommenders systems create this personalized experience:

> The newspaper … is undergoing the most momentous transformation.... Online versions are proliferating, ... yet so far, few newspaper sites look different from the pulp-and-ink papers that spawned them.... Often, the front page changes only once a day, just like the print version, and it shows the same news to all readers. There's no need for that uniformity. Every time a Web server generates a news page, ... it can generate different front pages, ... producing millions of distinct editions, each one targeting just one person – you.
> –*Greg Linden, creator of Findory news and Amazon recommendations* (2008)

> Last.fm connects you with your favorite music and uses your unique taste to find new music, people, and concerts you'll like.
> –*Last.fm website*

> TiVo, a television recording system, will automatically [create] your personal TV line-up. It will also learn your tastes, so that it can suggest other shows you may want to … watch.
> –*TiVo website, quoted in Sunstein* (2007)

Along with the benefits of personalization, however, a debate has emerged as to whether it has drawbacks. Personalizing websites means that we may no longer see the same newspaper articles, television shows, or books as our peers. Critics thus argue that recommenders systems will create fragmentation, causing users to have less and less in common with one another. An alternative view contends that recommenders may do the opposite: recommenders may have homogenizing effects because they share information among users who otherwise would not communicate. This paper presents the first empirical evidence for the debate on whether recommenders fragment versus homogenize users.

The motivation for this study is that fragmentation in consumption has implications for consumers, firms and society. For consumers, shared consumption often has an associated externality. For example, the value of discussing a movie or book with others presents a positive externality from shared consumption (Katz and Shapiro 1985) whereas the desire to signal a unique identity represents a negative externality from shared consumption (Berger and Heath 2008). If recommenders affect consumption similarity, these externalities are in turn impacted. For firms, the fragmentation question has marketing implications. Recommenders lower search costs. As a result, one interpretation of observing recommendations-influenced purchases is that it better reveals preferences. Observing that preferences are more versus less fragmented than previously thought could inform firms' marketing policies. To the extent more fragmentation occurs, narrow, targeted marketing policies appear more justified. To the extent less fragmentation occurs, consumers may prefer a range of experiences that narrow targeting does not deliver. Finally, from a policy perspective, the literature has expressed concern that fragmentation is a negative consequence for society. These critics suggest the media and government should do more to increase exposure to a variety of content. In contrast, finding evidence of homogenization would suggest that such policies and regulation of personalization on the internet are not warranted.

We note that it is possible to deliberately design systems with the goal of increasing commonality or similarly with the goal of decreasing commonality. However, commonality is not a design goal in practice and instead a side-effect of recommender use. Thus, our goal is to document the impact of a commonly used design rather than to investigate if there exists a design that can increase commonality or cause fragmentation. We find, in an empirical study of a music industry recommendation service, that recommendations are associated with an increase in commonality in the items consumed/purchased by their users. This increase in purchase similarity occurs for two reasons, which we term volume and taste effects. The volume effect is that consumers simply purchase more after recommendations, increasing the chance of having purchases in common with others. The taste effect is that consumers buy a more similar mix of products after recommendations, conditional on volume. When we view consumer purchases as a similarity network before versus after recommendations, we find that the network becomes denser and smaller, or characterized by shorter inter-user distances. These findings suggest that for this setting, concerns of fragmentation may be misplaced. Although our results are derived for one recommendation technology deployed in one setting, they clearly demonstrate that commonly used designs can have homogenizing effects and that the argument that these systems cause fragmentation is not universally true.

2. PRIOR WORK

A simplified taxonomy of recommender systems divides them into content-based versus collaborative filtering-based systems. Content-based systems use product information (e.g., genre, mood, author) to recommend items similar to those a user rated highly. Collaborative filters, in contrast, are unaware of a product's content and instead use correlations in sales or ratings to identify what similar customers bought or liked. Perhaps the best-known collaborative filter is Amazon.com's, with its tagline, "Customers who bought this also bought…" The design of these systems has been an active research area for at least fifteen years. An extensive review in the Information Systems literature is provided in Adomavicius & Tuzhilin (2005). Recent work in Marketing (e.g., Ansari et al. 2000; Bodapati 2008) considers how best to design these systems for business use.

Although a large body of work exists on designing recommenders systems, we know much less about how they affect the market and society. This is despite the thousands of papers that present new recommender algorithms and millions of transactions occurring through them. This paper continues a small stream of work in that direction. Recent work (Fleder and Hosanagar 2007, 2009; Hervas-Drane 2007; Oestreicher-Singer and Sundararajan 2009) ask how recommenders affect *products*: which products gain versus lose sales due to recommenders and whether recommenders increase the market for niche

goods, or "long tail". This paper asks the complementary question of how recommenders affect *consumers*: whether they cause consumers to consume more or less in common with one another.

A range of views exist as to whether recommenders will fragment versus homogenize users. Sunstein (2007) argues that recommenders create fragmentation by limiting users' media exposures to their predefined, narrow interests. These fragmentation effects, he argues, are undesirable. "In a democracy people do not live in echo chambers or information cocoons. They see and hear a wide range of topics and ideas, … even if they did not … choose to … in advance" (2007). While Sunstein is clear in explaining why fragmentation may be undesirable, the antecedent, that recommenders create fragmentation to begin with is ultimately an assumption. Pariser (2011) similarly argues that personalization on the Internet confines users' information consumption to algorithmically-computed interests and limits commonality in consumption. He indicates that "the filter bubble is the invisible, personal universe of information that results--a bubble you live in … the world you see online and the world I see may be very different" and this in turn limits "ability to put ourselves in other peoples' shoes" (Terdiman 2011). Another supporter of the fragmentation view is Pattie Maes, creator of one of the first recommender systems. Maes says that recommenders can have a "narrow-minded" and "hyperpersonalized" aspect. "You don't want to see a movie just because you think it's going to be good. It's also because everyone [else is] … talking about it, and you want to be able to talk about it too" (Thompson 2008). Consuming the same media and products "is a way of participating in society," and this could be lost on account of recommender systems (paraphrased in Thompson 2008).

Sunstein, Pariser and Maes all appear to view recommenders as causing fragmentation, but they differ in their views as to why this is undesirable. Sunstein argues that a democracy requires citizens to have a range of experiences and viewpoints. For example, in news programming, users should be exposed to multiple views on a topic, not just the one that reinforces their existing beliefs – which he believes will be the case as recommenders become more prevalent. Pariser argues that the information bubble that results from recommender use will limit creativity and also lead consumers to make poor decisions. Maes' has a different criticism: a product's popularity has a positive externality, and recommenders may cause us to forfeit this. If there is a benefit to reading the same books as others or seeing popular movies (e.g., by being able to discuss the experience with others) we should be wary of recommenders because these benefits could disappear.

A more moderate view is suggested by Nicholas Negroponte, co-founder of the MIT Media Lab. Negroponte coined the term "The Daily Me" (1995), referring to the ability of recommenders to create newspapers customized to each person's interests. The Daily Me might create fragmentation by showing users only the content that matches their viewpoints. However, Negroponte also discusses the "The Daily Us," suggesting that consumers may also turn to recommenders when they need help exploring areas

outside their interests, "learning about things [they] never knew [they] were interested in." Using recommenders this way would create commonality in knowledge among users, not fragmentation.

Van Alstyne and Brynjolfsson (2005) formalize this mixed view in an economic model. They ask whether internet technologies like recommender systems will lead to fragmentation versus homogenization – in their terms, a cyber-Balkans versus a global village. Fragmentation is measured both by physical interaction and consumers' knowledge overlap. They show that as technology lowers search costs and communication costs, either outcome can occur. Which outcome occurs in their model depends on a parameter representing consumers' taste for specialization. This parameter is difficult to specify, and so complementary empirical work is needed.

Similar mixed views were shared by the creators of early collaborative filters. At the Berkeley Collaborative Filtering Workshop in 1996, a time at which research on recommender systems was just beginning, Paul Resnick, then of AT&T Research, asked if the "global village [would] fracture into tribes" (Arnheim 1996). John Riedl, co-inventor of one of the first recommenders, asked if collaborative filtering would "democratize ... information ... or result in social fragmentation." More recently, Greg Linden, one of the developers of Amazon's original recommender system, states that "[critics] talk about personalization as narrowing and filtering. But that is not what personalization does. Personalization seeks to enhance discovery, to help you find novel and interesting things."

Lastly, Fleder and Hosanagar (2009) model how collaborative filters affect consumer choice. Their study is mainly concerned with the impact on products but one of their results does suggest that users become more similar to one another after recommendations. Collaborative filters, they show, are biased toward recommending products that others bought before. Thus when consumers accept recommendations, co-purchases are created with many other users and commonality increases. However, their results are based on a simulation study rather than real-world data and it is also unclear if their arguments are applicable for content-based designs.

The discussion reveals that there are mixed views as to whether recommenders will fragment users, but there is not yet any empirical evidence on the issue. The goal of this study is to provide the first empirical evidence on the impact of recommenders on purchase similarity.

3. PROBLEM FORMULATION

This section defines the problem formally. While many authors have discussed the fragmentation question qualitatively, the empirical question has not been posed in concrete terms. We view the formulation as one contribution of this work.

3.1. Research Questions

Throughout this paper, we operationalize the notion of fragmentation and homogenization in terms of commonality in items consumed by users. This is analogous to Van Alstyne and Brynjolfsson's analysis of how the internet affects knowledge overlap among users (2005) and is consistent with notions of fragmentation as suggested by Sunstein and Maes ("shared experiences").

Our goal is to study whether recommenders make users' consumption more or less similar to one another. We divide the question in two components:

1. Aggregate level: overall, are consumers farther or closer to one another?
2. Individual level: are there differential effects at the individual level, by which some users become closer and others farther?

The first question measures the overall effect of whether users become farther or closer to one another. The second question explains why. For example, effect (1) may show that users are less similar on average. Effect (2) explains why: for example, even though there is a net reduction in purchase similarity, it may be the case that the closest users became closer and the farthest became much farther.

A note on terminology: The meaning of "close" and "far" will be quantified in the next section. Qualitatively, throughout the paper we refer interchangeably to users who are "close" as exhibiting similarity, commonality, or homogeneity; opposite this, we refer to users who are "far" as exhibiting fragmentation and having little overlap in their purchases.

3.2. Two Group Design

The analysis design throughout is analogous to a two-group experiment. One group is "treated" with recommendations and their behavior compared before versus after. A control group is not treated with recommendations, and their behavior is compared over the same period. The data are in fact observational, as we will discuss, but the terminology of experiments simplifies the writing.

Let O_{it} denote an observation on group i during time period t. O_{it} is a list of tuples (user, artist, # songs purchased) for all users in group i during period t. Group $i = 1$ is the treated group, which is unexposed to the recommender during $t = 1$ but exposed to the recommender during $t = 2$. Group $i = 2$ is the control, which is unexposed to the recommender during both time periods. The time periods are the same for both groups. Figure 1 represents this setup, where X denotes exposure to recommendations.

Treated:	O_{11}	X	O_{12}
Control:	O_{21}		O_{22}

Figure 1. Schematic of the Two Group Design

Using this design, we can compare the treated group before and after recommendations. We can also compare the treated group to the control over the same period. The control accounts for factors such as time trends and maturation that might be confounded with recommender usage in a one group pre-post design (Campbell & Stanley 1963).

3.3. Hypotheses to Test

We wish to compare how the treated and control groups change over time. Let $T(O_{it})$ be some statistic of interest on O_{it} measuring fragmentation. As shorthand, we will write T_{it}. We define the following quantities of interest:

Difference in treated:	D_1	$\equiv T_{12} - T_{11}$
Difference in control:	D_2	$\equiv T_{22} - T_{21}$
Difference-in-differences:	D	$\equiv D_1 - D_2$

D_1 describes changes in the treated group. D_2 describes changes in the control. The difference-in-difference estimator, D, describes how much changes in the treated group exceed those in the control. For example, suppose that independent of recommendations, a time trend is occurring in the music industry that affects both groups. Thus observing $D_1 \neq 0$ does not mean recommendations have an effect on consumers because the same trend will affect D_2. However, the difference-in-differences estimator D can identify changes in the treated group beyond the time trend by subtracting the change in the control.

Let $\mu \equiv E[D]$, where D's distribution is not known to us. The central questions of this paper take the form

$$H_0: \mu \equiv E[D] = 0$$
$$H_a: \mu \equiv E[D] \neq 0$$

The above formulation is general for any underlying T, and many questions about similarity in consumer purchases can be posed in this framework. Several statistics of interest $T(\ )$ are defined in the next section. Each gives rise to a separate D and hence a separate hypothesis of the form above. The hypotheses are always stated as two-sided. This makes our tests more conservative, but it is necessary because the literature offers mixed views as to whether fragmentation versus homogenization will occur.

4. FORMULATION SPECIFICS

This section defines the quantities of interest $T(O_{it})$. To facilitate this, we take the intermediate step of defining a network $G(O_{it})$ among the firm's consumers and making $T(G(O_{it}))$ a function of that network.

At first glance, introducing networks appears to complicate the analysis by adding an extra step. In contrast, we will see this provides a great service for interpreting the data.

4.1. Motivation for Network Analysis

We define a network in which consumers are the nodes and edges represent similarity between consumers' purchases. This paper's goal of asking whether users' purchases become more or less similar after recommendations will become equivalent to asking how the consumer network changes pre-post recommendations.

For each O_{it} we will create a user network $G(O_{it})$. Then, we will define quantities of interest (e.g., density, path length) on the network $T(G(O_{it}))$ and study how these quantities change before versus after recommendations.

The consumer network is not a true social network because its edges do not represent physical relationships. Its edges instead represent similarity in purchases. Still, we find it useful to formulate the problem as a network one. First, the benefit of introducing networks is interpretation. Networks are a useful object for describing changes in user similarity. It is easy to conceive of a network expanding, shrinking, or becoming more dense. In contrast, such interpretations would be difficult if we instead studied a large correlation matrix of users' purchases. Second, network analysis is recently being applied to settings like ours in which edges represent similarity of interests or purchases. Huang et al. (2007) and Smith et al. (2007) use co-purchase and co-occurrence data to build an "implicit" network of individuals. In these examples, the network is not strictly necessary for measuring similarity, but it aids in interpretation.

4.2. Defining the Network

Mathematically, our network is a graph made of nodes and edges. Users are the nodes, and edge weights describe the similarity between user pairs, as defined by commonality in purchases.

For notation, we can interpret O_{it} as a users × artists matrix of purchase counts. An element $(O_{it})_{xy}$ is the number of songs user x purchased of artist y.[1] A row of this matrix is denoted $(O_{it})_x$. For each O_{it}, the corresponding network is $G(O_{it})$ which is denoted as simply G_{it}.[2] The network G_{it} is a users × users matrix of edge weights. An element $(G_{it})_{xy}$ is the edge weight between user x and user y. Defining the network is thus equivalent to defining the distance between any two users.

[1] The vector is defined in terms of artists rather than songs because the recommender used in our study operates at the artist level. That is, the input to the recommender is the artist being played. We discuss the recommender design in detail in Section 5.1.

[2] $G(\ )$ is a function that converts the purchase matrix into a network, or $G(O_{it}) \equiv G_{it}$.

Our main network has a simple construction. Within a given group and time period, users *x* and *y* have an edge between them if they purchase at least one artist in common.

Unweighted

$$(G_{it})_{xy} \equiv \begin{cases} 1, \text{ if users } x \text{ and } y \text{ have} \geq 1 \text{ artist in common } ((O_{it})_x \bullet (O_{it})_y \geq 1) \\ \text{Unconnected, otherwise} \end{cases}$$

This is an unweighted network in which any edge, if it exists, has weight 1. The • symbol indicates the vector dot product, showing how this definition might be generalized to other similarity functions.

There are many other ways to construct the network. In unweighted networks like the one above, there can be other definitions for when an edge should be present. Weighted networks are also possible, and there are many ways to define the edge weights. In the main sections of this paper, we focus our base case on the network above because its definition is simple and intuitive. In the electronic companion, we present results for other network definitions. We simplify the exposition in this way, since all of the networks tested yield nearly the same conclusions.

4.3. Defining *T*: Measures of the Network's Properties

With the network $G(O_{it})$ defined, we next define summary statistics of the network's properties, $T(G(O_{it}))$. *T* summarizes in one number a particular network property and thus facilitates comparisons of the network over time. We define three such measures below. As notation, let $d_x \equiv \sum_{y=1, y\neq x}^{n} G_{xy}$ be the degree of user *x* where *n* is the number of users in the network. Further, let ${}_nC_2$ denote the number of user pairs that can be formed from a set of *n* users (${}_nC_2 = n(n-1)/2$).

Measure	$T(G(O_{it}))$ =
Density	$\frac{1}{{}_nC_2}\sum_{x=1}^{n}\sum_{y<x} I(G_{xy}=1)$
Median Degree	$\text{Median}\{d_x\}_{x=1}^{n}$
Path Length	$\frac{1}{{}_nC_2}\sum_{x=1}^{n}\sum_{y<x} \textit{Shortest Distance}(x,y)$

Density. The density is the fraction of edges that exist out of the total number of edges possible. Higher density means users have more connections among them.

Median Degree. The median degree is the number of connections to other users that the typical (median) user has. The higher the median degree, the more similar users are to one another. This represents the median of the degree distribution, whereas the density is the degree distribution's average.

Path Length. The path length is the shortest distance between any two users, averaged over all users in the network. If users *x* and *y* are connected, the shortest distance is 1, the edge between them. Otherwise, the path is through other users. The shorter this distance, the "smaller" the network is said to be, using the terminology of Watts & Strogatz (1998), who popularized the study of "small world" networks. Mathematically, the shortest distance between users does not have a closed form expression, but it can be computed using Dijkstra's algorithm or, more efficiently, with the Floyd-Warshall algorithm (Papadimitriou & Steiglitz 1998).

To summarize the analysis setup, the data are in the form of a two group experiment (O_{it}). Each data set is converted to a network $G(O_{it})$. Summary statistics are computed on each network $T(G(O_{it}))$. Finally, these statistics are compared across the groups and time.

5. DATA

5.1. Data Source

We study the fragmentation question using data from an online music service, referred to here as Service.[3] Service is a free software add-on to Apple's iTunes. iTunes, in turn, is the music player that allows users to buy music from Apple's iTunes store, the largest music retailer in the U.S. (Apple 2008). Service personalizes the user experience in two ways. When users listen to music in iTunes, Service suggests other songs that the user may like. The suggestions appear in a window appended to iTunes, where the user can sample these songs and opt to purchase them. If a purchase results, Service earns a commission. Service also provides recommendations through a website where users can view the play histories of other Service users with similar taste. These play histories are uploaded automatically by the plugin to Service's website on a continual basis. Together, these two features comprise the personalization technology.[4]

Figure 2 shows a screenshot of the plugin. Apple's iTunes appears at left. The plugin, as appended to iTunes, is at right and displays a list of recommended songs. The song suggestions by the

[3] Service Inc. will be replaced with the firm's name upon publication. We have in writing legal permission to disclose this.

[4] It is common for most online firms to use multi-component recommendation systems. For example, Netflix and Amazon's "recommendations" page in fact have different types of recommendations generated from different algorithms all on the same page. In these environments, it is hard to isolate the impact of any one component. It is also debatable whether the researcher would prefer to analyze users exposed to just one component of Service's personalization technology. An analysis based on just one component would not be indicative of the real trend occurring online because users are typically exposed to all components. Accordingly, our study focuses on the net impact of the multi-component system.

plugin are based on the artist currently playing (i.e., the query to obtain recommendations is the current artist). Based on the current artist, Service identifies the 6 most similar artists and populates the window with this list. Artist-to-artist similarity is defined by a hybrid of content and collaborative data, although, at the time of data collection, the results are heavily weighted toward the content portion (90% versus 10%). Thus in the taxonomy of recommender systems by Adomavicius & Tuzhilin (2005), the plugin is effectively a content-based, item-to-item-based system.

Figure 2. Screen shot of the recommendation service.

5.2. Novelty of the Data

To study the effects of recommenders, a contrast is needed between users exposed and unexposed to recommendations. The data collected by most retailers (e.g., Amazon, Netflix) is inadequate because retailers only observe consumers after they arrive at their website and hence after exposure to recommendations. This may be the reason, we speculate, that others have not been able to study the fragmentation question. Our data are novel in this regard. When a user registers for Service, a history file is extracted from the user's iTunes player. This history file contains the names and timestamps of all songs ever added to that user's music library, and thus it provides a record of the user's behavior *prior* to joining Service. The user's post-registration purchases are also observed by Service because the plugin notifies Service via the internet of all songs added to the user's iTunes library, whether bought at the iTunes store or not. This combination of the history file and continued communication via the plugin thus gives us a before and after view of the user's behavior.

Besides comparing users' purchase histories before and after registering, we can also compare these users with a control group. The control data are obtained by again exploiting the history files of Service users. For users who register after our study, their history files allow us to look backward at their Service-uninfluenced behavior during the same time period. More detail is given in the next section. This use of eventual Service users for the control affords a measure of similarity between the groups. Thus the

new data source enables a before-after recommendations contrast as well as data on a control group for the same period.

5.3. Data Inclusion Criteria

This section describes the process for setting up the data in the two-group design introduced earlier. Figure 3 summarizes the details of this process. The data are collected via Service's plugin that is installed on each user's machine. The plugin relays to Service in near real-time the timestamp and product information of any song added to that user's iTunes library. For ease of writing, we refer to songs as purchases, but our data in fact capture all songs added to a user's library, whether purchased from Apple's iTunes store, purchased from another firm, or downloaded elsewhere online.

The original data comprise users who registered for Service between January – July 2007. We define the treated group as those users who registered sometime during March 2007. March is chosen because it is roughly in the middle and provides us with sufficient pre/post data. The time periods for the before-after comparison are the two month windows January-February and March-April.[5] The control group is defined by users who registered for Service sometime from May on. We observe this group's Service-unaffected behavior over January-April because upon their eventual registration, sometime from May on, we extract their iTunes history files and look backward at the January-April period.

A criterion for inclusion in the study is that each user began using iTunes in August 2006 or earlier. Upon installing iTunes or buying an iPod, users often load their CD collections onto their computers. We do not want to treat loading of old music as new purchases. Thus the criterion of installing iTunes in August 2006 or earlier creates a buffer of at least four months (September-December 2006) between installing iTunes and our analysis. This is conservative because the loading of old CDs typically occurs within the first month of iTunes/iPod use.[6]

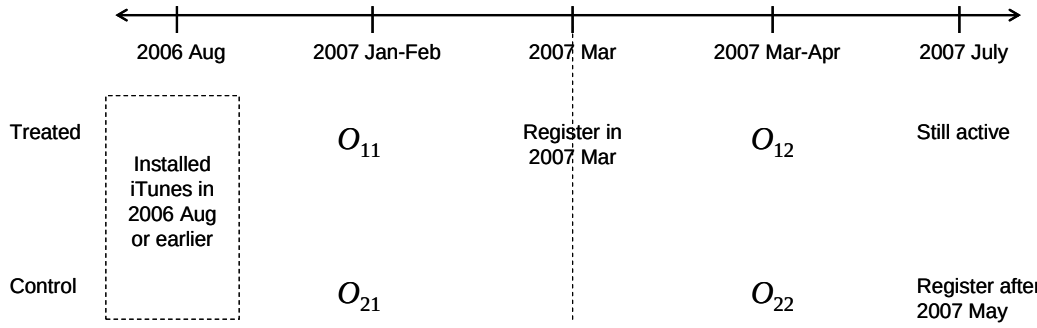

Figure 3. Data Setup and Analysis Design

[5] That some users registered in late March could dampen the results' magnitude because it allows some Service-unaffected data to enter the post-recommendations period. One cannot circumvent this by centering each user's before-after data exactly on his registration date: since each user differs in this date, there would be no well-defined period for constructing the control. We are conservative and accept this tradeoff of a possible dampening of results in order to have a well defined control group.

[6] The iPod/iTunes installation date is not recorded. We proxy it using the day the first song is added to each user's library.

The second criterion for inclusion is active user status. Some users uninstall the plugin before the study's end. So that our panel includes the same users before and after, which is required for our user-to-user before-versus-after comparisons, we adopt the criterion that users have the plugin installed for the study's entire duration.[7] The implications of these data-inclusion criteria are discussed below.

The data collection has two limitations. First, assignment to the treated versus control group is not randomized. Since registration is the user's choice, there can be a selection bias. For example, it is possible that registration is a response to changes in demand for music rather than a cause of it. A section later on sensitivity analysis shows this is unlikely. For example, we show that our results are qualitatively similar for a matched sample of users. Also, the control group demonstrates the same effects when these users later register for the service. We defer a detailed discussion of this to Section 9.1.

The second limitation involves users who uninstall the plugin. About half of the users in the treated group uninstall the plugin before the data collection ends (before the end of period $t = 2$). If the uninstallation decision is independent of music preference – for example, uninstalling the plugin to free up disk space or not liking the extra screen space occupied by the plugin – then the conclusions are unaffected because the selection is equivalent to our taking a random sample. If they are not independent, then the analysis of the non-attriting population may overstate the magnitude of the results but it will not change the direction of the results. This idea is discussed and empirically bounded in the section on sensitivity analysis.

The resulting data set consists of 1,794 users in the treated group and 858 users in the control group. Treated users purchased a total of 215,749 songs from 24,368 artists in the before period whereas control users purchased 106,431 songs from 14,785 artists in the before period. In terms of number of users, the treated group is larger than the control. We note that prior to recommendations, the total number of songs and artists purchased is higher in the treated group. This likely occurs because the treated group has more users and thus more song purchases and in turn a greater chance of covering more artists. In the experiments below, we will compare treated and control groups with an equal number of users to control for such differences.[8]

6. RESULTS ON THE OBSERVED DATA

This section shows how the consumer network changes when recommendations are introduced. Overall, we find consumers become more similar to one another: the consumer network becomes denser, more connected, and smaller.

[7] Un-installation is not observed, so we proxy this by including those users' whose plugin communicates with the Service at least once after the post-recommendations period.

[8] The average number of songs per user for the treated and control group are comparable at 120 and 124 respectively.

6.1. Aggregate Analysis on the Observed Data

Using the two group design, we construct the four networks – before and after recommendations for the treated and control – and calculate the summary measures $T(\ )$ on each. Then, for each summary statistic T, we calculate the changes over time $D_1 = T_{12} - T_{11}$, $D_2 = T_{22} - T_{21}$, and the difference-in-differences estimator $D = D_2 - D_1$.

Table 1 shows the results. Across the columns are the three T statistics: density, median degree, and path length. The rows present the statistics for the treated group (row "T") and control group (row "C"). The table's elements show the values of T before and after recommendations. The column D_i lists the difference for each group. Last, the column D/p lists the difference-in-differences estimate D with the p-value below it from a test that $D = 0$. (The "/" symbol indicates separate rows.) To test the hypothesis that $D = 0$, we use the non-parametric method of permutation tests. The testing procedure is described in the appendix.

Table 1. Summary Measures for the Unweighted Network – *Observed Data*

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	23%	46%	23%	22%	167	402	235	234	1.80	1.54	-0.26	-0.26
C	19%	19%	0%	<0.01	134	135	1	<0.01	1.86	1.86	0.00	<0.01

The results show that on all three measures, users' purchases are more similar to one another after recommendations. First, the treated network becomes denser, showing that users have more connections among themselves. Before recommendations, 23% of the edges are filled in, and after 46% are present, yielding $D_1 = 23\%$. This is a large increase in density. Over the same period, the control has no noticeable change and $D_2 \approx 0$. The difference-in-differences estimate is $D = 23\% > 0$, indicating that the treated network does become more similar relative to the control. This difference is significant, as the hypothesis $D = 0$ is rejected ($p < 0.01$). On the other two metrics, we also observe greater similarity after recommendations. The median degree increases, $D > 0$, indicating that the typical user has more connections to others. Similarly, the path length decreases, $D < 0$, indicating that on average users are fewer hops away from one another. All of the results are significant ($p < 0.01$).[9]

6.2. Individual-Level Analysis on the Observed Data

The above analysis showed that in aggregate users' purchases are more similar after recommendations. This section asks if there are differential effects at the individual level. For example,

[9] All the networks have one large, connected component containing nearly all users with few unconnected users outside it. Thus the density, degree, and path lengths are not biased due to changes in the size of the main component.

could close users become closer but far ones become farther – in such a way that the aggregate result masks this? If true, even though the network is more similar in aggregate, far users becoming farther would be evidence of fragmentation.

This question is similar to asking whether users form tighter clusters after recommendations. If so, the aggregate effect could mask a world in which within-cluster similarity increases but between-cluster similarity decreases. It is convenient to phrase this idea using the language of cluster analysis but difficult to draw firm conclusions using cluster analysis. The reason is that for every type of cluster analysis, there is no true number of clusters: it is a subjective quantity.[10] Thus one cannot say firmly whether there are fewer or more clusters after. Further, one cannot measure whether the clusters become tighter or looser, as this depends on determining the "right" number of clusters to begin with before and after.

To address this question, we compare the distance of all user pairs before versus after recommendations and examine whether there are sub-populations that become farther. Table presents these results. The table plots the path length between all ${}_nC_2$ user pairs. The horizontal axis is the number of hops before recommendations, and the vertical axis is the number of hops after recommendations. The values in the table are the percent of user-pairs falling in each cell. For example, 7.3% of the treated users were one hop away before recommendations and 2 hops away after recommendations. User-pairs becoming farther lie above the diagonal, while user-pairs becoming closer lie below it. A distance of infinity (∞) means there is no path between the given two users.[11]

The control group appears stable (right side), as it has roughly equal weight above and below the diagonal. In contrast, the treated group (left side) shows a different pattern. First, the aggregate effect toward similarity is evident: there are more user-pairs becoming closer (36.9% weight below the diagonal) than there are becoming farther (9.2% weight above the diagonal). This is consistent with the aggregate findings above. Second, the increase in similarity appears uniform: all types of users become closer to one another. Users who were close became closer, *and* users who were initially far became closer too. There does not appear to be evidence of a differential effect.

Note that some users do grow farther, but this is not a differential effect. We expect some chance fluctuation: users who were by chance closer revert to being farther, and users who were by chance

[10] There is no shortage of methods for "estimating" this subjective quantity. Two common ones, for example, are finding a scree plot's kink (the Gap statistic) or estimating the parameter governing the number of components in a mixture model (if one assumes a mixture model). However, this introduces two measures of subjectivity into the results: the choice of clustering method itself (e.g., vector quantization, hierarchical methods, mixture models, spectral, methods, etc) and the choice of method to estimate the number of clusters. Even in a mixture model that one generates himself this question is based on a subjective choice about resolution. Consider a 2 component Gaussian mixture where each component is a 4 Gaussian mixture. Depending on the resolution chosen, the number of clusters is 2 or 8.

[11] In Table 2, a very small number of pairs are four or five hops away. This number is so small ($\approx 0.04\%$) that for clarity we omit them from the presentation (but not the analysis) to avoid rows and columns of nearly all zeros.

farther revert to being closer. This is seen in the control group, where 11.3% went from 1 to 2 hops while 11.1% went from 2 to 1 hops. This level of mixing is roughly equal. In the treated group, some pairs do become farther – 7.3% go from 1 to 2 hops – but many more become closer – as 27.7%, went from 2 to 1 hops. This difference of 20.4% is large as well, since it is a fraction of ${}_nC_2 \approx 300{,}000$ user pairs. To summarize, the trend toward greater similarity exists at all initial path lengths, and so we do not see evidence of a differential effect.

Table 2. Path Lengths between all user pairs – *Observed data*.
Entries represent the percentage of all ${}_nC_2$ user-pairs.

Treated

# Hops After \ # Hops Before	1	2	3	∞
∞	0.2	1.2	0.2	0.3
3	0.0	0.3	0.1	0.1
2	7.3	37.8	3.0	3.8
1	15.5	27.7	1.0	1.3

Becoming closer (below diagonal)	36.9
Becoming farther (above diagonal)	9.2
No change (on diagonal)	53.7

Control

# Hops After \ # Hops Before	1	2	3	∞
∞	0.3	4.0	0.6	1.9
3	0.2	3.8	1.2	0.8
2	11.3	48.9	4.0	3.6
1	7.4	11.0	0.4	0.3

Becoming closer (below diagonal)	20.1
Becoming farther (above diagonal)	20.2
No change (on diagonal)	59.4

7. VOLUME EQUALIZATION

The results so far show that purchase similarity increases after recommendations. We next investigate the changes in volume and type of purchases to explore the drivers of this result. Table provides summary information on changes in purchase volume for the two groups. The table shows that purchases increase after recommendations for treated users. This increase was anticipated, although the size of roughly 50 percent is larger than expected. This increase in songs added is not seen in the control, where in fact the number of songs added decreases slightly. Consistent with the increase in songs purchased by treated users, the table also shows that the number of artists for whom at least one song was purchased increases considerably for the treated group, indicating that users explore a wider range of artists under recommendations. Again, no such increase is seen in the control.

Table 3. Summary statistics for the two-groups

	Treated		Control	
	Before	After	Before	After
Users	1,794	1,794	858	858
Songs purchased	215,749	326,640	106,431	97,553
Artists with at least one purchase	24,368	34,411	14,785	13,768

This fact raises the question of whether the volume alone is responsible for creating more edges and hence more similarity. After all, the more consumers purchase, the more likely they are to share *some* artist in common. We thus want to decompose the recommender's effects into *taste* and *volume* components. The taste component is the portion of *D* due to changes in the assortment of artists users buy, with volume held equal. The volume component is the portion of *D* due to a change in purchase volume, irrespective of a change in taste. Figure 4 illustrates this, showing that recommenders can change user similarity in one or two ways. Both are valid ways for recommenders to affect similarity, but we wish to distinguish them to improve our understanding of the phenomenon.

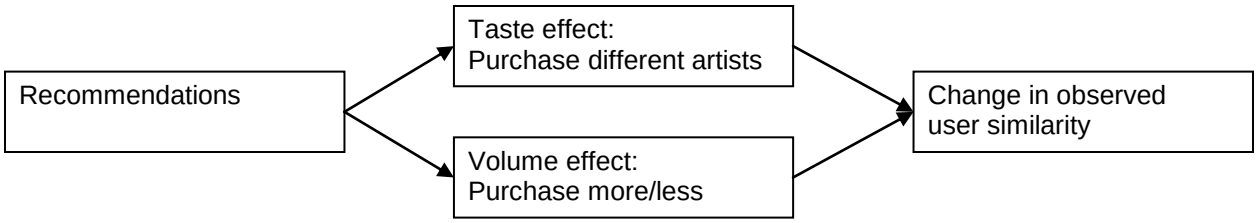

Figure 4. Changes in observed user similarity may have taste and/or volume components

We next decompose these effects. Until now, *D* was calculated on the observed data, for which volume increased after recommendations. This represented the combined taste and volume effects. Now, we equalize purchase volume before versus after but in a way that maintains the differences in the types of music users buy before versus after. Recalculating *D* on the volume-equalized data then identifies the standalone taste effect, if it is present.

To equalize the volume before versus after, we use the bootstrap (Efron & Tibshirani 1986). Instead of comparing O_{11} and O_{12}, we compare O_{11} and O^*_{12}, where O^*_{12} is sampled randomly with replacement from O_{12} and has sample size $|O_{11}|$. In other words, we are sampling for the empirical distribution of O_{12} and limiting the sample size. This procedure assumes the observations are i.i.d. over time, which is a common assumption in many statistical models of purchase data (e.g., latent-class multinomial models). For consistency, we also equalize the volume in the control group before versus after. (This is for consistency but likely unnecessary because in the control $|O_{21}| \approx |O_{22}|$ anyway.) Last, for consistency, we equalize the volume across O_{11} and O_{21}: before recommendations, we will see, the control has slightly more purchases per user than the treated group. To prevent this difference from affecting the results, we reduce $|O_{21}|$ to $|O_{11}|$ in the same manner. Thus in the volume-equalized case, we have four data sets O_{11}, O^*_{12}, O^*_{21}, and O^*_{22}, all with the number of purchases equal to $|O_{11}|$. This sampling introduces a source of variation in the results, and thus all results are averaged over repeated trials (1000 simulations).

7.1. Aggregate Analysis on the Volume Equalized Data

The aggregate analysis is repeated on the volume equalized data, and Table shows the results. The same conclusion of greater similarity after recommendations emerges. However, the magnitudes are smaller, as expected because of volume equalization. For example, the treated network's density increases from 23% to 27%. This magnitude is smaller than on the observed (unequalized) data, where it increased from 23% to 46%. Though the magnitude is smaller, it is still a significant increase compared with the control group (p = 0.03). The other measures show the same conclusions: the median degree increases, showing users have more connections to one another, and the average path length decreases, showing that users are closer to one another and the network is "smaller." In every case we reject $D = 0$ ($p \leq .05$), providing evidence of a standalone taste effect.

To summarize, when volume is held equal, purchase similarity increases after recommendations, revealing evidence of a standalone taste effect. When volume is allowed to reach its true, observed level, purchase similarity increases even more, revealing that both taste and volume effects are present.

Table 4. Summary measures for the unweighted network – *Volume-equalized data*

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	23%	27%	0.04	4%	167	213	46	43.35	1.80	1.74	-0.07	-0.06
C	12%	13%	0.00	0.03	79	82	3	<0.01	1.98	1.97	-0.01	0.05

7.2. Individual-Level Analysis on the Volume-Equalized Data

In the individual-level analysis under volume-equalization, there is again no evidence of a differential effect. Table shows these results. First, the aggregate effect toward similarity in the treated group is evident: there are more users becoming closer than there are becoming farther (24.6% weight below the diagonal versus 17.3% weight above it). This is consistent with the aggregate findings of greater similarity. The magnitude is again smaller, as expected, because volume equalization dampens the effect. Again, the control shows almost no change, with roughly equal weight below and above the diagonal. Second, the increase in similarity appears uniform: all types of users become closer to one another. Users who were close became closer, and users who were initially far became closer too. There does not appear to be evidence of a differential effect – which could have been masked by the aggregate result. This lack of differential effects exists for both the volume-equalized analysis here and the observed data analysis shown previously.

This section presented results for the unweighted network, and we focused on it because its definition is simple and intuitive. In the appendix, we present results for other network definitions, weighted and unweighted. All of the networks tested yield nearly the same conclusions.

Table 5. Path lengths between all user pairs – *Volume-equalized data*.
Entries represent the percentage of all ${}_nC_2$ user-pairs.

Treated

# Hops After \ # Hops Before	1	2	3	∞
∞	0.3	2.6	0.4	0.6
3	0.1	1.3	0.3	0.3
2	12.5	47.0	3.1	4.1
1	10.1	16.0	0.5	0.7

Becoming closer (below diagonal)	24.6
Becoming farther (above diagonal)	17.3
No change (on diagonal)	58.0

Control

# Hops After \ # Hops Before	1	2	3	∞
∞	0.4	5.6	1.8	3.6
3	0.4	5.6	1.9	2.1
2	8.1	45.7	5.9	5.4
1	3.6	8.3	0.5	0.5

Becoming closer (below diagonal)	22.8
Becoming farther (above diagonal)	21.9
No change (on diagonal)	54.8

8. SIMULTANEOUS TWO-GROUP ANALYSIS

In this section, we expand the analysis in two ways, examining changes between the groups and changes in the population as a whole.[12] Figure 5 shows this graphically. The previous analysis considered the *Treated* and *Control* regions of Figure 5. We enlarge the analysis to include the between group similarity (*Between*), which describes how close the entire treated group is to the entire control, and the overall similarity (*Overall*), which treats all users as a single population and describes the change in similarity within it.

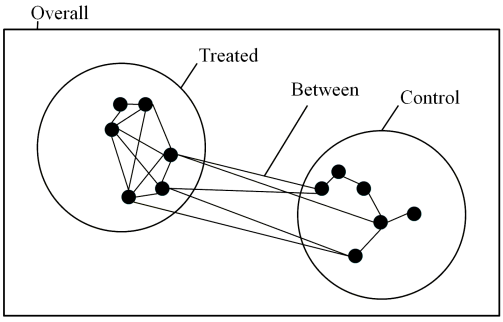

Figure 5. Edges in ***Overall*** can be partitioned into ***Treated, Contro***l, and ***Between***

Mathematically, consider a network built on the combined data $\{O_{1t}, O_{2t}\}$,which combines the treated and control groups. The set of all edges in the network, termed *Overall*, can be partitioned into three groups: *Treated*, *Control*, and *Between*. *Treated* is the set for which both nodes are in the treated group. *Control* is the set for which both nodes are in the control group. *Between* is the set for which one

[12] Population here means the treated and control groups combined, not the statistical sense of population versus sample.

node is in the treated group and the other is in the control. This section extends the analysis to *Overall* and *Between*.

The motivation is two-fold. First, although recommender systems are becoming increasingly common, it is possible that not everyone will be exposed to them at all times. In this case, the state of the world under recommendations may reflect *Overall* more than *Treated*. This situation is unlike many experiments, in which if a new method is effective we envision treating everyone. Second, the *Between* analysis tests for another type of fragmentation in which treated users become self-similar but distant from control users. For example, suppose half the population uses recommenders; if *Treated* users became more similar to each other but *Treated* and *Control* moved apart, this would be another form of fragmentation. The *Between* analysis tests for this.

Using the same network definition and same statistics *T*( ), we repeat the analysis on *Overall* and *Between*. Table presents the results. For ease of comparison, the *Treated* and *Control* results are reproduced from earlier. Examining *Between*, one sees that the treated and control groups do become closer to each other. There are more edges between the groups after recommendations than before, and the path length between users in different groups decreases. Thus the treated group has not moved away from the control; rather, they are becoming closer. Examining *Overall* shows a similar result: the items consumed by the population as a whole are becoming more similar after recommendations. The density increases after recommendations and the path length decreases. This could be expected for *Overall*, since if *Treated* and *Between* exhibit more similarity and *Control* shows little change, then *Overall* will show a weighted average of this trend.

As before, we examine whether this result is due solely to volume or has a standalone taste component. Table presents the results after equalizing the volume post-recommendations. As before, the magnitudes are dampened, but the results are the same: the treated and control groups move closer to one another and the overall population of users becomes more similar, as seen by the higher density, higher degree, and lower path length.

To summarize, in previous sections we found that treated users' purchases became more similar to one another, whereas the control showed almost no change. This section showed that the treated and control groups as a whole become closer too. This additional finding rules out another form of fragmentation in which the treated group, despite its becoming more self-similar, could have moved in entirety away from the control. Thus at several levels we observe a trend toward more similarity: within the treated group, between treated and control groups, and in the population as a whole.

Table 6. Overall analysis for the unweighted network – *Observed data*

	Density			Median Degree			Path Length		
	Before	After	D_i	Before	After	D_i	Before	After	D_i
Treated	0.23	0.46	0.23	167	401	235	1.80	1.54	-0.26
Control	0.19	0.19	0.00	134	135	1	1.85	1.83	-0.02
Between	0.21	0.30	0.09	147	235	88	1.83	1.71	-0.12
Overall	0.21	0.31	0.1	295	503	207	1.83	1.69	-0.13

All values of D_1 and D_2 are significantly different from zero ($p < 0.05$)

Table 7. Overall analysis for the unweighted network – *Volume-equalized data*

	Density			Median Degree			Path Length		
	Before	After	D_i	Before	After	D_i	Before	After	D_i
Treated	0.23	0.27	0.04	167	214	47	1.80	1.73	-0.07
Control	0.13	0.13	0.00	80	82	3	1.94	1.93	-0.01
Between	0.17	0.19	0.02	112	131	19	1.88	1.84	-0.04
Overall	0.17	0.19	0.02	230	275	45	1.87	1.83	-0.04

All values of D_1 and D_2 are significantly different from zero ($p < 0.05$)

9. SENSITIVITY TO THE LIMITATIONS OF DATA COLLECTION

In describing the data collection, we pointed out two limitations: non-randomized group assignment and uninstallation of the plugin. This section examines these limitations and why, in light of them, the above inferences can be drawn.

9.1. User Registration Decision

One limitation of the data collection is that assignment to the treated versus control group is not randomized. Registration is the user's choice, so the analysis cannot account for selection on unobservables. For example, both the registration decision as well as our observed changes in purchases similarity may be driven by changes in users preferences around the time of registration. In this case, the observed changes in purchase similarity might have occurred anyway even if users did not have access to the recommender. This cannot be ruled out, but the analysis in this section shows it is unlikely. We begin by presenting several arguments for why we believe this is unlikely. Next, we share results from a more formal investigation of selection bias along the following lines: (a) Ruling out a time trend among treated group, (b) Verifying impact of treatment on the control group, (c) Validating results for a matched sample of users.

We first note that both the treated and control users in this study are both eventual users of the recommender system. Thus, the selection issue is not as acute as is typical in many observational studies in which control users do not select the treatment. In our setting, the control users also select the treatment and do so only a few weeks later. This, by itself, ensures a significant level of similarity between the two

groups. Further, we attribute the small differences in adoption timing between the two groups to diffusion of product awareness as opposed to fundamentally different demand preferences. This is because Service was a new technology at the time of data collection and the very first iTunes plugin of its kind. Registration may thus be reasonably seen as a response to a change in supply rather than a change in consumers' own demand. And differences in registration timing among early users may similarly be viewed as arising due to spread of product awareness. This line of reasoning is the same as Waldfogel and Chen's study (2006) of how sales at unbranded retailers are affected by the introduction of information intermediaries on the web (comparison shopping engines), which were at the time a new technology that created a change in information supply.

To test this idea, Figure 6 shows the median number of songs treated users add to their libraries in the days before and after registration. The data are centered around each user's registration date. The figure shows that the change in behavior is sharp near registration and not part of a growing trend starting weeks before.

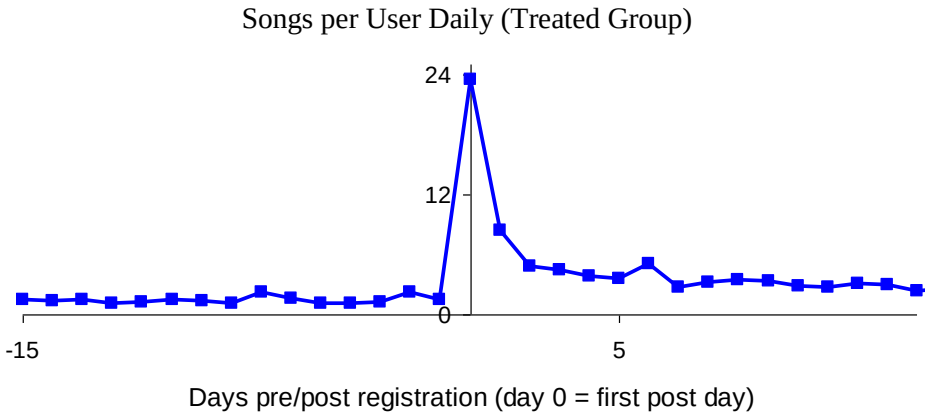

Figure 6. Daily songs added per user (average) centered on each user's registration date. Day 0 represents the time immediately after registration.

We now test the robustness of our results more formally.

Ruling out a time trend among treated users: One possibility is that the treated group had been experiencing changes in preferences in the days preceding registration and our results merely reflect these time trends rather than the impact of the recommender. Figure 6 suggests this is unlikely. We test this more formally by conducting a Difference-in-Difference (DiD) test of purchase similarity in multiple pre-treatment periods (Meyer 1995). The "before" period for this test is defined as January 2007 and the "after" period is defined as February 2007. Note that both groups had not been exposed to the

recommender system during this entire timeframe. However, if the treated users were experiencing change in preferences over time, then we expect these changes to show up in the DiD test. Table 8 shows that there are no significant changes in density, degree and path length for the treated users relative to the control users. Thus, we can rule out the possibility that our results reflect a time trend of increased purchased similarity among the treated users.

Table 8. Summary measures for pre-registration time periods

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	13%	11%	-2%	0%	66.66	53.98	-12.68	3.32	1.96	2.03	0.07	-0.02
C	11%	9%	-2%	0.96	55.00	39.00	-16.00	0.60	2.01	2.10	0.09	0.56
Volume Equalized Data:												
T	11%	11%	0%	1%	52.59	53.88	1.28	6.52	2.02	2.03	0.02	-0.03
C	8%	8%	0%	0.43	39.12	33.89	-5.24	0.25	2.10	2.15	0.05	0.50

Effect of treatment on control users: A unique aspect of our dataset is that the control users also registered for the recommender a few weeks after the treated users. If these control users do not demonstrate a similar change in purchase similarity upon registration, then it might suggest that the recommender system may not be driving the observed changes and that the treated users in our study are fundamentally different from the control users. This is similar to the analysis by Gruber (1994) in which a later federal mandate on maternity benefits (the treatment) resulted in some states that had not previously mandated such benefits (the original control states) to now be subject to the treatment. To evaluate the effect of the treatment on our original control group, we divide our control users into two groups. The first group, G1, registered for the recommender in May 2007 and the second group, G2, registered in July or August 2007. We consider March and April as the "before" period and May and June as the "After" period. Note that G1 users are exposed to recommendations in the after period whereas G2 users are unexposed throughout. Table 9 shows that we observe a significant increase in density and median degree and decrease in average shortest path length for G1 users relative to G2 users. Thus control users also experience an increase in purchase similarity when they are exposed to the recommender.

Table 9. Summary measures for the original control users

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
G1	22%	40%	18%	18%	155.05	340.85	185.80	191.30	1.82	1.61	-0.21	-0.22
G2	20%	19%	0%	0.00	141.00	135.50	-5.50	0.00	1.86	1.86	0.01	0.00
Volume Equalized Data:												
G1	22%	29%	7%	7%	155.23	230.10	74.87	70.38	1.82	1.72	-0.09	-0.09
G2	17%	18%	0%	0.00	117.59	122.08	4.50	0.00	1.90	1.89	0.00	0.00

Validating results with matched sample of users: It is possible that the changes in user preferences occurred simultaneously with the registration decision of users. If so, we will not observe such changes in user preferences in the results of Table 8. To control for this, we repeat our analysis for a matched sample of treated and control users. Propensity Score Matching (PSM) is a statistical matching technique for causal inference with observational data (Rosenbaum and Rubin 1983). PSM usually involves running a logistic regression for group membership to compute the probability of users belonging to the treated versus control group based on a set of observed predictors. Users in the treated group are then matched with users in the control group who have the same probability of treatment in order to control for confounding factors. A key weakness of PSM is that hidden biases may remain because matching cannot control for unobserved variables (Shadish et al. 2002; Pearl 2009). However, PSM works well with large samples and when a large number of pre-treatment covariates that are likely to influence group selection are available (Heckman et al. 1998; Shadish et al. 2002; Pearl 2009). As a result, PSM has been used as the primary technique for identification in several studies based on observational data (e.g., see Angrist 1998; Aral et al. 2009).

In order to create a matched sample, we first run a logistic regression for group membership against a number of behavioral covariates that specify users' taste in music. These variables include iTunes installation date, size of music library, average monthly music downloads, variance in music consumption, etc. Table 10 provides the results of a DiD test of purchase similarity on the matched sample of users. We continue to find a significant increase in density and degree (and decrease in average shortest path length) for the treated users relative to control users. The differences are significant at the 5% level. As before, the magnitude of the changes are smaller, yet significant, for the volume-equalized data. In other words, there is both a volume effect and a taste effect driving an increase in purchase similarity among treated users.

Table 10. Summary measures for matched sample of users

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	24%	46%	22%	23%	157.00	361.00	204.00	207.00	1.79	1.53	-0.26	-0.25
C	20%	20%	-1%	0.00	126.00	123.00	-3.00	0.00	1.86	1.86	0.00	0.00
Volume Equalized Data:												
T	23%	35%	11%	10%	150.20	253.77	103.57	92.70	1.80	1.65	-0.15	-0.12
C	18%	20%	1%	0.00	112.13	123.00	10.87	0.00	1.88	1.86	-0.03	0.00

9.2. Attrition

The second data limitation is attrition. About half of the users in the treated group uninstall the plugin before the data collection ends. The above analysis, as discussed, only considers those users who have Service installed for the study's duration. Although attrition is a common issue in all observational data and is not unique to our setup, we nonetheless provide a brief discussion of its impact.

The implication of this requirement is that we may overstate the magnitude of the results although not their direction. This conclusion requires the assumption that uninstallers return to pre-treatment behavior and resemble the control group. Thus to illustrate how attrition affects the magnitude, we can "average" the treated users who complete the study with control users as proxies for the drop-outs. From the previous results, we saw that the treated group's similarity increases and the control's shows almost no change, so "averaging" the results dampens the magnitude but not sign. This averaging is illustrated next.

To estimate the effect of attrition, suppose the treated group originally has n users and λn uninstall the service ($0 < \lambda < 1$). We observe the $(1 - \lambda)n$ users who remain with Service. Under the assumption that the drop-outs resemble the control, we can approximate the original treated group using all $(1 - \lambda)n$ treated users and λn control users. We refer to this group as *Composite*.

To estimate the change in similarity for *Composite*, three types of edges must be considered: edges among treated users, edges among control users (the surrogate dropouts), and edges between treated and control users (again, the surrogate dropouts). The maximum possible edges of each type is given in Table .

Table 11. Attrition sensitivity analysis: edge types and density for the *Composite* group.

Edge Type	Maximum possible edges	Density Before	Density After
Within Treated	${}_{(1-\lambda)n}C_2$	0.23	0.45
Within Control	${}_{\lambda n}C_2$	0.19	0.19
Between Treated and Control	$(1 - \lambda)n \times \lambda n$	0.21	0.30

The table also reproduces the density from the observed network in Section 8. We can thus estimate *Composite*'s change in similarity for a "typical" (average) user by taking a weighted average of the three densities using column 2 as the weights. For example, if half the users uninstall Service ($\lambda = 0.5$), then *Composite*'s density is estimated as

Composite's Density before $= ({}_{(1-\lambda)n}C_2 \times 0.23 + {}_{\lambda n}C_2 \times 0.19 + (1 - \lambda)n \times \lambda n \times 0.21) / {}_nC_2 = 0.21$

Composite's Density after $= ({}_{(1-\lambda)n}C_2 \times 0.45 + {}_{\lambda n}C_2 \times 0.19 + (1 - \lambda)n \times \lambda n \times 0.30) / {}_nC_2 = 0.31$

Composite's density increases from 0.21 to 0.31, which is positive but less than Treated's change from 0.23 to 0.45 (the magnitude is dampened). This holds for any λ and metric. The density metric was used for illustration, and results on the other metrics (e.g., path length) or other networks (e.g., volume-equalized) are similar.

10. RELATIONSHIP TO SERVICE'S RECOMMENDER SYSTEM

The results show that users' purchases appear more similar after recommendations. This section relates these findings to the recommendation system in use at Service.

As discussed, Service makes two components available to its users. The primary component is the iTunes plugin, which recommends songs based on the artist currently being played. The recommendation algorithm behind the plugin is a hybrid content and collaborative based system whose components have roughly 90% and 10% weight respectively. The second component of Service is a website in which users can browse other Service users' music purchase and play histories. With both the plugin and website, users can sample songs and purchase them if desired.

We believe similarity increases post-recommendations because Service makes users' choice sets more similar than if users were not members of the recommendation service. This appears true for both components of Service, the plugin and the website.

With the plugin, recommendations are based on the artist a user is currently listening to. When two people listen to the same artist, they receive the same list of recommendations. Because of this, users who are 1 hop away in the treated group should be more likely to remain 1 hop away than control users: having the same artist means they are more likely to see the same recommendations and thus more likely to purchase another common item. Table supports this. Treated users 1 hop away are 67% likely to remain 1 hop away afterward, whereas 1 hop away control users are only 38% likely to remain at 1 hop.[13] Seeing the same recommendations maintains the 1-hop position among treated users, whereas there is no such force maintaining the 1-hop position for control users.

Why do users not connected beforehand ($k \geq 2$) become closer? Such users do not own a common artist from which identical recommendations can be generated. Recall that Service provides a *list* of recommended artists in its plugin. When a $k \geq 2$ pair of users listens to related but different artists, their recommended lists can still include the same recommended artist. If both buy songs by this artist, the users now have a purchase in common. In this manner, treated $k \geq 2$ users should be more likely to connect than control users. As such, if this is the mechanism by which Service affects $k \geq 2$ users, we would expect this effect to be greater for $k = 2$ users than $k = 3$ and in turn $k = \infty$ users. To test this idea,

[13] The probabilities are approximated as the fraction of user pairs transitioning from k to 1 hops, and the data come from Table and Table .

one observes again in Table that Prob(1 hop away after | *k* hops away before) does show a primarily decreasing trend.

Table 12. Probability(User pair is 1 hop away after | *k* hops away before).

	Treated				Control			
Initial hops *k* =	1	2	3	∞	1	2	3	∞
Observed data	0.67	0.41	0.23	0.24	0.38	0.16	0.06	0.05
Volume equalized	0.44	0.24	0.13	0.12	0.29	0.13	0.05	0.04

At Service's website, a similar phenomenon creates co-purchases among users. When one examines another user's play history, those are songs the other user already owns. Thus any purchase of those songs creates a co-purchase. In turn, more co-purchases results in an increase in similarity on our summary measures *T*( ).

Ideally, one would vary the design of Service's components, such as the type of content-based recommender used, the type of collaborative filter used, and website layout to test how other design choices affect the results. This was unfortunately not possible. Two comments are in order. First, without variation in the components' design, one might argue that Service could design a perverse recommender to achieve any end it wanted, similarity or fragmentation. We do not believe we are observing this perverse case. Service's algorithm was designed to satisfy users and not for an explicit goal of creating or reducing fragmentation. Second, we believe Service's design is somewhat typical for the industry: a content based algorithm where songs in the same sub-genre are recommended; a collaborative algorithm where songs co-purchased are recommended; and a website where one can browse other users' profiles, as is common at many social networking sites. A large factorial design testing alternative designs for each component would certainly be desirable, and we hope future work will contribute to this.

11. CONCLUSIONS

This paper asked whether recommender systems fragment versus homogenize users. Using data from the music industry, we found that a network of users becomes more similar to one another after recommendations, as defined by purchase similarity. The trend toward similarity appeared in three ways: at the aggregate level in the treated group, at the individual level in treated group, and at the population level, which combined the treated and control groups.

At the aggregate level, we found that users' purchases appear more similar after recommendations. This finding occurred for two reasons. Users shifted their purchases toward more

similar items, the taste effect; and, users simply bought more under recommendations, the volume effect, which increased the likelihood of co-purchases with others.

At the individual level, we looked for fragmentation in terms of a differential effect – namely, close users becoming closer and far ones becoming farther – but did not find evidence of this. This helped rule out the possibility that users were fragmenting into groups. If users were splitting into groups, a form of fragmentation, we should have seen far users become farther, which we did not.

Last, at the population level, the combined network of treated and control users exhibited greater similarity afterward. Treated users became more similar to one another and more similar to the control. This ruled out the possibility that treated users might be becoming self-similar yet simultaneously separating from the control, a third way in which fragmentation could have but did not manifest itself.

These findings were observed for a variety of similarity measures and network definitions. For the setting of the music industry and our firm, it thus appears that recommender systems are associated with an increase in commonality rather than fragmentation.

The findings have policy and business implications. Each, in turn, introduces directions for future work. Regarding policy, we began with the question of whether recommender systems create fragmentation. The fragmentation outcome, should it exist, would be undesirable in the view of many thought leaders. Sunstein argued that the effects of recommender systems and personalization technologies have connections to democracy itself: "it is highly desirable for a democracy to contain a kind of 'social architecture' that offers both shared experiences and unanticipated exposures," (p. 206) and there is concern that recommenders could weaken this if they show people only what they already know. On balance, the internet allows people to access more sources of information than ever. However, to the extent technology design choices undermine the above goal, we might ask how we can build better recommender systems that offset such fragmentation effects. We did not, however, find evidence of fragmentation, despite multiple ways of trying to identify it. In the absence of such effects, there is not cause, based on this study, to modify the architecture of e-commerce or the web.

Regarding business, the study provides a window onto the ongoing trend of targeted marketing. Recommender systems lower search costs, so one interpretation of the post-recommendations data is that it better reveals preferences. Why then did we not observe consumers clustering into the hyper-specialized groups that some advertisers might expect? It is possible consumers are not yearning for narrowly targeted recommendations but looking to retain some commonality with others. If so, consumers may prefer a range of experiences that narrow targeting does not deliver. A difficulty in making this inference is separating the effects of reduced search costs from the influence of the recommender itself. The recommender lowers search costs, but it also influences users by choosing what items to show them. This influence is unavoidable and will exist for any recommender system. An interesting empirical

question is thus to separate these effects: when we observe greater commonality post recommendations, how much is due to consumers' preference for commonality versus the bias of what the recommender selects for the user. It is an important business question because evidence that consumers are seeking commonality appears to be under-emphasized in targeted marketing strategies.

A final direction for future work would be to study other recommender technologies and other application contexts. The results here are associated with the system employed at Service. Other technologies could differ; for example, a system that recommended the same product to everyone would clearly create commonality. As discussed, our aim in this paper is not to characterize the effect of hypothetical designs. Rather, we wish to characterize the impact of a major system online – which in turn shows that the current belief around fragmentation is not the case. That said, we hope future work will look at other designs and gradually catalog their effects. Similarly, the manner in which users respond to news or fashion recommendations may differ from the manner in which they respond to music recommendations. Investigating the fragmentation issue in other domains in which personalization has been deployed is important as well.

In *The Big Sort*, Bishop (2008) shows how over the last thirty years Americans have sorted themselves into like-minded groups in physical space (neighborhoods).. This paper asks a similar question about the virtual space of the web. While many predict these systems will further a trend of fragmentation, the evidence for the industry and firm studied here is to the contrary. As this is the first empirical study on the topic, we look forward to the perspective thirty more years will provide.

REFERENCES

Adomavicius, G. and A. Tuzhilin. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Enginering* 17(6):734-749.

Angrist, J. D. 1998. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica*, Vol. 66, No. 2., pp. 249-288.

Ansari A., S. Essegaier, and R. Kohli. 2000. Internet recommendation systems. *Journal of Marketing Research* 37(3):363-375.

Ansari A. and C. F. Mela. 2003. E-customization. *Journal of Marketing Research* 40(2):131-145.

Aral, S., Muchnik, L., & Sundararajan, A. 2009. Distinguishing Influence Based Contagion from Homophily Driven Diffusion in Dynamic Networks. *Proceedings of the National Academy of Sciences* 106 (51).

Arnheim, A. 1996. Summary of the proceedings of the U.C. Berkeley Collaborative

Filtering Workshop. March 16. Last accessed April 20, 2008. http://www2.sims.berkeley.edu/resources/collab/collab-report.html.

Apple. 2008. iTunes Store Top Music Retailer in the US. Company website. Last accessed November 27, 2008. http://www.apple.com/pr/library/2008/04/03itunes.html

Berger, J. and C. Heath. 2008. Who drives divergence? Identity-signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *J. Pers. Soc. Psychol.*, **95**(3) 593-607.

Bishop, B. 2008. *The Big Sort*. New York: Houghton Mifflin.

Bodapati, A. 2008. Recommendation systems with purchase data. *Journal of Marketing Research* 45(1):77-93.

Brynjolfsson, E., Y. Hu, and M. Smith. 2006. From niches to riches: the anatomy of the long tail. *Sloan Management Review* 47(4) 67-71.

Campbell, D. T. and J. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Clemons, E.K., G. G. Gao, and L.M. Hitt. 2006. When online reviews meet hyperdifferentiation. *J. of Management Information Systems* 23(2):149-171.

Conover, W. J. and R. L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* 35(3): 125-129.

Das, A., M. Datar, A. Garg, and S. Rajarm. 2007. Google news personalization: scalable online collaborative filtering. *Proc. of the 16th Int'l World Wide Web Conference*, p. 271-280.

Dellarocas, C. 2003. The digitization of word-of-mouth: promise and challenges of online reputation systems. *Management Science* 49(10):1407-1424.

Efron, B. and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1(1):54-75.

Fleder, D. and K. Hosanagar. 2007. Recommender systems and their impact on sales diversity. *Proceedings of the 8th ACM conference on Electronic Commerce*: 192-199.

Fleder, D. and K. Hosanagar. 2009. Blockbuster culture's next rise or fall: the impact of recommender systems on sales diversity. *Management Science* 55(5):697-712.

Good, P. 1994. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Gruber, J. 1994. The Incidence of Mandated Maternity Benefits. *American Economic Review* 84 (3): 622-641.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5), pp. 1017-98.

Hervas-Drane, A. 2007. Word of mouth and recommender systems: a theory of the long tail. *NET*

Institute Working Paper, No. 07-41. Available at SSRN: http://ssrn.com/abstract=1025123

Huang, Z., D. D. Zeng, and H. Chen. 2007. Analyzing consumer-product graphs: empirical findings and applications in recommender systems. *Management Science* 53(7):1146-1164.

Katz, M. and C. Shapiro. 1985. Network externalities, competition, and compatibility. *American Economic Review*, vol. 75, pp. 424-40.

Lamere, P. and S. Green. 2008. Project Aura: recommendation for the rest of us. Presentation at *Sun JavaOne Conference*. Slides last accessed 25 November 2008 at http://developers.sun.com/learning/javaoneonline/2008/pdf/TS-5841.pdf

Linden, G. 2008. People who read this article also read.... *IEEE Spectrum*. March.

Linden, G. 2011. Eli Pariser is Wrong. http://glinden.blogspot.com/2011/05/eli-pariser-is-wrong.html. Retrieved July 2011.

Manchanda P., J.-P. Dubé, K. Y. Goh, and P. K. Chintagunta. 2006. The effect of banner advertising on internet purchasing. *Journal of Marketing Research* 43(1):98-108.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

McGuire, M. and D. Slater. 2005. Consumer taste sharing is driving the online music business and democratizing culture. *Gartner Group and Harvard Law School Berkman Center for Internet & Society Report*. Report number G00131260.

Meyer B. D. 1995. Natural and Quasi Experiment in Economics. *Journal of Business and Economic Statistics*. 13(2) 151-161

Murthi, B. P. S. and S. Sarkar. 2003. The role of the management sciences in research on personalization. *Management Science* 49(10):1344-1362.

Negroponte, N. P. 1995. *Being Digital*. New York: Vintage Books.

Oestreicher-Singer, G. and A. Sundararajan. 2009. Recommendation networks and the long tail of electronic commerce. *SSRN eLibrary*, http://ssrn.com/abstract=1324064.

Papadimitriou, C. H. and K. Steiglitz. 1998. *Combinatorial Optimization: Algorithms and Complexity*. Toronto: Dover Publications.

Pariser, E. 2011. *The Filter Bubble*. Penguin.

Pearl, J. Understanding propensity scores. 2009. In *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Second Edition.

Rossi, P., R. E. McCulloch, and G. M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Science* 15(4):321-340.

Rosenbaum, P. R., D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41–55.

Shadish, W.R., T.D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shaffer, G. and Z. J. Zhang. 1995. Competitive coupon targeting. *Marketing Science* 14(4):395-416.

Smith, M., Giraud-Carrier, C. and Judkins, B. 2007. Implicit affinity networks. *Proc. of the 7th Annual Workshop on Information Technologies and Systems*. 1-6.

Sunstein, C. R. 2001. *Republic.com*. Princeton: Princeton University Press.

Terdiman, D. 2011. Why a hyper-personalized web is bad for you. CNET News. May 17.

Thompson, C. 2008. If you liked this, you're sure to love that. *The New York Times Magazine.* November 23.

Van Alstyne, M. and E. Brynjolfsson. 2005. Global village or cyber-Balkans? Modeling and measuring the integration of electronic communities. *Management Science* 51(6):851-868.

Waldfogel, J. and L. Chen. 2006. Does information undermine brand? Information intermediary use and preference for branded retailers. *J. Industrial Economics* 54(4):425-449.

Watts, D. J. and S. H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature* 393:440-44.

APPENDIX I. SIGNIFICANCE TESTING

The hypotheses tested in the aggregate analysis have the form

$$H_0: \mu \equiv E[D] = 0$$
$$H_a: \mu \equiv E[D] \neq 0$$

where $D \equiv (T_{12} - T_{11}) - (T_{22} - T_{21})$ and $\mu \equiv E[D]$. This is a statistical test of the null hypothesis that purchase records are distributed the same in the treated group and in the control group. The use of such test statistics is facilitated by permutation tests which allow us to calculate a null distribution for any test statistic. Statistical theory says that under the null hypothesis of equal distributions of purchase records (and conditional on the observed purchase records), all relabelings of the records as 'Treated' and 'Control' are equally likely. We obtain a null distribution and hence a *p*-value for *D* by repeatedly relabeling the purchase records, reconstructing the networks, recalculating *D*, and tallying the fraction of times these 'relabeled' values of *D* exceed the observed value of *D*. Enumerating all relabelings is not usually possible computationally, which is why one resorts to sampling a feasible number of relabelings that yields an approximate permutation *p*-value for *D*. Further details on the theory of permutation tests can be found in the appendix to Good (1994).

12. APPENDIX II. SENSITIVITY TO ALTERNATIVE NETWORK TYPES

The network examined in the base case is one type of network. In this section, we explore other network definitions, both unweighted and weighted. We find the conclusions of increased similarity generally hold across these other network definitions.

12.1. Alternative Network Definitions

This section defines several additional networks to test. Recall that defining the consumer network is equivalent to defining the distance, or edge weight, between all user pairs. We continue the notation from before in which $(G_{it})_{xy}$ is the edge weight between consumers x and y. As before, $(O_{it})_x$ is user x's vector of purchase counts, where the vector length is the number of artists.

The unweighted network used throughout the base case was defined

Unweighted

$$(G_{it})_{xy} \equiv \begin{cases} 1, \text{ if users } x \text{ and } y \text{ have} \geq 1 \text{ artist in common } ((O_{it})_x \bullet (O_{it})_y \geq 1) \\ \text{Unconnected, otherwise} \end{cases}$$

We generalize this definition to an arbitrary threshold of k artists in common:

Unweighted-k

$$(G_{it})_{xy} \equiv \begin{cases} 1, \text{ if users } x \text{ and } y \text{ have} \geq k \text{ artist in common } ((O_{it})_x \bullet (O_{it})_y \geq k) \\ \text{Unconnected, otherwise} \end{cases}$$

This definition allows us to test whether the findings are robust to the original choice of $k = 1$.

Weighted networks can also be defined. Perhaps the simplest starting point is to define the edge weight as the Euclidean distance between each user's vector of purchase counts.

Weighted

$$(G_{it})_{xy} \equiv \| (O_{it})_x - (O_{it})_y \|$$

We also define a weighted network in which the user vectors are first normalized to length 1. Let $(\tilde{O}_{it})_x \equiv (O_{it})_x / \|(O_{it})_x\|$. Each user is thus a point on the hyper-sphere of radius 1. The *Normalized-Weighted* network is defined by the Euclidean distance between normalized user vectors

Normalized-Weighted

$$(G_{it})_{xy} \equiv \| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|$$

In words, normalization forces distance to depend on the proportion of artists a user buys and not how much he buys. Geometrically, it amounts to comparing the angle between user vectors, regardless of the

vectors' lengths. This measure is proportional to the "cosine similarity" in the field of Information Retrieval.[14]

These networks span two characteristics that we wish to consider: sensitivity to purchases in common and sensitivity to purchase volume. Sensitivity to purchases in common, the first characteristic, captures how much users must overlap in their purchases to have a low edge weight between them. The *Unweighted*-1 network is not sensitive. Of the thousands of artists, users need overlap on only one to create an edge. In contrast, the *Normalized-Weighted* network, which amounts to measuring angles, is very sensitive. To have a low distance, it is generally not enough to have one or two purchases in common. Unless users have many artists in common, as we will see shortly, users are nearly orthogonal and hence far apart.

Sensitivity to purchase volume, the second characteristic, describes how much the total quantity a user purchases affects his distance to others, conditional on buying the same proportions of artists. An example makes this clear. Consider three users with the following purchases

User 1 buys	1 song of artist *a*	3 songs of artist *b*
User 2 buys	1 song of artist *a*	3 songs of artist *b*
User 3 buys	100 songs of artist *a*	300 songs of artist *b*

The *Normalized-Weighted* network says all three users are equidistant: volume is irrelevant, and distance is defined by the angle between user vectors (0 in this case). Users need only buy the same proportions of artists to be considered similar. In contrast, in the *Weighted* network purchase quantities are relevant, so these three users would not be equidistant. Users 1 and 2 would have distance 0 while user pairs 1-3 and 2-3 would be farther apart.[15]

12.2. Results for the Alternative Network Definitions

The results under the alternative network definitions generally yield the same conclusion of increased similarity. Because the results are so similar, we focus primarily on the points of departure.

For the *Unweighted-k* networks, we test three variants k = 1, 2, and 10 both with and without volume equalization. The results are shown in Table . In every case, users appear more similar after recommendations: density increases, the median degree increases, and path length decreases. We find evidence of both taste and volume effects for all of the unweighted networks. Table shows that the results are in the same direction on the volume-equalized data. All results are significant ($p \leq .05$) except one: in

[14] If x and y are vectors of length 1, $\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2x \bullet y = 2 - 2\ x \bullet y = 2 - 2\cos(x,y)$.

[15] Note, normalizing vectors to length one is not the same as volume equalization. Volume equalization controls for the overall change that occurs pre-post recommendations. Normalization controls for the within-period volume differences in volume across users, whether or not volume is equalized.

the *Unweighted*-10 network with volume equalization, the change in path length is not significant at conventional levels ($p = 0.32$), although the sign is consistent with previous results.

For the weighted networks, similar conclusions emerge, but there are nuances across the network definitions. These results are shown in Table and Table 1 for the observed and volume-equalized data respectively. Note, for the weighted networks density is not reported: all users are connected in the weighted network, albeit at varying distances, so the density is always one. Similarly, because all users are connected, the shortest path is no longer meaningful; the direct path is always the shortest one (by the triangle inequality). Thus instead of path length, we report the average distance between users

$$\text{Average Distance} = \frac{1}{{}_nC_2}\sum_{x=1}^{n}\sum_{y<x}(G_{it})_{xy}$$

The *Weighted* network (Table 1), which is based on Euclidean distance, exhibits greater similarity after recommendations: the median degree and average distance both decrease, indicating users are closer to one another (Note, in the weighted network, lower degree means greater similarity, in contrast to the unweighted networks.) So far, this is consistent with the previous findings. However, when volume is not equalized, users in *Weighted* are farther apart (Table). The reason can be seen by expanding the definition of Euclidean distance, on which *Weighted* is based

$$\|(O_{it})_x - (O_{it})_y\|^2 = \|(O_{it})_x\|^2 + \|(O_{it})_y\|^2 - 2(O_{it})_x \bullet (O_{it})_y$$

If purchase volume increases sufficiently (the first two terms), this can offset a trend toward commonality (the third term). Even if users purchase a more similar mix of artists, that higher quantity of purchases alone can cause the Euclidean distance to increase.[16]

The next weighted network is *Normalized-Weighted*, which normalizes each user's vector to length one and then applies Euclidean distance. On the observed data, users appear more similar: the median degree and average distance decrease ($p < .01$). On the volume-equalized data, there is little change and the differences are not significant. As before, the magnitudes fall under volume equalization, but here attenuation occurs twice: once due to volume equalization and once due to normalization.

The *Normalized-Weighted* network amounts to comparing angles between users, and now almost all user pairs are orthogonal. Figure illustrates this, showing that the distribution of average distances piles up at $\sqrt{2} \approx 1.41$. A user-to-user distance of $\sqrt{2}$ is equivalent to being orthogonal because

$$(G_{it})_{xy} \equiv \| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|$$

[16] A simple example shows this. Before, user x's vector is (1,4) and y's vector is (1,1). After, x's vector is (2,8) and y's is (2,2). The mix of artists they buy is unchanged, but using Euclidean distance the users are farther. As a more extreme case, if y's vector after were (2,4), the users are buying a more similar mix of artists after, but the Euclidean distance still increases.

$$= \sqrt{2 - 2\cos\left((\tilde{O}_{it})_x, (\tilde{O}_{it})_y\right)}$$
$$= \sqrt{2}\left(1 - (\tilde{O}_{it})_x \bullet (\tilde{O}_{it})_y\right)$$
$$\approx \sqrt{2} \Leftrightarrow (\tilde{O}_{it})_x \perp (\tilde{O}_{it})_y$$

Normalization makes each vector's element a small fraction; thus when users only overlap on a few artists, the product of these fractions is small and the dot-product is near zero.[17] (The figure shows the distribution for O_{11}, but the graph looks similar for the other groups and periods.)

We thus see that *Unweighted*-1 and *Normalized-Weighted* impose very different requirements for how much users must overlap in their purchases to be considered close. *Unweighted*-1 is a forgiving network: uses need only one artist in common to create an edge between them. *Normalized-Weighted* is the opposite: users must have many purchases in common to be considered close or else they will be near orthogonal, or $\sqrt{2}$ apart.

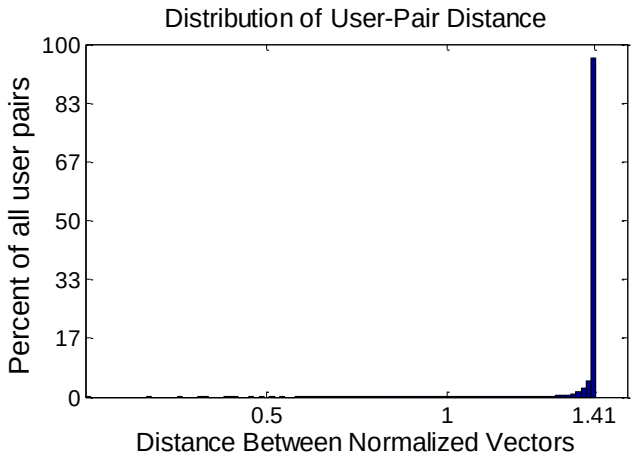

Figure 7. In the *Normalized-Weighted* network, almost every user pair is orthogonal ($\sqrt{2}$ apart).

The goal of the *Normalized-Weighted* network was to compare taste without taking into account differences in heavy versus light users. This network, though, is so strict in its definition – almost all users are orthogonal – that we introduce a more balanced measure. We test the additional network

Normalized-Weighted-Rank Transform

$$(G_{it})_{xy} \equiv \hat{F}\left(\| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|\right)$$

[17] This can also be seen by the argument, not proved here, that in high dimensions random vectors are nearly always orthogonal.

The *Normalized-Weighted-Rank Transform* network applies a rank transformation to the edge weights of the *Normalized-Weighted* network. The transformation $\hat{F}$ is the empirical CDF of the distribution of $\| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|$. This replaces $\| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|$ with its percentile rank among all user pairs. Whereas the distribution of $\| (\tilde{O}_{it})_x - (\tilde{O}_{it})_y \|$ piles up at $\sqrt{2}$, the rank transform spreads this out. The rank transformation has many applications in statistics (Conover & Iman, 1981). Here, we use it as a device to magnify differences among the user pairs that crowd at $\sqrt{2}$ .To apply this transformation, two CDFs are needed: one from the treated group and another from the control. In addition, for each group we use its period $t = 1$ CDF to transform both the $t = 1$ and $t = 2$ data. Comparisons would not be meaningful if we rescaled the data in period 2.

The results from the rank transformation yield the same conclusion that users appear more similar after recommendations. The median degree and average distance both decrease, and this holds for both the observed and volume-equalized data ($p < .01$).

These results show that the findings of greater similarity do not appear specific to our choice of network for the base case. The results hold for a variety of networks. We have analyzed these additional networks at the individual level too. The results show a similar pattern as before: a trend toward similarity, regardless of whether a user-pair's initial distance was close or far. For space reasons we omit the 24 additional tables ({treated versus control} × {before versus after} × 6 networks), but they are available on request.

Table 13. Summary measures for the unweighted networks – *Observed data*

Unweighted-1

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.23	0.46	0.22	0.23	167	402	235	234	1.80	1.54	-0.26	-0.26
C	0.19	0.19	0.00	<0.01	134	135	1	<0.01	1.86	1.86	0.00	<0.01

Unweighted-2

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.17	0.38	0.20	0.21	113	317	204	209.34	1.89	1.62	-0.26	-0.27
C	0.15	0.15	0.00	<0.01	97	92	-5	<0.01	1.94	1.95	0.00	<0.01

Unweighted-10

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.08	0.18	0.11	0.11	40	123	82	84	2.12	1.86	-0.26	-0.29
C	0.08	0.07	-0.01	<0.01	42	40	-2	<0.01	2.14	2.17	0.03	<0.01

Table 14. Summary measures for the unweighted networks – *Volume-equalized data*

Unweighted-1

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.23	0.27	0.04	0.04	167	213	46	43.35	1.80	1.74	-0.07	-0.06
C	0.12	0.13	0.00	0.03	79	82	3	<0.01	1.98	1.97	-0.01	0.05

Unweighted-2

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.17	0.23	0.06	0.06	112	176	64	61.20	1.89	1.79	-0.10	-0.09
C	0.11	0.12	0.00	<0.01	70	73	3	<0.01	2.02	2.00	-0.01	0.03

Unweighted-10

	Density				Median Degree				Path Length			
	Before	After	D_i	D/p	Before	After	D_i	D/p	Before	After	D_i	D/p
T	0.08	0.09	0.02	0.02	40	53	13	11	2.12	2.06	-0.05	-0.05
C	0.06	0.06	0.00	0.02	31	33	2	0.04	2.23	2.22	-0.01	0.32

Table 15. Summary Measures for the Weighted Networks – *Observed Data*

Weighted

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	21,472	28,250	6,778	8,854	32.42	41.65	9.23	12.20
C	26,417	24,341	-2,076	<0.01	38.90	35.93	-2.96	<0.01

Normalized-Weighted

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	1,135	1,132	-3.32	-2.79	1.41	1.40	0.00	-0.004
C	1,136	1,136	-0.53	<0.01	1.41	1.41	0.00	<0.01

Normalized-Weighted-Rank Transform

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	723	626	-97	-97.06	0.88	0.78	-0.10	-0.10
C	740	740	0	<0.01	0.90	0.90	0.00	<0.01

Table 1. Summary Measures for the Weighted Networks – *Volume-Equalized Data*

Weighted

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	21,460	20,209	-1,251	-1,695	32.46	29.16	-3.30	-3.82
C	24,536	24,980	444	0.05	36.11	36.63	0.52	<0.01

Normalized-Weighted

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	1,135	1,134	-0.99	-0.68	1.41	1.41	0.00	0.00
C	1,137	1,137	-0.31	0.16	1.41	1.41	0.00	0.91

Normalized-Weighted-Rank Transform

	Median Degree				Average Distance			
	Before	After	D_i	D/p	Before	After	D_i	D/p
T	723	695	-28	-26	0.88	0.85	-0.03	-0.03
C	766	764	-2	<0.01	0.94	0.93	0.00	<0.01