Management Science

Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation

Kartik Hosanagar, Daniel Fleder, Dokyun Lee, Andreas Buja

MANAGEMENT SCIENCE
Vol. 60, No. 4, April 2014, pp. 805–823
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1808

Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation

Kartik Hosanagar, Daniel Fleder, Dokyun Lee
Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104
{kartikh@wharton.upenn.edu, dfleder@wharton.upenn.edu, leedok@wharton.upenn.edu}

Andreas Buja
Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104,
buja@wharton.upenn.edu

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 hyperspecific to our interests, does it fragment Internet users, reducing shared experiences and narrowing media consumption? 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 that helps users widen their interests, which in turn creates commonality with others. This increase in commonality occurs for two reasons, which we term volume and product-mix effects. The volume effect is that consumers simply consume more after personalized recommendations, increasing the chance of having more items in common. The product-mix effect is that, conditional on volume, consumers buy a more similar mix of products after recommendations.

Keywords: information systems; electronic commerce; recommendation systems; collaborative filtering; filter bubble
History: Received August 4, 2011; accepted May 11, 2013, by Sandra Slaughter, information systems. Published online in *Articles in Advance* November 22, 2013.

Will the global village fracture into tribes?
—Paul Resnick (Arnheim 1996)

1. Introduction

Much of our time spent online is invisibly guided by recommendation algorithms. 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 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 and 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. As a result, personalization has also 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, Shaffer and Zhang 1995, Rossi et al. 1996), with its origins in targeted and customized marketing.

The following examples show how recommender 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. (Linden 2008, p. 46)

> Last.fm connects you with your favorite music and uses your unique taste to find new music, people, and concerts you'll like. (http://www.last.fm/)

Along with the benefits of personalization, however, a debate has emerged as to its broader impact on consumers. Personalizing websites means that we may no longer see the same newspaper articles, television shows, or books as our peers. Some 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.

Fragmentation in consumption has implications for consumers and society. For consumers, shared consumption often has an associated externality. For example, at the proverbial water cooler, people are able to discuss a shared book, artist, or news article. This is a positive externality from shared consumption (Katz and Shapiro 1985). If recommenders affect consumption similarity, these externalities are in turn impacted. Furthermore, a key promise of recommenders is that they can help consumers discover new and relevant items outside their sphere of interest. Understanding whether these systems aid such discovery helps us better understand whether recommenders are delivering on that promise. From a societal perspective, the literature has expressed concern that fragmentation is a negative consequence for society because social discourse suffers when people have a narrow information base with little in common with one another (Sunstein 2007). This could occur with many forms of personalization, but is most relevant for personalized news. 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. This paper presents the first empirical evidence for the debate on whether recommenders fragment versus homogenize users.[1]

We find, in an empirical study of a music industry recommendation service, that recommendations are associated with an increase in commonality. This increase in purchase similarity occurs for two reasons, which we term volume and product-mix effects. The volume effect is that consumers simply purchase more after recommendations, increasing the chance of having purchases in common with others. The product-mix 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 interuser distances. We find that this increase in commonality occurs because the system helps users explore and discover new items. These findings suggest that, for this setting, concerns of fragmentation may be misplaced.

We note that our results are derived for one recommendation technology deployed in one setting, and it is unclear whether the same results will arise for books, news, and other settings. Nonetheless, the results demonstrate that a common criticism that these systems cause fragmentation is not universally true and that commonly used designs can in fact increase commonality in consumption.

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...." Content-based recommenders do well when there is rich information about product attributes but cannot recommend across product categories or genres. Collaborative filters can recommend across product or genres and do not require knowledge of product attributes, but need sufficient purchase/ratings data for users and items in order to recommend. Hybrid systems combine the best of the two approaches. The design of these systems has been an active research area for almost 20 years. An extensive review in the information systems literature is provided in Adomavicius and Tuzhilin (2005).

Although a large body of work exists on designing recommenders systems, we know less about how they affect the market and society. This is despite

[1] In trying to understand the impact of recommenders on fragmentation, 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.

the large body of work on recommender algorithms and millions of transactions occurring through them. This paper continues a small stream of work in that direction. Recommenders can have a positive effect on sales and Web impressions (Ansari et al. 2000, Das et al. 2007, Bodapati 2008, De et al. 2010). For example, De et al. (2010) show that recommenders positively affect sales, and they expect this to be particularly important in industries with many stockkeeping units. The question of how recommenders affect *products* has recently been studied—which products gain versus lose sales due to recommenders and whether recommenders increase the market for niche goods, or "long tail" (Fleder and Hosanagar 2009, Oestreicher-Singer and Sundararajan 2012, Hervas-Drane 2013). It was commonly assumed that recommenders increase the long tail, and we now know this is not always true (Fleder and Hosanagar 2009). 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. He argues that although "some of the recommendations from Amazon.com and analogous services are miraculously good, ... it might well be disturbing if the consequence is to encourage people to narrow their horizons, or to cater to their existing tastes rather than to form new ones" (Sunstein 2007, p. 26). These fragmentation effects, he argues, can also have undesirable societal consequences in which people live in echo chambers and cannot relate to the views of others. Pariser (2011) similarly argues that online personalization, which includes recommenders, creates a filter bubble—an invisible, personal universe of information—due to which the world each user sees online may be very different (Terdiman 2011). "In an age when shared information is the bedrock of shared experience, the filter bubble is a centrifugal force pulling us apart," he argues (Pariser 2011, p. 9). Pattie Maes, who created one of the first recommender systems, also believed some recommenders could have a narrow-minded and hyperpersonalized aspect (Thompson 2008).

Sunstein (2007), Pariser (2011), and Thompson (2008) all believe that recommenders can fragment users. But they differ on why fragmentation is undesirable. Sunstein (2007) argues that society benefits when people have a range of viewpoints. Pariser (2011, p. 222) has a different critique: fragmentation would cause poor decision making because "the filter bubble confines us to our information neighborhood, unable to see or explore the rest of the enormous world of possibilities." Maes offers a third critique of fragmentation: lost externalities. A popular product can have the positive externality of allowing people to discuss it together. "You don't want to see a movie just because you think it's going to be good," Maes says. "It's also because everyone at school or work is going to be talking about it, and you want to be able to talk about it, too" (Thompson 2008). Under fragmentation, this benefit would disappear.

We agree with these critics that excessive fragmentation could be undesirable. However we believe that the antecedent, that recommenders create fragmentation, is ultimately an assumption. This paper tests that assumption. If recommenders do not create fragmentation, the proliferation of this view could be harmful to the adoption of otherwise valuable technologies. Part of the promise of recommenders is that they can help us discover new items outside our comfort zone and thereby expand our horizons. We believe that their view of the role of recommenders is rather narrow and a more moderate view is appropriate.

However, another camp offers a mixed view (Fleder et al. 2010). Some of the earliest researchers of recommender systems conceptualized both outcomes. Paul Resnick asked if the global village would fracture into tribes; John Riedl asked if recommenders would democratize information or result in social fragmentation (Arnheim 1996). Negroponte (1995) imagined that recommenders could cause both fragmentation or commonality, when discussing personalized news and "The Daily Me" versus "The Daily Us." Van Alstyne and Brynjolfsson (2005) ask whether Internet technologies will lead to fragmentation versus homogenization—in their terms, a cyber-Balkans versus a global village. They show that both outcomes can result depending on the value of a parameter representing consumers' preference for specialization.

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

Although many authors have discussed the fragmentation question qualitatively, the empirical question has not been posed in concrete terms. This section defines the problem formally.

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 (2005) analysis of how the Internet affects knowledge overlap among users and is consistent with notions of fragmentation as suggested by Sunstein (2007), Thompson (2008), and Pariser (2011) ("shared experiences"). We do not focus on underlying preferences of users and whether recommenders cause them to converge or diverge. Instead, consistent with the main thrust of the fragmentation debate, we focus exclusively on overlap in the items consumed. Changes in this overlap may arise due to changes in preferences or changes in users' product awareness due to recommendations.

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 from or closer to one another?

2. Disaggregate level: Are there differential effects at the individual level, by which some users become closer and others farther apart?

The first question measures the overall effect of whether users become farther apart 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 apart.

A note on terminology: The meaning of "close" and "far" will be quantified in the next section. Qualitatively, throughout this 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 set of tuples. For our music data, a tuple is of the form {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.

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 and Stanley 1963).

Figure 1 Schematic of the Two-Group Design

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

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:

$$\begin{aligned} \text{Difference in treated:}\quad & D_1 \equiv T_{12} - T_{11};\\ \text{Difference in control:}\quad & D_2 \equiv T_{22} - T_{21};\\ \text{Difference in differences:}\quad & D \equiv D_1 - D_2. \end{aligned}$$

Here, D_1 describes changes in the treated group, and D_2 describes changes in the control. The difference-in-differences 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

$$\begin{aligned} \text{H}_0\text{:}\quad & \mu \equiv E[D] = 0;\\ \text{H}_a\text{:}\quad & \mu \equiv E[D] \neq 0. \end{aligned}$$

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., median degree, average distance) 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 purchases. Huang et al. (2007) and Smith et al. (2007) use copurchase 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 $(O_{it})_{xa}$ is the number of songs user x purchased of artist a.[2] 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}.[3] 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.

Our main network uses a weighted network construction. Within a given group and time period, users x and y have an edge between them with weight given by the cosine distance.

The cosine distance network is as follows:

$$(G_{it})_{xy} \equiv 1 - \mathrm{Cos}((O_{it})_x, (O_{it})_y) = 1 - \frac{(O_{it})_x \bullet (O_{it})_y}{\|(O_{it})_x\| \|(O_{it})_y\|}.$$

The $\bullet$ symbol is the vector dot product. The cosine between two users' vectors is a measure of the angle between them. It reflects how similar these users are in the space of artists. Thus, 1 minus the cosine is how different they are. This measure is perhaps the most common similarity metric used for analyzing purchase data and the design of recommender systems. Using it in our context, it gives rise to a weighted, undirected network among users.

There are many other ways to construct the network. In weighted networks like the one above, there can be other way to define the edge weights such as Euclidean distance. A simple unweighted network definition is also possible where users x and y have an edge between them if they purchase at least one artist in common. In the main sections of this paper, we focus our base case on the cosine-based 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, $T(G(O_{it}))$. The statistic T summarizes in one number a particular network property and thus facilitates comparisons of the network over time. We define two such measures below. As notation, let $d_x \sum_{y=1,\, y\neq x}^{n} (G_{it})_{xy}$ be the degree of user x, where n is the number of users in the network. Furthermore, let ${}_nC_2$ denote the number of user pairs that can be formed from a set of n users (${}_nC_2 = n(n-1)/2$). Our statistics T are as follows:

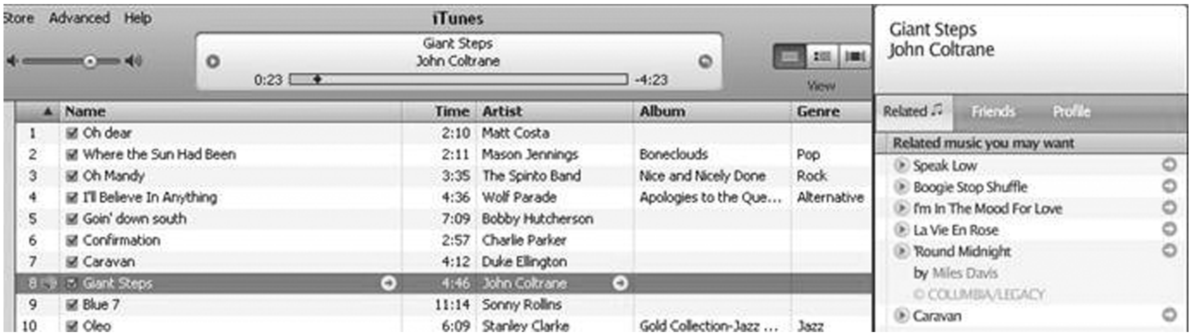
Measure	$T(G(O_{it})) =$
Median degree	$\mathrm{Median}\{d_x\}_{x=1}^{n}$
Average distance	$\frac{1}{{}_nC_2} \sum_{x=1}^{n} \sum_{y<x} (G_{it})_{xy}$

Median degree. For a weighted network, the median degree is the sum of distance (edge weights) to

[2] 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 §5.1.

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

Figure 2 Screenshot of the Recommendation Service



▲	Name	Time	Artist	Album	Genre
1	☑ Oh dear	2:10	Matt Costa		
2	☑ Where the Sun Had Been	2:11	Mason Jennings	Boneclouds	Pop
3	☑ Oh Mandy	3:35	The Spinto Band	Nice and Nicely Done	Rock
4	☑ I'll Believe In Anything	4:36	Wolf Parade	Apologies to the Que...	Alternative
5	☑ Goin' down south	7:09	Bobby Hutcherson		
6	☑ Confirmation	2:57	Charlie Parker		
7	☑ Caravan	4:12	Duke Ellington		
8	☑ Giant Steps	4:46	John Coltrane		
9	☑ Blue 7	11:14	Sonny Rollins		
10	☑ Oleo	6:09	Stanley Clarke	Gold Collection-Jazz ...	Jazz

Related music you may want

- Speak Low
- Boogie Stop Shuffle
- I'm In The Mood For Love
- La Vie En Rose
- 'Round Midnight
  by Miles Davis
  © COLUMBIA/LEGACY
- Caravan

other users that the typical (median) user has (Newman 2004).

Average distance. The average distance simply averages the pairwise distance of the users where the distances are measured by cosine distance metric.

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. 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 United States (Apple 2008). Service personalizes the user experience as follows. 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 libraries. These play histories are uploaded automatically by the plug-in to Service's website on a continual basis. Together, these two features comprise the personalization technology.[4]

Figure 2 shows a screenshot of the plug-in. Apple's iTunes appears at the left. The plug-in, as appended to iTunes, is at the right and displays a list of recommended songs. The song suggestions by the plug-in 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 six most similar artists and populates the window with this list. Artist-to-artist similarity is defined by a hybrid of content and collaborative data, though the results are heavily weighted toward the content portion (90% versus 10%). Thus in the taxonomy of recommender systems by Adomavicius and Tuzhilin (2005), the plug-in is for the most part a content-based, item-to-item-based system.

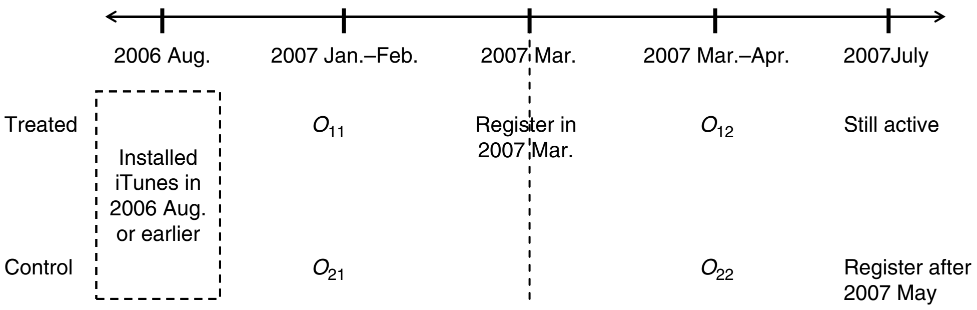
5.2. Novelty of the Data

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 time stamps 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 postregistration purchases are also observed by Service because the plug-in 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 plug-in 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

[4] It is common for most online firms to use multicomponent 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 multicomponent system.

Figure 3 Data Setup and Analysis Design



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 plug-in that is installed on each user's machine. The plug-in relays to Service in near real time the time stamp 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 and July 2007. We define the treated group as those users who registered sometime during March 2007. March was 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 as 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]

The second criterion for inclusion is active user status. Some users uninstall the plug-in 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 plug-in 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 effect. 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. We defer a detailed discussion of this to §9. However, we address the selection issue by conducting our main analysis for a matched sample of treated and control users using propensity score matching (PSM). PSM is a statistical matching technique for causal inference with observational data (Rosenbaum and Rubin 1983). Users in the treated group are matched with users in the control group who have the same probability of treatment[8] to adjust for confounding factors.

[5] That some users registered in late March could dampen the results' magnitude because it allows some Service-unaffected data to enter the postrecommendation 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 trade-off of a possible dampening of results 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.

[7] Uninstallation is not observed, so we proxy this by including those users' whose plug-in communicates with the Service at least once after the postrecommendation period.

[8] We match each control user to one treated user without replacement. Matching without replacement is essential to ensure that no user appears twice in the resulting data set. This is important

Table 1 Standardized Difference of the Covariates

	Standardized difference		*t*-test *p*-value	
Covariates	Before	After	Before	After
iTunes Installation Date	−0.14	0.01	<0.01	0.86
Library Size	0.23	0.07	<0.01	0.18
Downloads within 30 Days	0.23	0.09	<0.01	0.07
Avg Downloads per Month	0.10	0.09	0.01	0.05
Avg Monthly Change in Downloads	−0.09	−0.04	0.03	0.39
(*Library Size*)^2	0.07	0.04	<0.01	0.25
(*Downloads within 30 Days*)^2	0.15	0.07	<0.01	0.17

A 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 pretreatment covariates that are likely to influence group selection are available (Heckman et al. 1998, Shadish et al. 2002, Pearl 2009). We use the one-to-one caliper matching algorithm with rank-based Mahalanobis distance and utilize the Matchlt package by Ho et al. (2007).

To create a matched sample, we first run a logistic regression for group membership against a number of preexperiment behavioral covariates that specify users' music download behavior. These variables include *iTunes Installation Date*, *Library Size* (size of music library before start of experiment), *Downloads within* 30 *days of iTunes Installation*, *Average Monthly Downloads* and *Average Change in Monthly Downloads*. We find imbalances in many covariates before matching and particularly for covariates *Library Size* and *Downloads within* 30 *Days of Installation*. After matching, users from the two groups are no longer significantly different on any of these dimensions. Table 1 shows the standardized difference of covariates before and after the matching in which all unacceptable imbalance (absolute value greater than 0.1) have all been rendered acceptable (below 0.1 according to Austin et al. 2007). For the sake of completeness, we also report p-values for a difference in means t-test. All subsequent analysis is for the matched sample of users. We have also confirmed that our results are qualitatively similar for the original unmatched sample. Those results are available upon request.

A second limitation involves users who uninstall the plug-in. Several users in the treated group uninstall the plug-in 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 plug-in to free up disk space or not liking the extra screen space occupied by the plug-in—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 nonattriting population may overstate the magnitude of the results, but it will not change the direction of the results. This idea is discussed and bounded in the section on sensitivity analysis.

The resulting data set after matching and applying the above inclusion criteria consists of 858 users each in the treated and control groups. Treated users purchased a total of 97,226 songs from 31,395 artists in the before period, whereas control users purchased 106,431 songs from 32,163 artists in the before period.

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: in median degree and average distance.

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, $D_1 = T_{12} - T_{11}$ and $D_2 = T_{22} - T_{21}$, and the difference-in-differences estimator, $D = D_2 - D_1$.

Table 2 shows the results. Across the columns are the two statistics: median degree and average distance. 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. The column "D" lists the difference-in-differences estimate D, and the last column, "p," lists the p-value from a test that $D = 0$. To test the hypothesis that $D = 0$, we use the nonparametric method of permutation tests with 1,000 iterations to draw the two groups. We describe the test in detail in the appendix.

The results show that under both measures, users' purchases are more similar to one another after recommendations. First, median degree in the treated group decreases by −5.05 (i.e., the median treated user is closer to other users in the network) whereas the control only decreases by −0.58. The difference-in-differences D is −4.47, and it is significant ($p < 0.001$). The effect is sizeable. The standard deviation of the median degree for both the treated and control groups is around 0.28, and thus the D represents a change that is almost 16 times the standard

because our test statistic measures overlap in consumption across users. However, the order in which samples are matched introduces a source of variation. To ensure that our results are robust to the order in which matches are made, we also tried many different runs in which we shuffle the order of treated and control users in our sample. We also tried runs with different calipers. The results were qualitatively similar to the ones reported here.

Table 2 Summary Measures for the Observed Data

	Median degree					Average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	850.29	845.24	−5.05	−4.47	<0.001	0.9903	0.9842	−0.0061	−0.0052	<0.001
C	851.49	850.91	−0.58			0.9919	0.9910	−0.0009		

deviation. Similarly, the average pairwise distance in treated decreases more than the average distance in the control, giving $D = -0.0052$. The result is statistically significant ($p < 0.001$). It again represents a sizeable effect: the change is over 14 times the standard deviation of the average distance.[9] Users who see recommendations become closer, whereas control users do not.

6.2. Disaggregate Cluster Analysis on the Observed Data

The above analysis showed that users become closer after recommendations. This section asks if there are differential effects: could close users become closer but far ones farther in such a way that the aggregate result masks this? For example, users consuming jazz music may get more jazz recommendations and thus move closer to each other, but fragment away from users who consume more classical music. If true, even though the network is more similar in aggregate, far users becoming farther apart would be another type 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. We turn to cluster analysis to assess this.

We cluster users based on their artist-purchase vectors so that users with similar consumption are grouped together. We use the k-means clustering method with Hartigan's (1975) criterion and average silhouette width to determine the optimal number of clusters. In general, there is no "true" solution for clustering, and these are common approaches to determine number of clusters. Both Hartigan's criterion and average silhouette width indicated that the optimal number of clusters is 5.

By comparing the within-cluster and between-cluster user distance before and after, we assess whether it is close users (those in the same cluster), far users (those in different clusters), or both close and far users that become closer after recommendations. For users in each treated before (O_{11}) and control before (O_{21}) group, we run k-means clustering to obtain the cluster memberships. Then we ask what happens to the distance between users who were originally in the same cluster and between users who were originally in different clusters. We report the average distance between users, which was also used in the aggregate analysis, except now the measure has been separated into within-cluster and between-cluster distances. The median degree depends on the number of users in the network, which was comparable for the treated and control groups in the aggregate analysis because both groups had 858 users. After clustering, the two groups have different numbers of users within any cluster, and thus median degrees are no longer comparable. An alternative is to use median distance, which provides qualitatively similar results.

Table 3 presents these results. Both the within and between difference-in-differences estimates D are negative and statistically significant. Relative to the control, treated users within and across clusters are becoming similar in the music they consume. There is no evidence to suggest that within-cluster users are becoming similar while the clusters themselves are separating.

Note that the within average distance for the treated group increases from before to after. But this is not a differential effect. We expect some chance fluctuation: users who were by chance closer revert to being farther apart, and users who were by chance farther apart revert to being closer. This is seen in the control group, where within average distance increases considerably. Many users who were by chance close to each other regressed to being farther apart. In summary, cluster analysis supports the result from aggregate analysis and further shows that both close and far users come closer due to the recommendation system indicating the absence of differential effects.

7. Volume Equalization

The results so far show that similarity increases under recommendations. We next explore the mechanism by

[9] The standard deviations are low because the cosine measure, which is based on angles between the purchase vectors, is not a sensitive measure. Given that there are thousands of artists in our data set, most vectors are nearly orthogonal initially, and most distances are above 0.95. Furthermore, an increase in commonality by, say, 10–15 artists does not change the angle between two vectors by much. Nonetheless, the observed differences are both economically and statistically significant as highlighted above. The unweighted network presented in §9.1 offers more sensitive measures, and we observe that the standard deviations and the magnitude of changes are higher for that network.

Table 3 Summary Measures for the Cluster Analysis on Observed Data

	Within average distance					Between average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	0.9785	0.9809	0.0024	−0.0056	0.002	0.9937	0.9851	−0.0086	−0.0053	<0.001
C	0.9793	0.9875	0.0081			0.9951	0.9918	−0.0033		

Table 4 Summary Statistics for the Two Groups

	Treated		Control	
	Before	After	Before	After
Users	858	858	858	858
Songs purchased	97,226	173,088	106,431	97,553
Artists with at least one purchase	31,395	56,211	32,163	28,812

Figure 5 Illustrating Bootstrap with Three Users and Four Artists

$$\underset{O_{11}}{\begin{pmatrix} 02 & 02 & 06 & 07 \\ 00 & 04 & 01 & 04 \\ 00 & 01 & 01 & 02 \end{pmatrix}} \quad \underset{O_{12}}{\begin{pmatrix} 05 & 05 & 11 & 14 \\ 01 & 05 & 02 & 07 \\ 00 & 02 & 01 & 02 \end{pmatrix}} \quad \underset{O^*_{12}}{\begin{pmatrix} 02 & 04 & 09 & 06 \\ 00 & 02 & 00 & 03 \\ 00 & 01 & 01 & 02 \end{pmatrix}}$$

decomposing the result into volume and product-mix effects.

Table 4 suggests why this decomposition is needed. As the table shows, the recommender system appears to be working: users consume more after recommendations, whereas control users do not. (In fact, the number of songs purchased in the control group decreases.) The number of artists with at least one song 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.

These facts raise the question of whether the volume alone is responsible for creating more similarity. After all, the more consumers purchase, the more likely they are to share *some* artist in common. We thus decompose the recommender's effects into *volume* and *product-mix* components. The volume component is the portion of D due to a change in purchase volume. The product-mix component is the portion of D due to changes in the assortment of artists users buy, with volume held equal. Figure 4 illustrates this, showing that recommenders can change user similarity in two ways. Both are valid ways for recommenders to affect similarity, but we wish to distinguish them to understand the mechanism behind the overall result.

Until now, D was calculated on the observed data, for which volume increased after recommendations. This represented the combined product-mix 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 product-mix effect, if it is present. To equalize the volume before versus after, we use the bootstrap (Efron and Tibshirani 1986). Instead of comparing the original purchases O_{11} and O_{12}, we compare O_{11} and O^*_{12}, where O^*_{12} is sampled randomly 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 to be the same as the before period. This procedure assumes the observations are independent and identically distributed over time, which is a common assumption in many statistical models of purchase data (e.g., latent-class multinomial models).

Figure 5 illustrates the bootstrap procedure for a case with three users and four artists. Columns represent artists and rows represent purchases by users. For example, in O_{11}, user 1 purchases two songs by artists 1 and 2, six songs by artist 3 and seven songs by artist 4. The number of purchases is $|O_{11}| = 30$ and $|O_{12}| = 55$. Thirty purchases are randomly sampled from O_{12} to generate O^*_{12}, which has $|O^*_{12}| = 30$. By doing so, the total purchase volume is held constant across the before and after periods. 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}, reducing $|O_{21}|$ to $|O_{11}|$ in the same manner. Thus in the volume-equalized case, we

Figure 4 Changes in Observed User Similarity May Have Product-Mix and/or Volume Components

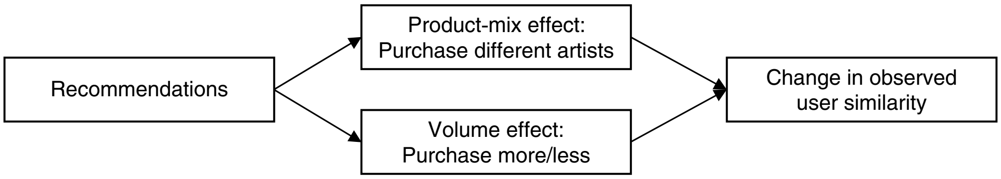


Table 5 Summary Measures for the Volume-Equalized Data

	Median degree					Average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	850.29	847.09	−3.20	−2.48	<0.001	0.9902	0.9867	−0.0034	−0.0023	0.002
C	851.63	850.91	−0.72			0.9921	0.9909	−0.0011		

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 (1,000 simulations).

7.1. Aggregate Analysis on the Volume-Equalized Data

The aggregate analysis is repeated on the volume-equalized data, and Table 5 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 difference in difference of the median degree is −2.48. This magnitude is smaller than on the observed (unequalized) data, which is −4.47. In short, product-mix effect accounts for roughly 55% of the difference in difference in median degree. Though the magnitude of the treatment effect is smaller under volume equalization, it is still significant ($p < 0.001$). The average distance shows the same conclusion. The difference-in-differences measure D is negative at −0.0023, showing treated users are closer to one another than the control users and the network is "smaller." Here again, we reject $D = 0$ ($p < 0.01$), providing evidence of a standalone product-mix effect. Comparing Tables 2 and 5, product-mix effect accounts for a little over 44% of the difference in difference in average distance. To summarize, when volume is held equal, purchase similarity increases after recommendations, revealing evidence of a standalone product-mix effect.

7.2. Disaggregate Cluster Analysis on the Volume-Equalized Data

In the cluster analysis under volume equalization, there is again no evidence of fragmentation. Table 6 shows these results. First, within-cluster average distance reduces as before ($D = -0.0002$), but is not significant, suggesting that after volume has been equalized, the recommender system does not have a significant effect on close users. However, far users (between-clusters) become closer with $D = -0.0029$ at p-value < 0.001. Combined with the previous results, we conclude that the recommender clearly has an effect of bringing all users together. Closer users are brought together by volume effect of the recommender, whereas far users are brought together by both volume effect and product-mix effect. Contrary to theories that recommenders keep far users far, these are the people whose similarity increases the most.

8. Product Discovery and Exploration

What accounts for this increase in commonality? Is the recommender simply suggesting the same item to everyone? Or is it exposing users to new types of content. The former, although demonstrating greater overall commonality, would nonetheless signal narrow media consumption and would be consistent with criticisms that these systems confine us to our information neighborhoods. In contrast, the latter would suggest that recommenders expand our horizons, help us explore and discover new types of content, and connect with other people. This, in our opinion, is the true promise of recommenders.

We examine this issue in two ways. First, we compare the changes in Gini coefficient for the two groups. The Gini is a common measure of purchase diversity, with a Gini of 0 indicating that all products have equal sales and a Gini of 1 indicating that one item generates all purchases. Higher values of Gini indicate limited diversity in the products consumed by users (for details on how to compute the coefficient, see Fleder and Hosanagar 2009). Table 7 lists the changes in the Gini coefficient for the two groups. The results indicate that there is a significant increase in purchase diversity after recommender use by treated users. One drawback of the Gini coefficient is that it is an aggregate measure of diversity and does not shed enough light on diversity at the individual level. Hence, we also compare the change in the number of unique artists

Table 6 Summary Measures for the Cluster Analysis on Volume-Equalized Data

	Within average distance					Between average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	0.9760	0.9832	0.0072	−0.0002	0.898	0.9939	0.9877	−0.0061	−0.0029	<0.001
C	0.9792	0.9866	0.0074			0.9952	0.9920	−0.0032		

Table 7 Summary Measures for Product Exploration and Discovery

	Gini coefficient					Number of unique artists				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	0.906	0.877	−0.03	−0.03	<0.001	9,709	13,741	4,032	4,679	<0.001
C	0.905	0.908	0.00			9,928	9,281	−647		
					Volume-equalized data:					
T	0.897	0.879	−0.02	−0.02	<0.001	9,458	10,488	1,030	1,065	0.02
C	0.897	0.898	0.00			9,316	9,281	−35		

consumed per user in the two groups. A t-test for paired observations reveals that treated users experience a significantly greater increase in artists consumed relative to the control group ($p < 0.001$). We also conduct a difference-in-differences test based on the average number of unique artists consumed by treated users versus control users (as before, the null distribution for the test is computed using a permutation test). The results are in Table 7. There is a significant increase in the number of artists consumed by treated users relative to control users, which suggests that the recommender helps them explore more artists. This is true even under volume equalization. In short, the recommender in our setting is increasing commonality in consumption, not by recommending the same items to many users, but by increasing the diversity of items consumed by the users.

9. Robustness Checks

Below, we discuss the robustness of our results to other network definitions. We also examine the impact of two data limitations: nonrandomized group assignment and uninstallation of the plug-in.

9.1. Results for Other Network Definition

The analysis thus far used a weighted network of users based on the cosine distance. In this section, we test sensitivity to other network definitions. This section presents one example, an unweighted network. In the online appendix (http://opim.wharton.upenn.edu/~kartikh/filteroa.pdf), we present results for other network definitions, weighted and unweighted. All of the networks tested yield nearly the same conclusions. We present the aggregate analysis on both observed and volume-equalized data using the unweighted network.

Our unweighted 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:

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

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.

In an unweighted network, the summary statistics are slightly different. With the network $G(O_{it})$ defined, we next define summary statistics of the network's properties, $T(G(O_{it}))$:

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. Unlike the definition from the weighted network, increase in median degree means higher similarity among users.

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 and 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 the Floyd–Warshall algorithm (Papadimitriou and Steiglitz 1998).

We present the aggregate analysis equivalent to that in §6.1 for observed data and in 7.1 for volume-equalized data. Table 8 shows the results for observed data using unweighted network. The results show

Table 8 **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	24	46	22	23	<0.01	157	361	204	207	< 0.01	1.79	1.53	−0.26	−0.25	<0.01
C	20	20	−1			126	123	−3			1.86	1.86	0.00		

Table 9 **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	24	35	0.11	10	<0.01	150	256	106	95	< 0.01	1.80	1.65	−0.15	−0.12	<0.01
C	19	20	0.01			112	123	11			1.88	1.86	−0.03		

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, 24% of the edges are filled in, and after, 46% are present, yielding $D_1 = 22\%$. This is a large increase in density. Over the same period, the control has no noticeable change, and $D_2 \approx -1\%$. The difference-in-differences estimate is $D = 23\% > 0$, indicating that the treated network does become more dense relative to the control. This difference is significant, as the hypothesis $D = 0$ is rejected ($p < 0.01$). 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 ($p < 0.01$).[10]

Table 9 shows the results for the volume-equalized data. The same conclusion of greater similarity after recommendations emerges. Again, the magnitudes are smaller, as expected because of volume equalization. For example, the difference in differences, D, in network density decreases from 23% to 10%, which implies that product-mix effect accounts for nearly 44% of the original D, and the volume effect accounts for the remaining 56%. 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 $D = 0$ ($p < 0.01$), providing evidence of a standalone product-mix effect. Product-mix effect explains roughly 47% of the difference in difference in median degree and 48% of that in path length.

Clustering analysis on the unweighted network yields qualitatively the same results as that for the weighted network. We do not replicate the results below, and they are available in the online appendix. Instead, we share results from complementary analysis that help illustrate the impact of the treatment on near versus far users for both the groups. Figure 6 presents the path length between all ${}_nC_2$ user pairs in the unweighted network. 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 figure are the percent of user pairs falling in each cell. For example, 8.01% of the treated users were one hop away before recommendations and two hops away after recommendations. User pairs becoming farther lie above the diagonal, whereas user pairs becoming closer lie below it. A distance of infinity (∞) means there is no path between the given two users.[11]

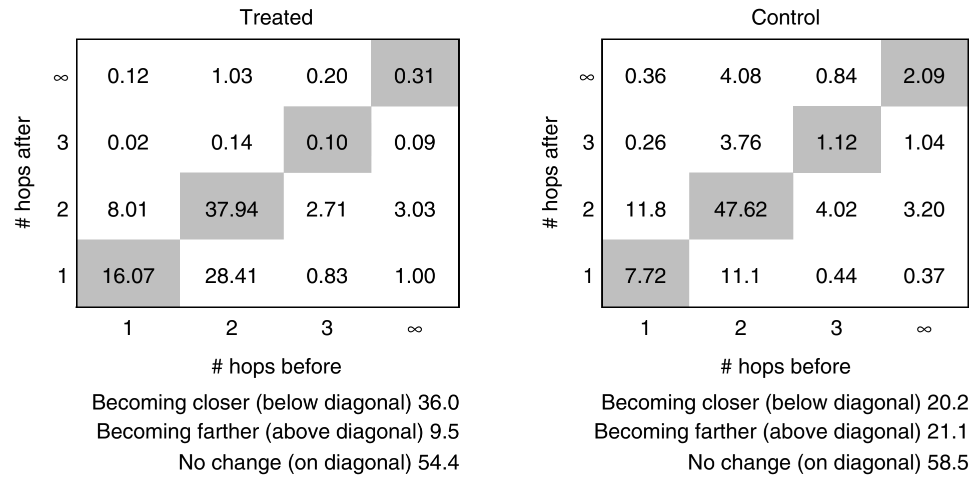
The control group appears stable (right side), because it has roughly equal weight above and below the diagonal. In contrast, the treated group (left side) has more user pairs becoming closer (36.07% weight below the diagonal) than becoming farther apart (9.51% weight above the diagonal). This is consistent with the 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.

As before, some treated users do grow farther apart, but this is not a differential effect. In the control group, 11.8% went from one to two hops, whereas 11.1% went from two hops to one hop. This level of mixing is roughly equal. In the treated group, some pairs do become farther apart (8.01% go from one to two hops) but many more become closer (as 28.41%

[10] All of 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 because of changes in the size of the main component.

[11] In Figure 6, a very small number of pairs are four or five hops away. This number is so small (< 0.04%) that for clarity we omit them from the presentation (but not the analysis) to avoid rows and columns of nearly all zeros.

Figure 6 Path Lengths Between All User Pairs—Observed Data



Treated

# hops after \ # hops before	1	2	3	∞
∞	0.12	1.03	0.20	0.31
3	0.02	0.14	0.10	0.09
2	8.01	37.94	2.71	3.03
1	16.07	28.41	0.83	1.00

Becoming closer (below diagonal) 36.0
Becoming farther (above diagonal) 9.5
No change (on diagonal) 54.4

Control

# hops after \ # hops before	1	2	3	∞
∞	0.36	4.08	0.84	2.09
3	0.26	3.76	1.12	1.04
2	11.8	47.62	4.02	3.20
1	7.72	11.1	0.44	0.37

Becoming closer (below diagonal) 20.2
Becoming farther (above diagonal) 21.1
No change (on diagonal) 58.5

Note. Entries represent the percentages of all ${}_nC_2$ user pairs.

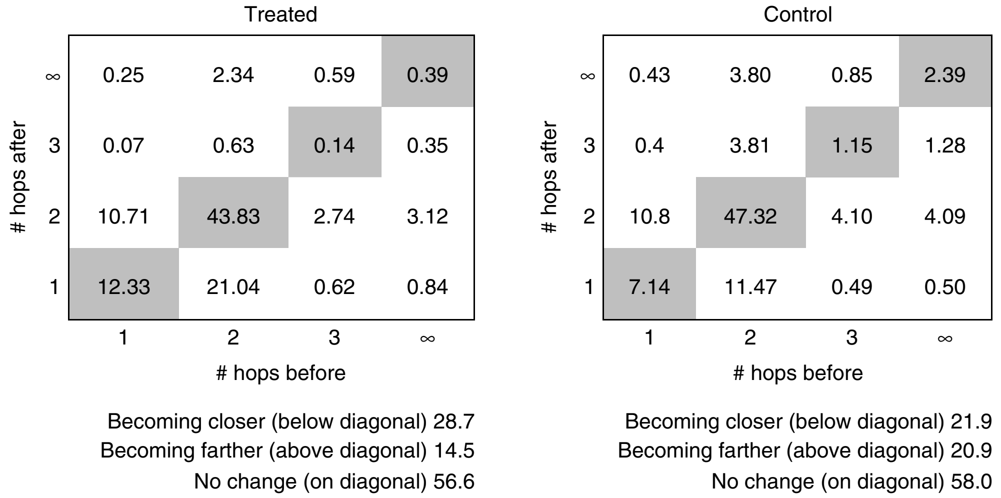
went from two hops to one hop). To summarize, the trend toward greater similarity exists at all initial path lengths, independent of whether users were initially close or far away.

Under volume equalization, there is again no evidence of differential effects. Figure 7 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 apart (28.71% weight below the diagonal versus 14.59% weight above it). The control group is roughly balanced. This is consistent with the aggregate findings. The magnitude is again smaller, as expected, because volume equalization dampens the effect. Second, the increase in similarity appears uniform: users who were close became closer, and users who were initially far became closer too.

9.2. 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 purchase similarity may be driven by changes in users' preferences around the time of registration. We addressed this issue by conducting all analysis on a matched sample of users, as is commonly done in the literature. In this section, we exploit some unique features of our data set to conduct additional tests that show that it is unlikely that our results are due to a selection bias. 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

Figure 7 Path Lengths Between All User Pairs—Volume-Equalized Data



Treated

# hops after \ # hops before	1	2	3	∞
∞	0.25	2.34	0.59	0.39
3	0.07	0.63	0.14	0.35
2	10.71	43.83	2.74	3.12
1	12.33	21.04	0.62	0.84

Becoming closer (below diagonal) 28.7
Becoming farther (above diagonal) 14.5
No change (on diagonal) 56.6

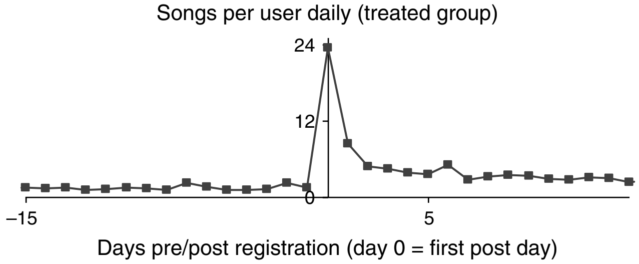
Control

# hops after \ # hops before	1	2	3	∞
∞	0.43	3.80	0.85	2.39
3	0.4	3.81	1.15	1.28
2	10.8	47.32	4.10	4.09
1	7.14	11.47	0.49	0.50

Becoming closer (below diagonal) 21.9
Becoming farther (above diagonal) 20.9
No change (on diagonal) 58.0

Note. Entries represent the percentages of all ${}_nC_2$ user pairs.

Figure 8 **Daily Songs Added per User (Median) Centered on Each User's Registration Date**



Note. Day 0 represents the time immediately after registration.

the following lines: (a) ruling out a time trend among treated group and (b) verifying impact of treatment on the control group.

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 high level of similarity between the two groups. Furthermore, we attribute the small differences in adoption timing between the two groups primarily 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 plug-in 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 primarily due to spread of product awareness. This line of reasoning is the same as in Waldfogel and Chen's (2006) study of how sales at unbranded retailers are affected by the introduction of comparison shopping engines on the Web, which were at the time a new technology.

To test this idea, Figure 8 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. We now test the robustness of our results more formally.

9.2.1. 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 8 suggests this is unlikely. We test this more formally by conducting a difference-in-differences test of purchase similarity in multiple pretreatment 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 difference-in-differences test. Table 10 shows that there are no significant changes in median degree and average distance 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.

9.2.2. Effect of Treatment on Control Users. A unique aspect of our data set 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

Table 10 **Summary Measures for Preregistration Time Periods**

	Median degree					Average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	655.70	656.00	0.3037	–0.3404	0.4020	0.9935	0.9933	–0.0002	–0.0007	0.3480
C	656.06	656.70	0.6441			0.9942	0.9947	0.0005		
					Volume-equalized data:					
T	655.99	655.99	0	–0.4650	0.2360	0.9940	0.9933	–0.0007	–0.0009	0.2320
C	656.38	656.85	0.4650			0.9948	0.9950	0.0002		

Table 11 Summary Measures for the Original Control Users

	Median degree					Average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
G1	809.36	805.34	−4.0190	−3.8929	<0.001	0.9912	0.9860	−0.0052	−0.0048	<0.001
G2	809.80	809.68	−0.1260			0.9912	0.9909	−0.0003		
					Volume-equalized data:					
G1	809.37	806.63	−2.7434	−2.4779	0.002	0.9912	0.9879	−0.0033	−0.0028	0.002
G2	810.08	809.81	−0.2654			0.9917	0.9913	−0.0005		

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 11 shows that we observe a significant decrease in median degree and average distance for G1 users relative to G2 users. Thus control users also experience an increase in purchase similarity when they are exposed to the recommender.

9.3. Attrition

The second data limitation is attrition. About half of the users in the treated group uninstall the plug-in 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 attrition is that we may overstate the magnitude of the results although not their direction. This conclusion requires the assumption that uninstallers return to pretreatment behavior and resemble the control group. To illustrate how attrition affects the magnitude, we can "average" the treated users who complete the study with control users as proxies for the dropouts. 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. A key question, however, is whether the increase in commonality continues to be significant even after the dampening. We test this below.

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 dropouts 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*. We simulate a composite group by randomly drawing users without replacement from the control group to replace treated users who uninstall. In our data, $\lambda \approx 0.5$, and Table 12 presents the results for the weighted network. For the observed data, the results show that there is a significant increase in purchase similarity among composite users even after accounting for attrition. For volume-equalized data, statistical significance is lost, but the direction stays the same for the weighted network. Two comments are in order. First, in several environments, firms deploy recommenders and users do not have an option to "turn off" personalization, and thus the dampening of magnitude for composite users may not occur. Second, the weighted network measure is not very sensitive. Even if a user adds many new artists in the after period, this may not significantly change the angle between the user's purchase vector and that of another user because of the large number of artists in the data set (and the resulting high-dimensional space). One way to test this idea is to repeat the analysis for an unweighted network. Table 13 shows that both observed data and volume-equalized data are still significant at $\lambda = 0.5$. Finally, we repeat the simulations for different values of λ and find that significance is lost for the weighted network (at $p = 0.05$) at $\lambda = 0.67$ (i.e., if 67% of treated users had uninstalled the service) for the observed data and $\lambda = 0.42$ for the volume-equalized data.

Table 12 Summary Measures for *Composite* Weighted Network ($\lambda = 0.5$)

	Median degree					Average distance				
	Before	After	D_i	D	p	Before	After	D_i	D	p
T	850.9983	848.329	−2.6693	−2.0891	<0.001	0.9912	0.9878	−0.0034	−0.0024	<0.001
C	851.4957	850.9155	−0.5802			0.9919	0.991	−0.0009		
					Volume-equalized data:					
T	851.1542	849.0389	−2.1153	−1.3949	0.06	0.9914	0.9888	−0.0026	−0.0015	0.116
C	851.6458	850.9254	−0.7204			0.9921	0.991	−0.0011		

Table 13 Summary Measures for *Composite* Unweighted Network ($\lambda = 0.5$)

	Density					Median degree					Path length				
	Before (%)	After (%)	D_i (%)	D (%)	p	Before	After	D_i	D	p	Before	After	D_i	D	p
T	22	32	10	11	<0.01	148	252	104	107	<0.01	1.84	1.70	−0.14	−0.14	<0.01
C	20	20	−1			126	123	−3			1.86	1.86	0.00		
						Volume equalized data:									
T	20	26	6	4	<0.01	143	206	63	47	<0.01	1.84	1.75	−0.09	−0.06	<0.01
C	17	19	2			118	134	15			1.88	1.86	−0.02		

10. Relationship to Service's Recommender System

This section relates these findings to the recommendation system in use at Service. We believe similarity increases after 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 plug-in and the website.

With the plug-in, 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 having the same artist are more likely to see the same recommendations and thus more likely to purchase another common item. In terms of the unweighted network from §9.1, users who are one hop away in the treated group should be more likely to remain one hop away than control users. Table 14 supports this. Treated users one hop away are 67% likely to remain one hop away afterward, whereas one-hop-away control users are only 38% likely to remain at one hop.[12] Seeing the same recommendations maintains the one-hop position among treated users, whereas there is no such force 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 plug-in. 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, one observes again in Table 14 that Prob(one hop away after | k hops away before) does show a primarily decreasing trend.

Table 14 Probability (User Pair Is One 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 copurchases 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 copurchase. In turn, more copurchases results in an increase in similarity.

Finally, 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 subgenre are recommended, a collaborative algorithm where songs copurchased 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

Much of our time spent online—whether reading news, listening to music, or purchasing books—is guided invisibly by recommendation algorithms. Despite the millions of hours guided by them and papers proposing new algorithms, we know much less about how they affect the market and society.

This paper asked whether recommender systems fragment versus homogenize users. Using data from the music industry, we found that users have more in common after recommendations, as measured by purchase similarity. The increase in commonality occurred for two reasons: the product-mix effect, in

[12] The probabilities are approximated as the fraction of user pairs transitioning from k to one hops, and the data come from Figures 6 and 7.

which users shifted their purchases toward more similar items, and the volume effect, in which users simply bought more under recommendations, increasing the likelihood of copurchases with others. Each effect contributed roughly equally to the overall result. Furthermore, when users are clustered by taste, both the within-cluster and between-cluster distances shrink after recommendations. Personalization thus appears to be a tool that helps users widen their interests and create commonality.

Regarding policy implications, many have criticized recommenders fearing that recommenders will fragment the online population. We agree that excessive fragmentation could be undesirable. However, our data show that recommenders appear to create commonality, not fragmentation. In the absence of such effects, there is no 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. Our study demonstrates that recommender systems can drive a significant increase in purchase volume and may further alter the mix of products users buy.

The results suggest several areas for future work. The first area is studying additional recommender technologies. This paper studied a major online recommender system. We hope future work will look at other designs and gradually catalog their effects. Second, one could study domains other than music (e.g., news, books, fashion). The manner in which users respond to news or fashion recommendations may differ from the manner in which they respond to music recommendations. The policy implications for news would be especially important. Finally, a third area is relating recommender design choices to commonality. For example, it is possible that some artists are boundary spanners (e.g., Elvis could be classified under both rock and country). Recommenders that explicitly promote boundary-spanning artists could play a key role in driving exploration and commonality. Future designs may want to consider this.

In *The Big Sort*, Bishop (2008) shows how over the last 30 years Americans have sorted themselves into like-minded neighborhoods. This paper asks a similar question about the virtual space of the Web. Although 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 30 more years will provide.

Acknowledgments

For valuable feedback, the authors thank seminar participants at the University of Michigan, the University of Utah, the University of Maryland, Temple University, Washington University, Stanford University, University of Texas, Purdue University, Emory University, Princeton University, Nanyang Technological University, and INSEAD. The authors also thank Mack Center and NET Institute for financial support.

Appendix. Significance Testing

The hypotheses tested in the aggregate analysis have the form

$$\begin{aligned} \mathrm{H}_0\colon &\quad \mu \equiv E[D] = 0; \\ \mathrm{H}_a\colon &\quad \mu \equiv E[D] \neq 0, \end{aligned}$$

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 purchases are distributed the same in the treated group and in the control group. The use of such test statistics is facilitated by permutation tests that 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). In our study, we used 1,000 iterations for permutation tests.

References

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

Ansari A, Essegaier S, Kohli R (2000) Internet recommendation systems. *J. Marketing Res.* 37(3):363–375.

Arnheim A (1996) Summary of the proceedings of the University of California at Berkeley Collaborative Filtering Workshop. Accessed April 20, 2008, http://www2.sims.berkeley.edu/resources/collab/collab-report.html.

Apple (2008) iTunes store top music retailer in the US. Accessed November 27, 2008, http://www.apple.com/pr/library/2008/04/03itunes.html.

Austin CP, Grootendorst P, Anderson GM (2007) A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statist. Medicine* 26(4):734–753.

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

Bodapati A (2008) Recommendation systems with purchase data. *J. Marketing Res.* 45(1):77–93.

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

Campbell DT, Stanley J (1963) *Experimental and Quasi-Experimental Designs for Research* (Houghton Mifflin Company, Boston).

Clemons EK, Gao GG, Hitt LM (2006) When online reviews meet hyperdifferentiation. *J. Management Inform. Systems* 23(2):149–171.

Das A, Datar M, Garg A, Rajarm S (2007) Google news personalization: Scalable online collaborative filtering. *Proc. 16th Internat. World Wide Web Conf.* (Association for Computing Machinery, New York), 271–280.

De P, Hu YJ, Rahman MS (2010) Technology usage and online sales: An empirical study. *Management Sci.* 56(11):1930–1945.

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

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

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

Fleder D, Hosanagar K, Buja A (2010) Recommender systems and their effects on consumers: The fragmentation debate. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 229–230.

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

Gruber J (1994) The incidence of mandated maternity benefits. *AER* 84(3):622–641.

Hartigan JA (1975) *Clustering Algorithms* (John Wiley & Sons, New York).

Heckman J, Ichimura H, Smith J, Todd P (1998) Characterizing selection bias using experimental data. *Econometrica* 66(5): 1017–1098.

Hervas-Drane A (2013) Word of mouth and sales concentration. Working paper, Universitat Pompeu Fabra, Barcelona, Spain. http://ssrn.com/abstract=1025123.

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal.* 15:199–236.

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

Katz M, Shapiro C (1985) Network externalities, competition, and compatibility. *Amer. Econom. Rev.* 75(3):424–440.

Lamere P, Green S (2008) Project Aura: Recommendation for the rest of us. Presentation, Sun JavaOne Conference. Accessed October 17, 2013, http://www.oracle.com/technetwork/systems/ts-5841-159144.pdf.

Linden G (2008) People who read this article also read… . *IEEE Spectrum* 45(3):46–60.

Meyer BD (1995) Natural and Quasi experiment in economics. *J. Bus. Econom. Statist.* 13(2):151–161.

Murthi BPS, Sarkar S (2003) The role of the management sciences in research on personalization. *Management Sci.* 49(10):1344–1362.

Negroponte NP (1995) *Being Digital* (Vintage Books, New York).

Newman MEJ (2004) Analysis of weighted networks. Accessed September 2008, http://arxiv.org/abs/condmat/0407503.

Oestreicher-Singer G, Sundararajan A (2012) Recommendation networks and the long tail of electronic commerce. *MIS Quart.* 36(1):65–83.

Papadimitriou CH, Steiglitz K (1998) *Combinatorial Optimization*: *Algorithms and Complexity* (Dover Publications, Toronto).

Pariser E (2011) *The Filter Bubble*: *What the Internet Is Hiding from You* (Penguin Press, New York).

Pearl J (2009) *Causality*: *Models, Reasoning, and Inference*, 2nd ed. (Cambridge University Press, Cambridge, UK).

Rossi P, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

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

Shadish WR, Cook TD, Campbell DT (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton-Mifflin, Boston).

Shaffer G, Zhang ZJ (1995) Competitive coupon targeting. *Marketing Sci.* 14(4):395–416.

Smith M, Giraud-Carrier C, Judkins B (2007) Implicit affinity networks. *Proc. 7th Annual Workshop Inform. Technologies and Systems* (Kluwer Academic Publishers, Hingham, MA), 123–134.

Sunstein CR (2007) *Republic.com 2.0* (Princeton University Press, Princeton, NJ).

Terdiman D (2011) Why a hyper-personalized Web is bad for you. *CNET News* (May 17), http://news.cnet.com/8301-13722_3-20063402-52.html.

Thompson C (2008) If you liked this, you're sure to love that. *New York Times Magazine* (November 21), http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html?pagewanted=all&_r=0.

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

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

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440–442.