“People Who Liked This Study Also Liked”: An Empirical Investigation of the Impact of Recommender Systems on Sales Volume and Diversity

Dokyun Lee
Kartik Hosanagar

The Wharton School
University of Pennsylvania

"People Who Liked This Study Also Liked": An Empirical Investigation of the Impact of Recommender Systems on Sales Volume and Diversity

Abstract

We investigate the impact of collaborative filtering recommender algorithms (e.g., Amazon.com's "Customers who bought this item also bought"), commonly used in e-commerce, on sales volume and diversity. We use data from a randomized field experiment on movie sales run by a top retailer in North America. For sales volume, we show that different algorithms have differential impacts. Purchase-based collaborative filtering ("Customers who bought this item also bought") causes a 25% lift in views and a 35% lift in the number of items purchased over the control group (no recommender). In contrast, View-based collaborative filtering ("Customers who viewed this item also viewed") shows only a 3% lift in views and a 9% lift in the number of items purchased, albeit not statistically significant. For sales diversity, we find that collaborative filtering algorithms cause individuals to discover and purchase a greater variety of products but push users to the same set of titles, leading to concentration bias at the aggregate level. We show that this differential impact on individual versus aggregate diversity is caused by users exploring into only a few 'pathway' popular genres. That is, the recommenders were more effective in aiding discovery for a few popular genres rather than uniformly aiding discovery in all genres. For managers, our results inform personalization and recommender strategy in e-commerce. From an academic standpoint, we provide the first empirical evidence from a randomized field experiment to help reconcile opposing views on the impact of recommenders on sales diversity.

Keywords: E-Commerce, Personalization, Recommender systems, Sales volume, Sales diversity, Consumer purchase behavior, Collaborative filtering, Gini coefficient

1 Introduction

Online consumers are constantly guided by some form of recommendation systems (RS)[1], be it for shopping or web browsing. For shopping, recommenders can be as simple as "the most popular" items sold on the site, or as sophisticated as a collaborative filter, CF, based on individual purchase history (e.g., Amazon.com's "Customers who bought this item also bought"). These systems are so pervasive in e-commerce and web services that a majority of consumers now expect and prefer websites that provide personalized recommendations (Accenture, 2012). At the same time, 94% of companies agree that "personalization of the web experience is critical to current and future success" (Econsultancy and Monetate, 2013). The benefits of RS for consumers include lower search costs and higher product-fit. In addition, with the rise of the long tail phenomenon, where in e-tailers offer a broad range of niche items (Anderson, 2008), recommenders help consumers manage a potential choice overload by reducing the consideration set (via showing the most relevant items). For firms, RS has been shown to promote sales, web usage, and customer retention (Das et al., 2007; Dias et al., 2008; Thompson, 2008; Monetate, 2013).

In addition to their positive impact on volume, recommenders are widely believed to affect sales diversity. This refers to the market share distribution of products sold at the firm level and the variety of items purchased at the individual consumer level. One school of thought believes that recommenders lower search costs and contribute to a long tail phenomenon in which consumers are exposed to more niche items, increasing both individual product consumption diversity and firm-level sales diversity. An opposing theory suggests that common recommender designs such as collaborative filters can lead to reduction in aggregate sales diversity. It argues that because collaborative filters recommend products based on sales and ratings, they cannot recommend items with limited historical data. This leads to popularity bias in these designs. Understanding which viewpoint is correct is important for retailers, consumers, and producers. For retailers whose strategy is to offer variety – under the premise that consumers will find better-suited products and thus purchase more – to the extent recommenders increase concentration, they may be at odds with such a goal. Similarly for consumers and niche producers, if there exist better product matches outside of the most popular titles, these groups may be better served (or underserved) if these recommender systems increase sales diversity (concentration).

[1]We use the terms recommender systems, RS, and recommender algorithms interchangeably.

There are several reasons for the divergent viewpoints on the impact of RS on sales diversity. First, a potential driver may be the lack of data that provides a contrast between users exposed and unexposed to recommendations. Most retailers observe consumers after they arrive at the website (i.e., after exposure to recommendations) and cannot observe the contrast needed to answer this question. This is one reason why the question remains empirically unanswered despite years of debate. Second, the majority of research to date has studied specific recommender algorithms in isolation rather than comparing different kinds of recommenders. In e-commerce, there are many types of recommenders (Schafer et al., 1999). Most commonly used designs tend to be collaborative filtering algorithms. There can be multiple flavors of collaborative filters, including those based on purchases ("Customers who purchased this also purchased") and views ("Customers who viewed this also viewed"). Other types of recommenders include Content-based ("Based on your consumption history") and Social-network-based recommenders ("Your friends bought"). One possible explanation for the divergent viewpoints is that these different recommendations have considerably different impact on sales volume and diversity. To the extent that this is true, it alters how firms must choose recommender designs. For example, one firm may prefer a design that maximizes sales, whereas another firm may prefer a design that better exposes consumers to its breadth of product assortment.

Our study attempts to address these gaps by utilizing a field experiment to examine the impact of the two most commonly used recommender algorithms on sales volume and diversity. By running different algorithms on the website of a top retailer in North America and randomly assigning recommender treatments to visiting consumers and tracking their views and purchases, we tease out the differential effects of recommender designs on both sales volume and diversity. Our findings help answer several critical questions related to recommender design. For example, is one recommender algorithm more effective for increasing the number of products consumers view or purchase? Do certain algorithms cause "popularity bias," wherein a top few bestsellers are promoted? How exactly do sales diversity shift occur? What algorithms should a manager use on an e-commerce site, given product assortment decisions and sales goals?

We find that recommender design affects both sales volume and diversity. In particular, a collaborative filtering algorithm based on purchase data, "Customers who purchased this item also purchased," was best at increasing the sales volume. Further, the algorithm increased individual consumption diversity but decreased aggregate consumption diversity. We show that this differential

impact in individual versus aggregate diversity is mainly caused by users exploring into only a few 'pathway' popular genres like comedy. The result highlights a potential limitation in the ability of traditional collaborative filtering to aid discovery of truly niche items and genres. We also show that not all collaborative filtering algorithms are equal by contrasting the results with that for View-based collaborative filtering.

Our results help reconcile opposing theories on the impact of recommenders using a randomized field experiment. Further, they unlock important managerial insights on which designs better address their goals. We find that collaborative filters, especially those based on purchases, are very effective at driving an increase in sales. However, to the extent that firms are interested in promoting a broader product assortment, we advocate that firms should modify traditional collaborative filtering algorithms to ensure that relevant items with limited historical sales can be discovered by consumers.

2 Prior Work

Given the significant influence of RS on e-commerce and consumer purchase behavior, the RS literature has been growing steadily in the last couple of decades. In the 1990s, which marked the rise of RS in e-commerce, researchers focused on developing different recommender algorithms (for an extensive survey, see Adomavicius and Tuzhilin (2005)). In the 2000s, a stream of research investigated the impact of RS on sales volume, or firm performance, looking at factors such as profit, revenue, and consumer retention (Das et al., 2007; Dias et al., 2008; Thompson, 2008). In the late 2000s and 2010s, studies of RS expanded to examining consumer consumption patterns and firm sales diversity (Fleder and Hosanagar, 2009; Hinz and Eckert, 2010; Oestreicher-Singer and Sundararajan, 2012; Jannach et al., 2013; Matt et al., 2013; Hosanagar et al., 2014). In this section, we first provide a taxonomy of recommender systems. Second, we review what the recommendation system literature tells us about recommenders' influence on sales volume and revenue, and then review the burgeoning literature on recommenders' influence on sales diversity. Lastly, we discuss the gap in the literature and position our paper in it.

2.1 Overview and Taxonomy of Recommendation Systems

Before the advent and growth of personalized recommenders, the primary recommendation approach used by most firms involved the use of simple signals such as the "most popular" or "highest rated" items. A well-known current example is *The New York Times*'s "most emailed articles" feature. Such signals have been shown to influence consumer learning and choice in consumer behavior literature and in studies tied to observational learning theory (Bikhchandani et al., 1992). Similar types of learning from the masses, whether based on positive views or reviews, have been well documented in the literature (Salganik et al., 2006; Muchnik et al., 2013; Tucker and Zhang, 2011; Lee et al.). While these kinds of recommendations based on aggregate signals are privacy preserving and serve the mass market, they may limit exploration by consumers and are less useful for consumers with niche tastes.

Personalized recommenders seek to recommend items based on individual-level data. Within *Personalized Recommenders* systems, a broad taxonomy distinguishes three types of algorithms: Content-based, Collaborative Filtering, and Hybrid, which combines the first two (Adomavicius and Tuzhilin, 2005). Content-based systems analyze product attributes to suggest products that are similar to those that a consumer bought or liked in the past. Collaborative filtering recommenders, unaware of product attributes, recommend products either purchased or liked by similar consumers, where similarity is measured by historical purchase (or like) data.

In recent years, social-network-based recommenders have emerged that can either simply signal to consumers that their friends have bought certain items or can be used as an extreme form of collaborative filtering that assumes that friends are similar in their tastes (Victor et al., 2011). From a marketing perspective, some recent papers have looked at how to incorporate consumers' expressed preference and learning (Ansari et al., 2000) and sensitivity of purchase probability (Bodapati, 2008) into recommender systems. For an extended survey from a computer science perspective, please refer to Adomavicius and Tuzhilin (2005).

2.2 Literature on the Impact of Recommenders on Sales Volume and Diversity

While there are a large number of studies on improving the actual algorithms mentioned above, we know little about how these algorithms affect consumers and markets. While the literature is unequivocal that recommender systems positively impact sales, it is unclear about which designs

have a greater impact. For example, industry reports have claimed considerable increase in revenue and/or usage due to the use of recommendation systems in the cases of Amazon, Netflix, and Google News (Thompson, 2008; Das et al., 2007; Marshall, 2006). Recommendation engines have been reported to increase revenue up to 300%, conversion rates up to 150%, and consumers' average order value up to 50%, according to a recent study (Monetate, 2013). De et al. (2010) show that the use of a recommendation system has a positive effect on the sales of both promoted and non-promoted products in contrast to the use of regular search engines, which increase sales only for promoted products and decrease sales for non-promoted products. Similarly, Hinz and Eckert (2010) show, via agent-based modeling, that the use of recommenders can increase profits for retailers. Dias et al. (2008) use a supermarket case study to show that RS increases both direct revenue and indirect revenue, in which indirect revenue is obtained by consumers' cross-buying behaviors. Jannach and Hegelich (2009) carry out a case study to evaluate the use of recommenders in a mobile app market and find that the use of RS increases click-through rates and overall sales. In summary, there is ample evidence to support the positive effect of RS on sales. However, the literature lacks field evidence comparing different types of recommender algorithms and which types of designs are more effective.

Unlike the *sales volume impact* of RS, the *sales diversity impact* of RS is a subject of much disagreement in the literature. On one side, Brynjolfsson et al. (2011) and Anderson (2008) indicate that the use of RS will contribute to a long tail phenomenon in which niche items gain more market share, increasing both individual- and firm-level product sales diversity. On the other side, using theoretical models and simulations, Fleder and Hosanagar (2009) argue that the use of collaborative filtering will lead to a decrease in firm-level aggregate sales diversity but it can lead to an increase in individual product consumption diversity. They suggest that collaborative filters can lead consumers to new items but aggregate diversity may still decrease because they push similar users to the same set of products. Similarly, Jannach et al. (2013) use a well-known MovieLens dataset and simulation study to show that many recommendation algorithms are biased toward broad-appeal items, causing a rich-get-richer situation that decreases aggregate product sales diversity. Celma and Cano (2008) examine the collaborative filtering recommendation of last.fm and find that the algorithm tends to reinforce the consumption of popular artists, supporting the hypothesis that collaborative filtering will decrease aggregate diversity while also showing that content-based algorithms are less biased toward popular artists. In contrast, Hinz and Eckert (2010) argue through agent-based modeling

that recommenders will drive a long tail phenomenon by reducing search costs and shifting demand from broad-appeal items to niche items. Echoing this idea, Oestreicher-Singer and Sundararajan (2012) argue that Amazon's collaborative filtering recommender (e.g., "Customers who bought this also bought") shifts demand from broad-appeal items to niche items, thereby increasing aggregate product sales diversity. Similarly, regarding YouTube, Zhou et al. (2010) find evidence that the recommender increases aggregate diversity. Via a lab experiment, Matt et al. (2013) argue that both collaborative filtering and content-based recommenders increase aggregate sales diversity and that the differences between different recommenders are small. Similarly, by studying a hybrid algorithm that is content-base heavy, Hosanagar et al. (2014) focus on individual consumption patterns and show that, while recommenders increase individual consumption diversity, they also increase commonality among consumers. They also find that aggregate diversity increases as a result. Lastly, Wu et al. (2011) use simulation and MovieLens data and find that content-based recommenders tend to increase the aggregate sales diversity, while collaborative filtering decreases it. Table 1 summarizes these academic papers and their main claims. In sum, there is no consensus among both popular and academic literature on how recommenders will affect sales diversity.

We believe this lack of consensus arises due to the following several reasons. Some studies evaluate one type of algorithm and are based on lab experiments or simulations calibrated to archival data, which makes generalization harder. Other studies measure non-purchase measure like purchase-intentions, use-intentions, and satisfaction rather than the actual views or purchases. The few based on field archival data are aggregate level data and are constrained by the limitations of observational data that make causal conclusions harder to derive. We carry out a randomized field experiment on a large e-commerce website using multiple recommender algorithms and investigate the differential effects on sales volume and diversity with direct individual-level view and purchase data. While our approach is not entirely free from mentioned problems, the strength of our approach is that 1) the field experiment conducted on a large e-commerce site allows us to observe recommenders' effects more realistically, 2) we directly measure individual-level view and purchase data, and 3) we have clean identification as a result.

Study	Method & Data	Sales Volume	Sales Diversity
De et al. (2010)	Archival Data & Econometrics	Increases sales	
Hinz and Eckert (2010)	MovieLens Data & Simulation	Increases sales and profit	Increase niche product consumption leading to **increase in aggregate sales diversity**
Dias et al. (2008)	Archival Data & Case Study	Increases direct and indirect revenue	
Jannach and Hegelich (2009)	Mobile App Market Data & Case Study	Increases sales	
Fleder and Hosanagar (2009)	Theoretical Models & Simulation		**Decrease in aggregate sales diversity** but **increase in individual sales diversity**
Hosanagar et al. (2014)	Archival Data & Econometrics	Increases individual consumption volume	Content-based RS **increase aggregate sales diversity** and **increase overlap/commonality in consumption**
Oestreicher-Singer and Sundararajan (2012)	Crawled Amazon Data & Econometrics	Increases revenue	Recommender shifts demand to niche item **increasing aggregate sales diversity**
Jannach et al. (2013)	MovieLens Data & Simulation		Different algorithms have different effects
Wu et al. (2011)	MovieLens Data & Simulation		Mixed result based on different algorithms. Collaborative filtering **decreases aggregate diversity** while content-based **increases** it
Celma and Cano (2008)	last.fm and Allmusic.com API data & Correlational Analysis		Collaborative filtering algorithm is linked to popularity bias suggesting **decreased aggregate consumption diversity**
Zhou et al. (2010)	Crawled YouTube Data & Correlational Analysis	Recommender accounts for 30% of video views	**Increases aggregate consumption diversity**
Matt et al. (2013)	Lab Experiments		**Increase in aggregate sales diversity** for variety of different recommenders except for bestseller list

Table 1: **Literature on Impact of Recommender Systems and Claims**

3 Problem Formulation

This section formally sets up research problems and hypotheses based on the previous literature.

3.1 Research Questions

We are interested in studying the impact of recommenders on sales volume and diversity. *Sales volume* is measured in terms of number of purchases and value of purchases ("wallet size"). In addition, we measure the *sales diversity* of the products sold with a measure called the Gini coefficient. The Gini coefficient has been widely adopted in the long tail and the RS literature as a measure of sales diversity (Brynjolfsson et al., 2011; Fleder and Hosanagar, 2009; Hosanagar et al., 2014). It is computed based on the Lorenz curve. Let $L(u)$ be the Lorenz curve denoting the percentage of the sales generated by the lowest $100u\%$ of items as shown in Figure 1. The Gini coefficient is defined as $G \equiv \frac{A}{A+B}$. It ranges from 0, representing the least amount of concentration or highest diversity, to 1, representing the highest amount of concentration or lowest diversity. A Gini coefficient of 0 means that all products have equal sales, while values near 1 mean that a few broad-appeal blockbuster items account for most of the sales.

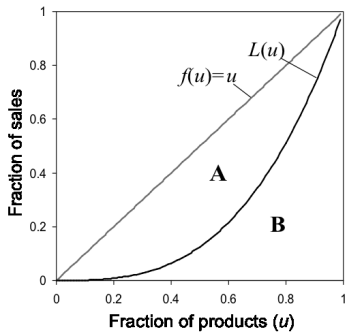


Figure 1: **Lorenz Curve**

We approach this problem with a field experiment in which consumers visiting a website are randomly assigned to a control or treatment group. The treatment group is shown a panel of different recommendations, much like Amazon's "Customers who bought this item also bought" recommenders. The control group is shown nothing. For each group, we analyze the following variables of interest.

1. *Sales Volume Impact*

   (a) Individual item **view** volume: This counts how many items individuals have clicked and viewed.

   (b) Individual item **purchase** volume: This counts how many items individuals buy. We

also look at individual wallet size, which measures the total individual purchase dollar amount.

2. *Sales Diversity Impact*

   (a) Aggregate firm-level item (and genre) **view and sales** diversity: This measures how the recommenders affect product view/sales diversity at the aggregate level (for each treatment group) and is measured by the Gini coefficient. We repeat this at the genre level to investigate genre cross-pollination.[2]

   (b) Individual item (and genre) **view and sales** diversity: This measures how the recommenders affect the diversity of products individuals view or purchase. Again, the Gini coefficient is used but it is computed based on an individual's purchases. The analysis is repeated at the genre level.

3.2 Treatment Groups

There are three groups in our study:

1. Control (no recommendations)
2. View-based collaborative filtering ("People who viewed this item also viewed")
3. Purchase-based collaborative filtering ("People who purchased this item also purchased")

We have two different treatment groups corresponding to two different recommender algorithms, plus a control group that was not shown any recommendations. The two treatments are two of the most commonly used types of collaborative filtering algorithms. One is based on views ("People who viewed this item also viewed"), while the other is based on purchases ("People who purchased this item also purchased"). Our data partner used the widely adopted open-source Apache Mahout framework (mahout.apache.org) for constructing the recommenders.

[2]Measuring genres consumed is a conservative and more robust way of measuring diversity, and it is included in the analysis because 1) it is easier to interpret the purchase diversity of genres and 2) mathematically, the Gini coefficient changes are more conservative at the genre level, making the results more robust. The results are the same at the item level, with more group comparisons statistically significant. We present both in the Results section.

3.3 Study Design

Let g_i represent group i and let f represent a function that calculates an aggregate measure of interest, D_i, for the given group (e.g., sales volume, Gini coefficient). We define the following quantity of interest:

Aggregate Measure, f, of Group 1	$D_1 \equiv f(g_1)$
Aggregate Measure, f, of Group 2	$D_2 \equiv f(g_2)$
Difference in Aggregate Measures	$D \equiv D_1 - D_2$

The difference in the aggregate measure, D, shows how different Group 1 is from Group 2. Let $\mu \equiv \mathbb{E}[D]$, with the distribution of D unknown. All hypotheses testing in this paper takes the form:

Null Hypothesis H_0	$\mu \equiv \mathbb{E}[D] = 0$
Alternate Hypothesis H_a	$\mu \equiv \mathbb{E}[D] \neq 0$

This null hypothesis tests if variables of interest, such as group revenues for Groups 1 and 2, are distributed equally. Since we have one aggregate measure (or statistic) for each group, in order to produce a p-value, we utilize a permutation test technique (Good, 2005) that allows us to calculate a null distribution for a given aggregate measure. If the null hypothesis of equal distribution is true, all relabelings of individuals as Groups 1 and 2 are equally likely. A permutation test involves repeatedly and randomly relabeling individuals into Groups 1 and 2 (e.g., control and treated) to produce a null distributions for any test statistics. By comparing statistics from null distributions to the actual test statistics from the real distribution and tallying how often null distribution statistics exceed the actual distribution statistic, we can determine the p-value. For more details, see Good (2005). Note that we carry out the hypotheses tests as two-sided tests (equal or not equal rather than greater than or less than) to stay conservative. In our study, we use 1000 iterations to get an accurate p-value up to 0.001.

3.4 Hypotheses

We organize our hypotheses on the impact of recommenders on sales volume and diversity in this section. Our hypotheses are informed by the extant literature discussed in Section 2. While the hypotheses on sales volume are clearly driven by unequivocal results of previous studies, we take a particular stance on sales diversity since the existing literature disagrees on this matter.

Sales Volume Hypotheses There appear to be many reasons why recommenders affect sales volume. Many long tail and recommender studies (Brynjolfsson et al., 2006, 2011; Oestreicher-Singer and Sundararajan, 2012; Hinz and Eckert, 2010) postulate that personalized recommenders reduce consumers' search costs by reducing the need to search for the right product. In doing so, recommenders make it easy for consumers to find the right products, which leads to more purchases. The impact on number of product views is less clear. On the one hand, a reduction in search cost can lead consumers to their desired product faster, resulting in fewer clicks and product views. On the other hand, recommenders may be effective in cross-selling, up-selling, or even in driving repeat visits, thereby resulting in increase in number of product views. We hypothesize that recommenders will contribute toward an increase in purchases as well as product views.

Hypothesis 1 *Collaborative filtering (CF) recommenders will* ***increase*** *the number of* ***product views*** *compared to no recommendation condition (Control).*

Hypothesis 2 *Collaborative filtering (CF) recommenders will* ***increase*** *the number of* ***product purchases*** *compared to no recommendation condition (Control).*

Although we have two different types of collaborative filters (Purchase-based or View-based), we do not make any hypotheses comparing the two. Intuition suggests that collaborative filtering based on purchase data will be more potent, since a purchase is a stronger signal than a view and is probably more accurate.

Sales Diversity Hypotheses Fleder and Hosanagar (2009) offer a clear conjecture on how collaborative filters affect product consumption diversity at both the aggregate and individual level. They argue that a collaborative filtering algorithm will show popularity bias by directing users to broad-appeal blockbuster items, leading to decreased aggregate firm-level product sales diversity. Likewise, Celma and Cano (2008) and Wu et al. (2011) also support the hypothesis that collaborative filtering will decrease aggregate product sales diversity. At the same time, Fleder and Hosanagar (2009) predict that individuals will be exposed to a greater variety of products, leading to increased individual product consumption diversity. Our hypotheses on sales diversity are based on the Fleder and Hosanagar (2009) paper, which presents a theoretical model specifically on collaborative filtering and provides conjecture on both the aggregate and individual diversity.

In summary, our hypotheses on the *sales diversity impact* of RS are as follows:

Hypothesis 3 *Collaborative filtering recommenders will* ***decrease*** *the* ***aggregate*** *product sales diversity compared to no recommendation condition (control), i.e., Aggregate Gini(CF) > Aggregate Gini(Control).*

Hypothesis 4 *Collaborative filtering recommenders will* ***increase*** *the* ***individual*** *product purchase diversity compared to no recommendation condition (control), i.e., Average Individual Gini(CF) < Average Individual Gini (Control).*

4 Data

Our dataset comes from a field experiment on a website of one of the top retailers in North America. The experiment was conducted for two weeks between August 8, 2013 and August 22, 2013. Focusing attention on a product category commonly used in the RS literature, this study examines the impact of recommenders on movie-related (Blu-ray disc and DVD) product views and purchases. The field experiment was run by the company using a state-of-the-art A/B/n testing platform. This platform implemented a session tracking technology whereby each visitor's IP address is recorded and given a unique visitor ID. Then visitors' behaviors are tracked over the period of the field experiment. This enables the website to track individuals' viewing logs and purchases over many days. Whenever new visitors access the website for the first time, they are randomly chosen to be in the control or one of the treatment groups. Upon clicking and viewing a particular item, the visitors are shown the appropriate recommender panel, as seen in Figure 2. Figure 2 is a collaborative filtering recommender based on views ("People who viewed this item also viewed"). Similarly, there is also a collaborative filtering based on purchases, as mentioned in Section 3.2. Users in the control group do not see this panel. At the end of the experiment, we have each consumer's view logs and purchase logs at the item level. The algorithms were retrained every three days to propagate the influence of users' purchase history multiple times over the period of the experiment. About half of the users in the dataset were returning users. The website provides its own movie categorization, but in order to make the genre categorization more robust, we categorize each movie using IMDB.com's[3] categorization. For each movie, we obtain IMDB.com's category information by asking at least three different users on Amazon Mechanical Turk ("Turkers") to provide primary genres from IMDB.com. Getting input from three Turkers ensures robustness, which is important in even this simplest look-

[3]World's top movie information website, according to Alexa rank.

up-and-copy/paste task. Table 2 shows each genre's product page views and purchase numbers. All data were anonymized to ensure privacy.

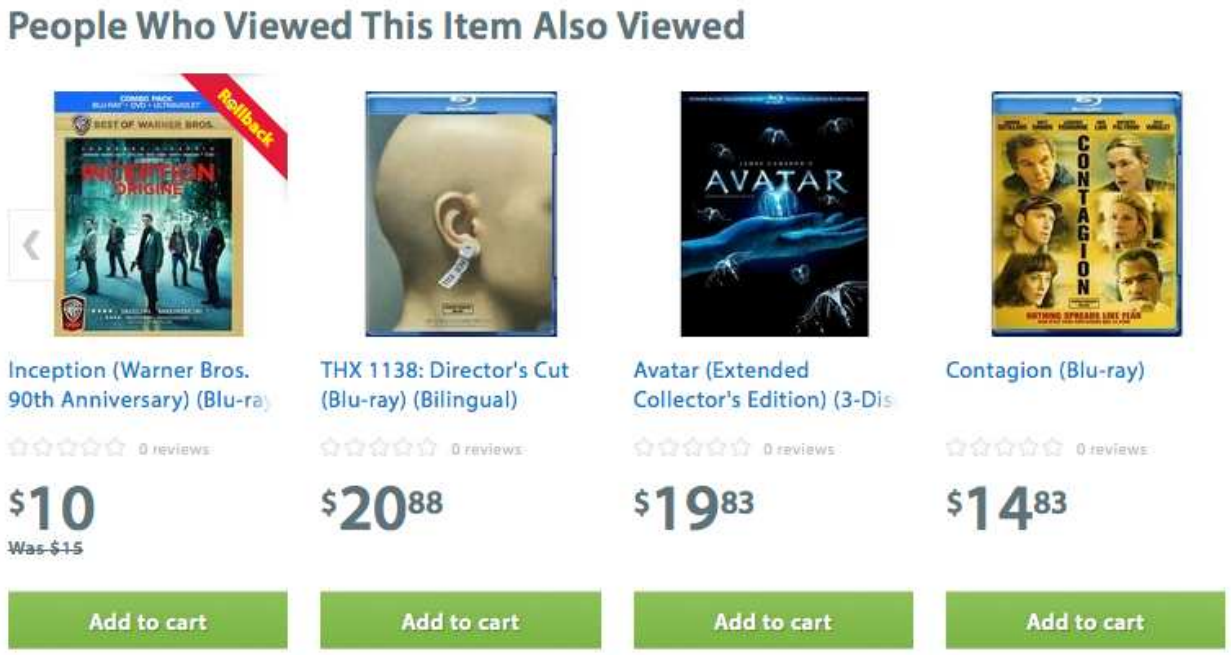


Figure 2: **Recommender Example:** Example of a recommender shown to a consumer. This consumer was in the treatment group of collaborative filtering based on views.

Viewed Genres						
Action	Adventure	Animation	Biography	Comedy	Crime	Documentary
625	114	557	21	565	41	48
Drama	Family	Fantasy	History	Horror	Music	Musical
465	321	65	1	96	4	29
Mystery	Romance	Sci Fi	Sport	Thriller	War	Western
9	38	169	181	39	17	17
Purchased Genres						
Action	Adventure	Animation	Biography	Comedy	Crime	Documentary
291	50	240	13	282	22	31
Drama	Family	Fantasy	History	Horror	Music	Musical
217	127	41	0	25	1	19
Mystery	Romance	Sci Fi	Sport	Thriller	War	Western
6	20	80	79	20	6	6

Table 2: **Movie Genres Viewed and Purchased**: This table shows the number of views and purchases in each movie genres in our dataset for those who've viewed or purchased movies (DVDs, Blu-ray discs) on this e-commerce site.

When the company ran the field experiment, it wanted to test the recommenders with a very

small fraction of its visitors since the company had never used recommenders on this particular website. Therefore, it randomly allocated 10% of its visitors to each collaborative filtering treatment group. Our analysis focuses on only those users who made at least one purchase during the study period. This is because it does not make as much sense to compute individual-level sales diversity measures (like the Gini coefficient) when no purchases have been made by the individual. There is no selection bias in this case since, ex ante, we randomized the treatment and control assignments for all visitors. A caveat to our study is that we investigate the effects on only those consumers who ended up making a purchase. Our resulting dataset has 572 unique users in the control group and about 70 to 90 who purchased movie-related products in each of the other groups. Multiple robustness checks that account for sample size differences and outliers are presented in Section 5.6, and all checks produced similar results.

Our field experiment dataset offers a clean way to tease out the causal impact of recommenders. However, it is lacking in that we do not consider all recommender designs used in practice, such as content-based recommenders. However, given that much of the debate in the recommendation systems literature relates to collaborative filters, we have the two most commonly used designs in our study.

5 Results

We have a total of three different groups in our field experiment, giving three unique pairs[4] to compare for each variable mentioned in Section 3.1. First, we present descriptive stats and summary results with visualization. Then, for each variable of interest, we present tables that show the difference in aggregate measures of interest and statistical significance associated with the differences in the measure. All results presented here are robust from sampling, outlier, and other influences, as will be discussed in Section 5.6.

5.1 Descriptive Results and Visualization

Table 3 presents descriptive summary and results of our data.

[4] 1) Control vs Purchase-based CF, 2) Control vs View-based CF, 3) Purchase-based CF vs View-based CF.

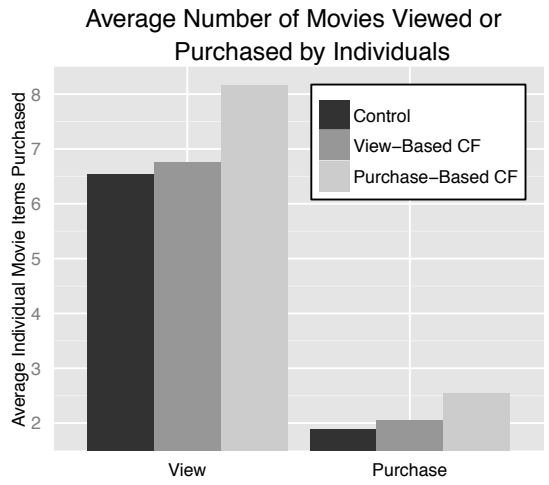
View	**Control**	**View-based CF**	**Purchase-based CF**
Unique Consumers Who Purchased Movie Products	572	73	95
Average Individual Movie Genres Viewed	1.65 (1.37)	1.67 (1.31)	1.84 (1.85)
Average Individual Movies Viewed	6.54 (8.64)	6.75 (7.39)	8.16 (13.70)
Number of Consumers Viewing More than 1 Movie	489	62	81
% of Consumers Viewing More than 1 Movie	0.851	0.849	0.852
Average Individual Movie Gini Coefficient	0.9988 (0.0016)	0.9988 (0.0011)	0.9983 (0.0045)
Average Individual Genre Gini Coefficient	0.9338 (0.0358)	0.9315 (0.0423)	0.9299 (0.0425)
Aggregate Genre Gini Coefficient	0.6262	0.6476	0.6772

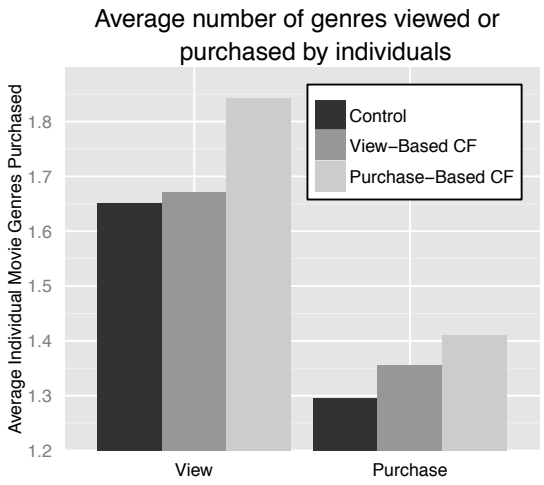
(a) **Summary Statistics for View**

Purchase	**Control**	**View-based CF**	**Purchase-based CF**
Unique Consumers Who Purchased Movie Products	572	73	95
Average Individual Movie Genres Bought	1.29 (0.85)	1.35 (0.75)	1.41 (1.20)
Average Individual Movies Bought	1.88 (2.29)	2.05 (1.41)	2.54 (7.83)
Average Individual Wallet Size	24.97 (32.01)	27.84 (24.59)	27.27 (51.67)
Number of Consumers Buying More than 1 Movie	186	35	34
% of Consumers Buying More than 1 Movie	0.325	0.479	0.357
Average Individual Movie Gini Coefficient	0.9985 (0.0018)	0.9984 (0.0011)	0.9979 (0.0062)
Average Individual Genre Gini Coefficient	0.9418 (0.0277)	0.9390 (0.0308)	0.9389 (0.0325)
Aggregate Genre Gini Coefficient	0.6076	0.6768	0.7044

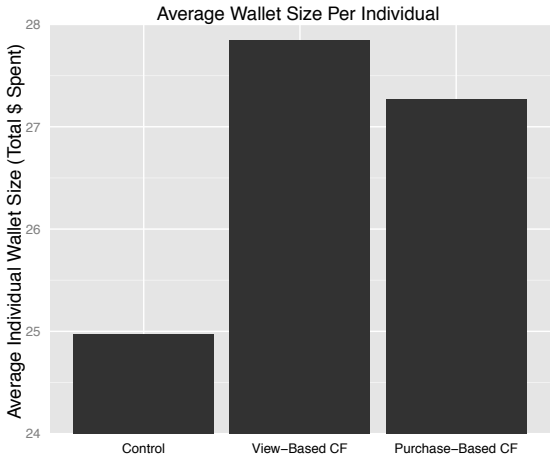
(b) **Summary Stat for Purchase**

Table 3: **Data Summary Statistics :** Standard deviation is in parentheses

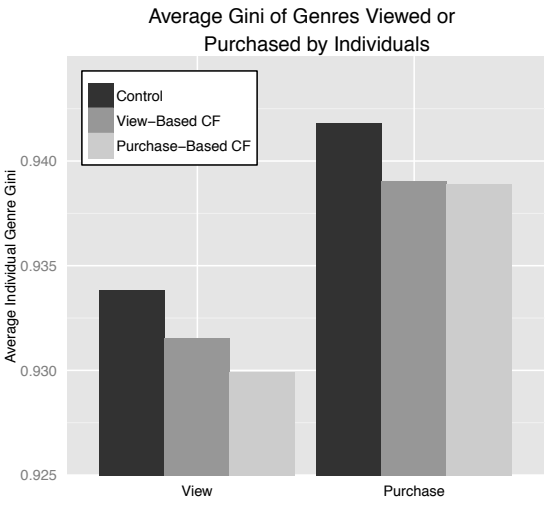




(a) **Average Individual Movies Viewed/Purchased**(b) **Average Individual Genres Viewed/Purchased**



(c) **Average Individual Wallet Size**



(d) **Average Individual Genre Gini**

Figure 3: **Average Individual Statistics:** These graphs visualizes the average individual number of movies viewed/purchased, wallet size (total $ spent), and the Gini measure of genres viewed/purchased.

Even at the summary data descriptive level, there are clear differences across the groups. Figure 3 shows summary statistics visually. Average individual movies and genres viewed are higher in the Purchase-based CF group (8.16 and 1.84, respectively) than in the other groups (around 6.5 and 1.65, respectively). This trend persists for purchases as well, with average individual movies at 2.5 and average individual genres being bought at 1.4, compared to other groups (ranging from 1.88-2.05 for individual movies bought and 1.29-1.35 for genres bought). The average individual wallet size is also different, with collaborative filtering groups spending more, at around 27-28 dollars per

person compared to 25 dollars per person in the control. Lastly, the summary-level Gini coefficients are also different. At the individual level, the Gini coefficient is lowest for the Purchase-based CF, suggesting that consumers' individual-purchase diversity is maximized with the Purchase-based CF. At the aggregate level, the Purchase-based CF Gini coefficient is the highest, suggesting a decrease in aggregate sales diversity. As we present later in this section, many of these differences are statistically significant. Figure 4 shows the Lorenz curve for aggregate firm-level sales diversity for movie genres, clearly illustrating the decrease in aggregate diversity for CF treatments. In the next section, we formalize the analysis with permutation test technique and provide group-level differences for each variable of interest with statistical significance.

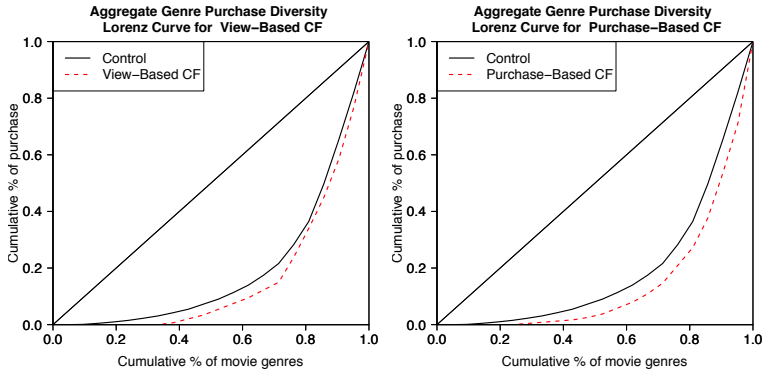


Figure 4: **Lorenz Curves for Movie Genres Purchased:** These Lorenz curves show that, on the genre level, firm-level aggregate sales diversity is decreased for both collaborative filtering algorithms.

5.2 Sales Volume Results

Individual Item Views Table 4 summarizes the results for the average number of items viewed per individual. The top rows present the average number of items viewed in each group and the difference relative to the control group. The remaining cells evaluate whether the difference between two groups is statistically significant or not. The test statistic is $D \equiv f(g_{row}) - f(g_{column})$ with f representing the average number of items viewed by individuals in a group; g_{row} refers to the group identified in that row. The test statistic is followed by the p-value associated with the value. The p-value is computed using a permutation test as outlined above. For example, consumers in the control group on average viewed 6.54 items, while View-based-CF-exposed consumers viewed 6.75 items. The difference between the two is -0.2070, and the p-value associated with this difference statistic D is 0.796. Consumers treated with the Purchase-based CF viewed 1.6 more items on average (a 25% lift) compared to the control, and this difference is statistically significant. The

Purchase-based CF also performs better than the View-based CF, which only lifts views by 3% compared to the control. It is clear that 1) CFs can increase exploration on e-commerce sites, and 2) there is a difference between CF algorithms in terms of their impact on product views. The success of the Purchase-based CF may be because they help broaden the consideration set by recommending other relevant alternatives or because they are effective at cross-selling other product categories. The former may not necessarily create new purchases even if it helps improve product match for consumers, whereas the latter should contribute towards new purchases.

Avg # of Items Viewed	6.5464	6.7534	8.1684
% Change from Control		3.1% ▲	24.7% ▲
	Control	View-based CF	Purchase-based CF
Control		-0.2070|0.796	**-1.6220|0.026**
View-based CF			-1.4149|0.402
		$Row - Column$|P-value	

Table 4: **Individual Item Views Comparison:** Individual items (movies) viewed averaged within each group. The top row shows the actual average items viewed in each group. Each cell shows a $ROW - COLUMN$ value with p-value (e.g., Average items viewed in control - average items viewed in View-based CF = -0.2070 and p-value obtained from permutation test is equal to 0.796).

Individual Item Purchases Table 5 presents the same information for item purchases. The impact of the treatments on purchases is directionally similar to their impact on views. The Purchase-based CF is the only group that is statistically significantly different from the control group. Consumers bought 0.66 more items on average (a 35% lift) under the influence of the Purchase-based CF than the control group. Consumers in the View-based CF group bought 0.17 more items on average (a 9% lift) than consumers in the control group (the difference is not statistically significant). The results again validate that CF can have a significant impact on purchases and that there are differences between CF algorithms. Further, we find that the Purchase-based CF is effective in creating new purchases.

Avg # of Items Purchased	1.8809	2.0547	2.5473
% Change from Control		9.2% ▲	35.4% ▲
	Control	View-based CF	Purchase-based CF
Control		-0.1738|0.342	**-0.6664|0.004**
View-based CF			-0.4925|0.932

Table 5: **Individual Item Purchases Comparison:** Individual items (movies) purchased averaged within each group.

Individual Wallet Size Table 6 shows that, while nothing is statistically significant in the individual wallet-size comparison, CF average wallet sizes were higher than the control's. Both CF groups yielded an average wallet size above $27 per person. The control group yielded the smallest average wallet size, coming up just below $25.

Avg Wallet Size	24.9761	27.8473	27.2794
% Change from Control		11.4% ▲	9.2% ▲
	Control	View-based CF	Purchase-based CF
Control		-2.8712|0.296	-2.3033|0.416
View-based CF			0.5679|0.946

Table 6: **Individual Wallet-Size Comparison:** Individual Wallet Size (total amount spent) averaged within each group.

Discussion Our results on the *sales volume impact* of recommenders are clear. Purchase-based CF exposes users to more items, and in fact, increases the sales volume. Specifically, Purchase-based CF drives a 25% lift in views and a 35% lift in the number of items purchased compared to the control group. In comparison, the View-based-CF group shows only a 3% lift in views and a 9% lift in the number of items purchased, which is not statistically significant. While the wallet-size analysis was not statistically significant, it suggests that CFs do increase the amount spent by consumers. It is highly likely that the lack of statistical significance is driven by the fact that purchases were infrequent and our data spans only a two-week period.

5.3 Sales Diversity Results

Aggregate View Diversity Table 7 shows the aggregate view diversity at the item and genre levels for the three groups. At the item level, both treated groups decrease in view diversity (i.e., increased Gini), and the results are statistically significant. Users as a whole view fewer items when shown collaborative filtering recommendations. At the genre level, only the Purchase-based collaborative filter decreased view diversity statistically significantly compared to the control group. The results stay the same when the analysis is repeated by first fixing the same number of randomly sampled users in each group instead of permuting with the entire sample.

Aggregate View Gini **Item**	0.6312	0.9374	0.8921
% Change from Control		48.5% ▲	41.3% ▲
	Control	View-based CF	Purchase-based CF
Control		**-0.3061|<0.001**	**-0.2608|<0.001**
View-based CF			-0.0452|0.1

(a) **Aggregate Item View Diversity Comparison**

Aggregate View Gini **Genre**	0.6262	0.6476	0.6772
% Change from Control		3.4% ▲	8.1% ▲
	Control	View-based CF	Purchase-based CF
Control		-0.0214|0.390	**-0.0509|0.042**
View-based CF			-0.0295|0.470

(b) **Aggregate Genre View Diversity Comparison**

Table 7: **Aggregate View Diversity**

Aggregate Purchase Gini **Item**	0.3972	0.8907	0.8235
% Change from Control		124.2% ▲	107.4% ▲
	Control	View-based CF	Purchase-based CF
Control		**-0.4935|<0.001**	**-0.4262|<0.001**
View-based CF			0.0672|0.306

(a) **Aggregate Item Sales Diversity Comparison**

Aggregate Purchase Gini **Genre**	0.6076	0.6768	0.7044
% Change from Control		11.3% ▲	15.9% ▲
	Control	View-based CF	Purchase-based CF
Control		**-0.0691|0.004**	**-0.0967|<0.001**
View-based CF			-0.0276|0.392

(b) **Aggregate Genre Sales Diversity Comparison**

Table 8: **Aggregate Sales Diversity**

Aggregate Sales Diversity Table 8 presents the first part of our main results on the *sales diversity impact*. Subtables 8a-8b show the impact of each recommender algorithm on aggregate sales diversity at the item and genre levels, respectively. At the item level, both treated groups show

a statistically significant reduction in aggregate sales diversity. In tandem with the results from the aggregate view diversity, we see that for both treated groups, consumers as a group explored and purchased a less wide variety of items. Users in the treated groups ended up buying the same broad-appeal items, leading to some level of herding. This popularity bias persists at the genre level. These results support the theoretical results presented in Fleder and Hosanagar (2009) that conjecture that typical collaborative filtering designs will show a popularity bias at the aggregate level because they make recommendations based on past purchases and/or views.

Tables 7 and 8 also show some differences in terms of how recommenders influence item diversity versus genre diversity. Collaborative filtering algorithms seem to cause heavier shifts in item diversity than in genre diversity both in percentage and absolute terms. This is in part due to the number of items, which are order of magnitude greater than the number of genres, thereby amplifying the effect of popularity bias of recommenders. Collectively, these tables suggest that CF algorithms change view and purchase diversity both at the item and genre level, but more so at the item level.

Individual View Gini **Item**	0.9988	0.9988	0.9983
% Change from Control		0%	0.05% ▼

	Control	View-based CF	Purchase-based CF
Control		0.00001|0.934	**0.0004**|<**0.001**
View-based CF			0.0004|0.482

(a) **Individual Item View Diversity Comparison**

Individual View Gini **Genre**	0.9338	0.9315	0.9299
% Change from Control		0.2% ▼	0.4% ▼

	Control	View-based CF	Purchase-based CF
Control		0.0022|0.408	0.0039|0.164
View-based CF			0.0016|0.814

(b) **Individual Genre View Diversity Comparison**

Table 9: **Individual View Diversity**

Individual View Diversity Table 9 presents results for individual view diversity. The only statistically significant result is at the item level with the Purchase-based CF algorithm. The Purchase-based CF causes individual view diversity to *increase* (i.e., the Gini coefficient is lower

than that of the control). Unlike aggregate view diversity, which *decreases,* individual view diversity *increases*. This means that individuals view a greater variety of items while consumers as a group view a more limited variety of items. While puzzling, this result is explained by Fleder and Hosanagar's (2009) theory that recommenders "can push each person to new products, but they often push users toward the same products" because one person's product views or purchases feed into recommendations made to another user.

Individual Purchase Gini **Item**	0.9985	0.9984	0.9979
% Change from Control		0.01% ▼	0.06% ▼

	Control	View-based CF	Purchase-based CF
Control		0.0001|0.426	**0.0005|0.002**
View-based CF			0.0004|0.946

(a) **Individual Item Purchase Diversity Comparison**

Individual Purchase Gini **Genre**	0.9418	0.9390	0.9389
% Change from Control		0.2% ▼	0.3% ▼

	Control	View-based CF	Purchase-based CF
Control		0.0028|0.184	0.0029|0.206
View-based CF			0.00008|0.996

(b) **Individual Genre Purchase Diversity Comparison**

Table 10: **Individual Purchase Diversity**

Individual Purchase Diversity Table 10 shows the results for individual purchase diversity. Similar to the individual view diversity, individual purchase diversity increases (i.e., the Gini coefficient decreases), suggesting that individuals buy a greater variety of items because of exposure to Purchase-based collaborative filtering recommenders. The directions for genre are the same but are not statistically significant.

Looking at Tables 7-10, we see a clear pattern emerging. Under the influence of the Purchase-based CF, individuals view and buy a greater variety of items than under no recommenders, but they collectively discover and buy a similar set of items, leading to a decrease in aggregate view and sales diversity. The View-based CF shows similar results but the effect sizes are smaller and sometimes not statistically significant. This suggest that, even within collaborative filtering algorithms, there are differences. Our results further show that recommenders influence both the consideration set

(views) as well as the eventual conversion (purchase).

5.4 Results Summary

Table 11 summarizes our results and hypotheses supported. Our results confirm anecdotal and industry reports of the *sales volume impact* of the recommenders, specifically for the collaborative filtering algorithms. Collaborative filtering increases both the number of product (in this case, DVD and Blu-ray) views and the number of items purchased, suggesting that these recommenders can successfully engage users in additional search and exploration and eventually additional purchases. Further, we show that not all collaborative filtering algorithms perform the same. In our setting, Purchase-based collaborative filters had a more significant impact on sales than View-based filters.

Regarding the *sales diversity impact* of recommenders, we have shown that collaborative filtering causes individuals to discover a greater variety of products but pushes consumers to the same set of titles, leading to concentration bias at the aggregate level. Here again, we find that not all collaborative filtering algorithms are the same. Purchase-based collaborative filters have a greater impact. This may be because: 1) the algorithm based on purchases might simply be better at delivering the best fit products, 2) consumers might be more influenced by the "purchased also purchased" signal than the "view also viewed" signal, 3) or both. However, we lack the access to detailed recommender data and consumer information to deeper investigate and quantify which of these effects are in play[5]. However, in the next section, 5.5, we investigate the source of diversity shift at the genre level, providing further insight into how the increase in individual diversity and the decrease in aggregate diversity can occur simultaneously.

Hypothesis	**Supported**	**Statistically Significant**	**Reference Table**
1: CF Increases product views	YES	YES	4
2: CF increases the # of product purchases	YES	YES	5
3: CF decreases aggregate product sales diversity	YES	YES	8
4: CF increases individual product consumption diversity	YES	YES	10

Table 11: **Hypotheses Tested**

[5]We believe that investigating and quantifying the two underlying recommender mechanism is a promising and interesting line of study.

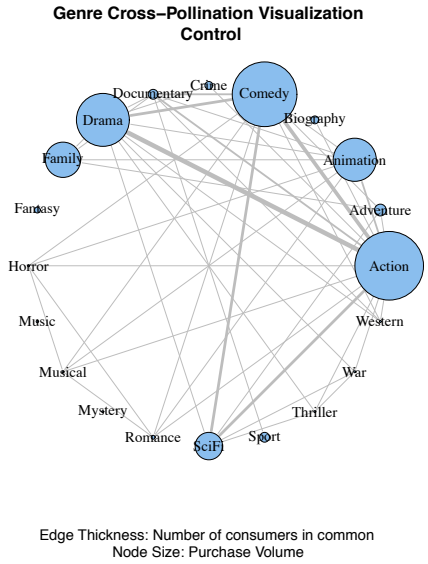
5.5 The Source of Diversity Shift: Genre Cross-pollination Investigation

In this section, we attempt to visualize our results and also understand from where the shift in aggregate and individual diversity stems. To do so, we construct sample-size normalized genre-level co-purchase networks for both Purchase-based CF and the control. We create network graphs in which each node in the graph represents a movie genre, and the size of the node is proportional to the *percent* of overall sales that went to the genre. An edge between two nodes indicates that there were users who purchased from these genres, and the thickness of the edge is proportional to the number of such users who exist. Thus, the relative sizes of the nodes convey the extent to which sales were (un)evenly distributed at the aggregate, and the edges convey the extent to which individuals explored content in diverse genres. Figure 5 compares two network graphs side by side.

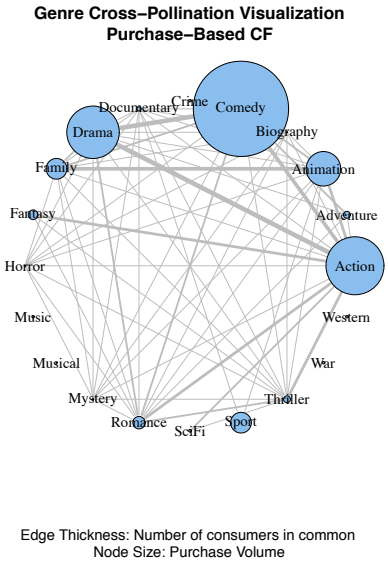
On visually comparing the two graphs, we note the following:

1. Relative size of the nodes show that the majority of purchases by the control is distributed across a few genres: action, drama and comedy. In the Purchase-based CF, however, comedy is much bigger than the rest, indicating that purchases were more concentrated in comedy. This might be because comedy titles were recommended more often by the recommendation algorithms or because consumers are more willing to explore and trust the recommender for comedy (perhaps due to less heterogeneity in taste across the users).

2. The Purchase-based CF graph is much better connected (i.e., denser) than the control. This indicates that there are more users who are buying different genres in the Purchase-based CF group, or, more specifically, there is greater individual cross-buying behavior. The connectedness of the Purchase-based CF graph reflects the increase in individual diversity that we noted previously. Individual users may be exploring more genres while sales may be simultaneously concentrated in a few genres at the aggregate level.



(a) **Control**



(b) **Purchase-based Collaborative Filtering**

Figure 5: **Co-Purchase Network Graphs of Genre Purchases under Control and Purchase-Based Collaborative Filtering.**

Figure 6 presents the frequencies of genre purchases in the control set against its frequencies in the treatment set (Purchase-based CF) to show the shift-to-comedy effect caused by the recommender. If a recommender has no influence on genre cross-pollination, we expect to see the genre on or near the line (which has slope = 1). The comedy genre is distinctively away from and above the line, showing that this recommender pushed consumers to buy more comedy. This is interesting since according to Table 2, action has the highest number of views and purchases, suggesting that this is not merely a volume effect.

	Top 1	Top 5	Top 10	All Genres
	Purchase-based CF Stat - Control Stat | P-value			
Top Genres Market Size Difference	**0.110|0.036**	**0.095|0.042**	**0.081|0.004**	Not meaningful

Table 12: **Permutation Test Results for Co-purchase Network Comparisons: Purchase-based CF vs. Control**

To formally test these differences, we use the same permutation technique to evaluate the market size of the top genres in each graph (market size of top N nodes). Table 12 shows the difference between the market-size statistics for the Purchase-base CF and control group as well as the corresponding p-values obtained via permutation tests. We replicated the analysis for top {1, 5, 10} genres. We see a clear shift to top genres in Purchase-based CF. Under the Purchase-based CF, the

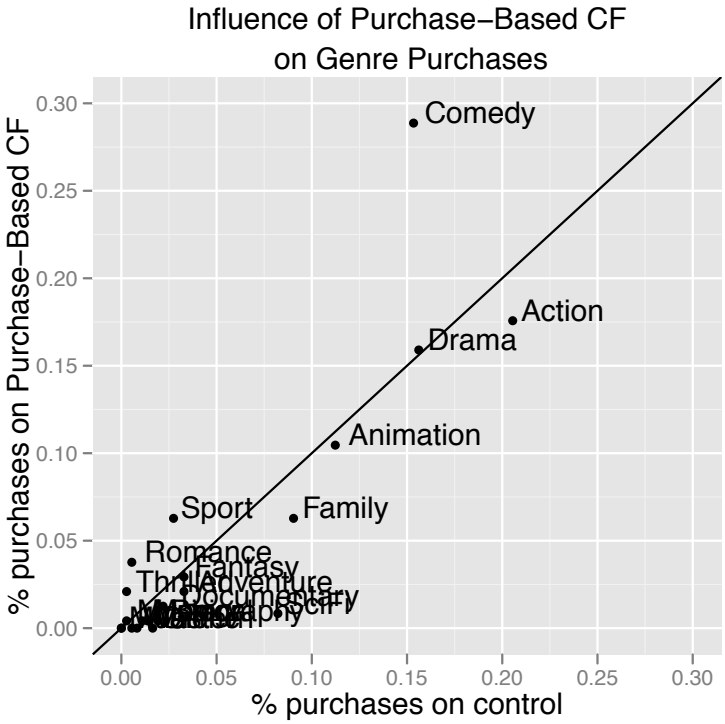


Figure 6: **Genre Purchase Share Comparison on Purchase-based CF vs. Control**

top genre, comedy, took 11% more market share compared to the top genre in the control group, action. We also see a decrease in top genres market size differences (i.e., 0.11 -> 0.095 -> 0.081) as we include more number of genres, suggesting higher concentration of purchases for the top genres. In summary, a Purchase-based collaborative filtering algorithm shifts users to buy a few top genres at the aggregate level while increasing individual diversity through a cross-buying behavior that is aided by a few 'pathway' genres.

5.6 Robustness Checks & Other Measures

We also ran a series of robustness checks in regards to consumer samples. They include the following:

R1 Our analysis thus far is based on treatment groups of different sizes. We replicated the analysis by randomly sampling a fixed number of users in each group, ex ante before the permutation test, so that each group has an equal number of consumers.

R2 We removed the outlier consumers in terms of number of views and purchased items.

We replicated the analysis excluding those users whose views or purchases exceeded 3 standard deviations from the mean.

R3 We replicated the analysis by using 1 or 0 (binary variable) for genre or items viewed (or purchased) instead of actual counts. This allows us to explore the notion of diversity from a perspective of number of unique items or genres viewed/purchased as opposed to proportion of sales.

R4 We replicated the analysis only on consumers who bought more than one item.

R5 Some groups (e.g. Purchase-based CF) see more views and purchases than others. In theory, this increase in volume should not affect sales diversity measures such as the Gini coefficient. To confirm this, we replicated the analysis after "Volume-Equalization" in which the number of items bought by individuals was normalized across groups to get rid of any effect from volume influencing the Gini coefficient.[6]

In all of these robustness checks, the findings are qualitatively similar to our main result. Table 13 presents the details of hypotheses supported under the different robustness checks.

	Hypotheses Supported under Robustness Checks				
Hypothesis	**R1**	**R2**	**R3**	**R4**	**R5**
1: CF Increases product views	BOTH	BOTH	DS-	BOTH	NA
2: CF increases the # of product purchases	BOTH	BOTH	BOTH	BOTH	NA
3: CF decreases aggregate product sales diversity	BOTH	BOTH	BOTH	BOTH	BOTH
4: CF increases individual product consumption diversity	BOTH	BOTH	BOTH	BOTH	DS+

Table 13: **Hypotheses under Robustness Checks** : "Both" means that both directions and statistical significances were identical to the presented main results. "D" represents that direction was reproduced. "S-" signifies the loss of statistical significances, and "S+" signifies the gain of statistical significance. Lastly, "NA" indicates not applicable due to the nature of the robustness test (e.g., "Volume-Equalization" necessarily gets rid of volume differences).

6 Discussion and Conclusion

With the advent of big data, recommenders and personalization technologies are fast taking over nearly every aspect of the web. Their use spans from the purchase of physical products (books,

[6]We followed the method used in Hosanagar et al. (2014).

DVDs, clothing, electronics, etc) to digital media (movies, news), and even online services such as dating and peer-to-peer lending. Despite their ubiquity, we still have much to learn about how different recommender algorithms influence markets and society.

Our study contributes to an emerging literature on the impact of personalization technologies by studying the impact of recommender algorithms on sales volume and diversity with movie products. We have three main findings. First, we show that recommenders have a positive impact on sales, which corroborates anecdotal evidence and prior findings in the literature. Second, we provide direct evidence from the field that collaborative filtering algorithms increase individual consumption diversity while decreasing aggregate consumption diversity and explain where this shift in diversity stems from at the genre-level. Third, for both sales volume and diversity, we show that different algorithms have different impact, extending baseline results for sales volumes while providing new insights for sales diversity. Furthermore, the result highlights a potential limitation in the ability of traditional collaborative filtering to aid discovery of truly niche items and genres. We reveal that users increase their purchase and view diversity by exploring into few top 'pathway' genres like comedy. Our study also helps resolve an ongoing debate among researchers on the impact of recommenders on sales diversity by providing an evidence from a randomized field experiment with individual view and purchase level data.

These results have significant managerial relevance. As the amount of consumer data available to firms grows exponentially, many retailers have aggressively adopted data mining and personalization technologies without deeply understanding how different designs may contribute toward (or deter) broader strategic goals. For example, a firm interested in exposing consumers to a broader assortment of products may prefer a different design from another simply interested in maximizing sales. To the extent that a firm is interested in pushing the "back catalog," it may seek to augment traditional collaborative filtering algorithms so that it is possible to identify relevant products with limited historical data (past views/purchases) and/or increase diversity, serendipity, or novelty of the recommended products using techniques from the extant literature (e.g., Adamopoulos and Tuzhilin (2013); Adomavicius and Kwon (2013); Oh et al. (2011)).

On a more general level, many firms have adopted theory-free predictive analytics approaches without trying to understand what drives changes caused by the online technologies that they have implemented. As data grows, this approach will increasingly be prone to issues tied to spurious correlations and a decrease in the signal-to-noise ratio. One promising alternative is to use theory-

driven causal inference techniques to discern true effects. This is one of the first randomized field experiment explicitly looking at the differential effects of different personalization algorithms on sales volume and diversity. We look forward to additional studies documenting these differences in other field settings.

We conclude by discussing some limitations of our study and opportunities for future work. Our study focused on the two most commonly used collaborative filtering designs. It is worth investigating the impact of other recommender designs, such as content-based and social-network-based recommenders. Second, we evaluated the impact of collaborative filters in one product category, namely, movies. A promising extension of our work will be in creating an empirical generalization by including other product categories such as apparel and electronics in the study and investigating specific product characteristics that influence the sales volume and diversity. Third, a valuable addition to our work will be studies that develop consumer behavior theories on how and why people react differently to different recommender systems and signals. Lab studies can be highly valuable in this regard.

References

Accenture: 2012, 'Today's Shopper Preferences: Channels, Social Media, Privacy and the Personalized Experience'. Technical report, Accenture Interactive.

Adamopoulos, P. and A. Tuzhilin: 2013, 'On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected'. *ACM Transactions on Intelligent Systems and Technology* **1**(1).

Adomavicius, G. and Y. Kwon: 2013, 'Optimization-Based Approaches for Maximizing Aggregate Recommendation Diversity'. *Journal on Computing* **26**(2), 351–369.

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

Anderson, C.: 2008, *The long tail: Why the future of business is selling less of more*. Hyperion Books.

Ansari, A., S. Essegaier, and R. Kohli: 2000, 'Internet Recommendation Systems'. *Journal of Marketing Research* **37**(3), 363–375.

Bikhchandani, S., D. Hirshleifer, and I. Welch: 1992, 'A theory of fads, fashion, custom, and cultural change as informational cascades'. *Journal of political Economy* pp. 992–1026.

Bodapati, A. V.: 2008, 'Recommendation Systems with Purchase Data'. *Journal of Marketing Research* **45**(1), 77–93.

Brynjolfsson, E., Y. J. Hu, and D. Simester: 2011, 'Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales'. *Management Science* **57**(8), 1373–1386.

Brynjolfsson, E., Y. J. Hu, and M. D. Smith: 2006, 'From niches to riches: The anatomy of the long tail'.

Celma, Ò. and P. Cano: 2008, 'From hits to niches?: or how popular artists can bias music recommendation and discovery'. In: *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. p. 5.

Das, A. S., M. Datar, A. Garg, and S. Rajaram: 2007, 'Google news personalization: scalable online collaborative filtering'. In: *Proceedings of the 16th international conference on World Wide Web*. pp. 271–280.

De, P., Y. J. Hu, and M. S. Rahman: 2010, 'Technology usage and online sales: An empirical study'. *Management Science* **56**(11), 1930–1945.

Dias, M. B., D. Locher, M. Li, W. El-Deredy, and P. J. Lisboa: 2008, 'The value of personalised recommender systems to e-business: a case study'. In: *Proceedings of the 2008 ACM conference on Recommender systems*. pp. 291–294.

Econsultancy and Monetate: 2013, 'The Realities of Online Personalization'. Technical report, Econsultancy and Monetate.

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

Good, P.: 2005, *Permutation, parametric and bootstrap tests of hypotheses*. Springer.

Hinz, J. D. O. and D.-K. J. Eckert: 2010, 'The impact of search and recommendation systems on sales in electronic commerce'. *Business & Information Systems Engineering* **2**(2), 67–77.

Hosanagar, K., D. Fleder, D. Lee, and A. Buja: 2014, 'Will the Global Village Fracture into Tribes: Recommender Systems and Their Effects on Consumers'. *Management Science* **60**(4), 805–823.

Jannach, D. and K. Hegelich: 2009, 'A case study on the effectiveness of recommendations in the mobile internet'. In: *Proceedings of the third ACM conference on Recommender systems*. pp. 205–208.

Jannach, D., L. Lerche, F. Gedikli, and G. Bonnin: 2013, 'What recommenders recommend – An analysis of accuracy, popularity and sales diversity effects'. *User Modeling, Adaptation, and Personalization - Lecture Notes in Computer Science* **7899**, 25–37.

Lee, Y., Y. Tan, and K. Hosanagar, 'Do I Follow My Friends or The Crowd? Information Cascades in Online Movie Rating'. *Working Paper*.

Marshall, M.: 2006, 'Aggregate Knowledge raises 5 Million dollar from Kleiner, on a roll'. *Venture Beat*.

Matt, C., T. Hess, and C. Weiß: 2013, 'The Differences between Recommender Technologies in their Impact on Sales Diversity'.

Monetate: 2013, 'Maximize Online Sales with Product Recommendations'. Technical report, Monetate.

Muchnik, L., S. Aral, and S. Taylor J.: 2013, 'Social Influence Bias: A Randomized Experiment'. *Science* **341**(6146), 647–651.

Oestreicher-Singer, G. and A. Sundararajan: 2012, 'Recommendation networks and the long tail of electronic commerce'. *MIS Quarterly* **36**(1), 65–84.

Oh, J., S. Park, H. Yu, M. Song, and S.-T. Park: 2011, 'Novel Recommendation Based on Personal Popularity Tendency'. In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. pp. 507–516.

Salganik, M. J., P. S. Dodds, and D. J. Watts: 2006, 'Experimental study of inequality and unpredictability in an artificial cultural market'. *Science* **311**(5762), 854–856.

Schafer, J. B., J. Konstan, and J. Riedi: 1999, 'Recommender systems in e-commerce'. In: *Proceedings of the 1st ACM conference on Electronic commerce*. pp. 158–166.

Thompson, C.: 2008, 'If you liked this, you're sure to love that'. *The New York Times* **21**.

Tucker, C. and J. Zhang: 2011, 'How does popularity information affect choices? A field experiment'. *Management Science* **57**(5), 828–842.

Victor, P., C. Cornelis, and M. De Cock: 2011, *Trust networks for recommender systems*, Vol. 4. Springer.

Wu, L.-L., Y.-J. Joung, and T.-E. Chiang: 2011, 'Recommendation Systems and Sales Concentration: The Moderating Effects of Consumers' Product Awareness and Acceptance to Recommendations'. In: *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. pp. 1–10.

Zhou, R., S. Khemmarat, and L. Gao: 2010, 'The impact of YouTube recommendation system on video views'. *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*.