When Does Social Network-based Prediction Work? A Large Scale Analysis of Brand and TV Audience Engagement by Twitter Users

Completed Research Paper

Shawndra Hill
The Wharton School
University of Pennsylvania
3730 Walnut Street, Philadelphia, PA
shawndra@wharton.upenn.edu

Adrian Benton
The Wharton School
University of Pennsylvania
3730 Walnut Street, Philadelphia, PA
adrianb@wharton.upenn.edu

Christophe Van den Bulte
The Wharton School
University of Pennsylvania
3730 Walnut Street, Philadelphia, PA
vdbulte@wharton.upenn.edu

Abstract

Social network-based prediction, more specifically targeting friends and contacts of existing customers, has proven successful in various domains like retail banking, telecommunications, and online advertising. However, little is known about for what types of product categories and brands social network-based marketing is especially effective at predicting brand engagement, both in absolute terms and compared to demographic targeting or collaborative filtering. In this work, we compare the performance of a social network-based recommendation engine against a product network-based recommendation engine of the kind used in collaborative filtering. We do so over 700 brands and 223,000 consumers a novel data set collected from Twitter. We compare the performance of the two approaches by product and user features. Preliminary results indicate that the variance in performance within and across methods is related to differences in brand and user popularity as well as brand audience. We believe that this is the first study to compare the effectiveness of social network-based marketing with traditional approaches to predict brand engagement over a large number of brands and product types.

Keywords: Social Media, Social Networks, Social TV, Recommendation Engines

Introduction

Recent increases in the quality and quantity of available social media data has enabled both product and social networks to be linked to user attributes such as demographics, and to business outcomes such as purchasing and responses to advertising and fraud. While the increasing size of the social media dataset is significant for research, the availability of this data through the Application Programming Interfaces (APIs) of sites such as Twitter, Facebook, and YouTube has also made it easier to derive value from social networks. This type of data enables researchers and practitioners to determine the features of users' social networks and use these to predict characteristics of other users. In this study, we explore the extent to which we can take user attributes, such as how many of a user's friends follow a particular brand, and use them to predict which brands the user will follow on Twitter. This research predicts behavior for hundreds of thousands of users regarding hundreds of brands.

Online social networks have been studied in previous literature to explore a wide range of research questions in a wide variety of fields that relate to our work. In Information Systems literature, the most notable work examines the spread of information and its influence on social networks (Aral et al. 2009). The difficulty of identifying influence in observational data is widely accepted. It is difficult to separate this influence from confounding factors of product and brand adoption such as homophily (McPherson et al. 2001), which is the concept that similar people cluster together due to influence, or contagion, the idea that people influence others to take certain actions (Shalizi et al. 2012). Recent work on influence is highly related to our study because it links social network features to product adoption outcomes, although researchers tend to focus on identifying influence, in particular, distinguishing it from homophily.

Rather than taking the same well-trodden path, we focus in this paper on social network-based prediction, which does not rely not on knowing how and why people are connected, but instead on the network structure and demographics of the brand audience. We exploit the fact that observers can know how brand preferences are correlated among friends because these preferences are visible online. Our goal is to highlight differences in our ability to predict for different brands based on both brand and user characteristics. In addition, we compare the social network (a network of friends) and the product network (a network of products connected through consumers who are not necessarily friends), both of which are in play for most online firms. Recent work in Information Systems and Marketing has examined the importance of both product and consumer networks in searching for information (Goldenberg et al. 2012) and content on the web. In this work, we compare the two networks' value for predicting brand preferences.

Large scale social network-based prediction — predicting individuals' attributes based on those of their friends (Domingos et al. 2001) — is a relatively new application of available data. Social network-based prediction has been successful in domains such as targeted marketing telecommunications services (Hill et al. 2006), online targeted advertising (Provost et al. 2009), the adoption of online services (Aral et al. 2009a), online searches (Goldenberg et al. 2012), and fraud detection (Hill et al. 2006). However, earlier studies have investigated only single products in one context at a time. In addition, although researchers did sometimes have access to information about different products or business outcomes, social network data for users was extremely limited. Therefore, there is very little knowledge about which product types and services are conducive to accurate social network-based predictions. The primary research question in this study is: *When does social network prediction work?*

To answer this question, we compare the predictive performance of a social network-based recommendation engine across multiple product and user categories. To achieve this comparison, we tasked this recommendation system with predicting which of more than 600 brands in 15 industry categories are followed by Twitter users. To ensure that the comparison is not overly specialized and is broadly applicable, we explore a broad range of categories, both of users and of brands, and apply a product network-based recommendation algorithm, used primarily as a benchmark, and a social network-based recommendation algorithm. We then compare the results to determine the effectiveness of both

these approaches' predictions of which brands and TV shows Twitter users follow. To do this, we first compare the results by size of the users' social network and by the size of the product's network. We then compare the results by the brands' industries and audience — in particular, whether the brands have a target "niche" demographic audience.

Our main contribution is an exploration of the performance of social network-based predictions across numerous brands and product categories as compared to more traditional approaches. We find that the difference in performance between the two can be explained by a number of factors, including the nicheness (how skewed the audience is to a particular demographic) and size of both the brand audience and users' social networks. Previous work on social networks explored homophily, the idea that similar people are connected. Our results confirm the importance of homophily in predicting a brand's followers, because brand audiences have particular demographics. The social network-based approach performs extremely well in this respect, as long as the brand has a significant number of users or followers, in the case of Twitter. In addition, we find that our social network-based approach makes more accurate predictions across industry categories than a traditional, collaborative filtering product network-based approach, which is better at making predictions within categories. More specifically, we find the following differences between the two approaches for user and product features:

1. ***User feature:*** Social network-based prediction dominates the product-based approach baseline when users have large social networks.

2. ***Product feature:*** Social network-based prediction dominates the product-based approach baseline when the product being recommended is popular.

3. ***Product feature:*** Social network-based prediction dominates the product-based approach baseline when the general product category being recommended has a "niche" demographic audience – for example, children's products that appeal to young parents.

4. ***Product feature:*** Social network-based prediction dominates the product-based approach baseline when the individual product being recommended has a "niche" demographic audience – for example energy drinks that appeal to young people.

5. ***Product feature:*** Social network-based prediction dominates the product-based approach when making predictions across product categories while the product network-based approach dominates for within category predictions.

6. ***User and Product features combined:*** We build a logit model to determine whether the social network-based prediction would perform best with respect to recall and find that the age skew of the brand of the product being predicted is positively correlated with the performance of the social network-based approach. We also find that strong, positive relationships exist between the predicted brand's age skew and the social network method's performance, as well as between the popularity of the output brand and the interaction between the user's and the output brand's number of followers.

Our work has both research and practical implications. To our knowledge, no other work has yet demonstrated the importance of context in making this type of prediction, when by context we understand the features of the item, brand or attribute which is being predicted for a networked user. , to a specific degree of accuracy. We find in our study that the social network-based prediction works extremely well for certain brands, while others benefit from the use of product network-based prediction. We also demonstrate that beyond this difference in brand performance, performance also varies by user type.

Background

Homophily

Recent work in network-based marketing (Hill et al. 2006) has summarized theoretical reasons that

might explain the correlation of social network neighbors' product preferences. These include both social influence and word of mouth. A social network may influence consumers in at least two ways: firstly, via the explicit advocacy of individual consumers, who by word of mouth spread information and comments about products, both online and offline. Secondly, there is the implicit advocacy of individuals who do not explicitly speak about products but instead reveal their engagement with a product by conspicuously adopting it. Examples of such adoption are consumers wearing specific brand logos, or revealing their interest in a brand by following it on Twitter. Both of these forms of advocacy may influence a consumer's friends or even the general population in their choice of product purchase. Beyond this, correlated preferences may be due to homophily (McPherson et al. 2001), the theory that similar people connect to one another. This theory alters the direction of influence: rather than being interested in a brand because their friends are, consumers may be friends because they are drawn to similar products -- that is, they share characteristic interests as well as other characteristics. The predisposition towards given products is seen as the cause of friendship, rather than friendship causing such a predisposition (Manski 1993). Regardless of whether connected users' preferences and characteristics are due to influence, homophily or both, social network-based prediction models rely on the similarity of these preferences and characteristics between friends.

In this paper, we explore to what extent the theory of homophily makes products targeting specific demographics are more amenable to social network-based prediction. If homophily is indeed a prime human characteristic, then social network-based targeting should perform best when a brand's product audience is of a specific demographic.

Social network-based Prediction

Network classification models use knowledge of the links between entities in a network to estimate a quantity of interest for those entities (Hill et al. 2006). Network-based classification methods rely on the fact that linked entities are similar. (Macskassy et al. 2007) provide a brief survey of different network classification methods. While there are many prediction algorithms that rely on making predictions based on the features of social network neighbors, most methods have not been applied to large consumer data. To remedy this absence of hard data, researchers have instead relied on building implicit networks of consumers based on shared interests (Domingos et al. 2001). Data limitations have made it difficult to test social-network based predictions both on real network social data and for many different products, brands or services. In this paper, we are able to move beyond this limitation, and compare the predictions of an existing social network-based approach across many contexts.

Recommendation Engines

Recommendation systems, used extensively by online firms incorporate different strategies to recommend products that consumers like. After surveying the RSs literature, (Adomavicius et al. 2005b) found that most RSs could be classified as one of three types: content-based (CB), collaborative filtering (CF), and hybrid (H). Content-based systems make recommendations by finding items with a high degree of similarity to consumers' preferred items, with those preferences generally being inferred through ratings or purchases (Pazzani et al. 2007). CF systems base item recommendations on historical information drawn from other users with similar preferences (Breese et al. 1998). Using collaborative filtering in a RS makes it possible to overcome some of the limitations of content-based systems because information about the products does not have to be codified at all, but this approach suffers from the new item problem - that is, the difficulty of generating recommendations for items that have never been rated by users and therefore have no history. It is known that collaborative filtering approaches favor recommending popular products. The hybrid approach combines collaborative- and content-based methods in various ways (Soboroff et al. 1999) to eliminate the shortcomings of both approaches and to improve prediction accuracy.

Improving prediction accuracy for recommendations is the subject of many papers (Adomavicius et al. 2005b) . Improvements in prediction performance have been achieved by tweaking both the algorithm and by incorporating new types of data on which to build models. For example, recent explorations have included contextual data to make recommendations (Adomavicius et al. 2005a) . However, little research has been performed to incorporate large-scale explicit social network data because the data was not

available for research in the past. In this work, we will explore the possibility of using social network data for recommendations across a wide variety of product types and large number of users. In the next section we will discuss the data that enables us to move beyond the limitations of prior research.

Data

We collected a large database of more than 700 brands and TV shows in 15 different industries. We then tracked all the Twitter followers of each brand in our database and created a sample of these followers by using their first-degree social networks. We then collected brand-related content from Facebook in order to link the brands to specific audiences based on demographics. In our results section, we will discuss how we assess the performance of recommendations for users based on different aspects of the brands, including audience demographics. The steps for data collection and the descriptive statistics and plots are described below.

Table 1. Demographic features collected from the Facebook Ad API and their mean proportion in the collection of brands

Demographic dimension	Demographic feature	Mean proportion
Gender	Men	37%
	Women	63%
Age	13–17	8%
	18–20	15%
	21–24	18%
	25–34	23%
	35–49	22%
	50–54	6%
	55–64	6%
	65+	3%
Education level	In high school	9%
	In college	13%
	Graduated college	78%
Family status	Is a parent	41%
Ethnicity	Hispanic	8%
	Non-Hispanic	92%

The dataset was collected through the following process: **(1)** A set of 734 widely recognized brands across 15 product types was identified from various online sources. **(2)** Next, 631 Twitter handles were found for brands that had a social media presence on Twitter. If a brand had multiple Twitter handles, we chose the handle with the most followers. **(3)** Using the Twitter API, the follower network of all 631 brands was collected, resulting in a network of approximately 18 million brand followers. **(4)** Random samples were created from each brand's follower network. Only users who had between 1 and 2,000 followers were considered in this network in order to avoid capturing celebrities', brands', and companies' handles, which tend to have massive numbers of followers. In addition, all users were required to follow at least

two of the 631 brands. For each brand in our dataset, a user meeting these criteria was selected at random from that brand's network. We selected approximately 500 users per brand. This process was repeated for all brands in our data, resulting in a sample of 223,517 users. **(5)** For all sampled users, the Twitter API was queried to collect the user identifications of their Twitter friends and followers. Thus, in addition to the network of brand followers, we also constructed the users' first-degree networks. **(6)** For each of these 631 brands, the Facebook Ads API was queried in order to retrieve aggregate-level demographic characteristics of the Facebook users who liked these brands. The demographic features were collected for 624 brands that had both a Facebook and Twitter following. During this process, we advertised our lab homepage on Facebook in order to gain access to the aggregate level demographic features. The aggregate-level demographic features and their mean proportions are listed in Table 1. The features were limited to those available from Facebook. For example, only 'Hispanic' was available as a race category, so we cannot use other races to compare recommendation systems using these data.

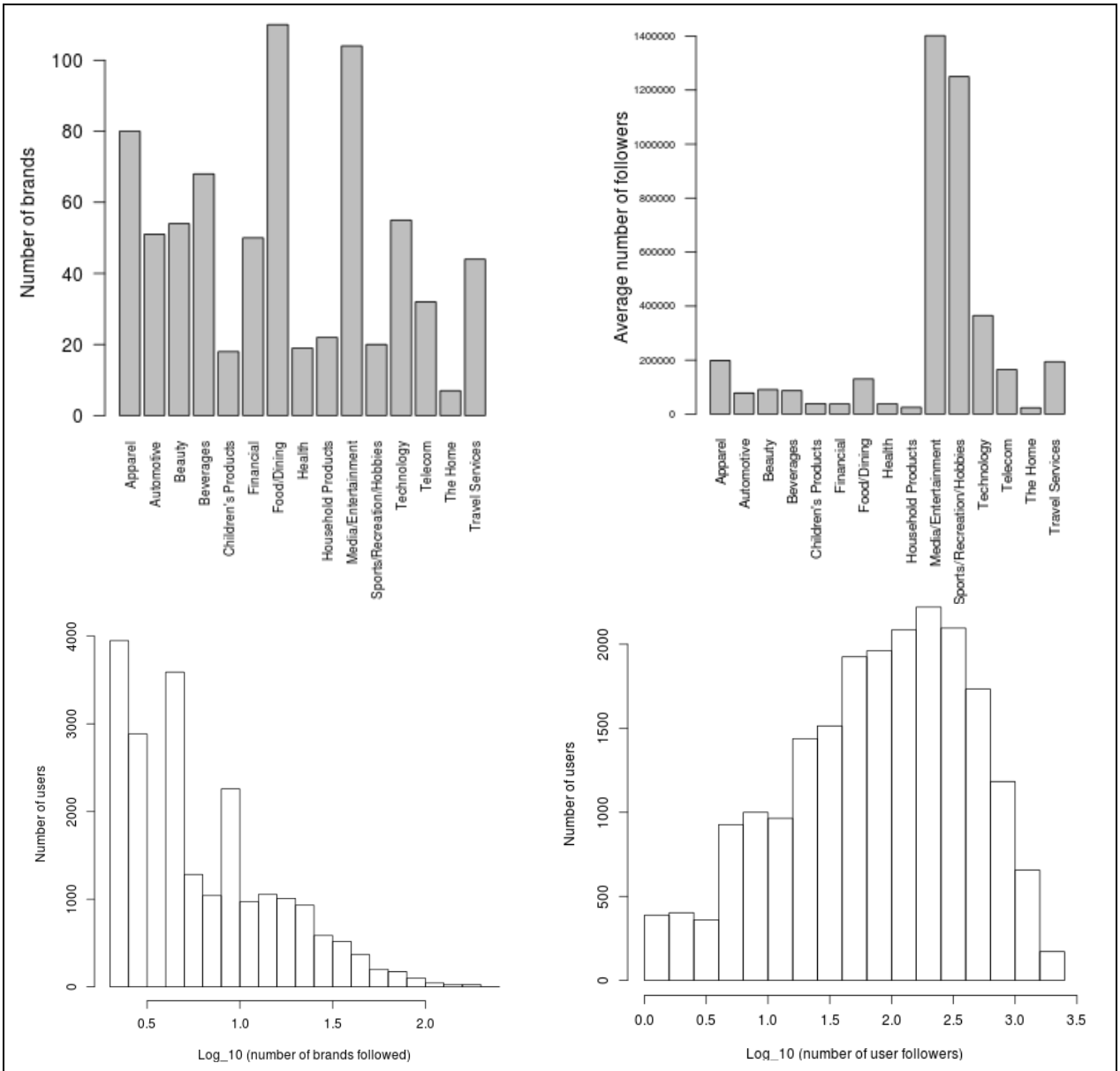

Figure 1. Product type and user features. (Top left) Number of brands, (top right) distribution of the average number of followers per brand type, (bottom left) distribution of the number of brands followed per user, and (bottom right) the number of followers per user, log (base 10) transformed.

The data we collected are complex and present two highly interesting dimensions: the brand type and the user type. Figure 1 focuses on characteristics of the brand type and on features of the users. The plots at the top of the figure display the distribution of the number of brands and the average number of followers per product type. Figure 1 reveals that there is very little correlation between the number of brands in a particular category and the average number of followers for brands in that category. The two bottom plots present the user characteristics. We plot the distribution of the number of users' followers and the number of brands followed by users.

The data we collected uniquely enables the testing of the recommendation systems using both types of networks (product and social). We can compare different approaches across many brands. In this paper, we focus on three different aspects of the brands described above: 1) popularity on Twitter, 2) product type, and 3) audience demographics (in particular, whether a brand has a specific audience). The third feature is based on the social science theory of homophily.

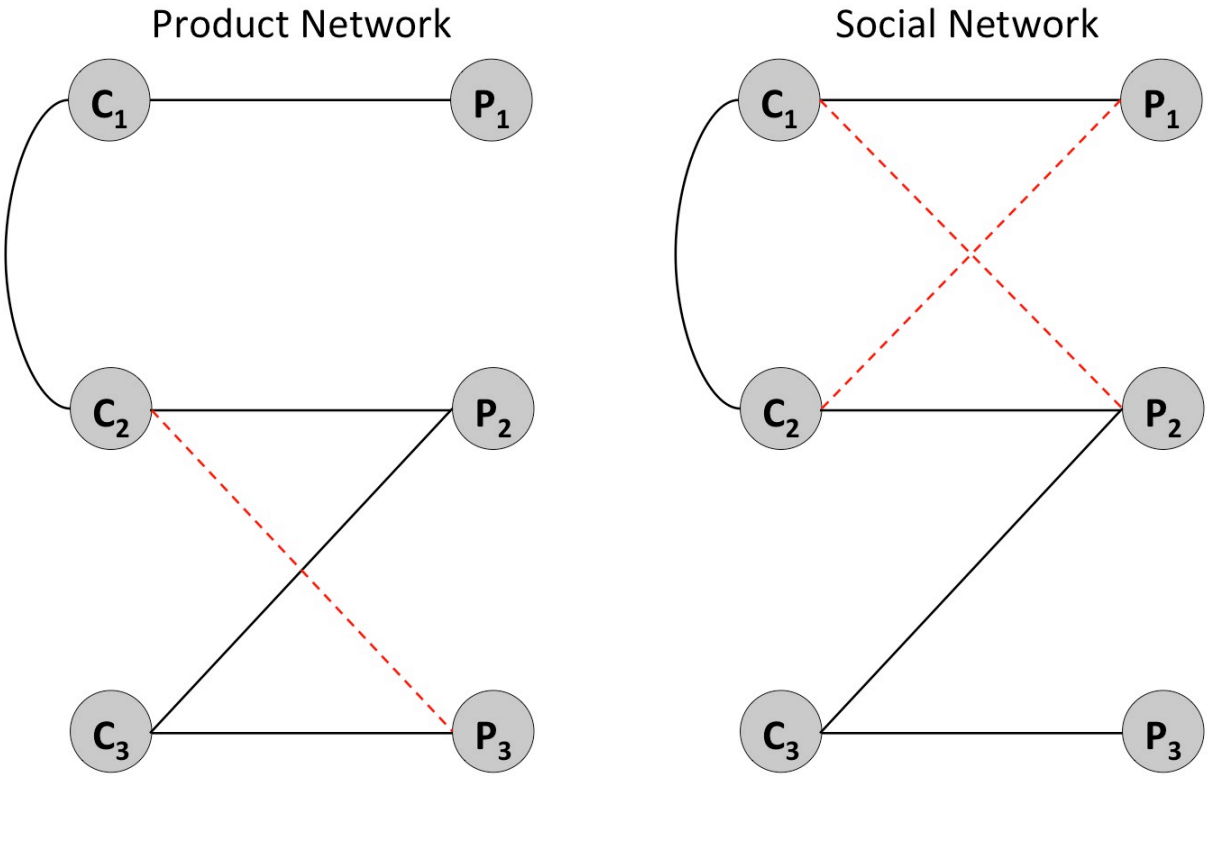

Figure 2. Illustration of differences between product and social network-based approaches. The red lines indicate connections between users (C) and products/brands (P). In the product network, consumers and products are connected when customers like two products. For example when C_3 likes P_2 and P_3. In the consumer network, consumers and products are connected when friends like products. For example because C_1 and C_2 are have a social tie then all products that C_1 is connected to can be recommended to C_2 (C_2 will be recommended P_1) and all products C_2 is connected to can be recommended to C_1 (C_1 will be recommended P_2)

Collaborative filtering system

Given a particular test user, u, who is known to follow the set of brands A, we calculate the similarity between this user and all training-set users as follows: For each training-set user, v, who follows the set of brands B, we calculate the similarity of u to v as sim(u, v) = |intersection(A, B)|/|A|. We then select the most similar users from the training set, K, and rank recommendations based on the popularity of brands

among similar users. We empirically set the parameter *K* to 20 because we found the performance of the system to plateau at that point. Brands in the set of input brands are omitted from the recommendations list.

Social network-based system

Given a particular test user, this approach finds all the followers of that particular test user, excluding any user in the test set. We rank recommendations based on the popularity of brands in the test user's local follower network. In other words, we consider all brands/TV shows friends follow and rank them by the frequency of the following. Brands that are followed by more friends will be recommended first. As in the product network collaborative filtering system, brands that belong to the test user's input brands are not included in the recommendation rankings.

The network data collected from Twitter enable both types of approaches to be explored and tested. Both approaches are built using the same Twitter consumers (C) and Twitter products and TV shows (P). In Figure 2, we illustrate the differences in how consumers and brands/products are connected using the two approaches. The red lines indicate connections between users (C) and products/brands (P) when the two approaches are used. In the product network, consumers and products are connected when consumers like two or more products. For example when C_3 likes P_2 and P_3, P_2 and P_3 become connected. They are connected not just for friends of C_3 but for everyone that we need to make a prediction for. In the consumer network, consumers and products are connected when friends like products. For example, because C_1 and C_2 have a social tie, all products that C_1 is connected to can be recommended to C_2 (C_2 will be recommended to P_1), and all products C_2 is connected to can be recommended to C_1 (C_1 will be recommended P_2). These two approaches, although built using the same data, can then result in vastly different connections between consumers (C) and products and brands (P). The product network, because it is based on all users, "should" be more general when making predictions, and the social network, because it is based on only a small subset of friends, "should" be more specific when making predictions.

Our proposition is that the social network-based approach will perform best for brands that cater to a "niche" demographic. This proposition is based on the sociology-based homophily theory that suggests that similar people are more likely to be friends (McPherson et al. 2001). If this theory holds, then the products that a consumer's social network likes should be able to reflect products that the consumer also likes, such as those products that are considered "niche" by a particular demographic.

Evaluation

Overall performance

First, the set of users was randomly split into training and test sets for 10-fold cross-validation, with 21,027 users in each test set. We used recommendation systems to make predictions for test-set users based on the training-set users. These systems were posed the following problem: given a particular test user who follows *N* brands, give the system *N-1* input brands that the user follows (omit one brand to test/predict). Given the user's local network and the *N-1* brands it follows, attempt to predict the N^{th} output brand the user follows. The held-out brands that the systems attempted to predict were selected in round-robin fashion from the brands that test users followed. The prediction task is simple: we feed the RS what a user follows and ask it to guess at what else they follow.

We identified two recommendation systems to assess: a collaborative filtering *K*-nearest neighbor product network approach and a social network-based approach, as described above. These systems were evaluated on their recall after a set number of recommendations. In other words, recall is calculated in the following way: Given the algorithm makes *K* recommendations, for what proportion of users in our test set could we recommend the N^{th} held-out brand they follow within the *K* recommendations? The collaborative filtering product based approach and social network based approach were then compared using the recall evaluation measure in two ways: firstly, across categories, considering all possible test users and held-out brand pairs, and secondly, within categories, considering only recommendations within one product category at a time.

Comparing across brand and user types

In addition to evaluating these methods' performance as a function of the number of overall recommendations made over test users, we also assessed their performance along various dimensions, including brands' popularity, product type, and audience skewness. We also assessed performance based on the users' popularity.

We divided the samples by one user feature (the number of followers), *N-1* input brands (number of input brands), and held-out brand (number of brand followers/popularity, product type, gender skew, age skew, education level skew). The skew features correspond to the symmetric *KL*-divergence from the observed distribution over the demographic groups for the held-out brand compared to the mean distribution for those demographic groups over all brands in our data. For brand demographic distribution, *D*, and mean distribution over all brands, *M*, over *n* groups, the symmetric *KL*-divergence between D and M is defined as $\sum_{i=1}^{n} \ln(\frac{D(i)}{M(i)})D(i) + \ln(\frac{M(i)}{D(i)})M(i)$. A higher *KL*-divergence thus corresponds to a less typical demographic make-up of the brand's followers. Suppose that, on average, 75% of brand followers are female and 25% of followers are male. In this case, a brand that is followed by an equal number of males and females would have a higher KL-divergence, or be considered a less typical demographic distribution than a brand with 90% female followers. When evaluating the two systems over these dimensions, the recall for each system is reported, with the number of recommendations fixed to 20, which is a reasonable number of recommendations for a firm to make. We present the results in the next section, in two parts: first, by comparing the results using the dimensions of brand and user popularity and, second, by using product category and audience skew.

Results

Overall performance

Figure 3 displays the overall performance of the social network-based and collaborative filtering-based methods across all brands. When considering a low number of recommendations, the social network-based method performs better than the collaborative filtering system. However, the higher the number of recommendations made, the more the social network-based system's performance dramatically degrades. This trend occurs, because the social network-based approach runs out of items to recommend. If a user's friends jointly follow only 5 brands, then 5 brands, at most, can be recommended.

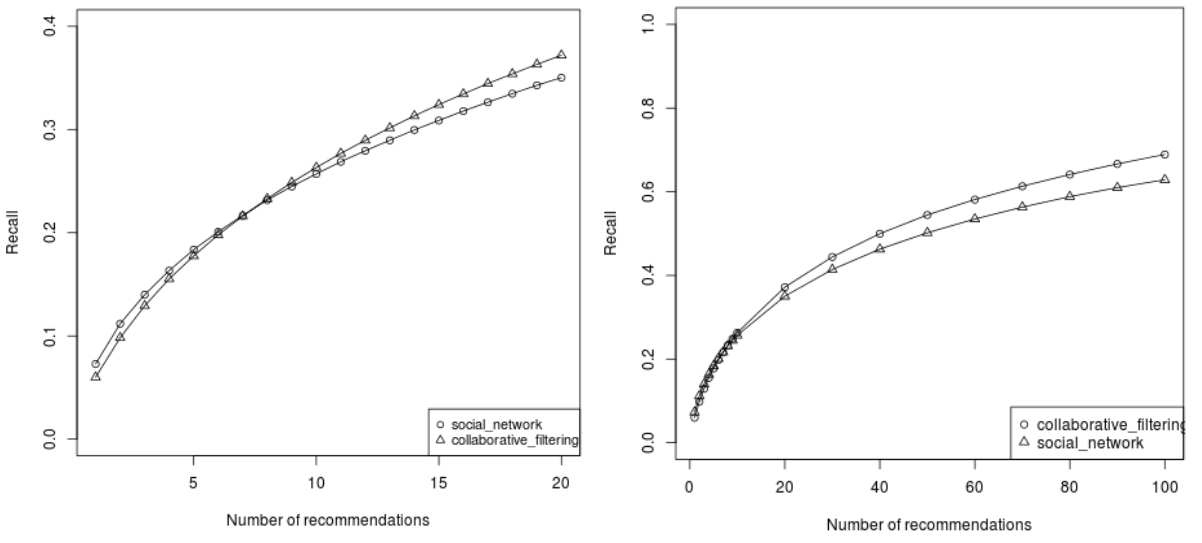

Figure 3. Across Categories. Overall recall of the *K*-nearest neighbor collaborative filtering and social network-based systems. The social network-based approach performs well at first, but the collaborative filtering approach overtakes it. The left plot focuses on the top 20 recommendations and the right plot shows results for 1 to 100 recommendations.

While the social network-based approach outperforms the collaborative filter approach when making predictions across categories, it does not do so when making predictions solely within categories. This is demonstrated in Figure 4, which displays the two strategies' performances when making only within-category recommendations. Though we highlight only three categories, a consistent pattern emerges: when making within-category recommendations, the collaborative filtering-based approach performs as well or better than the social network-based approach. This trend holds true in the plot included in the appendix, which shows that brands cluster by product category when using the product network to link brands.

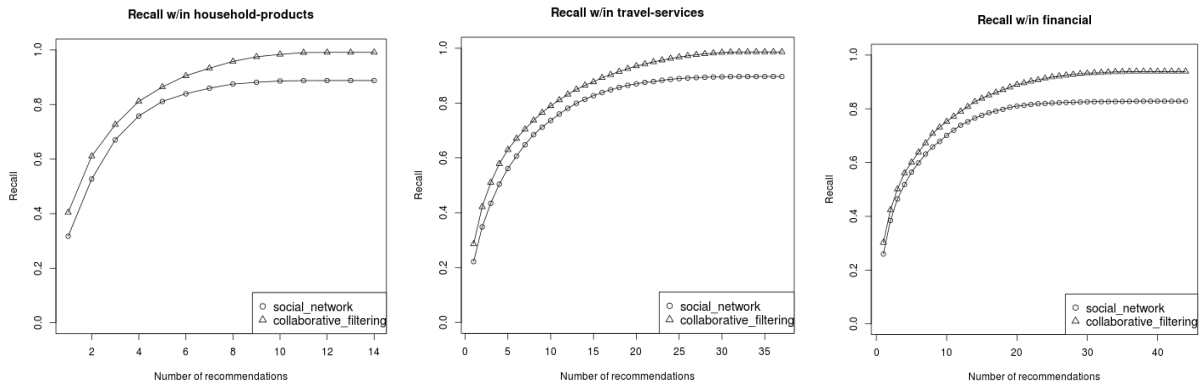

Figure 4. Within-Categories Analysis. Overall recall of the *K*-nearest neighbor collaborative filtering and social network-based systems. The collaborative filtering approach typically outperforms the social network system when making recommendations within categories. The household products, travel services, and financial categories are presented here.

To further display the difference in the two methods' strengths, we compare their performance when solely considering cases which look at within-category recommendation (when input brands are all in the same category and the output brand matches that category) and across-category recommendation (when input brands are all in the same category and the output brand does not match). The results can be found in Figure 5 (left side). We see that the collaborative filtering approach works best for within-category predictions and does much worse than the social network-based approach when making predictions outside of product type.

This is an important finding and has implications for practice. The result can inform firms of which strategy to use based on whether recommendations are to be made within a product category or across product categories. Therefore, we push on this result a bit further by breaking down the output brand by the likelihood that it will be recommended in general. In Figure 5 (right side), we evaluate the recall of social network and collaborative filtering product network by the probability that the "output product type" was drawn from the distribution over input product types. The brands are categorized into 5 bins with bin 5 reflecting the highest likelihood of being recommended and bin 1 reflecting the lowest likelihood of being recommended given the input brands. Those that are more likely to be recommended are generally more popular. The social network performance does not fluctuate much, but the collaborative filtering approach's performance is strongly dependent on this value. The product network-based approach only performs well when the product to be recommended across categories is popular or very likely to be recommended in general. This suggests that the social network based approach does a much better job of linking product categories. In short, we find that the social network-based approach generally performs better than the collaborative filtering approach if all categories are considered and there are enough brands to recommend. However, if we break down the prediction task to predicting within-category products versus across-category products, the collaborative filtering approach does much better at within-category predictions, while the social network-based approach performs much better when making predictions across categories. We believe this result is due to the nature of the differences in the two approaches. The product network, because it is based on all users, is more general when making predictions and therefore captures links between items that generally hang together and, as a result,

makes less diverse recommendations. The social network, because it is based on only a small subset of friends, is more specific to the user when making more diverse predictions. We have plotted the diversity of recommendations using the two approaches and have found the social network based approach makes far more diverse recommendations, where diversity is measured by both the number of unique recommendations made across users and the diversity of recommendations made across categories. Due to space constraints, however, we cannot include the results.

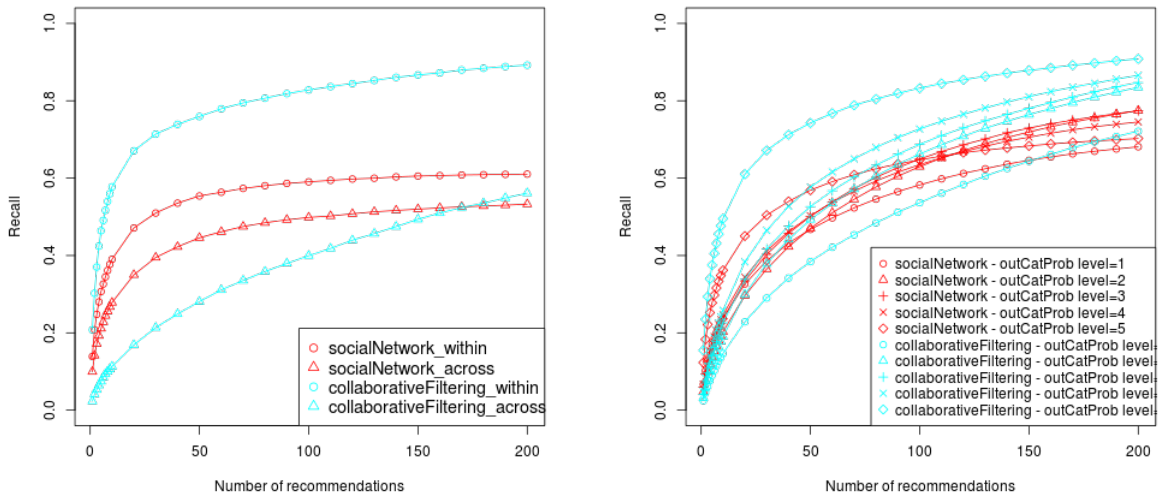

Figure 5. Across-Category Analysis. (left) Performance of each approach when within-category predictions are made and when across-category predictions are made. (right) Performance of social network and collaborative filtering product network by the probability that the "output product type" was drawn from the distribution over input product types. Level 1 corresponds to those products least likely to be predicted and Level 5 corresponds to those most likely to be recommended.

Comparing across brand and user types

We first compare the systems' performance by the popularity of both the brand and the user. Figure 6 displays the performance of both systems as a function of the number of followers for the held-out brand and the test user. For each approach, we present five lines illustrating the predicted brands divided by number of followers. The red lines present results from the social network-based approach, and the blue lines present those from the collaborative filtering approach. Bin 5 contains the brands with the most followers, and Bin 1 contains the brands with the fewest followers. On the horizontal axis, we divide users by their number of followers. The results suggest that users with large numbers of followers are crucial for the social network-based method to perform well and that both systems more easily predict more popular brands. However, the social network-based method appears to be more sensitive to brand popularity. The social network-based approach does not perform well with unpopular brands. To make this data easier to interpret, these results are also plotted in three dimensions (3D) on the right side of Figure 6.

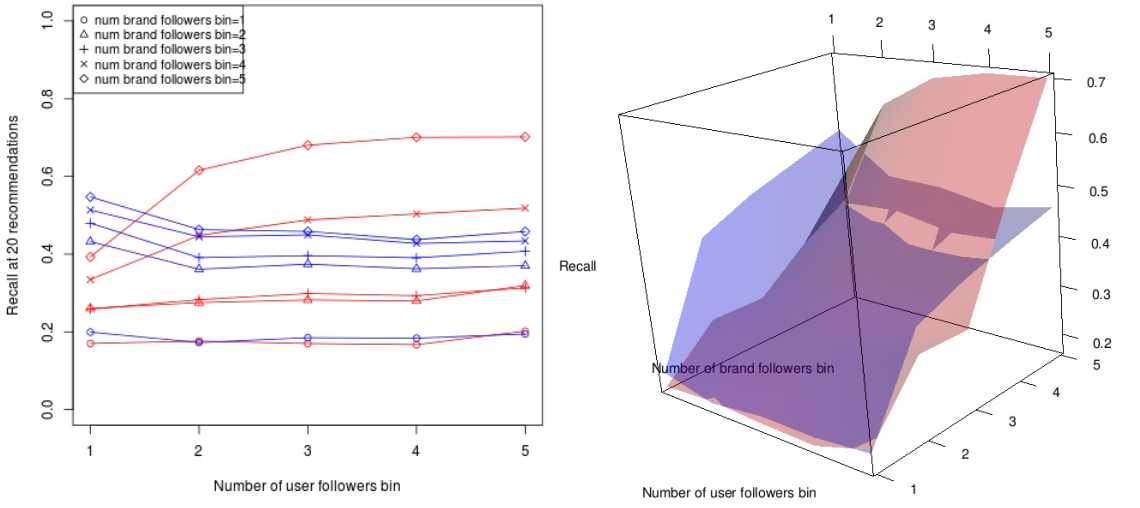

Figure 6. Performance of the social network and the collaborative filtering methods as a function of the held-out brand's popularity and the test user's number of followers. Users and brands were placed in equal-sized bins based on the number of followers. The number of recommendations was fixed to 20. (Left) The lines correspond to the different brands' bins. (Right) The blue plane corresponds to the collaborative filtering system, and the red to the social network-based system.

Secondly, we compare their performance by the product type of the held-out brand. These results are the most exciting because they demonstrate that the performance of the social network-based approach varies among different types of brands, differing by product type. Figure 7 displays the difference in recall between the two systems as a function of the product type of the held-out brand. The social network-based method outperforms the collaborative filtering method for a subset of categories, but as the number of recommendations increases, the collaborative filtering approach tends to overtake the social network method. Again, the social network-based approach's performs degrades, in part because it runs out of recommendations that can be made. The social network-based approach performs best in the categories of children's brands, home brands, and media and entertainment brands, which all have extremely high average *KL* divergence in demographic categories. These findings suggest that the social network-based approach is a superior method for calculating recommendations for brands and TV shows that have a demographic-based "niche" audience. This is a novel result because it shows that for some products, social network based prediction is not effective, while for others, it is.

While brands show some consistencies within a single category, there is still some variance in demographic skew by category. Therefore, we next evaluate performance by the held-out brand's demographic skew. Figure 8 displays the difference in recall between the two systems as a function of the demographic skew and the popularity of the predicted held-out brand (horizontal axis). Each plot has two lines: one line (red) represents the brands that have a higher than average skew, and the second line (blue) represents the brands that have a lower than average skew. The results suggest that the social network-method performs better when the predicted brand has an atypical age and/or education bias or an audience of a specific age. A gender bias does not necessarily yield better performance from the social network method. However, it is important to note that the social network-based approach does well with popular brands with both low and high demographic skews.

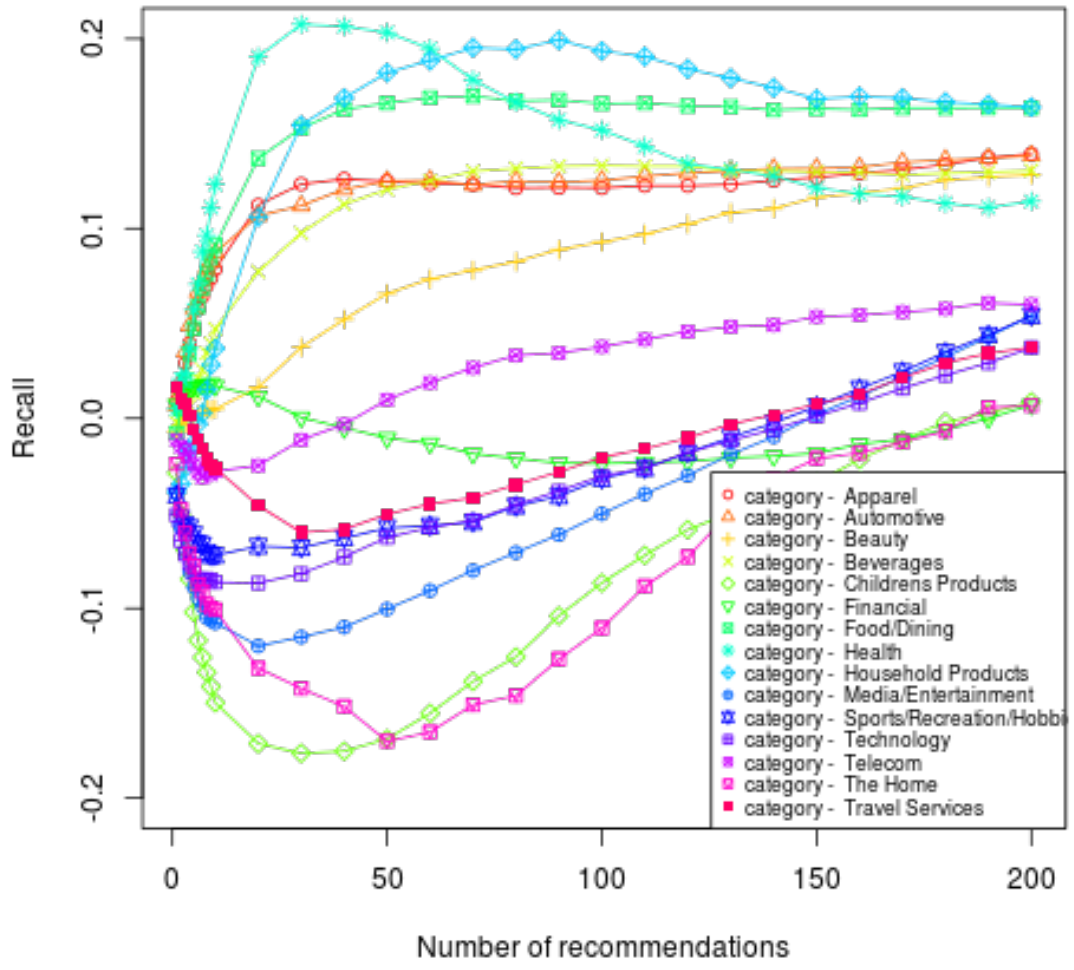

Figure 7. Difference in recall of the collaborative filtering and the social network systems. Negative values correspond to cases in which the social network system tends to outperform the *K*-nearest approach and vice-versa. Each line corresponds only to those cases in which the held-out brand is of a specific product type.

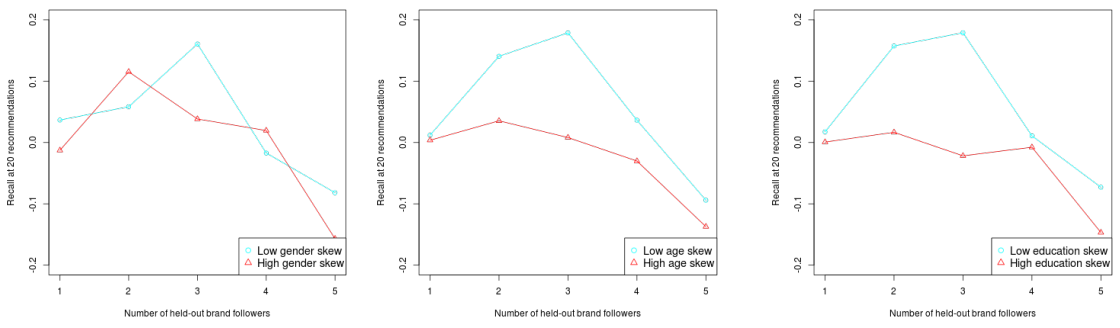

Figure 8. Difference in recall of the two systems as a function of the held-out brand's (left) gender skew, (center) age skew, and (right) education level skew. Negative values represent cases in which the social network system tends to outperform the collaborative filtering system and vice-versa.

To determine if the predicted brand's age skew was positively correlated with the performance of the social network-based approach, we fit a logit model to predict whether the social network approach could recommend the held-out brand within 20 recommendations (Table 2). We find that strong, positive relationships exist between the predicted brand's age skew and the social network method's performance. They also exist between the popularity of the output brand and the interaction between the user's and the output brand's number of followers. The interaction term of the predicted brand's popularity is both positive and significant, which is consistent with our evaluation using recall.

Table 2. Weights learned for predicting whether or not a method would correctly predict a held-out brand within 20 recommendations for both collaborative filtering and social network approaches. All weights are significant at the 0.0001 level. Due to space constraints, product type weights are omitted. Note that age skew, popularity of output brand, as well as interaction between output brand popularity and number of user followers are all large and positive.

Feature type	feature	Social network weight	Collaborative filtering weight
Intercept	**(Intercept)**	-5.252315	-7.704019
User	**log(user_followerCount, 10)**	-1.498872	0.212273
Brand	**log(outBrand_numFollowers, 10)**	0.427748	1.214396
	outBrand_categoryAutomotive	1.051923	0.783407
	outBrand_categoryBeauty	0.370927	-0.060016
	outBrand_categoryBeverages	0.256351	0.121779
	outBrand_categoryChildrens Products	1.197547	-0.317535
	outBrand_categoryFinancial	0.206902	-0.163115
	outBrand_categoryFood/Dining	0.390583	0.405314
	outBrand_categoryHealth	0.336513	0.756607
	outBrand_categoryHousehold Products	0.460982	0.576221
	outBrand_categoryMedia/Entertainment	-0.413795	-1.470283
	outBrand_categorySports/Recreation	-0.095063	-0.827611
	outBrand_categoryTechnology	0.139831	-0.74402
	outBrand_categoryTelecom	-0.260753	-0.958345
	outBrand_categoryThe Home	0.271268	-1.591411
	outBrand_categoryTravel Services	0.173285	-0.377128
	log(outBrand_genderKLDiv, 10)	-0.031902	-0.163921
	outBrand_meanAge	0.051345	0.029529
	log(outBrand_ageKLDiv, 10)	0.198198	-0.081729
In brands	**log(inBrands_count, 10)**	-0.334049	0.364556
User/ Brand interaction	**log(user_followerCount, 10): log(outBrand_numFollowers, 10)**	0.369481	-0.109223
Performance	**Accuracy**	69.60%	67.70%
	Baseline accuracy (guess misprediction)	65.00%	62.80%

Discussion and next steps

A number of recent studies in both information systems and business intelligence have investigated social and product networks' potential for deriving value for firms. Many successful companies go beyond these networks by using aspects of social networks to predict their customers' preferences and purchases, thus enabling the companies to better target their marketing and advertising efforts. There is however little

understanding of which circumstances allow the use of social networks to predict both products and users, and which contexts do not. To remedy this lack, this study demonstrates how performance varies across brands and product types, something no earlier research has examined. To explain differences in predictive performance, we investigate three features associated with brands (number of followers, brand type, and demographic skew) and one feature associated with users (number of followers). The results of this investigation demonstrate that, consistent with sociological literature on homophily, social network-based targeting works best when the brand audience has a demographic skew. In other words, if birds of a feather flock together (meaning that users share demographics with their friends), and some products appeal to a specific demographic (of friends), then predicting niche products for users liked by their friends (that are also assumed to be the same demographic) should be effective. The feasibility of social network-based targeting also depends on the number of connections held by both brands and users.

The research we have outlined presents many potential next steps. Our primary objective is to better understand which brand features are important for using social networks to make predictions, and which are irrelevant. This paper explores brand type, "nicheness" and popularity; in future work we will explore additional brand features such as how risky the brand is, whether it is an aspirational brand and features of the lifestyles of the brand audience. In addition, we will explore user features beyond simply their number of followers. In particular, we will measure user engagement levels in social media, especially as they relate to discussions of specific brands. We will also further explore other evaluation approaches that bring the method closer to real world applications. For example, instead of taking a round robin approach, where we hold out brands one at a time, we will try to predict the brand following over time, by holding out the brand that was last followed. and examine capacity for future predictions of the next brand the user will follow as opposed to taking a round robin approach to holding brands out. It is when making recommendations to users that prediction better matches firms' practices. Finally and most importantly, we will correlate our work with marketing and information systems literature on which factors determine the sociality of brands on the web.

We must acknowledge that this paper contains several assumptions, including the supposition that people follow brands they actually like and that following brands on Twitter is reflective of actual purchases. This assumption is significant because following a brand on Twitter is free of charge and, as a result, brands followed are often aspirational, i.e. luxury items that followers may not be in a financial position to ever actually purchase. Nevertheless, despite these assumptions, this paper is an important first step toward understanding the question: ***When does social network-based prediction work?***

References

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. 2005a. "Incorporating contextual information in recommender systems using a multidimensional approach," *ACM Transactions on Information Systems (TOIS)* (23:1), pp 103-145.

Adomavicius, G., and Tuzhilin, A. 2005b. "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), pp 734-749.

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

Breese, J. S., Heckerman, D., and Kadie, C. Year. "Empirical analysis of predictive algorithms for collaborative filtering," Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.1998, pp. 43-52.

Domingos, P., and Richardson, M. 2001. "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM: San Francisco, California, pp. 57-66.

Goldenberg, J., Oestreicher-Singer, G., and Reichman, S. 2012. "The quest for content: The integration of product networks and social networks in online content exploration," *Journal of Marketing Research* (49:4), pp 452-468.

Hill, S., Provost, F., and Volinsky, C. 2006. "Network-based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science* (21:2) May, pp 256-276.

Macskassy, S. A., and Provost, F. 2007. "Classification in networked data: A toolkit and a univariate case

study," *The Journal of Machine Learning Research* (8), pp 935-983.
Manski, C. F. 1993. "Identification of endogenous social effects: The reflection problem," *The review of economic studies* (60:3), pp 531-542.
McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. "Birds of a feather: Homophily in social networks," *Annual review of sociology*), pp 415-444.
Pazzani, M. J., and Billsus, D. 2007. "Content-based recommendation systems," in *The adaptive web,* B. Peter, K. Alfred and N. Wolfgang (eds.), Springer-Verlag, pp. 325-341.
Provost, F., Dalessandro, B., Hook, R., Zhang, X., and Murray, A. Year. "Audience selection for on-line brand advertising: privacy-friendly social network targeting," Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM2009, pp. 707-716.
Shalizi, C. R., and Thomas, A. C. 2012. "Homophily and contagion are generically confounded in observational social network studies," *Sociological Methods & Research* (40:2), pp 211-239.
Soboroff, I., and Nicholas, C. Year. "Combining content and collaboration in text filtering," Proceedings of the IJCAI1999, pp. 86-91.