This is a preprint of an article published in JASIST. http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1532-2890

Citation: Li, H., Bhowmick, S. S. and Sun, A. (2011), AffRank: Affinity-driven ranking of products in online social rating networks. Journal of the American Society for Information Science and Technology, 62: 1345–1359. doi: 10.1002/asi.21555

AffRank: Affinity-Driven Ranking of Products in Online Social Rating Networks

Hui Li, Sourav S. Bhowmick, Aixin Sun

Nanyang Technological University, Nanyang Avenue, Singapore 639798
herolee@pmail.ntu.edu.sg, {assourav,axsun}@ntu.edu.sg

Abstract

Large online social rating networks (*e.g., Epinions*, *Blippr*) have recently come into being containing information related to various types of products. Typically, each product in these networks is associated with a group of members who have provided ratings and comments on it. These people form a *product community*. A potential member can join a product community by giving a new rating to the product. We refer to this phenomenon of a product community's ability to "attract" new members as *product affinity*. The knowledge of a ranked list of products based on product affinity is of much importance to be utilized for implementing policies, marketing research, online advertisement, and other applications. In this paper, we identify and analyze an array of features that exert effect on product affinity and propose a novel model, called *AffRank*, that utilizes these features to predict the future rank of products according to their affinities. Evaluated on two real-world datasets, we demonstrate the effectiveness and superior prediction quality of *AffRank* compared to baseline methods. Our experiments show that features such as *affinity rank history*, *affinity evolution distance*, and *average rating* are the most important factors affecting future rank of products. At the same time, interestingly, traditional community features (*e.g.,* community size, member connectivity, and social context) have negligible influence on product affinities.

Keywords: Web mining, Web information systems, Information Retrieval, Filtering, Classification, Summarization, and Visualization, Business intelligence, Product ranks, Ratings, Recommendation

1 Introduction

Due to the proliferation of on-line communities in recent times, we are faced with the opportunity to analyze social network data at unprecedented levels of scale and temporal resolution. Consequently, this has attracted increasing research attention at the intersection of the computing and social sciences. For instance, large online social rating networks (e.g., *Epinions* [1] , *Blippr* [2]) have recently come into being containing information related to many categories of products. Within these websites, individual users are allowed to publish their comments or give ratings on different products. Besides, they can set up friendships by linking to each other. Figure 1 depicts the structure of such a social rating network. Observe that for a particular product (*i.e.,* p_2), there is a group of people who have given ratings and published their comments on it (*i.e.,* u_2, u_3). These people form a community (*i.e.,* c_2). In other words, each product in a social rating network is associated with a community [31]. In the sequel, we refer to such a community as *product community*. Clearly, these communities are a potential gold mine for all kinds of marketing and business analysts as users' comments and ratings toward a particular product may affect other consumers' purchasing behavior [31].

In social rating networks (SRN) a new user can join a product community by giving a *new* rating (review) to the product. Note that a review that is updated by an existing member is not considered as new. Hence, *the growth of a product community's size can be implicitly measured by the number of new reviews (from new users) it receives during a particular time slot.* We refer to this phenomenon of a product community to "attract" new users by giving new ratings as *product affinity*. Specifically, it is measured by the number of *new* ratings a product receives from *new* members during a particular time period. In fact, existing social rating networks often display a ranked list of products that received the most number of *new* ratings on a daily, weekly, or monthly basis. We refer to this ranked product list at a particular time slot as *affinity rank*. For example, consider Figure 2. It depicts two affinity ranks of top-5 movies extracted from the *Blippr* website during weeks $t-2$ and $t-1$, respectively. Observe that *Ninja Assassin* and *Sherlock Holmes* received the most number of new ratings.

It is important to compare the aforementioned notion of product affinity to *affinity* in marketing research, which refers to a marketing strategy [13]. In the latter, *affinity* involves two parties. The first party known as the "affinity group"; seeks to add value to its existing customers, members or donors by promoting products and services they do not currently sell (e.g., financial services). The second party known as the "product supplier"; seeks to acquire new customers by using the strength of another organization's relationship with its customers, through which it aims to distribute its product or service. In other words, the aim of affinity marketing is to build and develop new

[1] http://www.epinions.com
[2] http://www.blippr.com

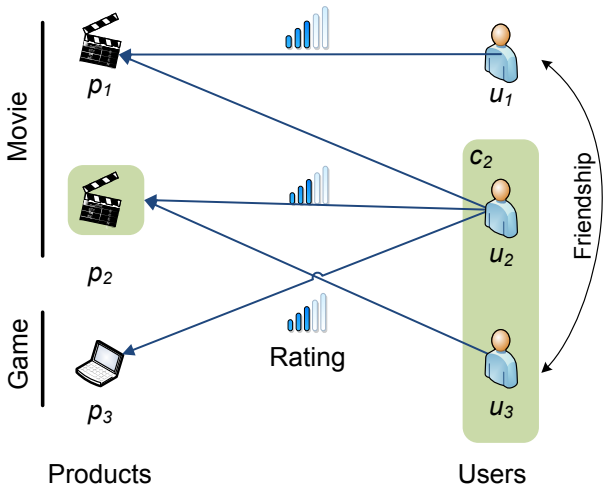


Figure 1: Social rating network structure.

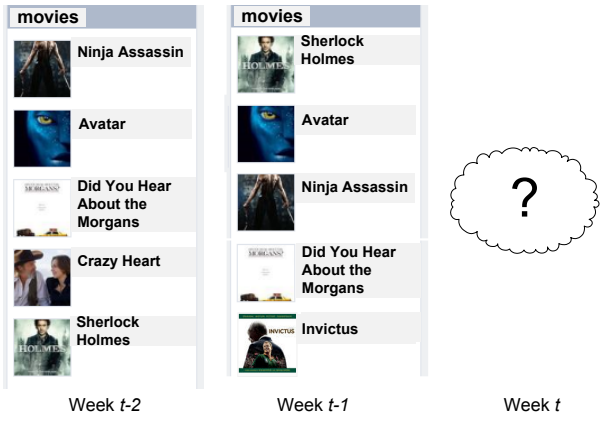


Figure 2: Product affinity and affinity ranks.

customer relationships through the existing distribution channels of a third party. In contrast in SRN, a product community is similar to an affinity group as the former promotes (or demotes) a product that they do not sell to its members by giving reviews or ratings. On the other hand, the "product supplier" in SRN seeks to attract new customers by exploiting the *affinity strength* of an "affinity group" (product community). In this paper, we undertake a quantitative study to analyze and predict the affinity strengths of product communities.

1.1 Motivation

The affinity ranks of products in the past weeks/months highlight the reviewers' affiliation of products in the (recent) past. Although such historical information is important for several applications, prediction of future affinity ranks of products is even more important to marketing and business strategists. For example, reconsider Figure 2. Suppose in week $t-1$ a company intends to put advertisements on 5 movie products during week t. Then, it makes sense to predict 5 most popular products at time t so that optimum benefit can be achieved. That is, it is desirable to predict the affinity ranks of products in the near future. In this context, instead of just listing most popular products, ranking them based on their affinity ranks makes more sense as a company may allocate different shares of their advertisement budget depending on the popularity of the products. Note that such top-k products may vary considerably at two different time points. For instance, consider the top-5 products during weeks $t-2$ and $t-1$ in Figure 2. Observe that *Sherlock Holmes* moved from rank 5 to the top rank in successive weeks. Further, *Invictus* first appeared in the top-5 list in week $t-1$ whereas *Crazy Heart* failed to remain in the top-5 list in this week.

Recent research on ranking products in SRN have primarily focused on evaluating a product by the strength of connections among its users [31] or by the features related to a particular product item (*i.e.,* price, released time etc.) [35]. However, these techniques are not designed to predict the ranks of products based on their affinities. In this paper, we propose a novel quantitative model called *AffRank* that utilizes historical and evolutionary affinity information as well as other features to predict the future ranks of products according to their affinities [3].

1.2 Overview

The problem of predicting and ranking product affinity is related to recent efforts [5, 26, 38, 39] in predicting community growth as the former is influenced by the number of new members joining the community. Specifically, these efforts reveal that the community size, connectivity between community members, number of friends a user has in a community [5], and similarity of interests a member has with a community have strong influence on the growth of communities [26]. Hence at a first glance, it may seem that these features should also strongly influence the product affinity. However, our study demonstrates that *the effects of these features are negligible in product community* as it is not strictly similar to traditional social network communities (*e.g.,* users are sparsely connected). Hence we propose three additional features, namely *affinity rank history*, *average ratings*, and *affinity evolution distance*, related to the product communities that may exert significant effect on affinity. In particular, *affinity rank history* represents the historical affinity ranks of products over time; *average ratings* measures the average of ratings received by a product at a given time point; and *affinity evolution distance* measures the distance between affinity evolution of a pair of different products. *To the best of our knowledge, these features have not been studied systematically in the literature.*

In Section 5, we evaluate the performance of all the aforementioned features and further compare the three new features with the traditional features for affinity rank prediction. Additionally, our investigation with two real-world datasets (*Epinion* and *Blippr*) revealed several interesting and novel findings related to these three features. For instance, we observe that product affinity is more likely to *increase spikily* and *drop down smoothly*. Further, there is $0 \sim 3$ days lag between the *peak*

[3] A shorter version of this paper appeared in [25]

time of affinity and users' ratings. Also, average *distances* between affinity evolution of products in the *same* category are always smaller than those from *different* categories.

Based on these findings, we propose the *AffRank* model to predict the future affinity rank of a product. Experiments conducted over real-world datasets demonstrate that our prediction and ranking scheme generates high quality results and outperforms several baseline methods. In summary, the main contributions in this paper are as follows.

- In Section 3, we investigate an array of features that exert effect on the evolution of product affinity and affinity rank prediction. To the best of our knowledge, we are the first to study the evolution of product affinities with the goal of predicting affinity ranks.
- In Section 4, we formulate the task of ranking product affinity as an autoregressive problem with exogenous input features. We propose a quantitative model called *AffRank* that utilizes historical ranks and evolutionary product affinity information to predict the future ranks based on product affinity.
- By applying *AffRank* to real-world datasets, in Section 5, we show its effectiveness and superiority of its prediction quality compared to baseline methods. Further, our experimental results show that traditional community features (*e.g.,* community size, member connectivity, and social context) presented in [5, 26, 38, 39] have negligible influence on product affinities. Instead, features such as affinity rank history, affinity evolution distance, and average rating exert significant influence on affinity rank prediction.

In the next section, we review related research.

2 Related Work

We address related work from a number of relevant research areas, including: affinity in marketing research; community affinity; community evolution and dynamics; ranking of product communities; collaborative filtering; and product sales and correlation between product sales and public sentiments.

Affinity in marketing research. Fock et al. [13] explored the importance of relationship, which exists between targeted users and their affiliated communities, for marketing. Affinity marketing has achieved breakthrough results in terms of advertising effectiveness, particularly in the credit card industry [30]. Research in this field mainly uses existing data and statistical analysis methods to show that the difference is obvious between the case where affinity marketing strategy is adopted and the case that does not. In this way, they prove that affinity marketing is an effective approach in many marketing field. In this paper, we also utilize the affiliation of users with product communities. However, we do not focus on investigating various statistical properties of marketing data. Instead, we employ machine learning models to investigate the behavior of product affinity. Specifically, we evaluate and predict behavior of users joining a product community associated with SRN. The results of this research can be utilized in advertising and marketing, especially affinity marketing where advertisers or marketers can determine which product community is the most valuable to target for advertisements in the near future.

Community affinity. More germane to this work is efforts in studying the factors that exert effect on users' inclination to join a community. Table 1 summarizes the differences between *AffRank* and existing approaches. For instance, Backstrom et al. [5] demonstrated that the community size, connectivity between community members, and number of friends a user has in a community have strong influence on community affinity. They modeled the affinity problem as a standard classification task: the nodes eventually join a group are denoted as positive while those do not are denoted as negative samples. They extract several features (friends in the group, clustering coefficient of the group etc.) for each nodes being examined and employed a decision tree to predict the sign of node samples. They showed that community affinity is highly affected by existence of friends in the target community.

Leskovec et al. [23] showed that an individual's probability of buying a DVD increases with the number of recommendations he has received. There is a *saturation point* at the value of 10, which means after a person receives 10 recommendations on buying a particular DVD, the probability of buying does not increase anymore. Further, they employed a logistic regression model to test the success of recommendation based on their findings on the affinity of buying a product. Cha et al. [9] conducted a study on *Flickr* over the same problem. They reported that the probability for a user to become a fan of a photo increases with the number of her friends who are already fans of the photo.

The aforementioned efforts did not undertake any systematic study on the effects of evolutionary properties of historical affinity and average ratings on community affinity. Further, they did not address the issue of predicting the future affinity ranks of products in a product community which is more valuable in many applications. In contrast, we take into account the evolution of community affinity and the difference of affinity evolution patterns between communities in order to rank communities according to their ability to attract new users in the near future.

Li et al. [24] analyzed a large publicly-available collections of blog information to predict which blogs are highly likely to join a *blog cascade* in the future. Specifically, they showed that four features, namely, number of friends, popularity of participants, number of participants, and time elapsed since the genesis of the cascade, played important roles in predicting *blog cascade affinity*. However, affinity problem in that work was targeted to blog networks whereas in this paper we explore this problem in social rating networks which has important applications in e-commerce and marketing strategies. Moreover, blog cascade affinity prediction did not involve investigation of the temporal and evolution characteristics of features that are important in SRN.

Community evolution and dynamics. Table 2 compares *Af-*

Table 1: Summary of community affinity research.

Models	***Features adopted***	***Method***	***Target network***	***Affinity evolution study***
Group formation [5]	friends in the group, group size, clustering coefficient	decision tree	collaboration network	No
Dynamics in VM [23]	#recommendations, price, #reviews	logistic regression	recommendation network	No
Propagation in Flickr [9]	#fans, #friends	statistic analysis	flickr network	No
Blog cascade [24]	#friends in a cascade, cascade size, elapsed time, cascade popularity	SVM regression	blogosphere	No
AffRank	affinity rank history, affinity evolution distance, average rating besides above features	ARX	social rating network	Yes

Table 2: Summary of community dynamics research.

Models	***Difference measure***	***Parameter-free***	***Evolution pattern comparison***
GraphScope [41]	minimum description length	Yes	-
MONIC [40]	cluster intersection	No	-
Stable Cluster [6]	jaccard similarity	No	-
FacetNet [27]	KL-divergence	No	-
AffRank	Δaffinity and Δaffinity rank	Yes	DTW distance

fRank with recent research on community dynamics. GraphScope [41] is a parameter-free algorithm where the Minimum Description Length (MDL) principle is employed to extract communities as well as their changes. MONIC [40] models the changes within each individual community. The authors defined a set of key events: survive, split, disappear, which are used to model the changes of clusters. Asur et al. [4] proposed another algorithm to study how communities are formed and dissolved using a group of microscopic events. Bansal et al. [6] used jaccard similarity to model community evolution. It is computed as the intersection of community members at different time points divided by the union of the members. Berger-Wolf et al. [7] used a generalized jaccard similarity to measure the change of a group of people over time. FacetNet [27] is a framework to identify communities as well as their evolutions. It employed a KL-divergence based method to measure the distance between consecutive temporal community structures. Specifically, it measures the distance between the partitions of community over time instead of the change of community itself.

As depicted in Table 2, existing techniques adopted different measures to analyze consecutive versions of a community at different timesteps. In *AffRank*, we analyze the evolution based on changes to affinity and affinity rank. Unlike *AffRank*, none of the aforementioned efforts explored the historical evolution pattern of a community or comparison between the evolutions of different communities. Evaluating the historical evolution pattern of a community enables us to comprehend the future trend of a community's evolution. Besides, comparing the evolution of different communities facilitates us to understand the significance of role a community may play compared to others in attracting new users. In this paper, we study the evolution of product communities not only from the aspect of network structure but also with respect to their historical affinity ranks and evolution distances between communities.

Ranking of product communities. In [31] the authors analyzed the effect that users' ratings exert on the trust between users and vice versa. Their research is conducted within an E-commerce website in Japan (*@cosme*) where users can bookmark their trusted users and post their own ratings toward different cosmetics. They proposed a measure called *Community Gravity*, which can be viewed as brand strength from user-interaction perspective. However, the result of the work is hard to evaluate as it is difficult to explicitly measure gravity. Moreover, they assumed that the gravity for a brand is constant. This may not be the true in SRNS as user-interactions are highly evolving. In this paper, we propose affinity rank which can be easily and explicitly evaluated in many SRNS. We also take into account the evolution of user-interactions in SRNS.

Collaborative filtering. Several existing works in product research belong to the field of collaborative filtering which mainly studies user preferences over products based on the relationships between users as well as interdependence among products. For instance, the work reported in [20] studies the factors and temporal dynamics in modeling user behavior based on user preferences and item features. There are several different user tasks in collaborative filtering, each is associated with some evaluation metrics [17, 42, 44, 37, 22]. These efforts primarily investigated the relationships between products which are always bought together by the same user or the similarity between users who often bought the same products. Hence, these models are able to recommend users a series of goods that match their interest or filtering discussion postings to determine which ones are worth reading.

Similarly, we also analyze the user ratings of products as well as historical popularity of products. However, in contrast

to these conventional approaches in collaborative filtering, we take an evolutionary view by studying the evolution patterns of ratings and product affinities. Furthermore, we predict the future trend of all products instead of recommending products to users at a specific time point.

Product sales and public sentiment. Public opinions and activities have been studied in several work in order to acquire better understanding of customer behaviors. [33, 34] discussed many models that are used to extract opinions and sentiments from a given document. Based on these models, a comment on a product can be summarized and quantified into a value indicating whether the author likes or dislikes that product. Consequently, many work have been proposed to analyze the correlation between public sentiment and product sales. For instance, the study in [28] analyzed sentiments from users' comments toward movies to predict box office trends of movies. It proved the existence of relationship between users' reviews and product sales, inspiring several recent works including the work reported in this paper. Although it took into account the historical sentiment in the regression model, the sentiment evolution pattern has not been explored. It did not propose any method to compare the sentiment evolution between different movies or to rank the movies. Lastly, it did not explore the source of the product sales: users' affinity towards a product.

A recent technique [3] evaluates the weight that customers place on individual product features. Additionally, several recent studies reveal the correlation between product sales and the public sentiment [15, 29, 14, 11]. These efforts only take into account the current sentiment and model the relationship between the sentiment and sales. However, these models do not take into account user interactions or the word-of-mouth effect. Moreover, the evolutionary pattern of product communities cannot be shown by considering only the temporal sentiment. In this paper we study the products from the aspect of their affinities and their evolution patterns. We predict the affinity rank of products so that marketers have an idea on which products are most valuable to invest. Note that we do not focus on sentiment analysis as the users' ratings of products are explicitly available in the representative SRNs. If such ratings are not explicitly provided then a preprocessing stage can be built on top of our model that uses existing techniques to extract and quantify sentiments.

3 Features for Affinity Prediction

In this section we describe the features that are used in our ranking model. We begin by introducing the real-world datasets we have used for our study. Then for the sake of completeness, we briefly describe existing community-based features proposed in the literature that influence the community size. Next, we propose new features for addressing the affinity rank prediction problem. All the values for these features are normalized into the interval [0,1] using Min-Max Normalization [16]. In the sequel, we shall use the notations shown in Table 3 to represent different concepts.

Table 4 describes two real-world datasets that are used in this paper. The *Blippr* dataset was crawled using *Blippr* API [4] till August, 2009. It includes user ratings toward 75 different products. The *Epinions* dataset was downloaded from TrustLet [5] . It contains ratings proposed during 2001. Additionally, we crawled the *Epinions* website to retrieve product category and user-user relationships information as TrustLet dataset does not provide them.

3.1 Traditional Features

Recall from preceding section, several previous work have demonstrated the existence of correlation between characteristics of a community and its affinity. These characteristics include the community size, connectivity between community members, number of friends a user has in a community [5, 21, 18], and the similarity of interests people have with a community [38, 39]. We refer to these features as traditional features as they are associated with communities in conventional social networks. In this section, we describe in detail how we utilize and compute these traditional features in the product community.

Community size. The relationship between community size and its affinity has been studied in several work [5, 45]. Therefore, we incorporate *community size* in our model to predict the future affinity ranks of products. Formally, the size of a product community at time t is calculated as $|\mathbf{C}_i^t|$.

Member connectivity. According to a recent study [5], people are attracted to a community not only because they have friends in it but also the close connections among friends. Thus, the connectivity within the community affect the community's affinity. Clustering coefficient [5, 21, 32, 10] is widely adopted to measure the connectivity within a community and is computed as follows.

$$CC_i(t) = \frac{|3 \times \text{closed triplets in } \mathbf{C}_i^t|}{|\text{triplets in } \mathbf{C}_i^t|} \quad (1)$$

We compute the clustering coefficient for each community and use them to predict the future affinity rank.

Social context. According to some existing studies [26, 38, 39, 12, 43], people are more probable to join a community he/she is interested in. *Social context* represents the similarity in the interest between the members of the community and their friends who are not in the community yet. The friends of users in $\mathbf{C}_i^t$ can be computed by the following equation.

$$\mathbf{S}_i^t = (\bigcup_{u_j \in \mathbf{C}_i^t} \mathbf{F}_j) \setminus \mathbf{C}_i^t. \quad (2)$$

To measure the similarity in the interests between $\mathbf{C}_i^t$ and $\mathbf{S}_i^t$, we first define *interest-similarity* between a pair of users. Let $Int(u)$ be the set of product categories user u is interested in.

[4] http://api.blippr.com/v2/
[5] http://www.trustlet.org/

Table 3: Symbols and Semantics.

Symbol	*Semantics*
$u_1, \ldots, u_n \in \mathbb{U}$	users
$p_1, \ldots, p_m \in \mathbb{P}$	products
$c_1, \ldots, c_m \in \mathbb{C}$	communities
$\mathbf{C}_i^t$	users in c_i at time slot t
$\mathbf{F}_j$	friends of user u_j
$\mathbf{S}_i^t$	friends set of $\mathbf{C}_i^t$
$\mathbf{U}_i^t$	set of new users on product p_i at time slot t
$\mathbf{R}_i^t$	bag of new ratings on product p_i at time slot t
$\overline{\mathbf{R}}_i^t$	average rating
a_i^t	number of new users for p_i at time slot t
$\alpha_i(t)$	affinity intensity of p_i at time slot t
$\rho_t(p_i)$	affinity rank of product p_i at time slot t
$r_t(p_i)$	predicted value of $\rho_t(p_i)$

Table 4: Statistics of the datasets.

Dataset	*#products*	*#categories*	*#ratings*	*#users*	*#user-user links*
Blippr	75	5	8,032	2,219	10,480
Epinions	678,725	12	13,362,381	132,000	242,831

Then, the *interest-similarity* between users u and v is defined as follows.

$$sim(u, v) = \frac{|Int(u) \cap Int(v)|}{|Int(u) \cup Int(v)|} \quad (3)$$

Using the above similarity measure we can compute the interest-similarity between two groups of users $\mathbf{S}_i^t$ and $\mathbf{C}_i^t$. The *interest-similarity* between two sets of users U and V is defined as follows.

$$simC(U, V) = \sum_{u \in U, v \in V} sim(u, v) \quad (4)$$

In our model, we calculate $simC(\mathbf{C}_i^t, \mathbf{S}_i^t)$ for each product community at time t and then normalize it to [0,1] as of other features.

3.2 Affinity Rank History

We now present three new features related to product communities that have been ignored in the literature, namely *affinity rank history*, *affinity evolution distance*, and *average rating*. We begin with the *affinity rank history*.

Recall that (Section 1) each product p_i is associated with an affinity rank at time t, denoted by $\rho_t(p_i)$. For example, in Figure 2 the affinity ranks of `Sherlock Holmes` movie is 5 and 1 in weeks $t-2$ and $t-1$, respectively. Since the affinity rank of a product depends on the number of new users (product affinity), here we characterize the evolutionary behaviors of product affinity and affinity rank. We first investigate how the affinity changes between consecutive time points in the history. We denote the number of new users for product p_i at time slot t as $a_i^t = |\mathbf{C}_i^t| - |\mathbf{C}_i^{t-1}|$. Figures 3(a) and 3(b) report the distribution of $\Delta a_i^t = a_i^t - a_i^{t-1}$ over all products p_i and time t. Observe that in Figure 3(a) the count of cases where $\Delta a \in [-1, -2]$ is more than that of $\Delta a \in [1, 2]$. Besides, the absolute value of power law exponent for the tail at negative side ($\alpha^- = -0.96$) is bigger than that of the positive side ($\alpha^+ = -0.76$). It indicates that the positive affinity change is more likely to have bigger Δa than the negative one. We observe the same phenomenon for *Epinions* dataset in Figure 3(b). Specifically, the count of negative changes is larger than the positive ones when $\Delta a \in [-8, 8]$; the absolute value of power law exponent for the tail at negative side ($\alpha^- = -3.62$) is bigger than that of the positive side ($\alpha^+ = -3.48$). Both of the above phenomena indicate that the product affinity for all the products discussed in this paper is more likely to *increase spikily* and *drop down smoothly*. As affinity is the number of new users associated to a product within a time interval, this phenomenon reflects the speed of growth of product community size. It suggests that the speed of growth tends to increase to a peak in a short time and diminishes slowly subsequently. The aforementioned phenomenon is generally applicable to most products, although we do acknowledge that in some specific categories (e.g., car batteries) this may not be true. However, investigating why such phenomenon occurs and to what extent it is applicable to different products is orthogonal to the problem addressed in this paper. In the following, we show that the change of affinity rank is symmetric while that of affinity is asymmetric according to above discussion.

We now investigate the change of affinity rank between consecutive time slots in the history for each product. In particular, the change of affinity rank for a product between times $t-1$ and t is measured as follows.

$$\Delta\rho_t(p_i) = \rho_t(p_i) - \rho_{t-1}(p_i). \quad (5)$$

We calculate the count for each $\Delta\rho_t(p_i)$ value over all products and time slots. The distribution of $\Delta\rho$ is reported in Figures 3(c) and 3(d). Clearly, it follows a long-tail distribution. If

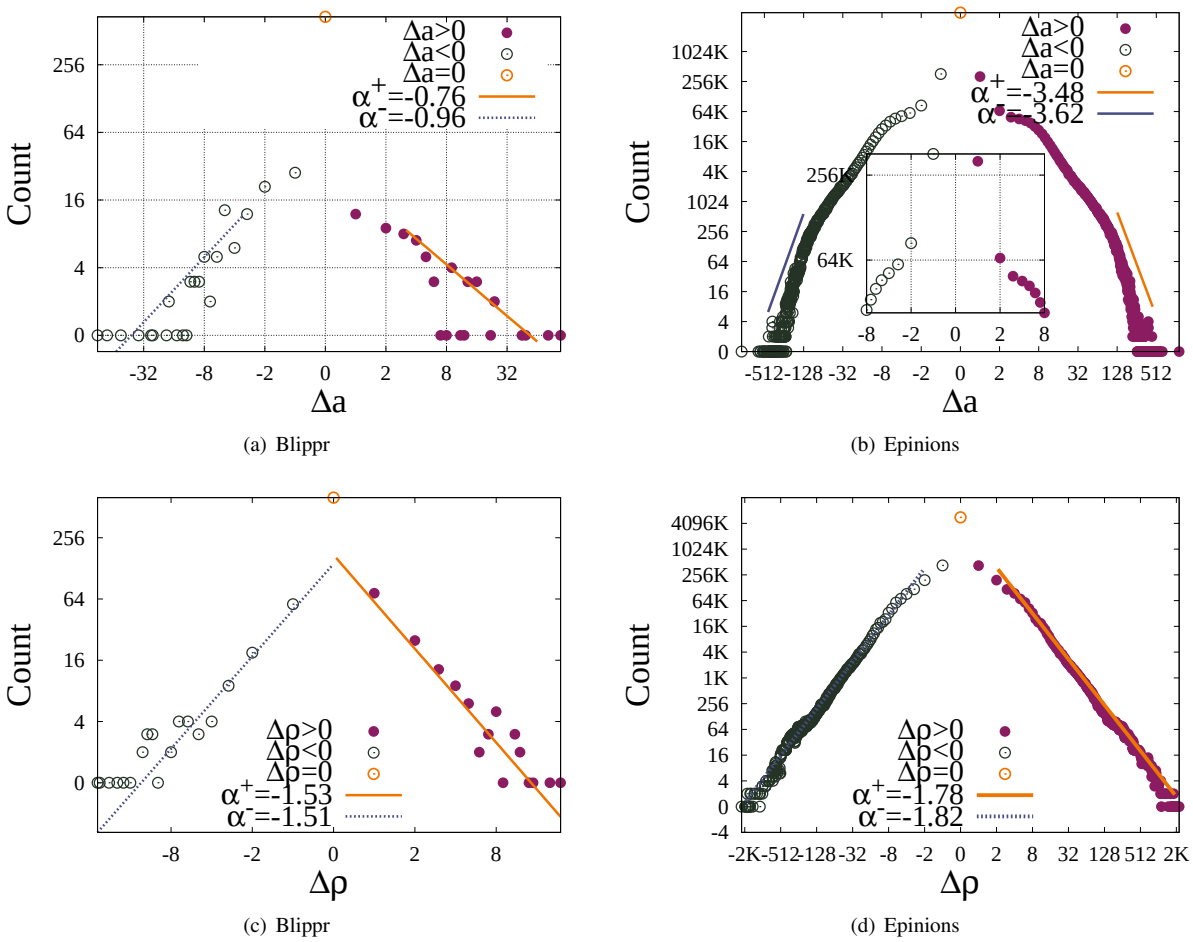


Figure 3: Distributions of Δa and $\Delta\rho$: $\Delta\rho$ show symmetrical patterns while Δa do not.

Table 5: Exponents in each specific category.

	Blippr					Epinions				
	M&T [6]	*Game*	*Music*	*Book*	*Appln* [7]	*Media*	*Cars*	*Elecs* [8]	*Games*	*H&G* [9]
α^+	-2.03	-1.50	-1.37	-1.30	-1.54	-1.80	-1.82	-1.78	-1.77	-1.78
α^-	-1.92	-1.51	-1.36	-1.35	-1.50	-1.81	-1.81	-1.83	-1.81	-1.82

we fit the curve in Figure 3(c) using power law models at both the left and right sides of $\Delta\rho = 0$ separately, the exponents equal to -1.51 for the negative changes where $\rho_{t-1}(p_i) > \rho_t(p_i)$ and -1.53 for the positive ones. The exponents are so close that the distributions of positive $\Delta\rho$ and negative $\Delta\rho$ are almost symmetrical. Figure 3(d) shows that *Epinions* dataset also exhibits the same behavior ($\alpha^+ = -1.78, \alpha^- = -1.82$). Besides, such a phenomenon also indicates that the change of rank is most probable to be within a small range. We observe that the aforementioned phenomenon also exists for each specific category of products in both datasets. Table 5 reports this phenomenon for a representative set of product categories. Thus, the rank of a product at any time slot is highly related to that of the previous time slot.

In summary, the distribution of Δa is asymmetrical over the positive and negative sides while the distribution of $\Delta\rho$ is symmetrical. In the next section, we shall exploit the symmetric property of $\Delta\rho$ instead of asymmetric Δa in our *AffRank* model. As we shall see in Section 5, our *AffRank* model outperforms the approach based on Δa.

3.3 Affinity Evolution Distance

We now analyze and compare the evolutionary nature of different product affinities. We begin by introducing the notion

[6] *Movie & TV*
[7] *Application*
[8] *Electronics*
[9] *Home & Garden*

of *affinity intensity*. Recall from Section 1, new users join a product community by giving new ratings.

Definition 1 ***[Affinity Intensity]*** Let $|\mathbf{U}_i^t|$ be the number of new users towards product p_i at time slot t. Then the *affinity intensity* of p_i at t is defined as:

$$\alpha_i(t) = \frac{|\mathbf{U}_i^t|}{\sum_{\tau=1}^{T} |\mathbf{U}_i^\tau|} \quad (6)$$

Note that $\mathbf{U}_i^t$ represents the set of new users towards product p_i at t. In the above definition, T is the number of time slots over which the affinity is normalized. The affinity intensity can be viewed as a normalized histogram where the values in each bin sum up to 1.

Figure 4 reports the evolution of affinity intensity values of five different products over time. Note that the label in front of a product name indicates the category of the product. The x-axis denotes the number of weeks since the product first appeared in the website, while y-axis represents the affinity intensity towards the product over different weeks. In the sequel, we refer to such curve as *Affinity Intensity Curve* (AIC). In the *Blippr* website (Figure 4(a)), *Gmail* and *Twitter* belong to the same category (`online applications`). *The Dark Knight* belongs to the `movie` category while the other two belong to the `game` category. Observe that out of the five AIC in Figure 4(a), the AIC of *Twitter* and *Gmail* look *similar*, the two products of `game` also exhibit similar AIC while that of *The Dark Knight* is quite different from the rest. We observe similar phenomenon for the *Epinions* website as well (Figure 4(b)). In other words, *these curves show that products in the same category tend to have similar affinity evolution patterns.* We now quantitatively measure the distance between affinity evolution patterns and investigate if this hypothesis holds.

As the AIC is a one-dimensional time series data, we can compute the distance between different curves using Dynamic Time Warping (DTW) distance. DTW distance is widely used to match similar time series data. In this paper, we adopt the DTW distance with Sakoe-Chiba band [36] which adds a window constraints w to the warping path found by DTW algorithm. As a result, the DTW algorithm will only match similar shaped data series that have small displacement within window w.

We compute the DTW distance between each pair of AIC. We fix the length of each data sample to be $T = 25$ weeks. We set $w \in \{T/4, T/5, T/6\}$. Interestingly, the average distances between products in the same category are always smaller than those from different categories for all values of w. Table 6 shows the average DTW distances between representative product categories in *Blippr* (for $w = T/5$). Similar phenomenon is observed in *Epinions* (Table 7).

We now elaborate on how we compute *affinity evolution distance* for a product p_i at time t (denoted by ϕ_i^t) using the notion of DTW distance. Since we intend to measure how similar an AIC of a product is compared to the product in the same category with most number of new users, we quantify ϕ_i^t by measuring the DTW distance between the AIC of a product p_i and the AIC of product p_j whose affinity rank is the highest in the same category. The procedure for computing ϕ_i^t is shown in Algorithm 1.

Table 6: Avg. DTW distance between different categories (Blippr)

	M&T				
M&T	**0.5376**	*Game*			
Game	0.6235	**0.5896**	*Music*		
Music	0.6891	0.6884	**0.6565**	*Book*	
Book	0.7324	0.6519	0.6765	**0.5993**	*Appln*
Appln	0.7088	0.6348	0.7414	0.6877	**0.4455**

Table 7: Avg. DTW distance between different categories (Epinions)

	Media				
Media	**0.4268**	*Cars*			
Cars	0.5975	**0.5226**	*Elecs*		
Elecs	0.7214	0.6194	**0.5837**	*Games*	
Games	0.7067	0.6322	0.6215	**0.6071**	*H&G*
H&G	0.6125	0.6571	0.6732	0.7025	**0.4455**

Algorithm 1: Affinity evolution distance.

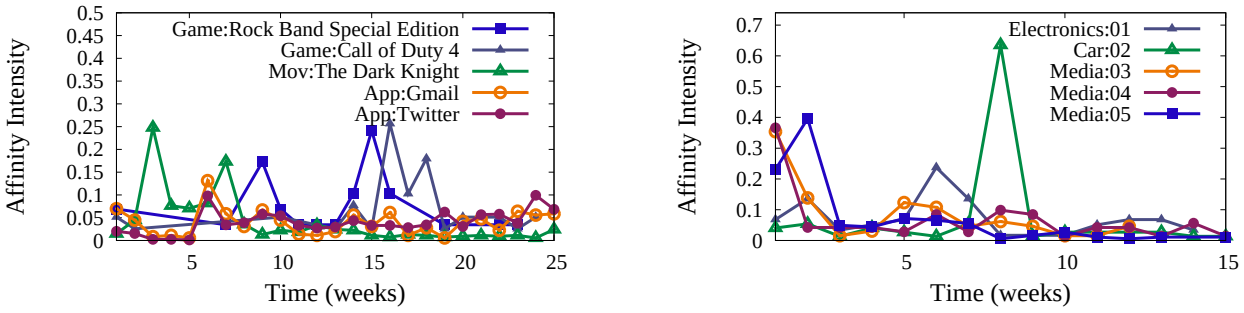
Input: affinity a_i^j for all products $p_i \in \mathbb{P}$ and $j \in [t-T, t-1]$,
Output: the value of *affinity evolution distance*: ϕ_i^t for products $p_i \in \mathbb{P}$ at time t

```
begin
    forall the p_i ∈ P do
        compute affinity intensity α_i(j) according to Definition 1
        for j ∈ [t − T, t − 1];
        λ_i = i;
        forall the p_j ∈ P and p_j.category = p_i.category do
            if ρ_{t−1}(p_j) < ρ_{t−1}(p_{λ_i}) then
                λ_i = j
    forall the p_i ∈ P do
        φ_i^t =
        DTW([α_i(t−T), ..., α_i(t−1)], [α_{λ_i}(t−T), ..., α_{λ_i}(t−1)]);
```

3.4 Average Rating

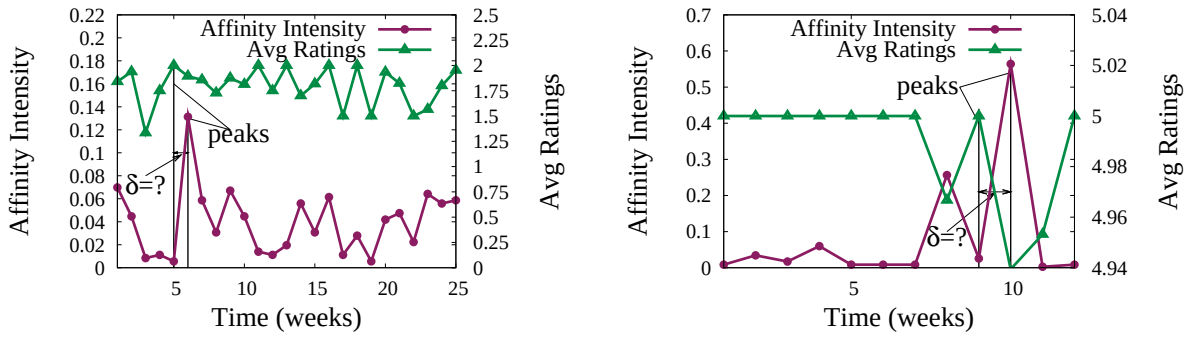
Observe that the affinity evolution distance feature describes evolutionary relationship between different products. We now focus on each individual product and investigate the relationship between a product's AIC and the evolution of *average ratings* it received. Note that users' ratings towards each product are explicitly provided in both datasets. In the case when the ratings are not explicitly provided in a specific dataset, an existing sentiment analysis model [33, 34] can be used to extract and quantify the users' comments into ratings. Extraction and quantification of such sentiments from users' comments is orthogonal to this work and hence it is not discussed here.

Definition 2 ***[Average Rating]*** Let $\mathbf{R}_i^t$ be the bag of ratings that product p_i received during time slot t. Then the *average rating*



(a) Blippr (b) Epinions

Figure 4: Affinity intensity evolution of different products.



(a) *Gmail* in Blippr (b) *Desitin Diaper* in Epinions

Figure 5: Lag between the average rating evolution and affinity intensity curve.

of p_i during t, denoted by $\overline{\mathbf{R}}_i^t$, is defined as:

$$\overline{\mathbf{R}}_i^t = \frac{1}{|\mathbf{R}_i^t|} \sum_{r \in \mathbf{R}_i^t} r. \tag{7}$$

Figure 5(a) depicts the AIC of *Gmail* in *Blippr* dataset as well as its average rating evolution. Observe that the evolutions of affinity intensity and average rating follow similar trend except that there is a certain delay δ between the peaks of the two curves. Similarly, Figure 5(b) shows the same phenomenon in *Epinions* dataset. In the following, we conduct a series of analysis on δ to characterize the relationship between users' ratings and affinity intensity.

Firstly, we study the correlation between the affinity intensity at time t and the average ratings at times $t - \delta$ (weekly). The correlation coefficients using different δ on the *Blippr* dataset are shown in Table 8. The affinity intensity is correlated with the average rating in the same week (week 0) with an average correlation coefficient of 0.7221. It is much higher than the correlation with average ratings in any of the previous weeks. The same phenomenon exists in *Epinions* as shown in Table 9. Thus, we can conclude that δ is less than a week.

Next, in order to find the exact value of δ, we conduct another set of experiments. For each product, we detect the first peaks [10] in both the AIC and average rating curve by days. After that, we compute the interval between the peaks of these two curves for each product. Finally, we calculate the probability distribution of δ according to the following equation.

$$P(\delta = n) = \frac{\left|\{p_i | Peak_\alpha(p_i) - Peak_{\overline{R}}(p_i) = n\}\right|}{|\{p_i\}|}. \tag{8}$$

where $Peak_\alpha(p_i)$ (resp. $Peak_{\overline{R}}(p_i)$) represents the day when the first peak appears in the AIC (resp. average rating evolution curve) of product p_i. The probability distribution of δ separated by category is reported in Figure 6. Observe that in *Blippr* the δ values for `music` are most likely to be 0 while the distribution of δ for `book` is stable across $[0, 2]$. Such a phenomenon may be due to the characteristics of products in various categories. Intuitively, the average ratings for musical product can take instant effect on future affinity intensity as potential users can listen to the music (often online) as soon as they see others' ratings. However, potential readers of a book need more time to purchase the book, read it, and present their ratings. In general, it is evident from Figure 6 that most of the δ values fall in the interval $[0, 3]$.

Based on the above observations, we incorporate the average users' ratings for $\delta \in [0, 3]$ days as a feature in our affinity rank

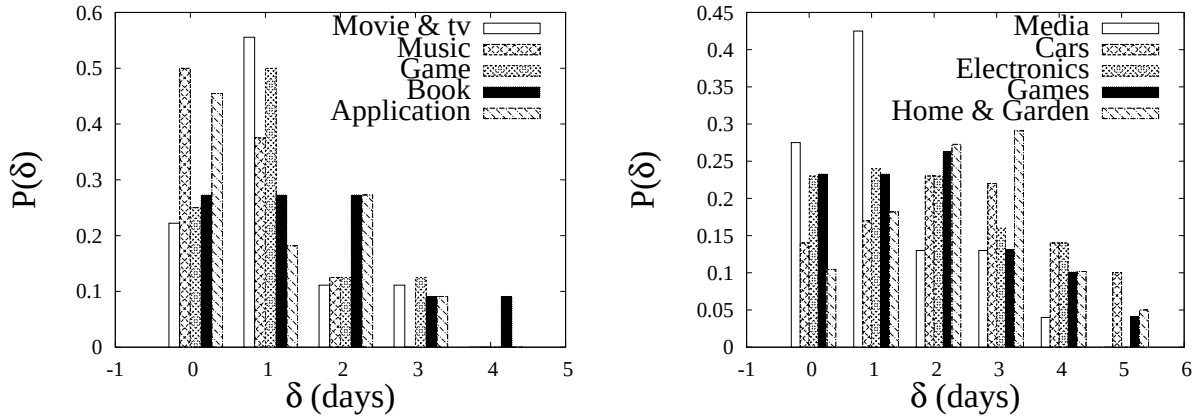
[10] We use the Matlab function "*peakdet*" downloaded from `http://www.billauer.co.il/peakdet.html` to detect the peak.

Table 8: Correlation between previous ratings and current affinity intensity (Blippr)

Lag time (in weeks)	*M&T*	*Game*	*Music*	*Book*	*Appln*
0	**0.6328**	**0.6737**	**0.8216**	**0.7347**	**0.7183**
1	0.1691	0.1745	0.1856	0.0860	0.4965
2	0.1132	0.0252	0.0484	0.0601	0.4121
3	0.0542	0.1460	-0.0600	0.0936	0.4319
4	0.0590	0.0608	0.0057	0.0418	0.3379

Table 9: Correlation between previous ratings and current affinity intensity (Epinions)

Lag time (in weeks)	*Media*	*Cars*	*Elecs*	*Games*	*H&G*
0	**0.4572**	**0.7101**	**0.7496**	**0.6883**	**0.6015**
1	0.2106	0.1061	0.1374	0.1140	0.1700
2	0.0892	0.0545	0.1007	0.1032	0.2093
3	0.0770	0.0819	0.1220	0.1529	-0.1059
4	-0.0600	-0.0940	0.1019	0.0429	0.1731



(a) Blippr (b) Epinions

Figure 6: Probability distribution of δ.

prediction model. If we denote the upper bound of δ as ℓ, then this feature can be computed as $\overline{\mathbf{R}}_i^{t-\delta-1}$ over all $\delta \in [0, \ell]$.

4 Affinity Ranks Prediction

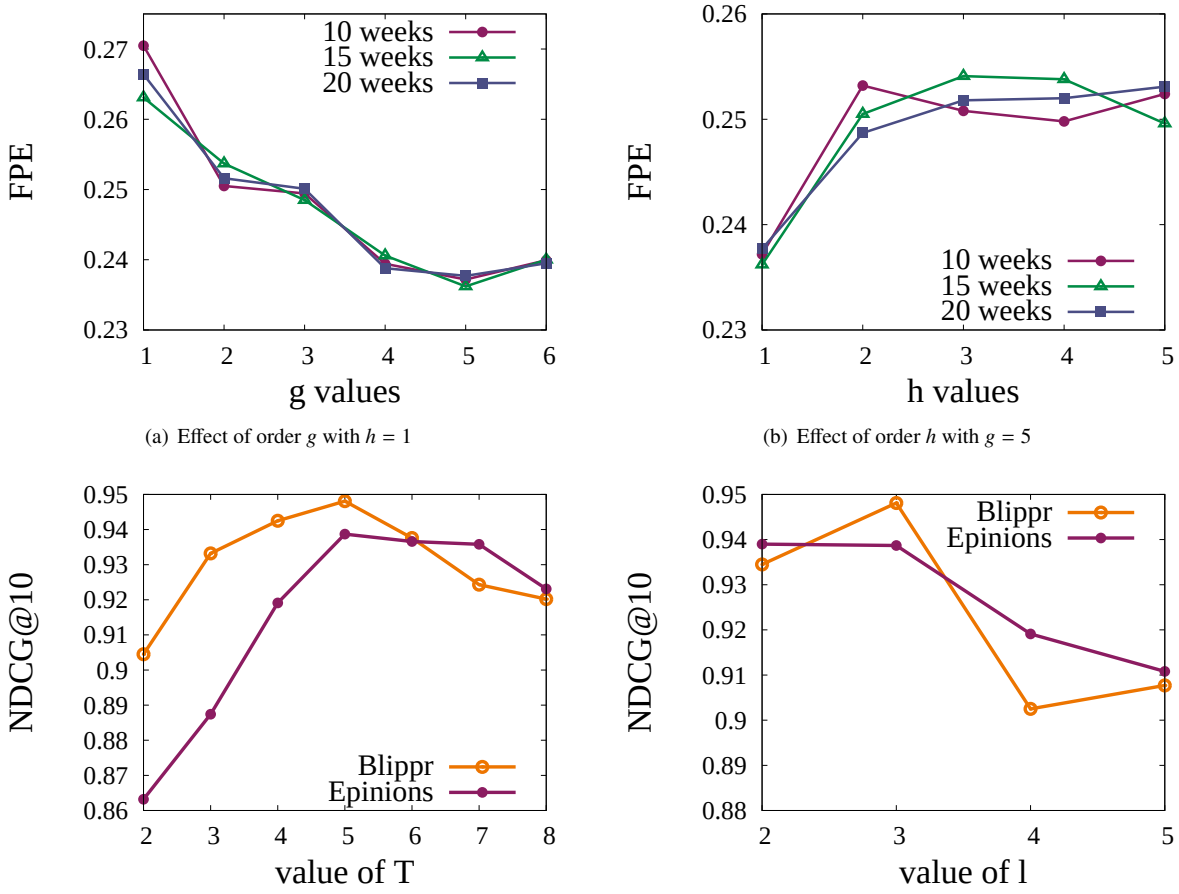
In this section, we propose the *AffRank* model in detail and present an algorithm to predict the affinity ranks.

It is evident from our earlier discussions that the affinity rank of a product is highly related to its ranks in the near past. However, how long of the past history should be taken into account is unknown yet. Thus, SVM-based regression technique (and its variants) cannot be adopted as it is difficult to define data samples with unknown dimensions. Instead, ARX (AutoRegressive model with exogenous inputs) is best suitable for this case as discussed in [8]. ARX model is capable of incorporating external inputs and is widely used in modeling various types of natural and social phenomena [8]. More importantly, it can find the best length of the period that should be taken into account by simply varying the *order* parameter in the model. Additionally, as ARX is a linear model, the weight of each feature clearly indicates how important that feature is. We can simply find from the features which are important and which can be ignored.

The ARX model of orders g and h is given in the following equation.

$$y_t = \varepsilon_t + \sum_{i=1}^{g} \varphi_i y_{t-i} + \sum_{j=1}^{s} \sum_{i=1}^{h} w_{i,j} b_{t-i,j} \tag{9}$$

In this equation, y is the time series data (*i.e.,* product ranks in various different time slots), s is the number of exogenous input features; $\varphi_1, \ldots, \varphi_g$ and $w_{1,1}, \ldots, w_{h,s}$ are the parameters to be estimated from the training data. Both g and h are the orders in the model to be manually determined before model estimation. In our context, g is the *order of product rank* which determines the number of previous product ranks to be considered in the modeling; h is the *order of feature* determining the number of past time slots from which the values of the corresponding features to be involved in the model estimation. The variable $b_{t-i,j}$ is the value for feature j at time $t - i$, and ε_t is white noise. The estimation of the ARX model is efficient as it solves linear regression equations in analytic form. Also, the solution is unique and always satisfies the global minimum of the loss function [8].



(a) Effect of order g with $h = 1$

(b) Effect of order h with $g = 5$

(c) Effect of time interval T in *affinity evolution distance*

(d) Effect of lag ℓ in *average rating*

Figure 7: (a)(b)ARX model order optimization; (c)(d)Feature parameter setting.

In the ARX model, the order of product rank g and the order of feature h need to be manually determined [28]. A common way of selecting g and h is to fix the value of g (or h) and try a range of values for h (or g); for each pair of g and h values, evaluate the accuracy of the model estimated using a measure called FPE [11] . A smaller FPE value means a more accurate model. Figures 7(a) and (b) report the FPE measures of varying values of g while fixing $h = 1$, and varying h while fixing $g = 5$, respectively. Based on the FPE measures, we set $g = 5$ and $h = 1$ for our experimental study (Section 5). Observe that in Figures 7(a) and (b), the size of the training data (*e.g.,* 10, 15, 20 weeks) has marginal impact on the accuracy of the estimated model for a given pair of g and h values. Hence in our experimental study, we use the latest 10 weeks data to estimate the ARX model.

We now present the algorithm to predict the affinity ranks using the learned ARX model. Detailed in Algorithm 2, for each product, the value of each feature j discussed in Section 3 at time $t - i$ (denoted by $b_{t-i,j}$) is derived and stored in feature vector Ψ_i^t. The target value y_t is predicted by the model using Ψ_i^t.

[11] More details of FPE measure can be found in [2].

Algorithm 2: Affinity rank prediction.

Input: ARX model parameters $\{\varphi_1, \varphi_2, \ldots, \varphi_p\}$, $\{w_{1,1}, w_{1,2}, \ldots, w_{h,s}\}$, test dataset Ψ_i^t for $p_i \in \mathbb{P}$
Output: the affinity rank $r_t(p_i)$ for all products $p_i \in \mathbb{P}$ at time t
begin
  while $j < \mathbb{P}.size$ **do**
    compute y_t^i according to Equation 9 using Ψ_i^t;
    $Y[j] = y_t^i$;
    $j++$;
  sort(Y) by *ascending order*;
  forall the $Y[j] = y_t^i$ **do**
    $r_t(p_i) = j + 1$;

5 Experiments

In this section, we report experimental results on *Blippr* and *Epinions* datasets. On *Blippr*, the experiments are conducted on all 75 products with statistics reported in Table 4. On *Epinions*, a subset of 1,311 products, each of which received at least 200 ratings, is used in our experiments. There are in total 343,154 ratings in this subset. The goals of the evaluation were to estab-

lish whether the proposed *AffRank* model can reliably predict product affinity ranks; compare the performance of the *AffRank* model and three baseline models; and seek to understand the importance of various features in the proposed model.

We begin by reporting the performance metric used to evaluate the accuracy of the predicted product affinity ranking.

Performance Metric. To evaluate the accuracy of the predicted product rank against the ground-truth rank at a given time slot, we adopted Normalized Discounted Cumulative Gain (NDCG) [19] measure which is commonly used in Information Retrieval (IR) to evaluate the effectiveness of ranking algorithms for Web search and other related applications. It is defined in the following equation.

$$NDCG@k = \frac{1}{Z} \sum_{i=1}^{k} \frac{rel_i}{\log_2 (i+1)} \tag{10}$$

In the above equation, i is the rank position and k is the number of top-ranked retrieved documents to be considered in NDCG computation. rel_i is the relevance of the ith retrieved document in IR setting. With a perfect ranking, a more relevant document shall be ranked higher than the less relevant document and so on. Lastly, Z is a normalizing constant that ensures the perfect ranking (i.e., the ground-truth rank in our case) achieves $NDCG@k$ of 1.0.

Note that we chose the aforementioned metric for the following reason. In viral marketing and online advertising context, logarithmic decay by positions is a reasonable assumption when products are presented in a list format [1]. For instance, consider the scenario discussed in Section 1.1. A company may allocate different investment budget for each movie in the top-5 list. Obviously, more investment should be put into higher ranked movies. Hence, accurate prediction of top ranked movies is much more important compared to lower ranked movies. Consequently, we use NDCG as it employs a log function to discount the positions and as a result the top positions are considered more valuable than the lower positions.

In our experiments, we vary k from 5 to 25 at the step of 5 to evaluate the accuracy of the rank involving top-k ranked products. We define the relevance of a product p_i to be the inverse of its ground-truth rank: $\frac{1}{\rho_t(p_i)}$. Observe that the value of NDCG heavily depends on the definition of relevance (i.e., rel_i). However, for a given relevance definition (e.g., $\frac{1}{\rho_t(p_i)}$), NDCG well reflects the accuracies of different ranking models.

Feature Parameter Setting. Following the approach reported in [28], we now evaluate the impact of the feature parameters. The two feature parameters are the number of time slots T involved in the computation of *affinity evolution distance* (see Section 3.3), and the time lag ℓ in computing *average rating* (see Section 3.4). With the ARX model fixed with g=5 and h=1, we evaluate the impact of T while fixing ℓ=3. As reported in Figure 7(c), the ranking model achieves the best performance when $T = 5$. Similarly, we fix $T = 5$, and report the effect of varying ℓ in Figure 7(d). On *Blippr dataset*, the best performance is achieved when $\ell = 3$; On *Epinions dataset*, the best performance is when $\ell = 2$ followed by a marginal drop in performance when $\ell = 3$. On both datasets, the performance drops significantly when ℓ is greater than 3. Hence, in the sequel we set $T = 5$ and $\ell = 3$.

Comparison to Baseline Methods. In this section, we compare the performance of the proposed *AffRank* with three other methods.

***LazyRank*.** This model predicts product rank at time slot t to be the same as the rank obtained in the last time slot $t-1$.

***AR*.** *AR* model refers to the AutoRegressive model without taking exogenous input features (see Equation 9). In another word, the product rank y_t is predicted solely by the ranks in the past g time slots, y_{t-g} to y_{t-1}, for a given product rank order g.

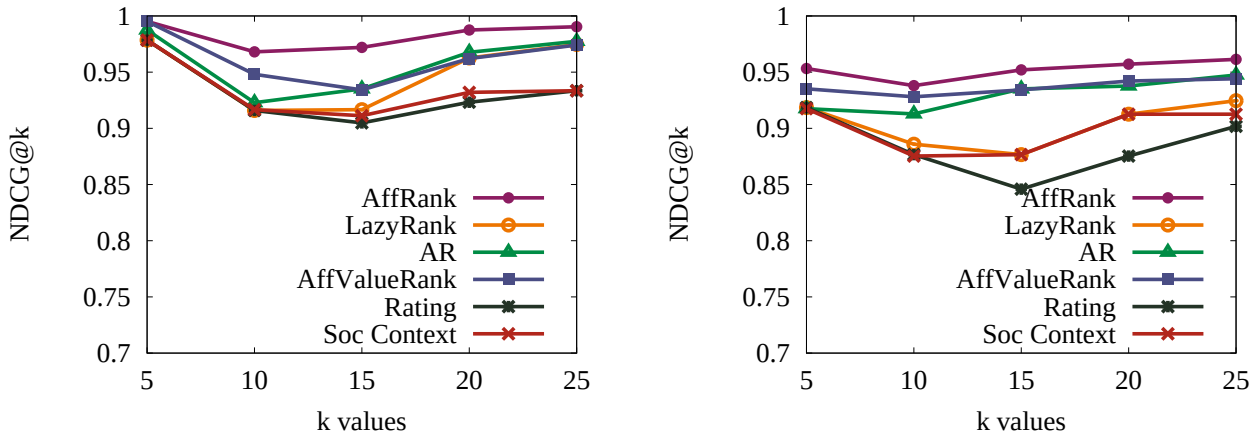
***AffValueRank*.** With *AffValueRank*, instead of predicting the affinity rank, the exact affinity value is predicted using the ARX model. That is, y_{t-i} in Equation 9 is set to be the affinity a_i^{t-i}. Using the learned model, the affinity values a_i^t is predicted; the products are then ranked accordingly.

As reported in Figure 8, the proposed *AffRank* model outperforms all three baseline models on both datasets for every k value. Overall, *AffValueRank* is the second best performing model followed by *AR*. The *LazyRank* model performs the worst. Considering the features involved in the four models, the experimental results show that, (i) product affinity rank can be better predicted using the past few product ranks than the single last rank (i.e., *AR* > *LazyRank*), (ii) the extra features besides the product rank lead to better prediction (i.e., *AffValueRank* > *AR*), and (iii), product affinity ranks can be better predicted than the affinity values (i.e., *AffRank* > *AffValueRank*) probably due to the smoother distribution of product affinity ranks than affinity values (see Section 3.2).

We also tested two other baseline methods: one is to rank products by actual user rating (*i.e., Rating*) and the other is to rank products by the *social context* feature (*i.e., Soc Context*). However, none of these approaches can outperform the simplest *LazyRank* model as shown in Figure 8.

Case Study. Table 10 shows an example of predicted affinity rank in *Blippr*. The second and third columns show the top-10 ranked products in weeks 12 and 13 of year 2009, respectively. The remaining 3 columns list the predicted ranks using different models. Obviously, our model *AffRank* accurately predicts the top 10 products in week 13 although the exact ranks for products *Google Earth*, *Google* and *Tweetie for iPhone* are not accurate. Besides, our model also detects that *Google Earth* will acquire more affinity than *Google* in week 13. Compared to week 12, there are four products newly listed in top-10 (i.e., *Tweetie for iPhone*, *The Dark Knight*, *The Shawshank Redemption*, and *Watchmen*). Our *AffRank* predicted all four accurately, *AffValueRank* missed one of them, *AR* missed two of them, and the *LazyRank* missed all four.

Affinity Feature Comparison. In the learning of the ARX model, all feature values were normalized into the same range of [0, 1]. Hence, the larger the learned weights (e.g., φ_i and $w_{i,j}$



(a) Blippr (b) Epinions

Figure 8: Comparison with other models.

Table 10: Top 10 popular products in weeks 12 & 13 (2009) and predicted ranks. (Blippr)

Rank	*Week 12 (LazyRank)*	*Week 13*	*Predicted rank in week 13*		
			AffRank	*AR*	*AffValueRank*
1	Twitter	Twitter	Twitter	Twitter	Twitter
2	Gmail	Gmail	Gmail	Gmail	Gmail
3	Mashable	Mashable	Mashable	Mashable	Mashable
4	Google	Tweetie for iPhone	Google Earth	Google Earth	Google
5	Google Earth	Google Earth	Google	Google	Google Earth
6	OK Computer	Google	Tweetie for iPhone	Google Reader	The Dark Knight
7	Google Reader	The Dark Knight	The Dark Knight	The Dark Knight	Tweetie for iPhone
8	Dropbox	The Shawshank Redemption	The Shawshank Redemption	Tweetie for iPhone	The Shawshank Redemption
9	WordPress	Google Reader	Google Reader	WordPress	Google Reader
10	In Rainbows	Watchmen	Watchmen	OK Computer	OK Computer

in Equation 9) in the ARX model the more strongly it influences product affinity ranking. Table 11 reports the average learned weights over both datasets. Observe that besides *affinity rank history*, *affinity evolution distance* has the biggest weight value which indicates that it plays an important role in the model. As *average rating* feature has a negative weight value, it suggests that a larger rating leads to a higher rank. Also observe that the *affinity rank history* (φ_1 to φ_5) follows a decreasing tend, suggesting that the more recent product ranks have more impact on its future rank. However, the three traditional features, listed in the bottom part of the table, all have very small weights indicating their negligible impact on product affinity ranking.

In summary, in social rating networks, community size, member connectivity, and social context do not have significant impact on the affinity of a product. Instead, growth of a product community is mainly because of its high affinity ranks in the past few weeks (which may make the product reach larger audience by appearing in the top ranked list on the website's homepage) and the received good ratings from users. Besides, a product showing similar affinity intensity evolution pattern with the most popular product is also likely to be popular in near future.

Table 11: Learnt weights of features with $g = 5, h = 1$.

Feature	*Weight* φ, w in Equation 9
Aff. Rank History (φ_1 - φ_5)	[0.720, 0.295, 0.267, 0.131, 0.113]
Aff. Evolution Distance	0.210
Average Rating	-0.124
Community Size	0.021
Social Context	0.007
Member Connectivity	−0.013

6 Conclusions and Future Work

In this paper, we analyzed two publicly-available social rating networks and proposed a predictive model called *AffRank*, that utilizes an array of features to predict the future rank of products according to their affinities. Informally, product affinity refers to a product community's ability to "attract" new members and is measured by the number of new ratings during a specific time slot. Such information plays an important role in several real-world applications such as online advertisement and marketing research.

We formulate the product affinity prediction problem as an

autoregressive model with exogenous inputs (features). We have identified in total six features, namely community size, member connectivity, social context, affinity rank history, affinity evolution distance, and average rating, for predicting product affinity. Our investigation revealed that the community size, member connectivity, and social context do have negligible influence on product affinities. Instead, the remaining features are the most important factors affecting future rank of products. Specifically, we discovered several interesting findings related to these features which we exploit in our model. Firstly, affinity of a product for most products tends to increase spikily and decrease smoothly. Secondly, the average DTW distances between products in the same category are always smaller than those between products from different categories. Thirdly, we studied the lag between the peak in the average rating curve and that in the affinity intensity curve. The affinity intensity is highly correlated with the users' average ratings. Particularly, as the average rating increases the affinity intensity increases accordingly within 3 days. Our exhaustive experimental study demonstrates the effectiveness and superior prediction quality of *AffRank* compared to three baseline methods.

As part of our future work, we intend to investigate prediction of future ranks of *newly-released* products (products that appear in the social rating networks less than a month ago). These types of products do not have sufficient historical information in the SRNS to make accurate prediction. Hence, *external* information (*e.g.,* online news media) may need to be exploited for predicting product affinities.

References

[1] D. Agarwal, E. Gabrilovich, R. Hall, V. Josifovski, and R. Khanna. Translating relevance scores to probabilities for contextual advertising. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1899–1902, 2009.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Aut. Contr.*, 19:716–723, 1974.

[3] N. Archak, A. Ghose, and P. G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65, New York, NY, USA, 2007. ACM.

[4] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *TKDD*, 3(4), 2009.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM.

[6] N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa. Seeking stable clusters in the blogosphere. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 806–817. VLDB Endowment, 2007.

[7] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, New York, NY, USA, 2006. ACM.

[8] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, March 2002.

[9] M. Cha, A. Mislove, and P. K. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, pages 721–730, 2009.

[10] C. Chen, F. Ibekwe-Sanjuan, and J. Hou. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *JASIST*, 61(7):1386–1409, 2010.

[11] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, August 2006.

[12] M. de Klepper, E. Sleebos, G. van de Bunt, and F. Agneessens. Similarity in friendship networks: Selection or influence? the effect of constraining contexts and non-visible individual attributes. *Social Networks*, August 2009.

[13] H. Fock, A. K. Chan, and D. Yan. Member-organization connection impacts in affinity marketing. *Journal of Business Research*, In Press, Corrected Proof:–, 2010.

[14] A. Ghose and A. Sundararajan. Evaluating pricing strategy using ecommerce data: Evidence and estimation challenges. *Statistical Science*, 21(2):131–142, 2006.

[15] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, New York, NY, USA, 2005. ACM.

[16] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Springer, Morgan Kaufmann Publishers Inc. 2005.

[17] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[18] P. Holme, C. R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174, May 2004.

[19] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[20] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data*, 4(1):1–24, 2010.

[21] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 741–750, New York, NY, USA, 2009. ACM.

[22] J. H. Lee. Analysis of user needs and information features in natural language queries seeking music information. *JASIST*, 61(5):1025–1045, 2010.

[23] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *TWEB*, 1(1), 2007.

[24] H. Li, S. S. Bhowmick, and A. Sun. Blog cascade affinity: analysis and prediction. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1117–1126, New York, NY, USA, 2009. ACM.

[25] H. Li, S. S. Bhowmick, and A. Sun. Affinity-driven prediction and ranking of products in online product review sites. In *CIKM '10: Proceeding of the 19th ACM conference on Information and knowledge management*, 2010.

[26] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM.

[27] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM Trans. Knowl. Discov. Data*, 3(2):1–31, 2009.

[28] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–614, New York, NY, USA, 2007. ACM.

[29] Y. Liu, X. Yu, X. Huang, and A. An. S-plasa+: adaptive sentiment analysis with application to sales performance prediction. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 873–874, New York, NY, USA, 2010. ACM.

[30] B. MacChiette and A. Roy. Direct marketing to the credit card industry: Utilizing the affinity concept. *Journal of Direct Marketing*, 5(2):34 – 43, 1991.

[31] Y. Matsuo and H. Yamamoto. Community gravity: measuring bidirectional effects by trust and rating on online social networks. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 751–760, New York, NY, USA, 2009. ACM.

[32] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155–163, May 2009.

[33] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 751–760, New York, NY, USA, 2010. ACM.

[34] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[35] K. H. Q. Feng and Y. Dai. Rainbow product ranking for upgrading e-commerce. *IEEE Internet Computing*, 13(5):72–80, 2009.

[36] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[37] B. Shapira and B. Zabar. Personalized search: Integrating collaboration and social networks. *JASIST*, 62(1):146–160, 2011.

[38] H. T. Shen, B. C. Ooi, and X. Zhou. Towards effective indexing for very large video sequence database. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 730–741, New York, NY, USA, 2005. ACM.

[39] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 655–664, New York, NY, USA, 2008. ACM.

[40] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Monic: modeling and monitoring cluster transitions. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 706–711, New York, NY, USA, 2006. ACM.

[41] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 687–696, New York, NY, USA, 2007. ACM.

[42] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *IEEE Trans. Knowl. Data Eng.*, 22(2):179–192, 2010.

[43] J. Tang and J. Zhang. Modeling the evolution of associated data. *Data Knowl. Eng.*, 69(9):965–978, 2010.

[44] M. Vojnovic, J. Cruise, D. Gunawardena, and P. Marbach. Ranking and suggesting popular items. *IEEE Trans. Knowl. Data Eng.*, 21(8):1133–1146, 2009.

[45] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(1):440–442, June 1998.