Unravelling the Dynamics of Online Ratings

Larissa Spinelli
Boston University
Email: lspinell@bu.edu

Mark Crovella
Boston University
Email: crovella@bu.edu

***Abstract*—Online product ratings are an immensely important source of information for consumers and accordingly a strong driver of commerce. Nonetheless, interpreting a particular rating in context can be very challenging. Ratings show significant variation over time, so understanding the reasons behind that variation is important for consumers, platform designers, and product creators. In this paper we contribute a set of tools and results that help shed light on the complexity of ratings dynamics. We consider multiple item types across multiple ratings platforms, and use a interpretable model to decompose ratings in a manner that facilitates comprehensibility. We show that the various kinds of dynamics observed in online ratings are largely understandable as a product of the nature of the ratings platform, the characteristics of the user population, known trends in ratings behavior, and the influence of recommendation systems. Taken together, these results provide a framework for both quantifying and interpreting the factors that drive the dynamics of online ratings.**

***Index Terms*—Recommender System; User Experience; Metrics**

I. Introduction

One of the ways that the Web has revolutionized society is through crowdsourced reviews. Almost any situation in which alternative choices may be evaluated is now supported by one or more review systems that record experiences and ratings that users have provided for items of interest.

From the standpoint of the review user, the value of a review system is to allow the user to assess the perceived quality of various alternatives before making a decision. However, there is considerable evidence that online reviews show considerable temporal dynamics [1], [2], so an important question concerns how to understand the dynamics of item reviews before using them to make decisions.

In this paper we seek to understand how item ratings change over time and what factors affect those changes. This is a complex question because there are many dimensions that can play a role in review dynamics. Of course, ratings may shift because the popular perception of an item is actually changing within the user population. However, many more factors come into play. For example, ratings can be affected by shifts in the nature of the population of users providing ratings. They can be affected by closed-loop effects in which previous ratings influence the set of users that are interested in and subsequently review the item. Ratings shifts can occur for some items in a manner that is different from other items. Furthermore, the dynamics of ratings can differ among different ratings platforms – even for the same set of items.

In this paper we show how to tease apart all of these effects, characterize them, and quantify their relative importance. Our goal is to form an integrated view of how the interplay of these effects ultimately determine the changes in item ratings over time. To do so, we make use of a variety of datasets, chosen for their ability to explore all of the questions above. To study platform effects, we study the dynamics of movie ratings across three major ratings platforms; and to study item category effects, we study various item categories on a single platform. We use clustering to distinguish items showing different rating dynamics on the same platform. And within a given platform and item category, we fit a nonlinear model that allows us to distinguish factors such as changes in user population, changes in user behavior, intrinsic changes in perceived item quality, and closed-loop interactions between previous ratings and changes in user population. This latter factor essentially captures the impact of the ratings platform as a recommender system.

Our multi-platform, multi-factor study goes beyond prior work by considering a much broader set of factors than previous studies. Using our methods, we show that there are consistent differences in rating dynamics that depend on the nature of the rating platform. These differences are not due to different sets of items being rated on different platforms – they persist even when looking at the same set of items on different platforms. We also show that on each platform, there are understandable shifts in the kinds of users that rate an item over time, and that in each case this population shift makes sense due to the nature of the platform. We show that there are consistent general trends in how perceived item quality changes over time, which are understandable in light of past studies. And we show that recommender systems play a role in affecting rating dynamics on some platforms, but not others, in a way that correlates with the nature of the rating platform.

II. Factors Affecting Rating Dynamics

Our goal is to form a holistic picture of the forces that combine to determine online rating dynamics. In particular, we seek to understand how the following factors interact in shaping online ratings:

Platform Characteristics. We consider first, does the platform explicitly support item sales, or is it purely informational? And second, does the platform provide a recommendation system as a service, or does it merely display ratings?

User Population. We want to evaluate whether their are different user types, and if so whether the balance among those types changes over time, and how those shifts affect ratings dynamics.

Item Perception. We seek to quantify the extent to which the popular perception of an item is shifting over time. This can reflect a shift in tastes in the population at large, or a tendency for a less-appreciated item to become better appreciated by the population over time.

Item Type. We seek to understand whether different types of items show different dynamics, and why. We also seek to understand the prevalence of non-trivial dynamics, i.e., the proportion of items within a category that typically show detectable dynamics over time, as opposed to the proportion of items whose ratings are approximately unchanging.

Closed-Loop Effects. Finally, we are interested in the extent to which online ratings or recommendations affect the set of users that subsequently consume an item, leading to shifts in dynamics of future ratings. This tells us the impact of "tuning" between items and the users that consume and rate the items, a tuning that is induced by recommendations.

To separate and evaluate these effects, we use the data and methods described in the next section.

III. Methods

In order to effectively disentangle all of these effects, we use a combination of carefully chosen datasets, unsupervised learning in the form of clustering, and supervised learning in the form of a model fitted to our various datasets.

A. Data

We make use of the following datasets to help distinguish the five factors above:

Movie Tweetings. This dataset is collected from well-structured movie evaluation tweets on Twitter from 2013 to 2017 [3]. This dataset represents a platform in which there is no explicit recommendation system, and there is no commercial entity providing the reviews for the purpose of commerce.

We selected a relatively dense subset of this dataset, namely movies that have at least 10 ratings and users that have at least 5 ratings. We denote this dataset `MT`. This dataset has 15632 users, 5780 movies and 521214 ratings. We centered the ratings in `MT` by rescaling them from [1:10] to [1:5].

Rotten Tomatoes. This dataset was crawled from the Rotten Tomatoes website [4] in late 2016, which we denote as `RT`. This dataset represents a platform in which there is a known distinction between two user types: *critics,* and *general users.* Like `MT` there is no explicit recommendation system or commercial role for the platform.

From the entire dataset we also selected the subset consisting of movies that have at least 10 ratings and users that have at least 5 ratings. The resulting dataset has 165585 users, 12122 movies, and 4845884 ratings. We centered the ratings in `RT` by rescaling them to the range [1:5].

Amazon. This dataset contains product reviews from Amazon spanning from May 1996 to July 2014 [5] and [6]. This dataset represents a platform in which there is an explicit recommendation system that makes personalized purchase suggestions to users. The platform also has a commercial role in support of sales in the Amazon store. Furthermore, the Amazon dataset contains items from multiple categories. In addition to movies, we use it to study electronics, home goods, CDs, mobile apps, and ebook (Kindle) titles. From the Movies and TV category, we first disambiguated movies names, including merging movies available in different media such as DVDs and BluRay which appeared as separate products. Next we select a dense subset of movies that had at least 5 reviews. We denote this dataset `AZ`, and it has 1957899 users, 53633 movies, and 4291173 ratings.

From the other categories, we selected their 5-core dataset - the dense subset of items with at least 5 reviews and users with at least 5 reviews. The resulting datasets are: Electronics (`AZ`-Ele) and with 192401 users, 63001 items, and 1689129 ratings; Home and Kitchen (`AZ`-Hom) and with 66518 users, 28237 items, and 551656 ratings; Kindle Store (`AZ`-Kin) and with 68222 users, 61933 items, and 982197 ratings; Apps for Android (`AZ`-App) with 87267 users, 13209 items, and 752832 ratings; and CDs and Vinyl (`AZ`-CDs) with 75256 users, 64443 items, and 1097555 ratings. Note that in what follows, `AZ` refers to Amazon movies, while the other Amazon categories have specialized names.

B. Modelling Temporal Dynamics

1) Definitions: In each application of our model, we consider a dataset having n users and m items. Items are objects over which the user provides a rating, e.g., movies. Each rating has an associated timestamp t (in units of days), and we denote a rating provided by user u for item i at time t as $r_{ui}(t)$. All ratings range from 1 (worst) to 5 (best).

For each rating, we define an associated *system time* which is the time since the item first appeared in the system. That is, if $t_0^{(i)}$ is the timestamp of item i's first recorded rating and t the timestamp of a given rating $r_{ui}(t)$, the system time for that rating is $t_s = t - t_0^{(i)}$.

In presenting our results, we are primarily concerned with *item progression.* This is defined as the index of where a review falls in the ordered set of reviews for an item. So item progression from 0 to 99 reflects the first 100 reviews of an item in order (regardless of how much real time elapsed between the first and last reviews in the sequence).

2) Model: To separate the factors at work in a single dataset, we fit the data to a predictive model we call `timeSVD--`. This model is a simplified version of the `timeSVD++` for collaborative filtering as proposed in [7] .

To model a rating $r_{ui}(t)$, `timeSVD--` incorporates three kinds of information. First, it uses properties of the user u: a term capturing the user's time-invariant average rating (bias), and a term capturing the evolution of the user's average rating over time. Second, it uses properties of the item i: a term capturing the item's time-invariant average rating, and a term capturing the evolution of the item's average rating over time. Finally, it incorporates latent factors for both the user and item, whose inner product models the personalization of the item to the user. This latter factor is essentially a matrix-factorization approach to personalization (as reflected by the 'SVD' in the name of the model).

Specifically, `timeSVD--` is parameterized as follows:

μ	Global mean of all ratings
b_i	Time-invariant bias (average rating) of item i
$b_{i,Bin(t)}$	Time-varying bias of item i at timebin $Bin(t)$
b_u	Time-invariant bias of user u
$\alpha_u \text{dev}_u(t)$	Time-varying bias of *user* u
q_i	k-dimensional latent factor of item i
$p_u(t)$	Time-varying k-dimensional latent factor of user u

The model reflects the assumption that user preferences may change over time ($p_u(t)$) while item features are time-invariant (q_i).

The `timeSVD--` model is then:

$$r_{ui}(t) = \mu + b_i + b_{i,Bin(t)} + b_u + \alpha_u \text{dev}_u(t) + q_i^T p_u(t) \quad (1)$$

`timeSVD--` incorporates various strategies to capture time evolution of model components without unduly expanding the set of parameters to be learned. In the case of item bias, time is discretized into bins of seven days. For time-varying user parameters, the model fits a symmetrized polynomial:

$$\text{dev}_u(t) = sign(t - t_u)|t - t_u|^\beta$$

where t_u is the mean date of rating of *user* u. This function is used in time-varying user bias as well as in the time-varying user latent factor:

$$p_{u\ell}(t) = p_{u\ell} + \sigma_{u\ell}\text{dev}_u(t) \qquad \ell = 1, \ldots, k$$

In the rest of the paper, we will refer to $q_i^T p_u(t)$ as the *interaction* score between u and i, and the rest of the terms in (1) as the *baseline* score between u and i.

We train `timeSVD--` on each dataset using system time t_s as the value of t for each rating. To learn model parameters we apply stochastic gradient descent to a risk function incorporating a regularization to (1):

$$f(\theta) = \textstyle\sum_{\text{all ratings}} (r_{ui}(t) - (\mu + b_i + b_{i,Bin(t)} + b_u + \alpha_u \text{dev}_u(t) + q_i^T p_u(t)))^2$$
$$+\gamma(\textstyle\sum_i (b_i^2 + ||q_i||^2 + \sum_{Bin(t)} b_{i,Bin(t)}^2) + \sum_u (b_u^2 + \alpha_u^2 + ||p_u||^2 + ||\sigma_u||^2))$$

We set model hyperparameters γ and β by cross-validation.

4) Clustering: Within a particular dataset, we expect different items to show different dynamics over time. In order to efficiently separate items by the properties of their ratings dynamics, we use a clustering algorithm well-suited to work on timeseries: `k-Shape` [8]. To study factors at work for different kinds of items, we apply both `timeSVD--` and `k-Shape,` and take averages of the `timeSVD--` results over clusters identified by `k-Shape`.

IV. Analysis

We divide our analysis into two parts: first we characterize the range of observed phenomena in review dynamics, and then we decompose those phenomena to gain understanding of how they arise.

A. Characterizing Ratings Dynamics

Our basic tool for studying review dynamics is the *relative rating score*. This is the average value of ratings on a daily basis, offset by a constant that makes the first set of average ratings equal to zero. We call the average value of the first item ratings the *initial score* and the average across the study period (usually 200 reviews) the *average score*. We focus on *item progression* which, as described above, is the ordered sequence of reviews for an item.

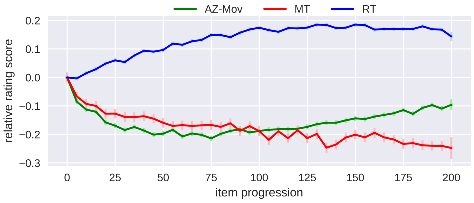


Fig. 1: Relative ratings progression

Throughout our analysis consider only movies with at least 200 reviews, and analyze the first 200 reviews. This means that the set of movies contributing to each average rating is not changing over time.

How do item ratings change over time? We start by addressing this basic question in Figure 1. This figure shows that each dataset shows distinctive behavior. The `RT` dataset shows a generally increasing trend; the `MT` dataset shows a generally decreasing trend; and the `AZ` dataset shows a trend that first decreases, and then increases.

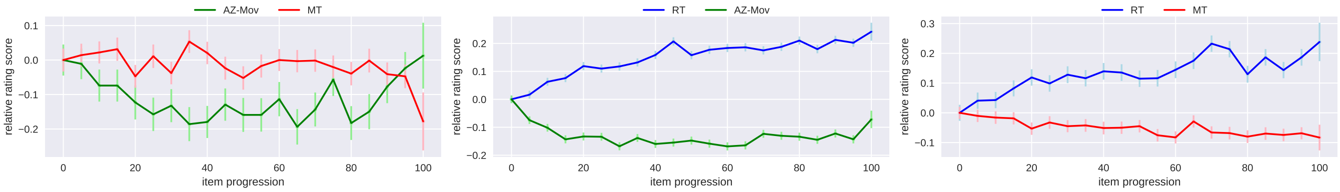
Are platform-specific ratings dynamics consistent? One possible explanation for the platform-specific differences in rating dynamics shown in Figure 1 could be that they are due to the fact that the set of movies rated on each platform is different. First, we show that differences shown in Figure 1 in platform-specific dynamics are *not* due to the different sets of movies rated.

For each pair of platforms, we select the set of movies that are rated at least 100 times on both platforms (we use a smaller window of 100 reviews to increase the size of the sets being analyzed). We match movies based on title and year (where available), discarding any cases in which duplicate matches occur. Figure 2 shows the item progression for movies in common between each pair of datasets, and that in each case, platform-specific trends are preserved. Specifically, Figure 2a shows the 127 movies in common among `AZ` and `MT`, Figure 2b shows the 1451 movies in common among `RT` and `AZ`, and Figure 2c shows the 387 movies in common among `RT` and `MT`.

In each case, the platform-specific trends shown in Figure 1 a preserved (although due to the smaller dataset sizes, there is more variability and trends are correspondingly weaker in some cases.) We conclude that the differences shown in Figure 1 are consistently present when studying the same sets of movies on different platforms.

We also note that the relationships between initial and average scores across platforms are preserved when restricting attention to common movies, with `AZ` $>$ `MT` in Figure 2a ((4.24, 4.12) $>$ (3.87, 3.86)), `AZ` $>$ `RT` in Figure 2b ((4.29, 4.16) $>$ (2.93, 3.08)), and `MT` $>$ `RT` in Figure 2c ((3.59, 3.54) $>$ (2.87, 3.00)).

Another way to assess whether platform-specific ratings dynamics are consistent is to ask whether the same dynamics are seen across multiple item categories on a given platform.



(a) AZ-Mov and MT: 127 common movies (b) RT and AZ-Mov: 1451 common movies (c) RT and Movie MT: 387 common movies

Fig. 2: Common movies: relative ratings progression.

To confirm this, we look at relative ratings scores across the six categories of Amazon data, shown in Figure 3a . This Figure shows that the general behavior of declining followed by increasing ratings is widespread across most of the item categories on the Amazon platform.

The above results suggest that the platform-specific ratings dynamics we observe are not due solely to differences in items rated on the different platforms, but rather that these effects are relatively consistent.

Do all items change in the same way within each platform? A final characterization question concerns how the platform-wide effects seen in Figure 1 are produced from the individual contributions of each movie. We explore this question by clustering the movies individual item progressions using the `k-Shape` algorithm [8], and studying cluster-wide averages. We use a default of five clusters in each case, which we observe to balance clear separation of classes against noise introduced due to small samples.

Figure 4 shows the results of this clustering for all platforms. The Figure shows that the characteristic dynamics on each platform are not always present for all movies. In the case of `AZ`, the characteristic decrease/increase pattern is primarily present in a cluster 1, comprising about 12% of all movies. The other clusters primarily show a simpler decreasing trend. In the case of `RT`, the characteristic increase is primarily present in clusters 0 and 4, comprising about 45% of all movies. Finally, in the case of `MT`, in general all movie clusters show the platform's characteristic downward trend, with the strongest trends in clusters 0, 2, and 3.

We conclude from Figure 4 that not all items are showing strong dynamics in each dataset and that, furthermore, dynamics are not occurring uniformly in each. As a result, in what follows we will generally distinguish between "large effect" movies (`AZ` cluster 1, `MT` clusters 0, 2, 3, `RT` clusters 0,4) and "small effect" movies (movies in the remaining clusters).

B. Decomposing Ratings Dynamics

To develop an understanding of the forces driving the effects seen in Section IV-A, we decompose ratings using `timeSVD--`. The components of the model bear direct relationship to various factors of interest as described in Section II. In particular, we can study the impact of the user population by looking at the user time-varying and invariant components of the model ($\alpha_u \text{dev}_u(t)$ and b_u), we can study the impact of item perception by studying the item time-varying and invariant components of the model ($b_{i,Bin(t)}$ and b_i), and we can study the impact of closed-loop effects by studying the model's interaction score ($q_i^T p_u(t)$).

What are the main model factors affecting ratings dynamics? We start by decomposing the three datasets according to our model, and according to movie type (small effect vs. large effect as described above). The results are shown in Figure 5.

We start first with Figure 5a, Figure 5d, and Figure 5g which show relative contributions of factors, respectively, for `AZ`, `MT`, and `RT`. There are a number of high-level observations. First, user invariant components and item time-varying components are the largest and primary contributors to ratings dynamics. Furthermore, the only platform in which interaction score shows significant dynamics is `AZ`.

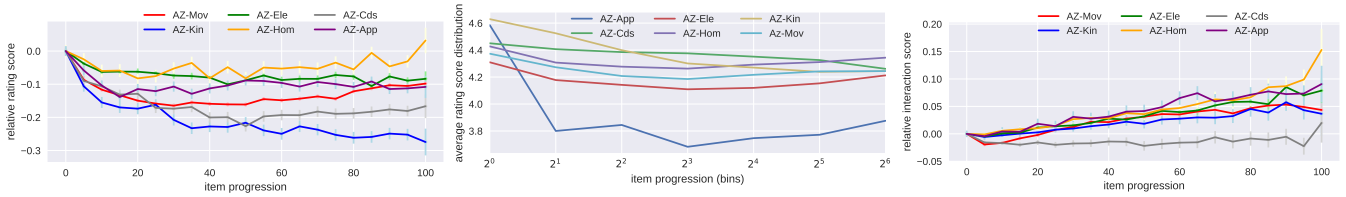
Figures 5c, 5f, and 5i show the corresponding breakdowns for the large-effect movies, and the results there confirm the conclusion that user invariant and item time-varying components are the main contributors to the respective platform dynamics. (Figures 5b, 5e and 5h show the small-effect movies – note the difference in scale on the y-axes).

We now explore each of the factors in turn.

How do users contribute to rating dynamics? We first examine the role of users in rating dynamics. We note from Figure 5 that the user time-varying component (purple line) does not show significant contribution to rating dynamics, but the user time-invariant component (green line) does show significant contribution. This means that while users individual ratings averages are not changing over time, users' contribution to changes in ratings are nonetheless significant. In other words, *changes in the user population* – in a consistent way – are a major driver of ratings dynamics (on all three platforms).

For `AZ` the contribution of changes in user population (green line) reflects the overall platform pattern of initial decline followed by increase. This component contributes about 50% of the overall change at the end of the 200 review period. For `MT` the contribution of changes in user population has a decreasing trend of similar range for the whole dataset analysis (Figure 5d) as well both subsets (Figures 5e and5f). This also covers above 50% of the relative changes in ratings for the large-effect set (Figure 5f). For `RT`, we also see that changes in user population play a significant role, contributing about 50% of the change in the large effects subset.

To understand how this significant shift in user population comes about, we turn to the `RT` dataset. In that dataset, we have the advantage that users are classified as either (professional) critics or general reviewers. We use this classification to achieve a better understanding of role of user population in rating dynamics.

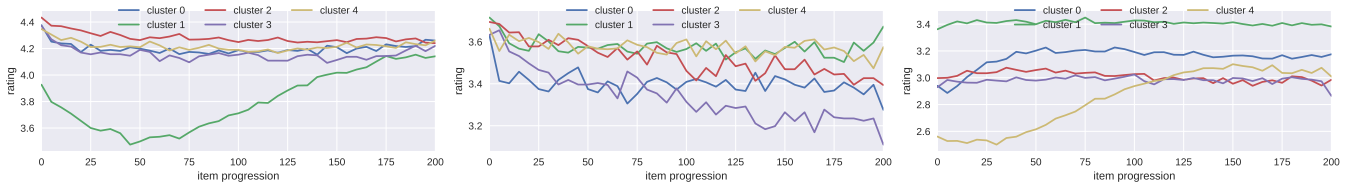


(a) Relative Ratings Progression

(b) Amazon early user effect

(c) Amazon Interaction Score - all items

Fig. 3: Relative progression across Amazon categories.

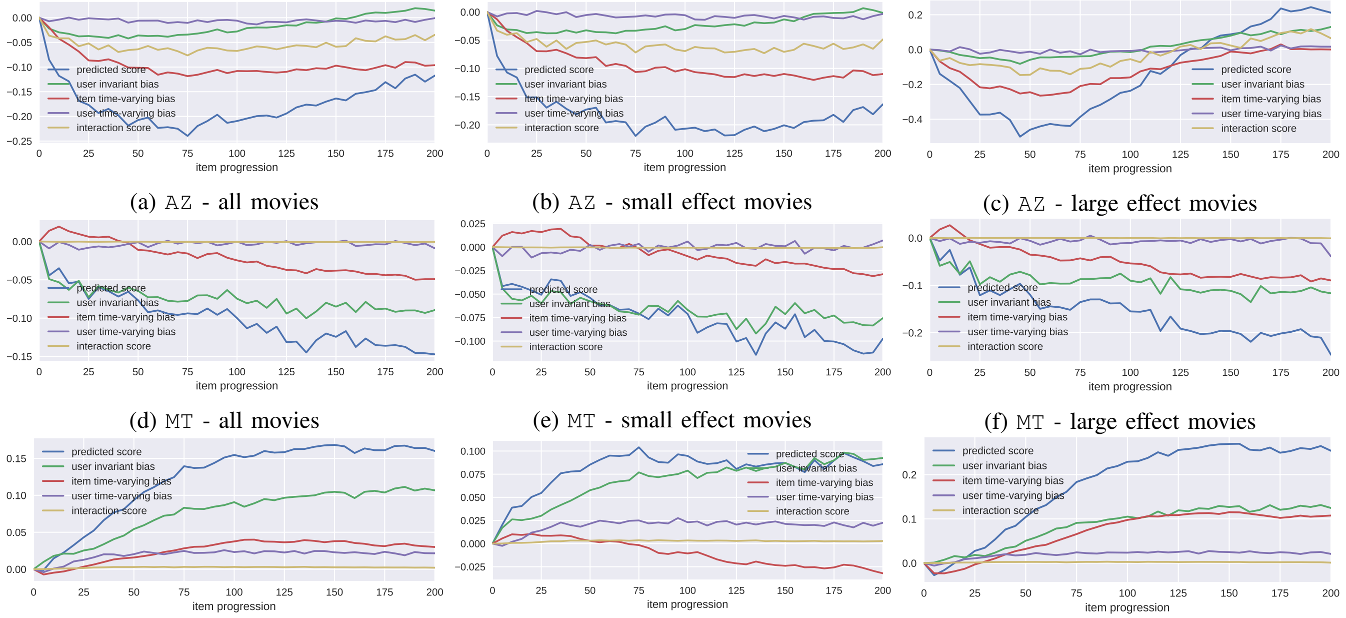


(a) Amazon (AZ)

(b) Movie Tweetings (MT)

(c) Rotten Tomatoes (RT)

Fig. 4: Relative ratings progression by cluster.



(a) AZ - all movies

(b) AZ - small effect movies

(c) AZ - large effect movies

(d) MT - all movies

(e) MT - small effect movies

(f) MT - large effect movies

(g) RT - all movies

(h) RT - small effect movies

(i) RT - large effect movies

Fig. 5: Relative timeSVD-- Components Progression

Figure 6 breaks down relative ratings score according to user type in RT. In each plot of that figure the blue line represents the critic's reviewers contribution, the green line represents the general's reviewers contribution, the red line the general contribution of all reviewers. The grey line (with y-axis scale on the left side) represents the proportion of critics that are reviewers contributing to the average across movies. The proportion (grey line) of critics is the same in all plots.

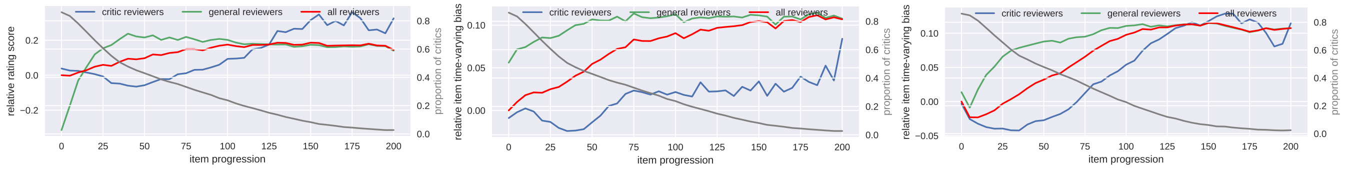
Figures 6a and 6b shed considerable light on the user population component of ratings dynamics. It shows that critics are responsible for most of the initial reviews, and that critics on the whole tend to have lower average reviews than general users. (The facts that the user line in each figure drops at the beginning, and the critic lines rise at the end, are due to small-sample effects.) The effect is particularly clear when extracting just the user time-invariant component in Figure 6b.

The contrast to MT is interesting, because there the user population shift has a decreasing effect on ratings. We note that the MT platform is quite different from the other two, because of the absence of a well-defined critic population, as well as the fact that previous ratings of a movie are not as easily accessible. We hypothesize that this means that users whose average ratings are lower will be more likely to review movies later in time.

Overall, this analysis goes a long way to explaining how user population shifts contribute to ratings dynamics. In RT it can help explain the entire dynamics of the user time-invariant component of the model. In AZ it can help explain the eventual increase in the user time-invariant component; we will explore the initial decrease later in the paper.

How do items contribute to ratings dynamics? The second significant component exposed by timeSVD-- in Figure 5 is the item time-varying contribution (red line).

For AZ in the all movies case (Figure 5a) we can ob-



(a) Ratings - All (b) User time-invariant component - All (c) Item time-varying component - large

Fig. 6: Rotten Tomatoes: population change between critic reviewers and general reviewers

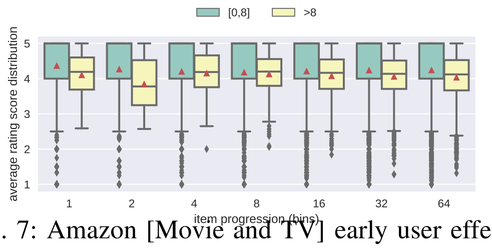


Fig. 7: Amazon [Movie and TV] early user effect.

serve that the item time-varying component accounts for a substantial proportion of the relative change in ratings score, reaching almost 80% at the end of the progression. For MT the item time-varying component always has a decreasing trend – i.e., when the items lose value while aging. We note that this is consistent with previous work (eg, [9]) showing that, in the absence of other factors, online ratings tend to decline when prior reviews are hard to access or evaluate. The fraction of contribution is considerable – reaching up to 40% of the relative score – in the large effect subset (Figure 5f), and it is present across essentially all clusters within the RT dataset (Figure 4b). Finally, in RT the item time-varying component has an increasing trend – i.e. items get a higher score when time progresses – for the all movies case (Figure 5g) and for the subset of large effect moives (Figure 5i) where it accounts up to 40% of the relative predicted score. The difference in the case of RT can also be understood in the context of [9] due to the presence of a large set of reliable reviews (reviews that are labeled as coming from critics).

Movies with a strong increasing time-varying component are those that show significant improvement in rating over time; they can be thought of as "sleepers" that take time to become well-liked. A more detailed analysis including the separation of the item time-varying component among critics and general reviewers in Figure 6c shows that general reviews tend to view "improving" movies earlier in time, while critics tend to view "improving" movies later. This suggests that users are quicker to identify "sleeper" movies and that critics follow. Overall, our analysis shows that on a platform like Twitter, movies with declining ratings over time are more likely to accumulate subsequent reviews than on platforms like Amazon and Rotten Tomatoes, where previous reviews are more accessible and more easily interpreted.

Why do ratings initially decline on the Amazon platform? One of the striking properties of ratings on the Amazon platform is the initial decline followed by subsequent increase. Figure 1 shows this effect, Figure 4a shows that it primarily derives from about 12% of all movies (although most movies show an initial decline) and Figure 5c shows that the effect has contributions from both item time-varying and user time-invariant components. This behavior constrasts starkly with the case for MT and RT.

In investigating this we note that AZ has a numerous quantity of users that just provided a small number of reviews; this, combined with the fact that Amazon is an e-commerce platform, raise the questions of whether initial reviews are intentially inflated in some way. This could be a strategy to attract buyers when a product is first introduced.

We investigated this hypothesis by analyzing the users in the AZ dataset. We conjecture that if large numbers of early reviews were artificially inflated, then there should be a subpopulation of users who are providing almost exclusively early reviews for items.

Hence, fo each user, we compute their average movie rating time (i.e., how early in the item progression time the user provides a review), the user's average rating score, and the number of reviews that that user provided. We summarize the results in Figure 7.

In the Figure 7, we show the distribution of average rating score of a user versus the average item progression time for that user's ratings. We separate users that contributed less than eight reviews from those that gave more than eight. The figure shows that users that proffered less than eight reviews have a higher average score than users that provide more than eight. This can be observed by comparing the user's results (green over yellow) at each bin time of the item progression. Furthermore, by observing the distribution of those users that provided less than eight reviews overtime (green box), we can see that their average score declines over time.

These results suggest that the initial drop in ratings seen on the Amazon platform is driven at least in part by a subpopulation of users who provide few reviews overall and who provide inflated ratings for a product early in its lifetime. We hypothesize that this arises due to the nature of the Amazon rating system's existence in support of product purchases. Figures 3a and 3b confirm this effect and explanation across Amazon categories.

We note that if this explanation holds, then it should be a consistent property across the Amazon platform. Indeed, we find that this is the case, as shown in Figure 3a. All categories from Amazon present an initial drop in ratings and most of them – except for AZ-App – have an average rating increase afterward. That figure shows that the initial-decline of ratings is a fairly common feature across categories on the Amazon platform.

We can likewise explore our hypothetical explanation – that a subset of reviewers provide early, inflated ratings – for each of the Amazon categories. The results are shown in Figure 3b. The figure shows that the early-reviewer effect is present in every Amazon category, and that it is particularly pronounced in certain product categories (Apps and Kindle books).

How do recommendations contribute to ratings dynamics? The final factor to consider, as discussed in Section II, is the presence of a recommendation system on a given platform. We expect that if a recommendation system is suggesting items to users, then subsequent ratings for the item should show a higher interaction score because this would reflect an improved 'match' between the preferences of users and the features of the item.

We can assess this effect in two ways: we can ask whether individual items show higher interaction scores over time, and we can ask whether items that have high interaction scores receive more ratings. In the latter case, this may be because more ratings allows the system to do a better job of forming recommendations, and it may be because items that are successfully recommended will garner more ratings.

To ask whether individual items show higher interaction scores over time, we recall from Figure 5 that `RT` and `MT` show essentially zero variation in interaction score (yellow lines). This is consistent with the observation that those platforms are not actively providing users with recommendations that affect which items a user consumes or chooses to rate.

However, that Figure 5 shows an interaction score effect for the `AZ` platform. To augment that result, we perform `timeSVD--` decomposition of each of the other Amazon categories. The results are shown in Figure 8. Interestingly, it appears that individual items do not show an increased interaction score over time (yellow lines). In general, interaction scores decline somewhat over time.

However, the effect on an individual item may be subtle over time. A more likely effect of a recommender system would occur between the number of ratings an item receives and the interaction score of the item. For this analysis, we return to looking at all items in the dataset (not just those having 100 or more ratings). The results (looking only at interaction score) are shown in Figure 9. This figure shows that on the Amazon platform, there is a strong positive correlation between the interaction score (a measure of the effectiveness of the recommendation system) and the number of ratings that an item receives.

V. Related Work

In this section we review work that relates to our study. A number of previous studies have looked at temporal dynamics in online reviews but to the best of our knowledge we are the first that addresses the role played by the complete set of factors listed in Section II.

One starting point for our analysis is [7], which proposes a recommender system based on collaborative filtering that incorporates temporal dynamics, and splits prediction score between various factors. The authors in [10] present a temporal rating model that additionally incorporates review text; we focus just on review scores as a function of time.

McAuley and Leskovec propose a latent factor recommender system in [1] that models user development caused by the consumption of products over time. They show the role of user experience and expertise through analysis of beer, wine, food, and movie reviews; we do not find a significant impact of user evolution in our study. The authors in [11], [12], and [13] also model temporal dynamics as a strategy to improve recommendation accuracy, and use models similar in spirit to our model; however their purpose is not to understand ratings dynamics. Likewise, the authors in [14] study how positive and negative movie reviews change over time and propose a recommender system model that takes into account time-varying and temporal effect of positive and negative reviews for future behavior.

While all of these studies propose new methods for improving recommendations, none seeks to understand a broad set of factors underlying the evolution of rating dynamics observed in practice such as platform differences or population shift.

The authors in [9] analyzed the evolution of online ratings over sequence and time for a book ratings dataset. They show that, on average, ratings in sequence and time decrease, although there are distinct dynamic processes occurring. Although they provide some explanations for those dynamic processes their analysis is limited to a specific platform and item type.

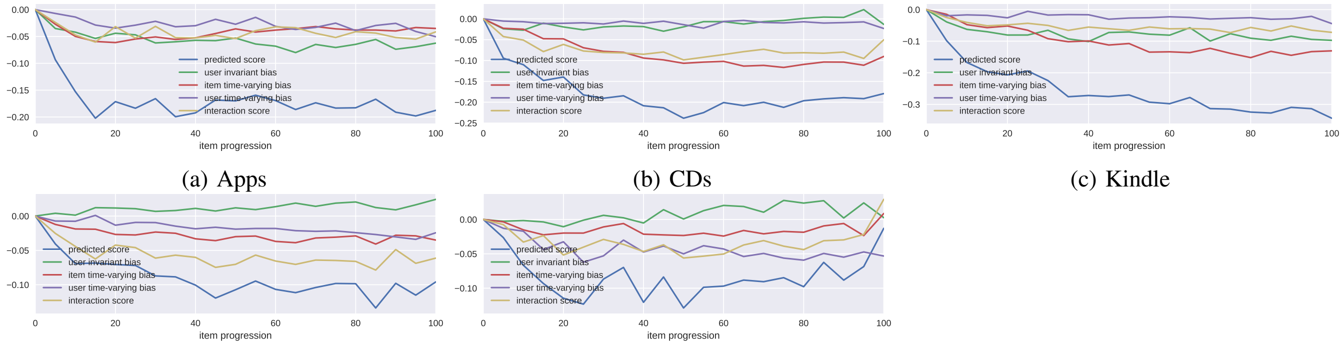
Finally, we point out some studies that are looking at the dynamics of online reviews focused in some particular correlations. Tha authors in [15] model the positive feedback mechanism between online word-of-mouth (WOM) and retail using a movie dataset. Authors in [16] create a model to understand online product ratings from a consumer perspective and comparing evaluations of products from consumer magazine and online ratings they observed that besides the product quality online ratings reflects the customer's satisfaction. Authors in [17] analyzed the role of social dynamics in cultural markets. In a similar perspective [18] and [19] analyze in online systems the effect of conformity or social influence bias – the inclination to conform to the observed norm of a community. Furthermore, [20] proposes a recommender system that mitigates this conformity effect while [21] a system to embrace it.

VI. Conclusion

In this paper we've taken a broad look at the factors that drive changes in item ratings in online review systems.

Our results take two parts. First, we characterize the range of ratings dynamics and show how platforms differ. Importantly, different platforms have different and distinctive dynamics. These are preserved when looking at the same sets of items across platforms, and they are preserved when looking across different types of items on the same platform.

Next, we use our model to unravel the factors affecting rating dynamics. First and foremost, we show that changes in user populations are a significant driver of ratings dynamics. In general we observe a trend for user population shifts to increase ratings over time and our `RT` analysis suggests that



(a) Apps (b) CDs (c) Kindle

(d) Eletronics (e) Home

Fig. 8: Relative TimeSVD- Components Progression per Category

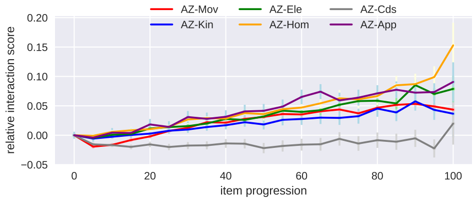


Fig. 9: Amazon Interaction Score per Category -all items

an important factor is the shift from critics to general users over time. Next, we show that there is in general significant variation in the perceived quality of items over time. This suggests a general trend that may be due to presence of accessible, well characterized reviews (eg in RT) or the lack thereof (in MT). Then, we show that in the case where ratings are in support of an e-commerce platform (ie, AZ) there is a significant tendency for a subset of users who provide few reviews overall to provide early, inflated ratings for items. This is consistent across categories of Amazon products but does not occur in ratings-only sites like Rotten Tomatoes and Twitter. Finally, we find that the presence of a recommendation system on a site like Amazon helps explain the tendency for items (across all categories) that show higher interaction scores to acquire more ratings overall.

Taken as a whole, we show both the complexity behind the dynamics of online reviews and a set of understandable factors that interact to generate that complexity. Hence, we believe that these results provide a framework for interpreting item reviews and how they may be expected to change over time.

VII. Acknowledgments

This material is based upon work supported by the National Science Foundation under grant numbers IIS-1421759 and CNS-1618207.

References

[1] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," in *Proceedings of the 22nd WWW*. New York, NY, USA: ACM, 2013.

[2] Y. Zhang, T. Lappas, M. Crovella, and E. Kolaczyk, "Online ratings: convergence towards a positive perspective?" in *Proceedings of ICASSP*, Florence, Italy, 2014.

[3] S. Dooms, T. De Pessemier, and L. Martens, "Movietweetings: a movie rating dataset collected from twitter," in *CrowdRec at RecSys 2013*, 2013. [Online]. Available: https://github.com/sidooms/MovieTweetings

[4] *Rotten Tomatoes*, 2016 (accessed Setember 22, 2016). [Online]. Available: https://www.rottentomatoes.com

[5] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of ACM SIGIR*. New York, NY, USA: ACM, 2015.

[6] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proceedings of WWW*, ser. WWW '16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016.

[7] Y. Koren, "Collaborative filtering with temporal dynamics," *Communications of the ACM*, vol. 53, no. 4, pp. 89–97, 2010.

[8] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," *SIGMOD Rec.*, 2016.

[9] D. Godes and J. C. Silva, "Sequential and temporal dynamics of online opinion," *Marketing Science*, vol. 31, no. 3, pp. 448–473, 2012.

[10] Y. Liu, Y. Liu, Y. Shen, and K. Li, "Recommendation in a changing world: Exploiting temporal dynamics in ratings and reviews," *ACM Trans. Web*, vol. 12, no. 1, pp. 3:1–3:20.

[11] C. Zhang, K. Wang, H. Yu, J. Sun, and E.-P. Lim, *Latent Factor Transition for Dynamic Collaborative Filtering*, pp. 452–460.

[12] N. N. Liu, M. Zhao, E. Xiang, and Q. Yang, "Online evolutionary collaborative filtering," in *Proceedings of RecSys*, ser. RecSys '10. New York, NY, USA: ACM.

[13] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, *Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization*, pp. 211–222.

[14] W.-J. Li, Q. Dong, and Y. Fu, "Investigating the temporal effect of user preferences with application in movie recommendation," *Mobile Information Systems*, 2017.

[15] W. Duan, B. Gu, and A. B. Whinston, "The dynamics of online word-of-mouth and product sales – an empirical investigation of the movie industry," *Journal of Retailing*, vol. 84, pp. 233 – 242, 2008.

[16] T. H. Engler, P. Winter, and M. Schulz, "Understanding online product ratings: A customer satisfaction model," *Journal of Retailing and Consumer Services*, vol. 27, pp. 113 – 120, 2015.

[17] M. J. Salganik and D. J. Watts, "Web-based experiments for the study of collective social dynamics in cultural markets," *topiCS*, vol. 1, no. 3, pp. 439–468, 2009.

[18] X. Li and L. M. Hitt, "Self-selection and information role of online product reviews," *Information Systems Research*, vol. 19, no. 4, pp. 456–474, 2008.

[19] Z. Yang, Z.-K. Zhang, and T. Zhou, "Anchoring bias in online voting," *EPL (Europhysics Letters)*, vol. 100, no. 6, p. 68002, 2012.

[20] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg, "A methodology for learning, analyzing, and mitigating social influence bias in recommender systems," in *Proceedings of RecSys*. ACM, 2014.

[21] Y. Liu, X. Cao, and Y. Yu, "Are you influenced by others when rating?: Improve rating prediction by conformity modeling," in *Proceedings RecSys*, ser. RecSys '16. New York, NY, USA: ACM, 2016.