The Netflix Recommender System: Algorithms, Business Value, and Innovation

PRASHANT GOEL and ROHAN SINGH ASWAL,

DCE Gurugram

Abstract

This article describes the various algorithms that make up the Netflix recommender system and explains their business purpose. It also describes the role of search and related algorithms. This also makes it a recommendation issue for us. Explaining the rationale, we combined her A/B tests, which focused on improving member retention and medium-term engagement, with offline experiments using past member her engagement data to improve our recommendation algorithm. Review the approaches you use to improve. Discusses some issues in designing and interpreting A/B tests. Finally, we will discuss some areas of recent innovation, such as globalization and language support of recommendation engines.

Discusses some issues in the design and interpretation of A/B tests. Finally, we discuss recent focused innovation areas, such as making the recommendation engine global and language aware.

Category and Subject Descriptors: C.2.2 [Recommender System]: Machine Learning
General Terms: Algorithm, Recommendation System, A/B Testing, Product Innovation
Additional Keywords and Phrases: Recommendation System

1. INTRODUCTION

Storytelling has always been at the core of human nature. Major technological breakthroughs that changed society in fundamental ways have also allowed for richer and more engaging stories to be told. It is not hard to imagine our ancestors gathering around a fire in a cave and enjoying stories that were made richer by supporting cave paintings. Writing, and later the printing press, led to more varied and richer stories that were distributed more widely than ever before. More recently, TV has exploded the use and proliferation of video for storytelling. Today, we are all fortunate to witness the changes that the Internet brings. As with previous major technological breakthroughs, the Internet has had a major impact on his storytelling.

Netflix sits at the intersection of the Internet and storytelling.

Internet TV. invention. Our main product and source of revenue is our subscription service, which allows our members to stream videos from our collection of movies and TV shows on

anytime on a variety of internet-connected devices. As of this writing, 4,444 of her over 65 million members stream over 100 million hours of her movies and 4,444 TV shows a day.

The Internet TV space is young and the competition is maturing, so innovation is key. A key pillar of our product is a recommendation system that helps members find her

videos to watch in each session. Our recommender system is not a single algorithm, But it's a collection of different algorithms that serve different use cases and together create the complete Netflix experience. Section 2 provides an overview of the various

algorithms for recommendation systems, and Section 3 discusses their business value. The process used to improve the algorithm in Section 4 is described in Section 5, and conclusions are presented in Section 6

2. NETFLIX RECOMMENDER SYSTEM

Internet TV is a matter of choice: what, when, where, compared to linear broadcast and cable systems that offer it all, there are probably 10 things played on a

20 favorite channels from. However, people are surprisingly bad at choosing among many options, and can quickly become overwhelmed and choose "none of the above" or make the wrong decisions (e.g., Schwartz [2015]). At the same time, the advantage of Internet

TV is the ability to broadcast videos from a broader catalog, including niche titles that appeal to a wide range of

demographics and tastes and are of interest to only a relatively small

user group.

According to consumer research, the typical Netflix member probably spends 60 to 90 seconds of It has been suggested that they later lose interest. Either the user will find something interesting or the risk of the user leaving our service is greatly increased. The issue of referrals is to ensure that every member of our diverse pool finds something interesting to see on these two screens and understands why it is of interest.

Historically, the Netflix recommendation problem has been thought of as equivalent

to the problem of predicting the number of stars that a person would rate a video after

watching it, on a scale from 1 to 5. We indeed relied on such an algorithm heavily when

our main business was shipping DVDs by mail, partly because in that context, a star

rating was the main feedback that we received that a member had actually watched the video. We even organized a competition aimed at improving the accuracy of the rating prediction, resulting in algorithms that we use in production to predict ratings to this day [Netflix Prize 2009].

But the days when stars and DVDs were the focus of recommendations at Netflix have long passed. Now, we stream the content, and have vast amounts of data that describe what each Netflix member watches, how each member watches (e.g., the device, time of day, day of week, intensity of watching), the place in our product in which each video was discovered, and even the recommendations that were shown but not played in each session. This data, and the experience we've had as a result of improving the Netflix product, suggests that helping viewers find videos to watch requires much more than just focusing on videos with the expected high star ratings. It turns out there is a better way.

Well, our recommendation system is made up of different algorithms that together define the Netflix experience, most of which converge on the Netflix homepage.

This is the first page Netflix members see when they log into their Netflix profile on any device (TV, tablet, phone, or browser).

An instance of our contemporary TV homepage is proven in Figure 1. It has a matrixlike layout. Each access withinside the matrix is a advocated video, and every row of motion pictures carries tips with a similar “topic.” Rows are categorized in line with their topic to make the topic obvious and (we think) greater intuitive to our members.

2.1. Personalized Video Ranker: PVR

There are normally approximately forty rows on every homepage (relying at the abilities of the device), and as much as seventy-five motion pictures in keeping with row; those numbers range incredibly throughout gadgets

due to hardware and person enjoy considerations. The motion pictures in a given row usually derived from a single algorithm. Genre series, such as suspense movies, shown at

on the left side of Figure 1, are driven by a Personalized Video Ranker (PVR) algorithm.

As the name suggests, this algorithm ranks an entire catalog of videos (or a subset, selected by genre or other filters) for each member's profile in a personalized way.

The resulting order is used to select the order of the videos in the genre and other rows. This is why there are often completely different videos of her in the same genre line shown to different members. We use PVR a lot, so it should be good for overall relative ranking across the catalog. This limits the range that can actually be personalized. Similarly, PVR works well when you combine a personalized signal with a fairly healthy (non-personalized) popularity rating. We will use this to drive the recommendations in the popular row on the left side of Figure 2. Watch Amatoria Inn and Basilico [2012] Learn more about personalized video rankings here.

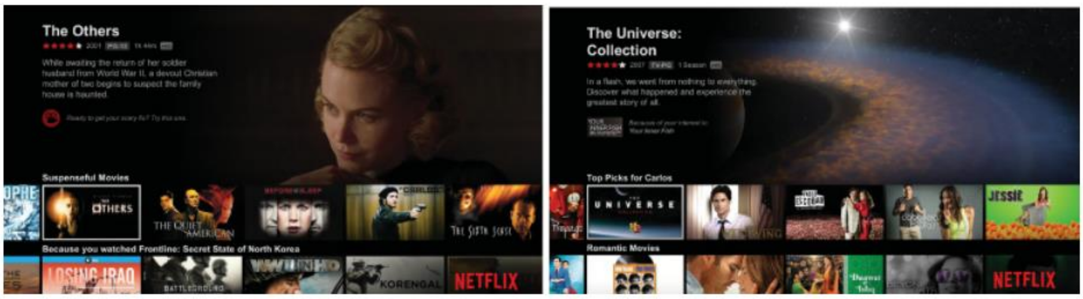

Fig. 1. (Left) An example of the page of recommendations, showing two of the roughly 40 rows of recommendations on that page. Suspenseful Movies is an example of a genre row driven by the PVR algorithm (Section 2.1). The second row is a Because You Watched row driven by the sims algorithm (Section 2.5). (Right) A homepage showing the Top Picks row driven by the Top N algorithm (Section 2.2). Romantic Movies is a genre row driven by PVR.

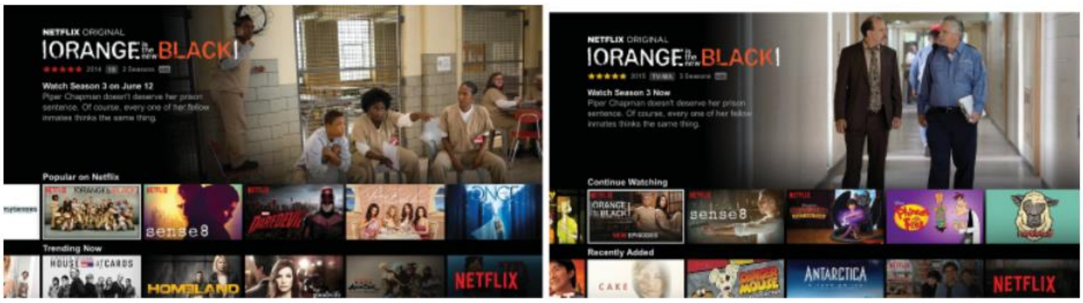

Fig. 2. (Left) Two more rows of recommendations on a homepage. The popularity-heavy Popular row and the Trending Now row (Section 2.3) focus on the latest viewing trends. (Right) A homepage for a Continue Watching session with a Continue Watching row (Section 2.4).

2.2. Top N Video Ranker

There is also a Top N Video Ranker that generates recommendations for the top pick rows as shown in Figure 1 on the right. The purpose of this algorithm is to find her most personalized recommendations across each member's catalog. So is only focused on the top of the rankings. Any subset of the catalog. Therefore, Top-N rankers are optimized and ranked using metrics and algorithms that refer only to the top of the catalog rankings that the algorithm generates, rather than ranking the entire catalog (similar to PVR). in the case). Otherwise, Top N Ranking and PVR share similar values attributes, for example, combining personalization with popularity, and identifying and incorporating viewing trends over different time windows ranging from a day to a year.

2.3. Trending Now

We have also found that shorter-term temporal trends, ranging from a few minutes to perhaps a few days, are powerful predictors of videos that our members will watch, especially when combined with the right dose of personalization, giving us a trending ranker [Padmanabhan et al. 2015] used to drive the Trending Now row.There are two types of trends that this ranker identifies nicely: (1) those that repeat every several months (e.g., yearly) yet have a short-term effect when they occur, such as the uptick of romantic video watching during Valentine`s Day in North America, and (2) one-off, short-term events, for example, a big hurricane with an impending arrival to some densely populated area, being covered by many media outlets, driving increased short-term interest in documentaries and movies about hurricanes and other natural disasters.

2.4. Continue Watching

Given the importance of episodic content viewed over several sessions, as well as the freedom to view non episodic content in small bites, another important video ranking algorithm is the continue watching ranker that orders the videos in the Continue Watching row (see right of Figure 2). Most of our rankings rank unwatched titles at and we only have extrapolated information from there. In contrast, the Continue Watching Ranking shows the number of recently watched titles based on the member's best estimate of whether they intend to continue watching or rewatching, or whether the member has given up on something less interesting than expected. Sort the subset. The signals we use include time elapsed since viewing, end time (whether the program is in the middle or at the beginning or end), whether another title was viewed after
, and the device used. In general, our different video ranking algorithms use different mathematical and statistical models, use different signals and data as input, and use different rankings designed for the specific purpose of each rank. requires extensive model training.

2.5. Video to Video Similarity Lines

“Because you saw it” (BYW) is another method of classification. BYW line anchors recommendations to a single video watched by a member. The video-video-similarity algorithm (just called "Sims") drives her recommendations in these lines. An example line is shown on the left in Figure 1. The Sims Algorithm is, a non-personalized algorithm that calculates a video ranking (similar to) for each video in the catalog. The Sims rankings are not personalized, but the selection of which BYW series are featured on the home page is personalized, and the subset of BYW videos recommended for a particular BYW series is compared to a subset of similar videos. Based on personalizations we think members will enjoy (or people have already seen).

2.6. Page Building: Series Selection and Ranking

Videos selected in each series represent an estimate of the best selection of videos presented to a particular user. However, most members have different moods from session to session and many accounts are shared by multiple members of the household. By providing with a variety of line choices, members can easily skip videos that may be a good choice for another time, occasion, or family member and quickly identify what is relevant. want to be. The page generation algorithm creates individual pages of recommendations using the output of all previous algorithms, taking into account each row's relevance to members and page diversity. A typical member has tens of thousands of rows that may appear on the home page, and the calculations required to evaluate them are difficult to manage. For this reason, prior to 2015, we used a rule-based approach that defined the type of row (genre row, BYW row, popular row, etc.) that should appear at each vertical position on the page. This page layout was used by to create all home pages for all members. We now have a fully personalized mathematical algorithm that can select and order rows from a large pool of candidates to create an order optimized for relevance and diversity. The current algorithm doesn't use templates, so you're free to fine-tune the experience, such as not picking the BYW row on a particular homepage and having half of the page be her BYW row on another homepage. A recent blog post [Alvino and Basilico 2015] about this algorithm explains it in more detail.

2.7. Evidence

Together, these algorithms make up the complete Netflix recommender system. But there are other algorithms, such as evidence selection ones, that work together with our recommendation algorithms to define the Netflix experience and help our members determine if a video is right for them. We think of evidence as all the information we show on the top left of the page, including the predicted star rating that was the focus on the Netflix prize; the synopsis; other facts displayed about the video, such as any awards, cast, or other metadata; and the images we use to support our recommendations in the rows and elsewhere in the UI. Evidence selection algorithms evaluate all the possible evidence items that we can display for every recommendation, to select the few that we think will be most helpful to the member viewing the recommendation. For example, evidence algorithms decide whether to show that a certain movie won an Oscar or instead show the member that the movie is similar to another video recently watched by that member; they also decide which image out of several versions use to best support a given recommendation support our recommendations in the rows and elsewhere in the UI. Evidence selection algorithms evaluate all the possible evidence items that we can display for every recommendation, to select the few that we think will be most helpful to the member viewing the recommendation. For example, evidence algorithms decide whether to show that a certain movie won an Oscar or instead show the member that the movie is similar to another video recently watched by that member; they also decide which image out of several versions use to best support a given recommendation.

2.8. Search

Our recommendation system is used on most Netflix product screens outside of the home page and overall influences choices about 80% of the time streamed on Netflix. The remaining 20% of are from searches that require proprietary algorithms. Member often searches the catalog for videos, actors, or genres. Use information retrieval and related technology to find and display relevant videos to members. However, members often search for videos, actors, or genres (Fig. 3, left) or general concepts (Fig. 3, right) that are not included in his catalog, so even searches are not recommended. It becomes a problem. In such cases, Search will recommend video for a specific search query as an alternative result for a failed search. The extreme rawness of text entry on TV screens means that it is also particularly important to interpret 's two- to three-letter subrequests in the context of what we know about search member 's preferences. increase.

The search experience is based on multiple algorithms. The algorithm tries to find videos that match the given query. For example, get the frenemies of the subquery fren. Another algorithm predicts interest in a concept given a partial query, for example by identifying the concept "French cinema" for the query "fren". A third algorithm finds video recommendations for a specific concept. For example, enter videos recommended under the concept of French Cinema. Our search algorithm combines

game data, search data and metadata to provide results and recommendations.

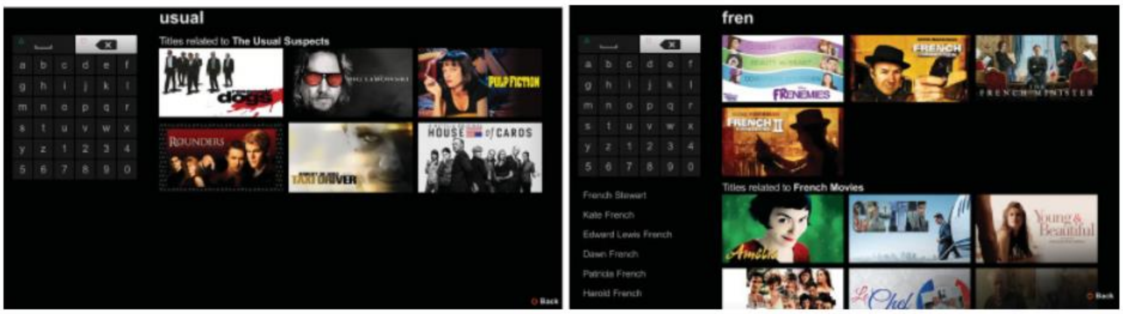

Fig. 3. (Left) Search experience for query “usual,” presumably for the movie “The Usual Suspects” which was not available at Netflix at the time of the query. The results are instead recommendations based on the query entered. (Right) Search experience for the query “fren,” showing standard search results at the top for videos with names that contain the substring “fren,” people results on the lower left, and search recommendations based on the guess that the intent was searching for French Movies.

2.9. Related Works

Each algorithm in the recommendation engine relies on statistical and machine learning techniques. This includes both supervised (classification, regression) and unsupervised approaches (dimensionality reduction with clustering or compaction, such as with topic models). [2011] and Murphy [2012] provide excellent reviews of such techniques, Blei et al. [2003] and Teh et al. [2006] is a good example of a useful topic model, as well as a specialized adaptation in the area of recommendation systems, especially matrix factorization. A good introduction to the factorization approach is Koren et al. [2009], with more detailed material

found in Koren [2008]. Useful generalizations of more traditional factorization approaches include factorization machines [Rendle 2010], methods to reduce the number of

parameters in a model (e.g. Paterek [2007]), and probabilistic graphical models Includes links to (e.g. Mnih and Salakhutdinov [2007]).),

can be easily extended to address different problems.

3. Business value

We aim to grow our business on a tremendous scale. That means being a producer and distributor with a full global reach of

shows and movies. For

reasons, we develop and use his

recommendation system because we believe it is core to our business. Our referral system helps us win the moment of truth: when a member

starts a session and helps that member find something exciting within seconds, we know the service is Prevents being abandoned for alternative entertainment options. Personalization keeps TV broadcasts from getting too small an audience to support significant advertising revenue, or around broadcast or cable channels to prove. You can find a relatively niche audience for videos that are not meant for models. This is very evident in our data, showing that our recommendation system

evenly distributes views across more videos than the non-personalized

system. To make this more precise, we introduce a specific metric next.

Effective Catalog Size (ECS) is a measure of how views are distributed across her

items in the catalog. If most of the views are from one video,

will be close to 1. Otherwise it's somewhere in between. ECS is described in detail in

Appendix A.

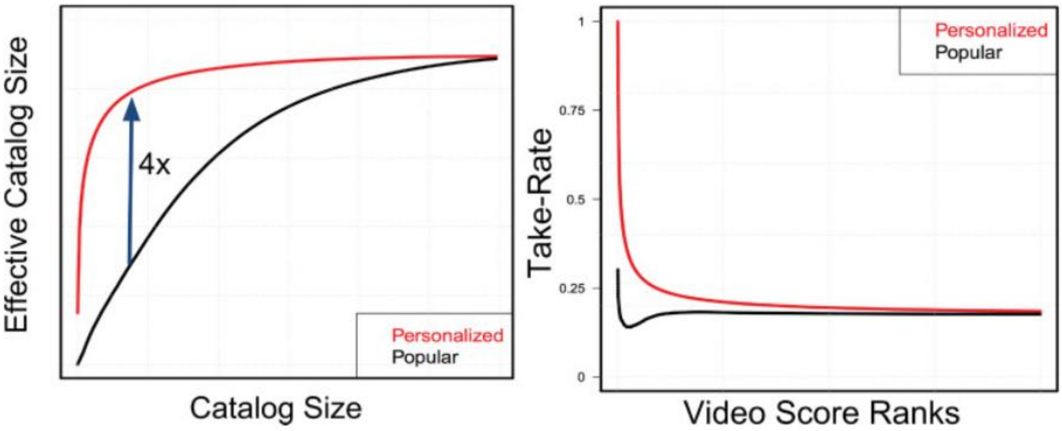

Fig. 4. (Left) The black line is the effective catalog size (ECS) plotted as a function of the number of most popular videos considered in the catalog, ranging from 1 through N (the number of videos in the catalog) on the x-axis. The red line is the effective catalog size for the first k PVR-ranked videos for each member. At a PVR rank corresponding to the median rank across all plays, the ECS in red is roughly 4 times that in black. The values in the x and y axis are not shown for competitive reasons. For more details, see Appendix A. (Right) The take-rate from the first k ranks, as a function of the video popularity rank in black, and as a function of the PVR rank in red. The y-values were normalized through division by a constant so that the maximum value shown equalled 1.

Without personalization, the same video of her is recommended for all members. The black line in the left graph of Figure 4 shows how his ECS without personalization
increases as the number of videos included in the data increases. It starts with the most popular videos first and then adds the next most popular videos.
x-axis. On the other hand, the red line in the same graph shows that
increases as a function of PVR rank included to measure personalization, rather than
increasing as a function of videos with ECS included. The difference in catalog search numbers with and without personalization is noticeable, but not convincing enough by itself. Finally, we can monitor
more evenly by providing completely random recommendations for each session.
More importantly, personalization greatly increases the likelihood of success in providing recommendations. One metric to achieve this is take rate. This is the percentage of recommendations served that lead to games. The two lines in the graph
on the right side of FIG. 4 show take rates, one as a function of the video's popularity
and the other as a function of the video's PVR rank. The increase in take rate from
referrals is significant. Most importantly, properly crafted and used referrals can significantly increase overall engagement with product
(e.g. streaming time) and lower subscription cancellation rates.

Subscriber monthly churn rates are in the low single digits, mostly due to payment defaults
and not subscribers' explicit choice to cancel service. He spent
years developing personalization and recommendations, resulting in a few percent reduction in churn for
. Lowering the monthly churn rate increases both the lifetime value of her existing
subscribers and the number of new subscribers he needs to acquire
to replace terminated members. We believe the combined effect of personalization
and referrals will save us over $1 billion annually.

4 Algorithm Improvement

Good companies pay attention to what their customers say. But what customers want (as many choices as possible, comprehensive search and navigation tools, etc.) and what actually works (easily presenting a few compelling choices) are very different.

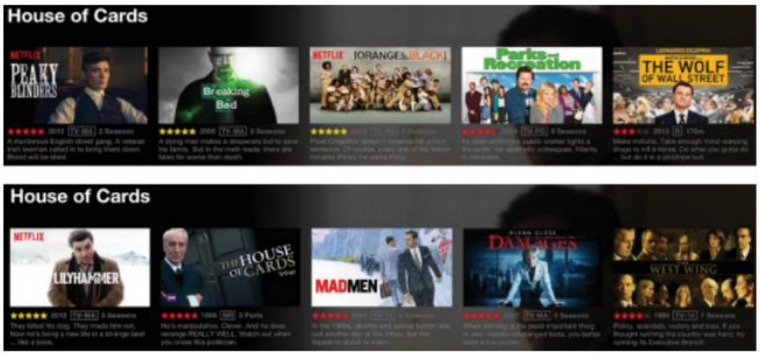

Fig. 5. Two sets of video similars for "House of Cards." The bottom ones seem more relevant, but turn out to be worse than the ones shown on top that have a stronger popularity influence.

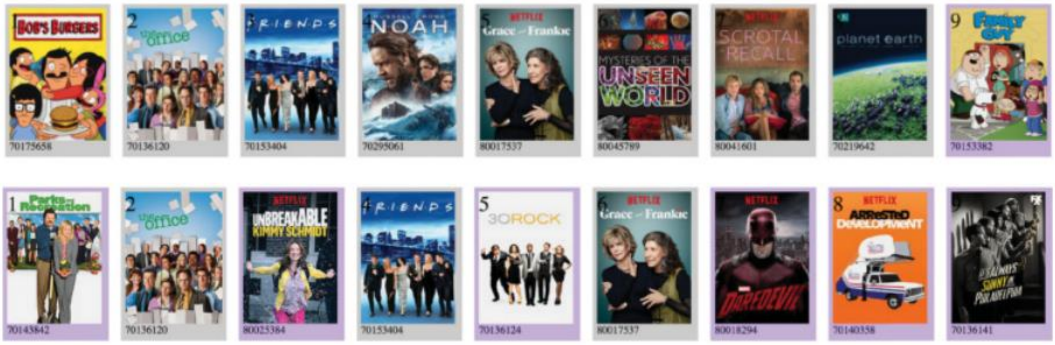

Fig. 6. The 9 highest ranking videos in the catalog according to two variants of the PVR algorithm, evaluated for one of the authors.

Using one's own intuition, or even collective intuition, to select the best variant of a recommendation algorithm often leads to wrong answers, trying to distinguish between particularly good and good recommendations. The case is often simply not possible. As an example of the

intuition error, Figure 5 shows two sets of his videos similar to "House

of Cards". People often intuitively think that lower is better because lower seems more relevant. For example, it contains the original version of "House of Cards". However, according to A/B testing, other similar statements above are better (see Section 4.1).

Another example, Figure 6, shows the author's highest-ranking PVR video in her

catalog. Estimating other people's rankings is

even more difficult. So how do you know if one algorithm variant is better or worse than another?

4.1. Choosing A/B Testing Metrics

Our subscription business model offers a framework for finding answers. Our

revenue is derived solely from the monthly subscription fees paid by current

members, with the ability to easily unsubscribe at any time to maximize

revenue through product changes feels pretty much the same as maximizing the value of

. Our members are derived from our service. Revenue is proportional to the number of her members, and three processes directly affect this number: new member acquisition rate, member churn rate, and former member rejoin rate.

Creating a more engaging service by offering more personalized recommendations will keep members on the fence longer and increase retention.

Plus have an improved experience All members (not just those on the fence) are likely to be more enthusiastic when explaining Netflix to their friends, and this has a big impact on word-of-mouth new subscriber acquisition. Both are reminiscent of the better

Increased experience and word of mouth may encourage former members to rejoin soon. While it is possible to directly measure retention (and changes in retention from A/B testing), there is no reliable way to measure word of mouth for different algorithm variants. Anyone who has experienced her variant on Netflix. Changes to the product directly affect current members only. Therefore, the primary measurement goal of any change in the recommended algorithm is an improvement in member retention

. needs significant improvement.

However, we found that increased engagement, the amount of time a member spends watching her Netflix content, is highly correlated with increased customer retention. Therefore, we designed a randomized controlled experiment, often called an A/B test, to compare medium-term engagement with Netflix and member churn rates of across algorithm variants. Algorithms that improve test metrics are considered better. Therefore, we are developing algorithms with the goal of maximizing medium-term engagement with Netflix and member retention.

Specifically, our A/B test randomly assigns different members to different experiences (called cells). For example, each cell in an A/B test can be assigned a different video similarity algorithm. One of them reflects the standard algorithm (often called "production") and serves as the experimental control cell. The cell has been tested with a test cell. We then let the members in each cell interact with the product over a period of months, typically 2 to 6 months. Finally, we analyze the resulting data to answer several questions about member behavior from a statistical perspective, including:

—Are members finding the part of the product that was changed relative to the control

more useful? For example, are they finding more videos to watch from the video
similars algorithm than in the control?
—Are members in a test cell streaming more on Netflix than in the control? For example, is the median or other percentile of hours streamed per member for the duration
of the test higher in a test cell than in the control?1
—Are members in a test cell retaining their Netflix subscription more than members
in the control? If the
test cell showed a significant improvement over the current experience, we know that members are more engaged with the changed parts of the product (increased local engagement index) and more engaged with the Netflix product as a whole. (increased overall engagement). His retention rate is
higher (a clear overall advantage). Year after year, we see some distinct increases, but more of an increase in overall engagement that isn't big enough to impact retention.
simply cannibalize the streaming of other parts of the product, or the overall retention or increase in retention is too small to be detected with reasonable confidence given the sample size of the test).

A/B testing is designed to ensure that each member of the
test has a consistent product experience for its duration. A more traditional alternative is to randomly select the algorithmic experience to offer each Netflix session, which improves the statistical performance of local metrics (see e.g. Chapelle et al. [2012]) but is less dynamic. Design without the ability to measure change overall engagement across the product, or retention rate over many sessions.

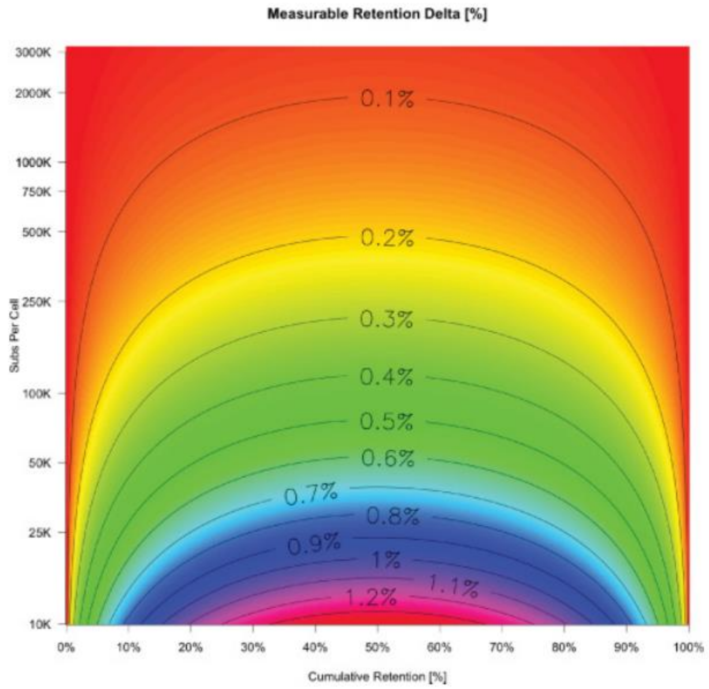

Fig. 7. A plot for the minimum retention delta that can be measured with statistical confidence, as a function of the average retention per cell and the cell size.

4.2. Test Cell Size for Statistical Validity

Use the statistic as a guide to determine if there is enough data to conclude that there is a difference in her A/B test metrics between cells . As an example, suppose that after 2 months we found that some pc and pt members of the control and test cells in a 2-cell A/B

test are still Netflix members, where = pt − pc > 0. increase. . Intuitively, the more members participating in the test, the more confidence we should have in the observed

Delta. But how many members are enough to trust the test results? and use that model to estimate how much the metric is expected to change with repetition. Repeat the experiment (with the same sample size) many times. For example, the smaller the percentage of repeated hypothetical experiments that the probabilistic model considers to be negative, the greater the confidence that the test cells actually increase retention. See Appendix B for examples of such probabilistic models.

or Siroker and Koomen [2013], Deng et al. [2013] and Pekelis et al. [2015]

for details on A/B test statistics. 4464 Probabilistic models can also be used to determine the sample size required to measure an increase or decrease of a given magnitude with reasonable confidence. As an example of

, Figure 7 shows the size of the measurable binding delta across two test

cells with the same number of members as a function of both the average percent binding (x-axis) across the two cells and the number of members. I'm here. Simple connection probability model in Appendix B for each cell (y-axis). Measure the combined delta from 50.05% to 49.95% = 0.1%.

4.3. Nuances of A/B Testing

A/B test results are the primary source of information for product decisions. Our tests are very informative in most cases. But despite the statistical sophistication of A/B test design and analysis, interpretation of A/B tests is still an art. For example, you might pass statistical test
, but see retention gains that aren't supported by increases in overall or local engagement metrics. In such cases, we tend to assume random fluctuations that are not due to testing experience. Our common practice is to rerun such A/B tests. You will usually find that retention gains are not repeatable. This is in contrast to the larger improvements supported by local and overall engagement metrics.

Other cases show an increase in overall engagement without an increase in local metrics. Repeat them often. The number of tests with apparently misleading results can be reduced with more sophisticated experimental design and analysis. [2013]) to make the cells in a test even more comparable to each other, for
instance, in terms of attributes that are likely to correlate highly with streaming and retention rates, such as the method of payment or the device of sign-up.

4.4. Alternative Metrics

There are many other possible metrics that we could use, such as time to first play, sessions without a play, days with a play, number of abandoned plays, and more. Each of these changes, perhaps quite sensitively, with variations in algorithms, but we are unable to judge which changes are for the better. For example, reducing time to first play could be associated with presenting better choices to members; however, presenting more representative supporting evidence might cause members to skip choices that they might otherwise have played, resulting in a better eventual choice and more satisfaction, but associated with a longer time to first play.

4.5. Test Audience

We typically test algorithm changes on two groups of members: existing members and new members. The advantage of testing with existing members is a
larger sample size. However, an existing member, in the past he has experienced

different versions of the product. A sudden change in their experience to reflect that of the test cell can result in behavior influenced by previous experiences.

Such tests often measure the impact of immediate product changes rather than the medium-term impact of the new experience itself. For example, how to find actors, measures often change negatively. When changes only appear in other alternatives, novelty often reveals previously undiscovered titles, leading to positive, non-representative measurements of better alternatives in the medium to long term.

We prefer to test on new members because they have not experienced a different version of the product before; thus, their responses tend to be indicative of the effectiveness
of the alternative versions of the algorithm rather than the change from old to new,
yielding cleaner measurements. A disadvantage is that we have fewer new members,
only as many signups as we get during the time period when we allocate new members
into a test. Another disadvantage is that we offer new members a one-month-free trial,
so we see few cancellations before this free month expires and cannot measure accurate
retention rates until one month after the last new member in the test joined Netflix.

4.6. Faster Innovation Through Offline Experiments

The time scale of our A/B tests might seem long, especially compared to those used by many other companies to optimize metrics, such as click-through rates. This is partly addressed by testing multiple variants against a control in each test; thus, rather than having two variants, A and B, we typically include 5 to 10 algorithm variants in each test, for example, using the same new model but different signal subsets and/or parameters and/or model trainings. This is still slow, but too slow for finding the best parameter values for a model with many parameters, for example. For new her
members, more test cells means more days to allocate new enrollments to the test so that the
has the same sample size in each cell. Another way to speed up your testing is to run many different A/B tests against the same member population at the same time. Each new member can be assigned to several different tests at once, as long as we assume that the test experience variations are compatible with each other and that they do not couple non-linearly with the experience. For example, a similar test, a PVR algorithm test and a search test. Therefore, a single her
member can receive version B of the Similars algorithm, version D of the PVR algorithm, and version F of the
search result. Over perhaps 30+ sessions over the course of the test, the members'

experiences are accumulated into metrics for each of the three different tests.

But to really accelerate innovation, we also rely on another type of experimentation based on analysis of historical data. This offline experiment varies for each algorithm

, but always consists of computing metrics for each algorithm variant

tested. This metric shows how well the algorithm variant matches previous user interactions.

For example, for PVR, you might have 100 different variants trained on data from two days ago, differing only in the

parameter values used.

Then use each algorithm variant to rank the sample member's catalog using his

data up to 2 days ago to find the rank of the videos played in the last 2 days by the sample

member . We then use these ranks to calculate her

metrics for each user across variants. For example, mutual rank mean, precision, recall, etc. These are averaged across members in the sample, possibly with some normalization. See Alvino and Basilico [2015] for another detailed offline metric example used in the construction algorithm on page

. Offline experiments allow rapid iteration of algorithm prototypes and shortened candidate variants for use in her real A/B experiments. A typical innovation flow is shown in Figure 8.

Offline experiments are attractive, but have one major drawback:

When a new algorithm is evaluated and recommendations are generated, we assume that members behaved similarly, such as playing the same videos.

For example, a new algorithm resulting in a very different production algorithm recommendation

found that its recommendation

was played more frequently than the corresponding production algorithm recommendation

that actually provided recommendations to its members. unlikely to discover. This suggests that the offline experiment

should be interpreted in the context of how the tested algorithm

differs from the production algorithm. However, it is unclear which distance metric across the

algorithm leads to better interpretation of offline experiments that correlate better with

A/B test results. Because the latter is what we are looking for. So we rely heavily on offline experimentation, but for lack of better options,

decides when to A/B test new algorithms and which ones to test. Results as we like.

4.7. Estimating Word of Mouth Effectiveness

As mentioned earlier, improving the member experience can be expected to generate more reviews. This is by definition hard to measure as it affects beyond the range of the A/B test cell. By taking advantage of several natural experiments that have been able to examine long-term changes in experience localized in one country and not in another, the rate of acquisition between pairwise estimates of countries can be discerned from variations in , and approximate limits can be derived. About the degree of word of mouth for such changes. Estimates are based on many assumptions and are fairly unreliable, but if a change results in retaining more existing members for a particular period

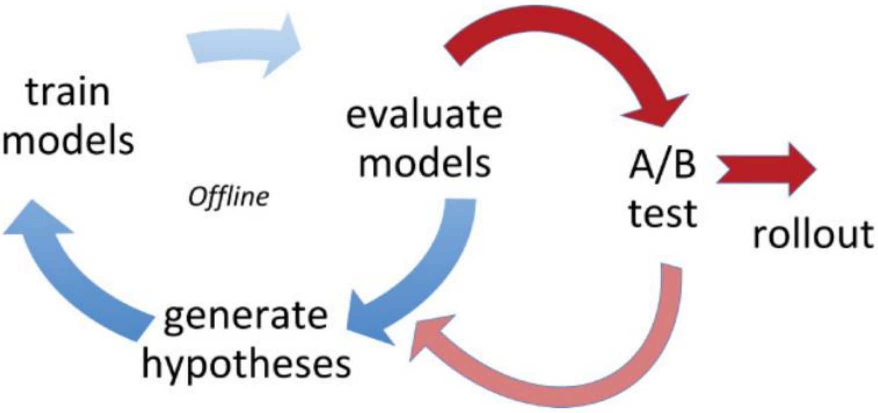

Fig. 8. We iterate quickly to prototype an algorithm through offline experimentation by analyzing historical data to quantify how well a new algorithm can predict previous positive member engagement, such as plays. The key underlying assumption, which is not always true, is that members would have engaged with our product in exactly the same way, for example, playing the same videos, had the new algorithm been used to generate recommendations. Once we see encouraging-enough results in offline experiments, we build an A/B test to use the new algorithm to generate recommendations for members. If the A/B test succeeds, we change our product to use that new algorithm by default. If the A/B test is flat or negative, we either abandon the research direction or go back to the offline experimentation world to try to make the new algorithm even better for a possible future A/B test.

of time, it could generate an increase in word of mouth, which It concludes that it may inspire comparable order.

new members.

5. Major Open Issues

Netflix has invested in developing our recommendation system for over a decade and we still believe in it. Our recommendations could be significantly better than they are now. Some of the most important open questions today are about A/B testing, and others about the recommendation algorithms themselves.

5.1. Better Experimentation Protocols

We want a better alternative to offline experimentation, one possibility he in
we're investigating is focusing on metrics from local algorithms such as click-through rates That's his A/B test for Interleave his base. It remains to be seen if the results of these tests can determine situations where they correlate well with the
total increase in streaming and retention in a standard A/B test.Another option is A/B testing to develop new offline experimental indicators that better predict the outcome of I'm also interested in improving A/B testing in general. For example, effective variance reduction methods for running experiments with higher resolution and less noisy results, and a new A/B engagement metric that further increases retention correlation. The challenge associated with engagement metrics is finding the right way to balance long-form and short-form content. Since we cover both movies (usually 90-120,
minute views) and multi-season TV shows (sometimes 60 hour episodes), a single
Discovery event can cover a night or several weeks.
views can attract customers. His
show credit for multiple seasons is just too big to just count streaming hours. Counting the "novel plays" (different titles detected) could be overcorrected in favor of his one-off movie of.

5.2. Global Algorithm

We plan to offer Netflix worldwide by the end of 2016. Our industry relies on Content License
. This is often exclusive and specific to a region or country. This creates
in different Netflix video catalogs in different countries. Today, we group our
countries into regions that share very similar catalogs, but with a member base of
sufficient to generate enough data to fit all the models we need. still have Then run a copy of all
algorithms in isolation in each region. Rather than scaling this approach as
we offer our service around the world, we are developing a single global recommender
system that shares data across countries. The data shared include not only the relevant
engagement data, such as plays, but also what the catalog of videos is in each country.
Our goal is to improve the recommendations for smaller countries without affecting
larger ones. We are thus interested in approaches that generalize many of the standard
mathematical tools and techniques used for recommendations to reflect that different
members have access to different catalogs, for example, relying on ideas from the
statistical community on handling missing data [Schafer 1997].

We are also interested in models that take into account how the languages available for the audio and subtitles of each video match the languages that each member across the world is likely to be comfortable with when generating the recommendations, for example, if a member is only comfortable (based on explicit and implicit data) with Thai and we think would love to watch "House of Cards," but we do not have Thai audio or subtitles for it, then perhaps we should not recommend "House of Cards" to that member, or if we do have "House of Cards" in Thai, we should highlight this language option to the member when recommending "House of Cards."

Part of our mission is to commission original content across the world, license local content from all over the world, and bring this global content to the rest of the world.

I would like to introduce the best French drama in Asia, the best Japanese anime in Europe and so on. Translating all her

titles into every other language would be too cumbersome and costly, so based on a sample of the content viewed by each member and how they viewed it, we would need to know. This allows us to suggest appropriate subsets of tracks based on what members enjoy.

5.3. Controlling for Presentation Bias

We have a system with a strong positive feedback loop, in which videos that members engage highly with are recommended to many members, leading to high engagement with those videos, and so on. Yet, most of our statistical models, as well as the standard mathematical techniques used to generate recommendations, do not take this feedback loop into account. In our opinion, it is very likely that better algorithms explicitly accounting for the videos that were actually recommended to our members, in addition to the outcome of each recommendation, will remove the potential negative effects of such a feedback loop and result in better recommendations. For example, a problem in this area is finding clusters of members that respond similarly to different recommendations; another is finding effective ways to introduce randomness into the recommendations and learn better models.

5.4. Page Building

Page building is relatively new and unexplored territory for us. It took us several years to find a fully personalized algorithm to build

recommendation pages that are better A/B tested than pages based on templates (which themselves have been A/B tested

. B test was optimized). We believe there are endless ways to improve this algorithm. We do not yet know that page building problems are the main focus of the
academic recommender community, but we believe that many recommender problems share similar characteristics.
General problem of placing items in a
catalog individually for each person relative to each other.
5.5. Member Cold Start
We know the referral system does a satisfying job of helping members with
's great Netflix history, but not for new members we know very little about. For example, our PVR algorithm tends to rank much higher before playing videos discovered by members before being played by existing members than by new members. As new members get her 1month free trial, the cancellation rate is the highest among them, she's and then drops off rapidly. This shouldn't come as a surprise, as her new member has to decide whether to pay for her Netflix. Meanwhile, the long-time member has already paid her to Netflix the previous month, and in another month, she just has to decide whether to pay. Therefore, we are always interested in finding better models and her signals to improve our recommendations to new members and increase engagement and her retention rate. Today, the cold-start approach to members has evolved to vote conducted during the sign-up process. During this time, ask the new member to select video from the algorithm input set to use as input to all algorithms.

5.6. Account Sharing

We market Netflix subscriptions to families; in many cases, several individuals with
different tastes share a single account. We allow our members to create up to 5 different
profiles for every account, and we personalize the experience for each profile. However,
a large percentage of profiles are still used by multiple people in the household. Our
recommender system has, by necessity, evolved through years of A/B testing to deliver
a mix (union) of suggestions necessary to provide good suggestions to whichever member of the household may be viewing (owner, spouse, children) at any time, but such amalgamated views are not as effective as separated views. We have lots of research and exploration left to understand how to automatically credit viewing to the proper profile, to share viewing data when more than one person
is viewing in a session, and to provide simple tools to create recommendations for the
intersection of two or more individuals` tastes instead of the union, as we do today.

5.7. Selecting the Best Evidence to Support Each

Recommendation There are several images, summaries, and other evidence that can be used to present each recommendation. These can be selected to emphasize different aspects of a video such as B. Actors or directors involved, awards won, settings, genres, etc. For us, the area of evidence selection involves finding the best evidence for each recommendation. We are currently investigating to what extent these choices can be personalized.

6. Conclusion

We have described the various algorithms that make up the
Netflix recommender system, the process we use to improve it, and some open issues. Humans are facing an increasing number of choices in every aspect of their lives—certainly around media such as videos, music, and books, other taste-based questions such as vacation rentals, restaurants, and so on, but more importantly, around areas such as health insurance plans and treatments and tests, job searches, education and learning, dating and finding life partners, and many other areas in which choice matters significantly.
We are convinced that the field of recommender systems will continue to play a pivotal role in using the wealth of data now available to make these choices manageable, effectively guiding people to the truly best few options for them to be evaluated, resulting in better decisions. We also believe that recommender systems can democratize access to long-tail products, services, and information, because machines have a much better ability to learn from vastly bigger data pools than expert humans, thus can make useful predictions for areas in which human capacity simply is not adequate to have enough experience to generalize usefully at the tail.

APPENDIXES

A. THE EFFECTIVE CATALOG SIZE

Assume that we have N items in the catalog, ordered from the most popular in terms of hours streamed to the least popular and denoted by v1,...,vN. The vector p = [p1,...,pN] gives the probability mass function (p.m.f.) corresponding to the percentage of the

hours streamed by the popular ranking videos in the catalog. H. pi is the
fraction of all streaming time coming from video vi, the i-th streamed
video. Note that pi ≥ pi+1 for i = 1,..., N − 1 and N
i=1 pi = 1. Print range [1, N] as output
. In a way, this tells you how many videos you need to stream in a typical
hours. This metric returns a value slightly higher than 1 if the most popular
video v1 accounted for the most streamed time, or 1 if all videos in the
catalog accounted for the same amount of streaming. Returns the value of N. One such metric is Effective Catalog Size
(ECS), defined as:

$$ECS(\mathbf{p}) = 2\left(\sum_{i=1}^{N} p_i i\right) - 1.$$

Equation (1) simply computes the average of the video indices under p.m.f. p and the
rescales it to the appropriate range. It is easy to check that for all i, the ECS has a minimum value of 1 when p1 = 1 and a maximum value of N when pi = 1/N.

ECS can access all p.m.f. We start by calculating the ECS
of the reference, p.m.f. This takes into account the hours of the most popular k videos only when you increase k from 1 to N. In particular, define p(k) = α[p1,..., pk]. where α = 1/(ki=1 pi) is the normalization constant and records ECS(p(k)). , varying k to obtain the black line in the left plot of Figure 4. This line is below the identity line (not shown).), because not all videos are equally popular. The red line on the same graph is the result of applying the -ECS formula to another p.m.f. Varying k from 1 to N yields q(k). p.m.f. q(k) is the percentage of time from each PVR rank above k out of all streaming time from the top k PVR ranks. To form q(k), take the k highest ranked PVR videos of her for each member, find all the streaming times generated by those member-video pairs, and find the i-th Define entries as shares of these streaming times. from PVR rank i. Note that for each member q(k) contains only her
k videos, but p(k) contains more videos. It contains videos, probably all N, across the member sample. This is because PVR is personalized. At her PVR rank of, which is the average rank for all games, the effective catalog size is approximately four times the corresponding non-personalized effective catalog size.

B. Example of A/B Test Statistics

A simple standard retention model assumes that each member in control cell
tosses a coin and the probability of that coin coming up heads is μc. Each member of test cell
similarly tosses and keeps a coin, but with probability μt. We want to estimate the retention difference = $\mu t - \mu c$. Applying the maximum likelihood to the retained data
in each cell gives an estimate where the chance variable is set to 1 if the member u is retained in the control cell c, and 0 otherwise. increase. Xut similarly describes the result if the test cell holds member u, where nc and nt are the number of members of the control. test cell. Then estimate by ˆ = pt − pc. Then the variance of the estimate of
pc is simply $\mu c(1 - \mu c)/nc \approx pc(1 - pc)/nc$, and a similar formula gives the variance of the estimate of pt as
. Simply die sum of
variances of pc and pt estimates, i.e. H. σ 2 = $pc(1 - pc)/nc + pt(1 - pt)/nt$. standard deviation σ is much smaller than ˆ If so, you can be confident that the high retention in your test cells is not due to a finite or small sample of members in each cell. Roughly speaking, the standard approach is to assume that it follows a
Gaussian distribution with mean ˆ and variance σ 2 and declare the test cell positive for retention if ˆ ≥ 1.96σ. A plot of 1.96 σ for the decision limit
is shown in FIG. 44 as a function of cell size and retention when the two cells are of equal size
and approximately equal retention. This type of chart can be used as a guide for selecting a sample size of cells for testing. For example, to detect a retention delta of 0.2%, the sample size indicated by the black line labeled 0.2% is required.
is tracked. Varies as a function of average retention when the experiment was stopped, with
being the maximum (south of 500,000 members per cell) at 50% retention.
Different probabilistic models give different results. For example, previous test results can be used to construct different previous distributions for different parameters such as μc and μt
. We can consider the parameters
of this underlying beta distribution for each cell, or
stratified samples if used to create the test cells.

ACKNOWLEDGMENTS

We thank the many algorithms teams at Netflix for their work to run and improve our recommender system. We are also grateful to Angadh Singh for the effective catalog size and take-rate plots, and to Justin Basilico for helpful suggestions on this manuscript.

Figure 1 (left): The Others c 2001, Miramax. The Quiet American c 2003, Miramax. Before I Go to Sleep c 2014, Relativity Media, LLC. Carlos c 2010, IFC. The Sixth Sense c 1999, Buena Vista Pictures and Spyglass Entertainment Group, LP. Frontline: Losing Iraq c 2014, WGBH Educational Foundation. Battleground Afghanistan c 2013, National Geographic Channel. All Rights Reserved. WWII in HD c 2009, A&E Television Networks. All Rights Reserved. Korengal c 2014, Virgil Films.

Figure 1 (right): La Prepago c 2013, Sony Pictures Television Group. All Rights Reserved. The Universe c 2007, A&E Television Networks. All Rights Reserved. The West Wing c 2006, Warner Bros. Entertainment Inc. Escobar, el Patron del Mal ´ c 2015, Caracol. Los Caballeros Las Prefieren Brutas c 2010, Sony Pictures Television Group. All Rights Reserved. Jessie c Disney, All Rights Reserved, Disney Channel. High Fidelity c 2000, Touchstone Pictures. All Rights Reserved. Daawat-e-Ishq c 2014, Vista India. Beyond the Lights c 2014, Relativity Media, LLC.

Figure 2 (left): Transformers c 2007, Paramount Pictures. Orange Is the New Black c 2015, Lionsgate Television Inc. All Rights Reserved. Sense8 c 2015, Georgeville Television, LLC. Marvel's Daredevil c 2015, MARVEL & ABC Studios. Once Upon a Time c ABC Disney. Pretty Little Liars c 2015, Warner Bros. Entertainment Inc. House of Cards c 2015, MRC II Distribution Company L.P. All Rights Reserved. Homeland c 2015, TCFFC. All Rights Reserved. The Good Wife c 2015, CBS Corp. Avatar: The Last Airbender c 2013, Viacom International Inc. Total Drama c 2008, Cake.

Figure 2 (right): Scooby Doo c Hanna-Barbera and Warner Bros. Entertainment Inc. Orange is the New Black c 2015, Lionsgate Television Inc. All Rights Reserved. Sense8 c 2015, Georgeville Television, LLC. Dragons: Race to the Edge c 2015, DreamWorks Animation LLC. All Rights Reserved. Phineas and Ferb c Disney, All Rights Reserved, Disney Channel. Notbad c 2013, Anthill Films. Cake c 2014, Turtles Crossing/Freestyle. Danger Mouse c Fremantlemedia. Antarctica: A Year on Ice c 2013, Music Box. Some Assembly Required c 2015, Thunderbird.

Figure 3 (left): Reservoir Dogs c 1992, Miramax. The Big Lebowski c 1998, Universal Studios. All Rights Reserved. Pulp Fiction c 1994, Miramax. Rounders c 1998, Miramax. Taxi Driver c 1976, Columbia Pictures, a Sony Corporation. All Rights Reserved. House of Cards c 2015, MRC II Distribution Company L.P. All Rights Reserved.

Figure 3 (right): Frenemies c Disney, All Rights Reserved, Disney Channel. French Connection c 1971, TCFFC. All Rights Reserved. The French Minister c 2013, IFC. French Connection II c 1975, TCFFC. All Rights Reserved. Amelie c 2001, Miramax. Capital c 2012, Cohen Media Group. Young & Beautiful c 2013, IFC. Le Chef c 2012, Cohen Media Group.

Figure 5: Peaky Blinders c 2014, The Weinstein Company. Breaking Bad c 2013, Sony Pictures Television Group. All Rights Reserved. Orange is the New Black c 2015, Lionsgate Television Inc. All Rights Reserved. Parks and Recreation c 2015, Universal Television LLC. All Rights Reserved. The Wolf of Wall Street c 2013, Paramount Pictures. Lilyhammer c 2014, SevenOne International. House of Cards c 2015, MRC II Distribution Company L.P. All Rights Reserved. Mad Men c 2014, Lionsgate Television Inc. All Rights Reserved. Damages c 2012, Sony Pictures Television Group. All Rights Reserved. The West Wing c 2006, Warner Bros. Entertainment Inc.

Figure 6: Bob's Burgers c 2015, TCFFC. All Rights Reserved. The Office c 2012, Universal Television LLC. All Rights Reserved. Friends c 2004, Warner Bros. Entertainment Inc. Noah c 2014, Paramount Pictures. Grace and Frankie c 2015, Skydance Productions. Mysteries of the Unseen World c 2013, Virgil Films. Scrotal Recall c 2014, BBC. Planet Earth c 2006, BBC. Family Guy c 2015, TCFFC. All Rights Reserved. Unbreakable Kimmy Schmidt c 2014, Universal Television LLC. All Rights Reserved. 30 Rock c 2012, NBC Universal, Inc. All Rights Reserved. Marvel's Daredevil c 2015, MARVEL & ABC Studios. Arrested Development c 2013, TCFFC. All Rights Reserved. It's Always Sunny in Philadelphia c 2015, TCFFC. All Rights Reserved.

REFERENCES

Chris Alvino and Justin Basilico. 2015. Learning a Personalized Homepage. Retrieved December 6, 2015 from http://techblog.netflix.com/2015/04/learning-personalized-homepage.html.

Xavier Amatriain and Justin Basilico. 2012. Netflix Recommendations: Beyond the 5 stars (Part 2). Retrieved December 6, 2015 from http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5- stars.html

David M Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. ACM Transactions on Information Systems 30, 1. DOI:http:// dx.doi.org/10.1145/2094072.2094078

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In WSDM.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2011. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd. ed.). Springer. Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY).

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 8, 30–37.

Andriy Mnih and Ruslan Salakhutdinov. 2007. Probabilistic matrix factorization. In Advances in Neural Information Processing Systems. 1257–1264.

Kevin P. Murphy. 2012. Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge MA. Prasanna Padmanabhan, Kedar Sadekar, and Gopal Krishnan. 2015. What's trending on Netflix. Retrieved December 6, 2015 from http://techblog.netflix.com/2015/02/whats-trending-on-netflix.html.

Arkadiusz Paterek. 2007. Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD Cup and Workshop. 5–8. Leo Pekelis, David Walsh, and Ramesh Johari. 2015. The New Stats Engine. Internet. Retrieved December 6, 2015 from http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf. Netflix Prize. 2009. The Netflix Prize. Retrieved December 6, 2015 from http://www.netflixprize.com/.

Steffen Rendle. 2010. Factorization machines. In 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 995–1000.

Joseph L. Schafer. 1997. Analysis of Incomplete Multivariate Data. CRC Press, Boca Raton, FL.

Barry Schwartz. 2015. The Paradox of Choice: Why More Is Less. Harper Perennial, New York, NY.

Bryan Gumm. 2013. Appendix 2: Metrics and the Statistics Behind A/B Testing. In A/B Testing: The Most Powerful Way to Turn Clicks into Customers, Dan Siroker and Pete Koomen (Eds.). Wiley, Hoboken, NJ. Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association 101, 476.