Who's Watching?
De-anonymization of Netflix Reviews using Amazon Reviews

Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng
{marchie, sgershon, akatcoff, a2z}@mit.edu

Abstract— **Many companies' privacy policies state they can only release customer data once personal identifiable information has been removed; however it has been shown by Narayanan and Shmatikov (2008) and reinforced in this paper that removal of personal identifiable information is not enough to anonymize datasets. Herein we describe a method for de-anonymizing the Netflix Prize dataset users using publicly available Amazon review data [3], [4]. Based on the matching Amazon user profile, we can then discover more information about the supposedly anonymous Netflix user, including the user's full name and shopping habits. Even when datasets are cleaned and perturbed to protect user privacy, because of the sheer quantity of information publicly available through the Internet, it is difficult for individuals or companies like Netflix to guarantee that the data they release will not violate the privacy and anonymity of their users.**

I. INTRODUCTION

When users sign up for online services, individuals give companies the right to analyze and share their data, but only after their personal identifiable information (e.g., name, address, email, date of birth) has been scrubbed from the dataset. However, it has been shown in several studies [1], [2] that despite the removal of an individuals personal identifiable information (PII), it is possible to identify the individual using publicly available data. Not only is the confidentiality of the individual compromised, but the individual may be targeted based on the information released in the dataset.

This research paper presents a case study to identify users from an anonymized Netflix[1] dataset using product reviews from Amazon[2]. Re-identification of individuals could reveal possibly sensitive information about their interests and purchasing habits. By comparing reviews written on the same date with the same rating in both datasets, we were able to recover likely matches for **NUMBER OF USERS** Netflix users. Using a similarity score calculated based on the rarity of a movie or TV show title, difference between review dates and difference between ratings, we established likely identities of **NUMBER OF USERS** other Netflix users. Thus, despite Netflix's attempts to protect users in the dataset before publishing, the identities of some of its users were revealed.

II. RELATED WORK

The idea of de-anonymizing Netflix data using Amazon data was greatly influenced by prior work by Narayanan and Shmatikov (2008), who de-anonymized Netflix data using data from the Internet Movie Database[3] (IMDb). They developed a formal model for privacy breaches in anonymized micro-data, e.g. recommendations. Narayanan and Shmatikov also proposed an algorithm that predicts if ratings between datasets are correlated (by date and numerical rating). Using publicly available data from IMDb, they were able to identify several users in the "anonymized" Netflix dataset and learn potentially sensitive information about them, including political affiliations [2].

We aim to extend these results to show we can identify users from the "anonymized" dataset using publicly available Amazon reviews. As a result, we can learn about Netflix users' spending habits and reveal possibly private information about them.

III. DATASETS

For our study, we used two datasets: the Netflix Prize Dataset [5] and Amazon Product Data [3], [4]. Because of the size of these datasets, we only used data from 2005. We also standardized the product titles between the datasets to maximize the number of matches between them.

A. Netflix Prize Dataset

Today, Netflix is famous for providing online video-on-demand and streaming services. However, up until 2013, it focused on DVD sales and rentals. In 2006, the company hosted a competition called the Netflix Prize to improve its movie recommendation system. As part of the competition, Netflix released over 100 million ratings from almost 500,000 randomly selected customers. The online movie rental service published data collected between 1998 and 2005. Users rated DVDs on a scale from 1 to 5, with 5 being the best. Before publishing this subset of data, Netflix attempted to remove all personal identifiable information from the dataset, such that all that remained was the rating, date of rating, customer ID, title and year of release of each movie. Further attempts to abide to the Privacy Policy included replacing customer IDs with randomly generated ones and perturbing the rating scores. Based on these measures and the fact that the company only released less than one-tenth of its data, Netflix claimed that it was unlikely for an adversary to reliably identify someone in the dataset, even if the adversary somehow had the person's rating history [5], [6], [7].

[1] Netflix: https://www.netflix.com

[2] Amazon: https://www.amazon.com/

[3] Internet Movie Database: https://www.imdb.com/

B. Amazon Product Data

Amazon product reviews are public and therefore can be web-scraped. The Amazon Product Data we used was collected by McAuley et al. (2015) and contains product reviews between the years 1996 and 2014. McAuley et al. (2015) web-scraped almost 150 million reviews from Amazon. In particular, we used the reviews corresponding to the *Movies and TV* product category, a dataset containing approximately 1.7 million reviews. Each product review consists of the Amazon Standard Identification Number (ASIN) of the product, a rating on a scale from 1 to 5, the review date and other metadata. In addition, each product review features the reviewer's basic information, such as Amazon User ID and name (if specified by the user). Not all users compose multiple product reviews and thus, this contributes to the sparsity of the dataset. To reduce this sparsity, we only looked at users and movies that have at least 5 reviews each [3], [4].

C. Standardizing Movie Titles

The movie titles in both datasets are inconsistent. Unlike Netflix, product listings on Amazon are not only created by the company itself, but also by other sellers. Since Amazon does not have a set naming convention for movie titles, it possible that several product titles map to the same movie. By comparing the unstandardized titles of movies in both datasets, few matches were found. To confidently identify a user in the Netflix dataset, we want to maximize the number of movies appearing in both datasets, and thus, several rules were created to standardize the titles. Table I lists these rules in the order in which they were applied to the dataset. Using the standardized titles and movie release dates, we produced 8376 matching titles between the datasets.

IV. TECHNICAL APPROACH

To identify possibly linked Netflix and Amazon accounts, we explored two methods - exact matches and similarity scores.

A. Approach 1: Exact Matches

Our initial approach in identifying overlapping users between the two datasets involved finding *exact matches*. An exact match between a Netflix review and an Amazon review is defined as a pair of reviews that were written on the same date with the same rating for the same movie. The number of exact matches for each pair of users between the two datasets was used as a measure of the likelihood that a Netflix user and an Amazon user were the same person. This count was useful because a higher count would indicate that the two users watched the same movies around the same time (since they reviewed it on the same day) and had the same opinion about it (based on the rating). The strictness of this metric gave a small number of likely matches, which were straightforward to investigate. For this reason, this approach acted as a proof of concept to see if we could match users between the datasets.

The exact match count relies on the assumption that a user reviews a movie on the same date on both sites with exactly the same rating. However, this guideline is overly strict, and it does not account for users who might review movies on both sites a few days apart. In addition, users who review popular movies or who review many movies are likely to have a higher number of exact matches with other users. Therefore, a high number of exact matches does not in itself imply a user overlap.

B. Approach 2: Similarity scores

Because exact matches were relatively uncommon, we decided to implement a score which reflects how similar two users are based on the rarity of the movie, difference between review dates and the difference between the rating values.

One important piece of information which our exact match approach did not encompass was the movie's popularity. A Western from the 1930's is typically less commonly reviewed than a Marvel superhero movie, so a shared rating for the Western should support our hypothesis of a Netflix user and Amazon user being the same person more so than the popular movie.

For each movie m, we defined a rarity score function $R(m)$ determined by the the number of reviews the movie received in each of the datasets. If we let f_m^A be the fraction of reviews containing a movie m in the Amazon database and f_m^N be the fraction of reviews containing a movie m in the Netflix database, we can calculate a rarity score as follows:

$$R(m) = -\log_2\left(f_m^A\right) - \log_2\left(f_m^N\right) = -\log_2\left(f_m^A f_m^N\right) \qquad (1)$$

Note that this score reflects the sum of the entropies of the occurrences of each movie in each of the datasets. Given a qunatifier for the rarity of each movie, we can use it as part of a similarity score between two users, i and j. Denote M_{ij} as the set of movies reviewed by both, r_{mi} as user i's rating of movie m, and d_{mi} as the day of that rating.

$$\text{Similarity}(i,j) = \sum_{m \in M_{ij}} \frac{R(m)\left(2 - |r_{mi} - r_{mj}|\right)}{\left(\frac{|d_{mi}-d_{mj}|}{10} + 1\right)^2} \qquad (2)$$

The similarity of two users with no movies in common defaults to 0. Note that the difference in ratings creates negative terms in our similarity score if two users rated the movie more than 2 stars apart, reflecting that the same person is unlikely to rate the same movie with significantly different ratings. Additionally, we expect the same user to rate the movie on both sites within a reasonable time frame; for example, if an Amazon account rates a movie a 5 on January 4, 2005, and we find two Netflix accounts with ratings for the same movie on January 6 and September 23 of the same year, we would expect the Netflix account with the rating on January 6 to be more likely to be the same person as the Amazon account. The similarity score's denominator increases exponentially with the difference in dates, while allowing similar ratings 30 days apart to still

Rule	Example
Lowercase letters only	The Little Mermaid → the little mermaid
&amp; (HTML Encoding) → and	Beauty &amp; The Beast → beauty and the beast
& → and	The Princess & The Frog → the princess and the frog
&quot; (HTML Encoding) → ”	&quot;The Lion King&quot; → “the lion king”
Remove punctuation (non-alphanumeric characters)	Tangled: Before Ever After → tangled before ever after
Remove phrases *vhs set, box set, dvd set, disc set*	Disney Princesses Box Set → disney princesses
Remove phrases *vhs, box, dvd, disc, bluray, imax, edition*	Hercules [VHS] → hercules
vol → volume	Lizzie McGuire Vol. 2 → lizzie mcguire volume 2
Number words → Integers	Toy Story Two → toy story 2
Roman numerals → Integers	Pocahontas II: Journey to a New World → pocahontas 2 journey to a new world
Remove multiple whitespaces	Out of the Box [VHS]: Season 2 → out of the box season 2

TABLE I

TITLE STANDARDIZATION RULES LISTED IN ORDER OF USE.

have some impact on the score; the denominator is equal to 1 at 0 days and 16 at 30 days.

Even though we had a score to compare any two users, we still needed heuristics to guide which users to compare to avoid an excessive amount of computation. We therefore calculated the similarity score for (1) pairs of accounts with at least 1 exact match, as described in the previous section, and (2) pairs of users with at least 5 movies with the same exact rating. Each of these subsets included about 360,000 pairs. Thus we were able to narrow our search considerably based on the users we thought were more likely to be the same.

V. RESULTS

§ IV-A describes a heuristic to identify user pairs likely to correspond to the same person. We were able to recover at least one such user pair after manual verification; a case study of that user pair follows.

The simple matching metric and variations thereof tend to produce a large number of false positives. With the heuristic matches as a starting point, we filtered the user pairs using the similarity scoring metric to determine which candidate user pairs were *significant*, i.e., those that were least likely to be coincidences.

A. Overall Findings

Exact matches	User pairs
0	3 462 285 189
1	355 909
2	1 649
3	133
4	48
5	14
6	12
7	1
8	1
9	2
11	1
12	1

TABLE II

NUMBER OF USER PAIRS WITH n EXACT MATCHES

The Amazon review data contained reviews from 7,905 distinct users, while the Netflix data contained reviews from 438,032 users. Out of a total of 3,462,642,960 possible user pairings, 357,771 of them had at least one exact matching review. 366,913 had at least 5 movies in common with the same rating (regardless of date). Table II and Figure 1 show the number of pairs with the given count of exact matches and same-rating reviews, respectively. 32 exact match user pairs had at least 5 such exact matches between them.

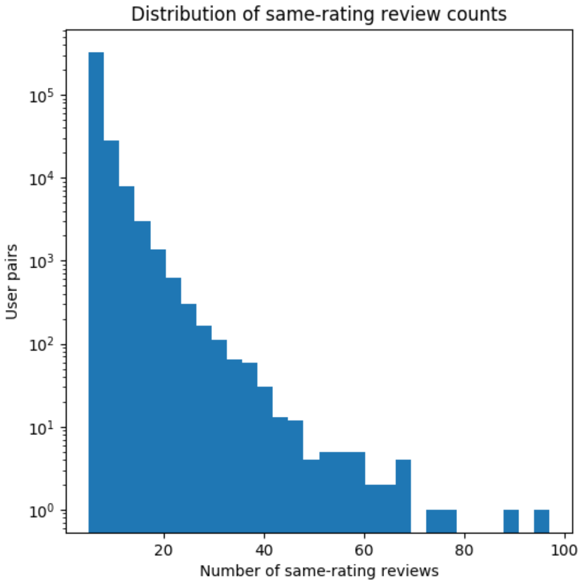

Fig. 1. Number of user pairs with $n \geq 5$ same-rating reviews

A majority of the user pairs with exact matches likely did not correspond to the same person. For example, of the user pairs with 5 or more exact matches, pairs with Amazon user AY69ZK7G6CNYJ (the DVD Report blogger mentioned previously) constituted 12 of 32. This user had 122 total reviews and had a maximum of 12 exact matches with any single Netflix users. Additionally, based on account information, this Amazon account was a writer for “thedvdreport.blogspot.com,” indicating that the user most likely reviewed popular movies. This increased the blogger’s likelihood of getting exact matches. Moreover, the exact matches metric does not account for users that might have reviewed the same movie a day apart on the two different platforms. One user pair stood out from this heuristic search.

We expected the overall count of same-rating reviews to provide a better metric, since it avoided the overly narrow date-matching criterion of the previous attempt; nevertheless, many spurious results appeared, similar to the ones found by the exact matches heuristic.

Using the similarity score formula, the user pair candidates could be narrowed to those for whom the matching movies were significant. Fig. 2 shows the distribution of similarity scores for the subset of user pairs with at least 5 same-rating reviews between them. Fig. 3 shows the top similarity scores for the same subset. From these user pairs, we were able to determine the identity of another Netflix user. That user had only one exact matching review with the corresponding Amazon identity, but the uncommon products reviewed by the user pair yielded a high similarity score, which made the pairing significant.

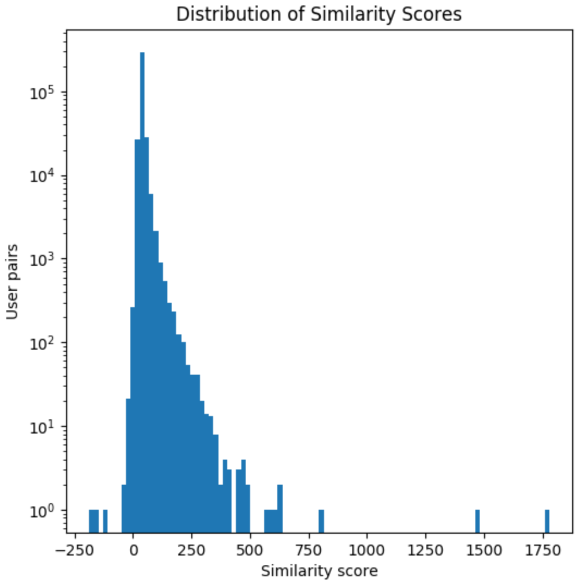

Fig. 2. Distribution of similarity scores

B. Case Studies

After analyzing possibly linked Netflix accounts with Amazon accounts based on exact matches and similarity scores, we have identified at least two likely matches. This section describes two case studies, Steven H. and Ronnie C., chosen because of their high similarity scores and uncommon movie preferences. These findings are summarized in Table III. By presenting these case studies, we challenge Netflix's claim that it sufficiently distorted the dataset to preserve its customers' privacy. While these matched accounts may not leak harmful information beyond additional movies reviewed on Netflix that weren't publicly reviewed on Amazon, a match between an innocent-looking Amazon account and a Netflix account with several porn reviews could be damaging to a user whose Netflix ratings were leaked in the "anonymous" Netflix Prize dataset.

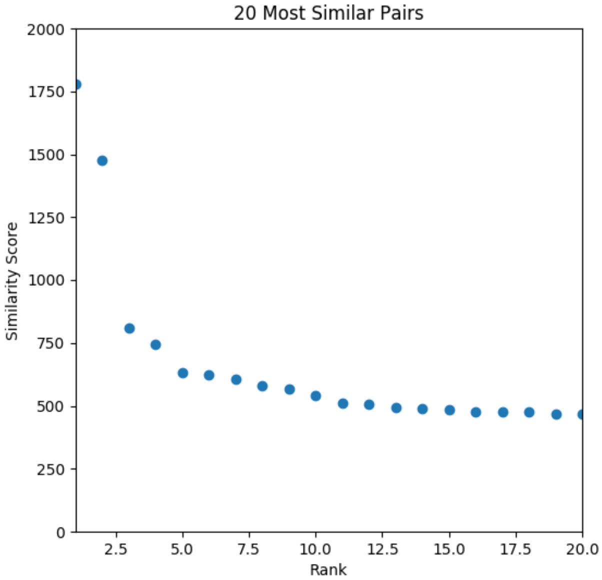

Fig. 3. Top similarity scores

Note that we list here the users' names only with a last initial, but the users' full names were available to us in the Amazon data, along with other profile information that was publicly available online and which would have allowed us to identify the individual precisely.

1) Steven H.: The Amazon account AK61LQI92GTCH (Steven H., based on the Amazon account information) and Netflix account 1764595 pair reviewed 43 titles in common. Of these 43 titles, 35 were given the same ratings and the remaining 8 were only off by 1 star. Almost all overlapping reviews in the datasets were made during the same month. In addition, there were 11 exact matches for the given Amazon and Netflix pair. Furthermore, the two accounts had the highest similarity score we found, 1778.3. This was heavily influenced by the rareness of the movies this user rated in both datasets. The user tended to rate old (mid-20th century) action movies.

2) Ronnie C.: Netflix account 1664010 reviewed over 7500 movies and TV shows on the popular entertainment company's website compared to the 155 written by Amazon account A1VCLTAGM5RLND (also known as Ronnie C.). With a similarity score of 581.3 and 141 titles reviewed by the pair in both datasets, it is highly likely that these two accounts are linked and owned by Ronnie C. 76 of the 141 movies and TV shows were given the same rating. Even though only 3 reviews were written on the same date, more than half of the overlapping ratings were made within a month of each other. By analyzing the common movie and TV show titles, we learned that the Ronnie C. liked to watch anime and salacious content.

	Steven H.	**Ronnie C.**
Amazon ID	AK61LQI92GTCH	A1VCLTAGM5RLND
Netflix ID	1764595	1664010
Similarity Score	1778.3	581.3
Amazon Reviews	46	155
Netflix Reviews	78	7597
Reviewed on Both	43	141
Same Ratings	35	76
Same Dates	12	3
Exact Matches	11	2

TABLE III

A SUMMARY OF THE NUMBER OF MOVIES REVIEWED BY TWO POSSIBLE USERS IDENTIFIED BY ANALYZING THE RATINGS, REVIEW DATES AND MOVIE RARENESS.

VI. LIMITATIONS AND FUTURE WORK

While we believe we succeeded in recovering users present in both datasets, our approach had some limitations.

Many of our limitations stemmed from our limited computational resources. To reduce the amount of data to use, we limited the range of dates we considered in the two datasets to just 2005. We certainly could have recovered more common users had we used a larger range, since we would have had a larger set of reviews for each user. Our limited resources also prevented us from computing similarity scores for every pair of users, perhaps keeping us from finding important overlaps.

Although we found some accounts with significant overlap in their ratings, our approach never gives us 100 percent certainty that a Netflix account and an Amazon account are owned by the same person. To achieve this, we would either need to obtain private Netflix and Amazon data or contact the suspected account owners.

More work can be done evaluating different types of similarity scores. In particular, our similarity score currently does not take into account movies the two users did not have in common. The movies which are not shared in the two accounts may reflect differences in the users' tastes. Many functions can be used to quantify the distance between two feature vectors; trying some of these may result in better outputs.

VII. CONCLUSION

Removing personal identifiable information from datasets is not enough to anonymize data. In this project we were able to confirm this idea by doing a membership attack on the Netflix Prize dataset, even though all personal identifiable information was removed and the data was slightly perturbed. Through the use of exact matches and similarity scores, we were able to identify Netflix users in the dataset (by first and last name), using only publicly available Amazon reviews. A main takeaway of this project is that companies need to be careful when releasing any user data, and must take into account all data available on the internet in addition to the data being released. When companies release compromising information, it can harm their trustworthiness, which could have serious implications for their business.

A concern we have after doing this project is that on Amazon, a surprising amount of account information is public. From a particular user's review, we can click on the user's name (often a full legal name) and view all reviews the user has made on Amazon along with the user's wishlist [8]. The items someone buys (even without text reviews) can reveal a significant amount of information about a person, including the region they live, their religion, their gender, their medical condition, and more. We believe some users may not realize how public their Amazon profiles are, meaning they may feel their privacy is violated on finding out the amount of information Amazon reveals.

Overall, we advise people to be cautious about what they put on the Internet; some companies may inadvertently reveal information in a format which they believe is anonymous but is not (Netflix) or reveal information a user may not realize is public (Amazon).

ACKNOWLEDGMENTS

We would like to thank our TA Akshay Degwekar for help in developing the project idea and Professor Adam Smith for getting us interested in dataset privacy and membership attacks.

REFERENCES

[1] Barth-Jones, Daniel C. The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now, 2012.

[2] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. 2008.

[3] R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016.

[4] J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015.

[5] Netflix Prize Dataset: https://www.kaggle.com/netflix-inc/netflix-prize-data. Downloaded on April 26, 2018.

[6] Netflix Prize Rules: https://www.netflixprize.com/rules.html. Accessed on May 12, 2018.

[7] Netflix Prize FAQ: https://www.netflixprize.com/faq.html. Accessed on May 12, 2018.

[8] Amazon: https://www.amazon.com/. Accessed on May 12, 2018.