Anonymization and Subsequent De-Anonymization of Letterboxd Ratings Data

Jordan Bass, Elijah Hansen, Albert Lu, Nicket Mauskar

Department of Computer Science, Northwestern University, Evanston, IL, USA

Introduction

Project Aims:

- **Anonymization of sensitive microdata**
- **Build off Arvind Narayanan and Vitaly Shmatikov's work**
- **Our project aims to continue the exploration of data anonymization and deanonymization**
- **We hope to draw conclusions that link the hyperspecificity of modern online usage to possible privacy issues, such as unwanted identification**

Letterboxd

Your life in film.

Methodology

Anonymization

Review_ID	Movie_Title	Rating	User_ID
5fc57c5d6758f69	feast-2014	7	deathproof
5fc57c5d6758f69	loving-2016	7	deathproof
5fc57c926758f69	the-age-of-inno	6	kurstboy
5fc57c926758f69	alice-doesnt-li	9	kurstboy
5fc57c966758f69	the-hunger	7	davidehrlich
5fc57c966758f69	siberia-2020	4	davidehrlich

Data Generalization via Binning of Ratings
Data Masking of Usernames

Review_ID	Movie_Title	Rating	User_ID
5fc57c5d6758f69	feast-2014	7-10	1
5fc57c5d6758f69	loving-2016	7-10	1
5fc57c926758f69	the-age-of-inno	4-6	2
5fc57c926758f69	alice-doesnt-li	7-10	2
5fc57c966758f69	the-hunger	7-10	3
5fc57c966758f69	siberia-2020	4-6	3

Addition of Synthetic Data

Review_ID	Movie_Title	Rating	User_ID
5fc57c5d6758f69	feast-2014	7-10	1
5fc57c5d6758f69	loving-2016	7-10	1
5fc57c926758f69	the-age-of-inno	4-6	2
5fc57c926758f69	alice-doesnt-li	7-10	4
5fc57c966758f69	the-hunger	7-10	3
5fc57c966758f69	siberia-2020	4-6	3
5fc57c926738f69	the-hunger-gam	0-3	2

Data Swapping

De-Anonymization

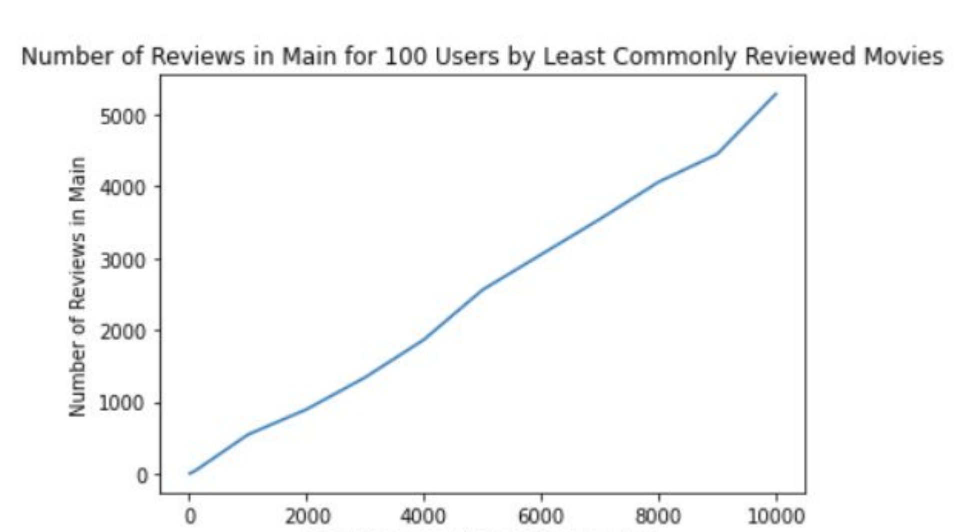
We instantiate the scoring function as follows:

$$\mathsf{Score}(\mathsf{aux}, r') = \sum_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{wt}(i)\left(e^{\frac{\rho_i - \rho'_i}{\rho_0}} + e^{\frac{d_i - d'_i}{d_0}}\right)$$

where $\mathsf{wt}(i) = \frac{1}{\log |\mathsf{supp}(i)|}$ ($|\mathsf{supp}(i)|$ is the number of subscribers who have rated movie i), ρ_i and d_i are the rating and date, respectively, of movie i in the auxiliary information, and ρ'_i and d'_i are the rating and date in the candidate record r'.[5] As explained in section 4,

-Narayanan and Shamtikov (2008)

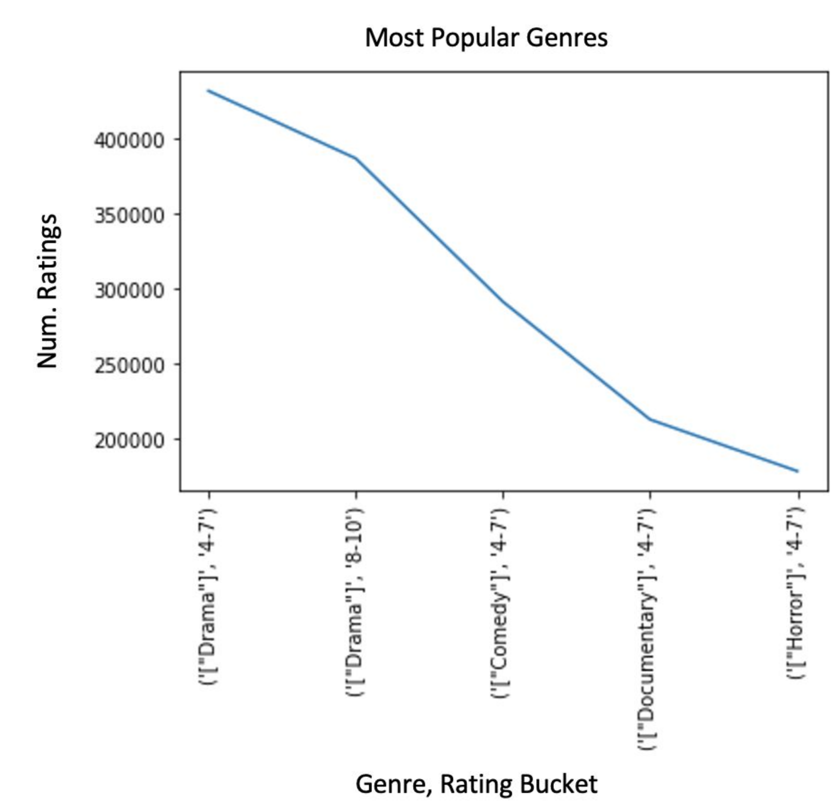
- **Altered deanonymization by filtering both datasets for more rare records**

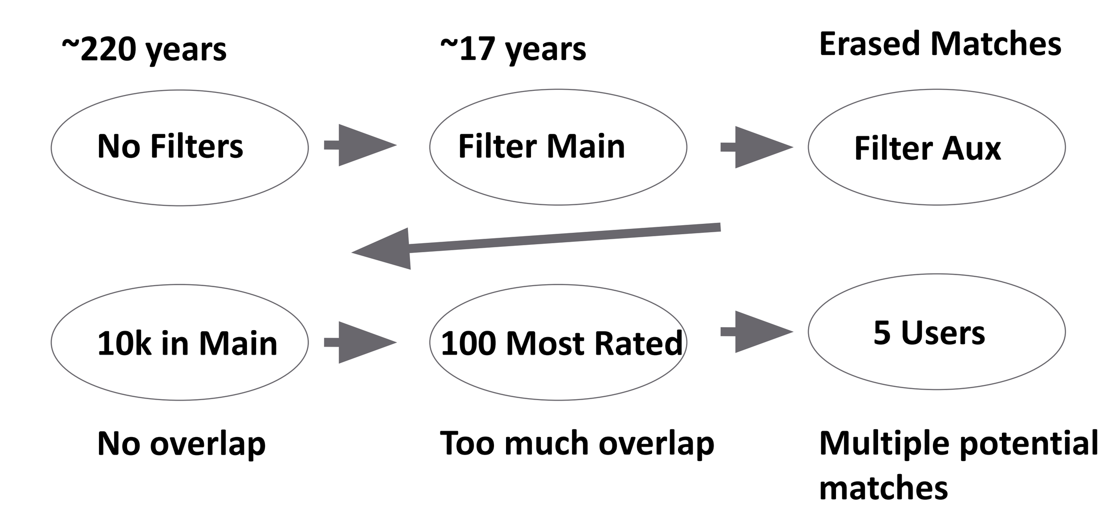


Results

Here are the results for average k-anonymity for the various anonymization techniques:

Control	Data Generalization	Data Swapping (one million swaps)	Synthetic Data (ten million swaps)	Improved Data Generalization
12.38	527.65	11.91	6.92	24030.73





Proof of Concept: For users 1-5
['deathproof', 'superpulse', ' lilfilm', 'ianamurray', 'punchdrunklizzy'] (username in Letterboxd)

[138990, 72315, 24610, 83426, 60656] (user_id in MovieLens)

Conclusion and Implications

- **We quickly realized that when a dataset is properly anonymized, it is very hard for an attacker to deanonymize a sparse dataset with such a large size like MovieLens**
- **We also came to the conclusion that an algorithm like Scoreboard-RH would take nearly 38 years to complete when applied to our datasets**
- **Due to the combination of size and anonymization of the dataset, an algorithm would have to be very efficient in order to deanonymize**

Implications for Future Work:

- **Use MovieLens dataset as main dataset and IMDb as auxiliary**
- **Would need better hardware for parallel processing when executing deanonymization**

References & Acknowledgements

Narayanan, Arvind, and Vitaly Shmatikov. "Robust De-Anonymization of Large Sparse Datasets." 2008 IEEE Symposium on Security and Privacy (Sp 2008), 2008, doi:10.1109/sp.2008.33.

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. https://doi.org/10.1145/2827872.

Learner, Sam. "LETTERBOXD Movie Ratings Data." Kaggle, 22 Mar. 2022, www.kaggle.com/datasets/samlearner/letterboxd-movie-ratings-data.