Differential Privacy Framework for Movie Recommendation System

Mert Kosan & M. Mucahid Benlioglu

Supervised by: Yucel Saygin

Sabancı Üniversitesi

Privacy and why it is needed

- Consider a trusted 3rd party that holds a statistical database of sensitive and private information (e.g. movie ratings, medical records, e-mail patterns) that would like to provide global, statistical information about the data for helpful applications and researches.

- Danger of revealing information about the individuals.
    - Usage of anonym identities for information.

- Deanonymization techniques using two or more separately innocuous databases.

An Example of Deanonymization

- Imagine a medical institute keeps records of certain illness and provides public and anonym statistical data to help researchers for possible cures. An important public figure is going to test for this illness in said institute. An adversary learns about this and by deanonymizing the research data he hopes to gather the private information about the public figure and use it against him, armed with this knowledge the adversary asks the database the query *"How many people have diagnosed with this disease so far?"*, after getting the answer the adversary asks the question again soon after public figure's test, and if result increases by one, he will conclude that the public figure is diagnosed with the disease.

- *How to solve this issue?*

Differential Privacy

- Differential privacy aims to provide means to maximize the accuracy of these statistical queries while minimizing the chances of identifying its records.

- It introduces noise to real data so that, adding or removing one user to database does not make noticeable difference in the data, thus preventing to identify his/her private information.

- It is probabilistic concept, therefore, any differentially private mechanism is necessarily randomized with Laplace mechanism, exponential mechanism etc.

Differential Privacy

- A more formal definition:

- Let ε be a positive real number and A be a randomized algorithm that takes a dataset as input (representing the actions of the trusted party holding the data). The algorithm A is *ε-differentially private* if for all datasets D_1 and D_2 that differ on a single element (i.e., the data of one person), and all subsets S of image of A.

$$Pr[A(D_1) \in S] \leq e^{\varepsilon} \times Pr[A(D_2) \in S]$$

where the probability is taken over the randomness used by the algorithm.

How would this help to previous example?

- In our previous example, if the database was differentially private, the adversary would get different randomized results each time he/she run the query, which would prevent him to reach a conclusion about the public figure’s test results, thus preventing his/her privacy.

Our project and the issue we tried to solve

- In the scope of this project we have considered a movie ratings database and using the ratings of given by users we calculated the most similar movies to user’s movie and recommend suitable ones among them.

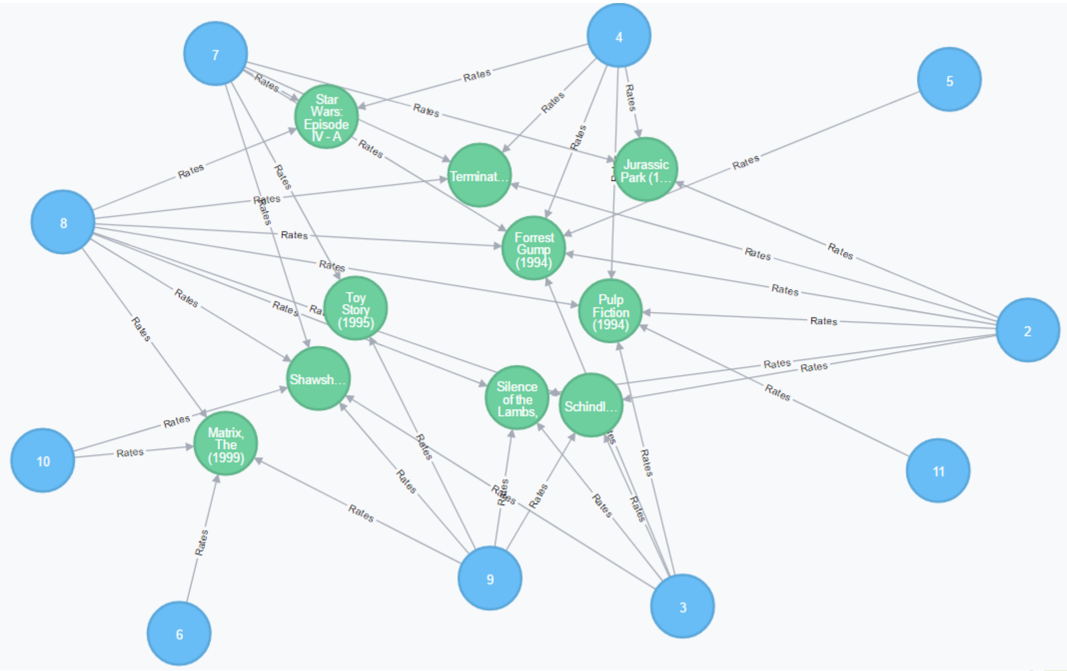
- We have used a graph database to represent users and movies as nodes, and placed edges between them to represent the rating the user gives.

- Our purpose was applying differential privacy to our recommendation system to guarantee the privacy of individual user ratings.

Graph Database

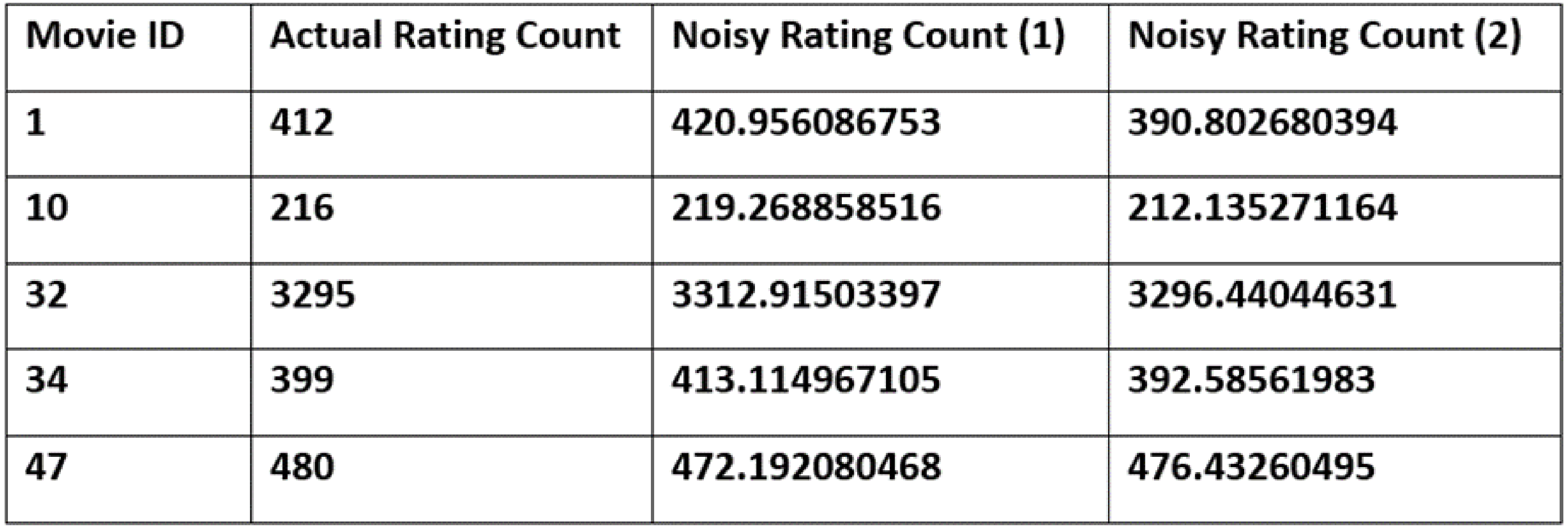
- Here is a snapshot from shrunk version of our database to give insight about the representation of the data.

- Blue nodes are anonymized users, green nodes are movies and the edges are rating relations between users and movies.



Example Results

- Query: *What is the number of ratings given to a movie? (movie name as parameter)*



Movie ID	Actual Rating Count	Noisy Rating Count (1)	Noisy Rating Count (2)
1	412	420.956086753	390.802680394
10	216	219.268858516	212.135271164
32	3295	3312.91503397	3296.44044631
34	399	413.114967105	392.58561983
47	480	472.192080468	476.43260495

References

- Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science: Vol. 9: No. 3–4*, 211-407.
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi, & T. Rabin, *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876*. Springer, Berlin, Heidelberg.
- *Elasticsearch*. (n.d.). Retrieved from http://db-engines.com/en/ranking/search+engine
- Ghosh, A., Roughgarden, T., & Sundararajan, M. (2012). Universally Utility-maximizing Privacy Mechanisms. *SIAM Journal on Computing, 41*(6), 1673-1693.
- Greenberg, A. (2017, March 23). *Apple's 'Differential Privacy' Is About Collecting Your Data*. Retrieved from https://www.wired.com/2016/06/apples-differential-privacy-collecting-data
- Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2016, September 9). Privacy-Preserving Learning Analytics: Challenges and Techniques. *IEEE Transactions on Learning Technologies*, pp. 68-81.
- Inan, A., Gursoy, M. E., Esmerdag, E., & Saygin, Y. (2016). Graph-based modelling of query sets for differential privacy. *Proceedings of the 28th International Conference on Scientific and Statistical Database Management* (p. 3). ACM.
- *Item-item collaborative filtering*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Item-item_collaborative_filteringKosan, M., & Benlioglu, M. M. (2017). *Differential Privacy Framework for Movie Recommendation System (Progress Report II).*
- Kosan, M., & Benlioglu, M. M. (2017). *Differential Privacy Framework for Text Analytics (Progress Report-II).*
- Kosan, M., & Benlioglu, M. M. (2017). *Differential Privacy Framework for Text Analytics (Final Report).*
- *MovieLens*. (n.d.). Retrieved from https://grouplens.org/datasets/movielens/
- Norton, R. M. (1984). The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Satisfaction, 38*(2), 135-136.
- Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web* (pp. 285-296). ACM. Xu, J., Zhang, Z., Xiao, X., Yang, Y., & Yu, G. (2013). Differentially private histogram publication. *The VLDB Journal*, 797-822.