POPULARITY BIAS IN SEQUENTIAL RECOMMENDATION

COMPARISON AND EVALUATION OF USER PRIVACY RISK IN RECOMMENDATION SYSTEMS

An Undergraduate Thesis Portfolio
Presented to the Faculty of the
School of Engineering and Applied Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Shivaen Ramshetty

May 6, 2021

SOCIOTECHNICAL SYNTHESIS

The pervasiveness of recommendation systems within online platforms has created broad privacy risk for users that access and utilize these applications as well as the interest in improving their experience. In the technical work, our focus was creating manners in which models are less likely to be biased based on the data's popularity. In doing so, users are given better recommendations that are more accurate to them as an individual, without the need for more data collection. In regard to the STS work, the research highlighted issues of privacy within recommendation systems and the importance of finding ways to mitigate user risk. Furthermore, the STS paper also expatiated on the fact that recommendation systems are not the only such online system to pose such a threat; therefore, industry wide solutions/procedures must be implemented. While the technical work improves user experience through better recommendations without additional data, the STS work stresses and introduces ways to lessen the risk of users sharing their data to these systems.

Recommendation models have grown in both complexity and scope in the recent past due to their growing demand and the accompanying research. An important component of such models is their accuracy and performance to users, otherwise they are ineffective and costly to the employers of the technology. Hence, our research and implementation of better models that meet the needs of more users than those interested solely on popular items, means better user experience as well as a more diverse set of recommendations. By using the same data available across various platforms, we were able to show that simply reducing the bias within the models can greatly improve the system. Finally, users were not placed under greater risk to receive better recommendations, meaning that privacy was not compromised in place of performance.

Through the technical research we were able to recognize popularity bias within systems and evaluate the extent to which the system was biased. First, we extrapolated published methods for popularity bias analysis in other machine learning applications to sequential models and visualized the results. We found that bias was also worse during testing, the time where a smaller subset of the data is used to validate the model's performance. Then, we developed a way of evaluating popularity bias across time periods, which allows for sequential models to utilize previous metrics while accounting for their unique design. In summary, our technical work was able to extend prior work to a new application, which will allow future research to develop models that are less biased and more suitable for a wide-range of users.

User privacy has become a large concern within the online ecosystem; however, it is unclear whether a select few systems can be blamed for the growing risk. In reality, most systems that comprise the online paradigm lack necessary security procedures or components. By analyzing the data collected and used across recommendation, notification, and advertising systems it became clear that the data was very similar across all three. Therefore, susceptibility to attack or failure of the system in securing data would apply to all three. Finally, using system design documents as well as reported issues within the industry, we also found that a majority of systems, including the three mentioned before, do not have a well-defined data security protocol or practice in place.

In order to illuminate the needs of the industry in regards to data privacy, we first had to examine the data used. In recommendation, notification, and advertising systems, data such as clicks or actions taken are used to identify groups of users. Then, using Mesthene's technical complexity model, we were able to visualize the effects of the industry focusing on model design rather than privacy protection. The aforementioned lack of data security in the design of the

system was seen through the emphasis of product-user experience in the system architectures. That is, the components of the system maintaining and utilizing the data were stagnant, while models and methods of communicating to the user were variable. Accordingly, many online systems provide great products without considering the effects of their data consumption and storage.

Alongside the development of recommendation systems is the increasing supply of data and ways in which that data can be abused. Users must be willing to trust systems such as recommendations in keeping their information secure in order to continue using them. Therefore, it is the responsibility of various agencies and the industry itself to improve today's data practices to better meet consumer expectations.

TABLE OF CONTENTS