Statistical Analysis of Network Data with applications in Marketing
Project proposal

Benjamin DONNOT
Thibault LAUGEL

November 2014

Recommendation engines are commonly used by online marketplaces such as Amazon or streaming platforms such as NetFlix to help their customers find the right product and thus highly increase the company revenue. However, with the recent progress in "Big Data" and statistical models conception, ethical issues concerning the respect of the most basic privacy have arisen. A few examples are :

- In 2010, NetFlix had to cancel their "Recommendation Engine contest" because some of the partipants had been able to identify precisely some NetFlix users thanks to their movies ratings. *(NetFlix Cancels Recommendation Contest After Privacy Lawsuit*[1])
- In 2012, Target, a discount store brand, figured out a teenage girl was 1 month-pregnant thanks to the products she was buying, before her parents even knew about it (*How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did* [2])

For most of these problems, the solutions found were unsatisfying : either forcing the model to make mistakes (Target), or start over (NetFlix). However, some companies have found ways to create privacy-respecting models that are today being used. Most of them rely on data aggregation to find similarities in the different behaviors of their customers and model them as a group, hence the usage of clustering methods.

In this project, we will focus on the first of these examples, the NetFlix prise. Following the contest in 2010, two Data Scientists from Miscrosoft, Frank McSherry and Ilya Mironov, published a paper explaining the construction of a recommender engine algorithm for NetFlix respecting the privacy of its customers[3]. This algorithm relies on a recent definition of privacy, the differential privacy, based on the fact that for any outcome of a randomized computation of the model, that outcome should be nearly equally likely with and without any one record. Using this definition, McSherry and Mironov have showned that introducing a noise parameter in the computation of the average ratings and the covariance matrix of the graph is enough to guarantee privacy. Our goal is to implement this algorithm and show that it respects privacy while giving good predictions.

[1] available at `http://www.wired.com/2010/03/netflix-cancels-contest/`
[2] available at `http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregna`
[3] available at: `http://research.microsoft.com/pubs/80511/NetflixPrivacy.pdf`

Using the NetFlix Prise data, For each movies the following information is available:

- Its title
- All its reviews, each containing :
  - The (unique) ID of the reviewer
  - The rate (from 1 to 5)
  - The date

In the whole data set, 17 770 are rated from 480 189 user, making 100 480 507 different reviews. We will certainly focus on a subset of these, considering only the reviews before December, 31st 2014 for example.

To build this model, we will first start by measuring the average ratings for each user and object, introducing a noise parameter in the computation. Using these measurements, we will then build a covariance matrix of the user rating vectors, introducing more noise. Given the covariance matrix and the other measurements, we will finally apply the k-NN method, followed by the SVD mechanism, to get our final results. The accuracy of the model will be measured through the RMSE.

Showing that it is indeed possible get accurate results with a recommender engine without being too intrusive is extremely important today, as people are starting to worry more and more about the respect of their privacy and that in the mean time, recommender engines have become a must-have for every commercial website.