CS670 Project Proposal
Online Lockstep Behavior Detection

Team Name: **End of Fraud**

Team Members (alphabetically ordered):
Majid Alfifi, Parisa Kaghazgaran, Xing Zhao

March 31, 2017

1 Introduction

How can we detect if a politician has purchased fake followers on Twitter or if a product's reviews on Amazon are not genuine?

A common method has been to represent *users* and *items* as a matrix where in the simplest case a cell can take on a binary value of 1 if there is a relationship between the corresponding user and item or 0 otherwise. The problem can then be transformed to finding dense regions in this matrix [1]. Moreover, this method has been lately extended from matrix to tensor representation to incorporate more dimensions from the domain such as timestamp, Twitter followers count, or number of stars of an Amazon product along with a scalable MapReduce-based implementation **D-Cube** [2]. Extraordinary dense blocks in the tensor correspond to groups of users with lockstep behaviors both in the products they review and along the additional dimensions (for example, multiple users reviewing the same products at the exact same time).

2 Project Goals

We intend to use an existing MapReduce-based implementation of the D-Cube algorithm[1] (Section 3.3) to our own datasets and explore the effectiveness of different dimensions in detecting fraudulent behavior hopefully informing our own research. In particular we will apply the algorithm to the following datasets:

- Twitter dataset: Can we identify tweets promoting/undermining certain hashtags? How do those suspicious tweets temporally differ from ordinary tweets (timestamp dimension)? etc.

[1] https://github.com/kijungs/dcube

- Amazon dataset: Can we find the hidden relation among reviewers in Amazon who aim to promote a product by writing fake reviews using additional dimensions such as temporal, rating and so on.

- Yelp dataset: ...

Feature engineering: We will extract set of dimensions useful in detecting the lockstep behavior for each of the datasets.

The paper [2] defines several density measures appropriate for anomaly detection (Section 2.2). We aim to build on those functions and propose a flexible density measure which gives different weights to different features (e.g., In Amazon temporal features are more informative than rating).

3 Tools/Resources

We will make use of the following resources:

- A local Hadoop cluster for running the algorithm
- Large dataset of Twitter data
- Amazon reviews
- Yelp dataset?

4 Project Outcome

- A set of users/items potentially participating in a lockstep behavior for each of the datasets under study.
- Evaluation of different dimensions.
- Evaluation of different density measures.

References

[1] Hooi, Bryan, et al. "Fraudar: Bounding graph fraud in the face of camouflage." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[2] Shin, Kijung, et al. "D-cube: Dense-block detection in terabyte-scale tensors." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.