Systematic detection of fraudulent account registration

Yun Zhang, Long Sun, Mia Mao and Lin Tao
Uber Technologies, Inc.
{yunzhang, longs, mia.mao, lin.tao}@uber.com

With the fast expansion in the global markets, internet companies become the targets of various financial crimes, including those migrating from older technologies and techniques. In order to provide a safe and user friendly platform for the clients, it requires blocking fraudulent accounts at the earliest stage. However, early actions on fraudulent account registration are not always feasible due to information sparsity and technical difficulties. Today's fraudsters try to take advantage of that by creating fraudulent accounts that hibernate to build account age before being used to attempt fraud. To combat this, it is important to identify fraudulent account registration at a large scale from the earliest stage. In this paper, we present a two-dimensional ML system that can not only detect the fraudulent account during its signup period, but also remove those hibernated fraudulent account clusters in a scalable and efficient way.

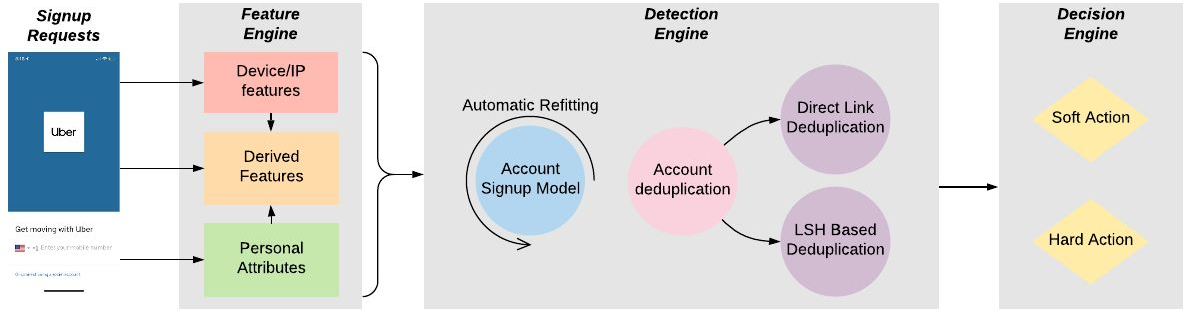

Figure 1. The detection system of fraudulent account registration on the Uber platform.

The proposed system consists of three components - feature engine, detection engine and decision engine as shown in Fig.1. The feature engine collects customer input personal attributes and device/IP information, and computes various derived features including email/name gibberish scores and historical statistics. These features will be transferred to the ML based detection engine that provides a two-dimensional protection through an account signup model and a large-scale account deduplication process. The account signup model is a tree-based gradient boosting classifier identifying single fraudulent account in a real time fashion, while it will automatically refreshed to boost the model performance. The account deduplication process aims to identify duplicate signups by comparing account pairs. Since the brute force algorithm (N^2) for pairwise comparisons is impossible due to the scalability issue, a new locality sensitive hashing (LSH) based method is utilized to enable account deduplication at scale. The LSH will hash the accounts into buckets so that similar accounts are located in the same buckets with high probability. The system will only compare accounts that share direct links (phone number, device, etc.) or at least one LSH bucket (Fig. 2) which will tremendously reduce the computational costs. At last, the final decision engine can take advantage of the signals from the detection engine and react accordingly.

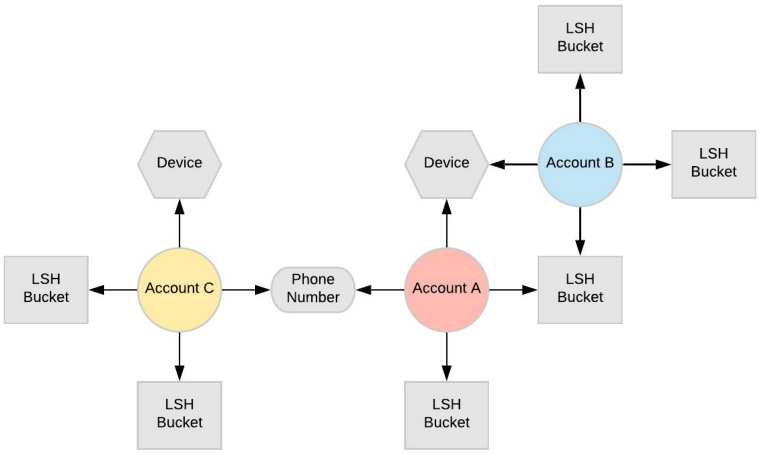

Figure 2. An example of the account deduplication through direct links and LSH buckets.

To summarize, a scalable and powerful detection system has been proposed and implemented, which successfully identifies fraudulent account registration on the Uber platform. This system stands out by considering both single fraudulent account detection and account deduplication and is generic enough to be applied on any other platforms.