Call for Talks: Joint PayPal/BGU Seminar

On March 21 the Members of PayPal's Global Risk Data Sciences Club will be holding a joint seminar with BGU researchers working related areas. The goal of the seminar is to strengthen the relationships between PayPal and BGU following similar collaboration with Stanford University, University of California at Berkeley, University of California at Santa Cruz, Fudan University at Shanghai. PayPal would like to use this opportunity to identify potential research project that PayPal and BGU can collaborate on in the near future.

The seminar will consist of three talks by PayPal researchers and a number of talks by BGU researchers. If you are interested in given a talk about your research area, please send a short abstract to DL-PayPal-BGU-Seminar-Abstracts@paypal.com

The PayPal team consists of about 10 core members – all with advanced degrees and expertise in data science, machine learning and platform engineering. Our Team provides advanced feature engineering solutions and builds models using advanced machine learning algorithms such as Gradient Boosting, Deep Learning, and Natural Language Processing. We employ big data platforms such as Hadoop and Spark. The primary domain we are focused on is payment fraud prevention. We are also taking responsibilities in other domains such as consumer & merchant marketing and credit.

Talk Abstracts:

Model implementation and online computing service (real-time machine learning predictions)

Speaker: Tu Fangbo, Shanghai

The anti-fraud system from Paypal can make a rapid decision on signup, login, checkout, withdrawal and other risk checkpoints to save Paypal's loss. One of the major components of the system is online computing service, which based on modelling technology and big data processing- coupled with offline streaming and batch system – provide risk analytic evaluation for decision engine, also known as real-time machine learning predictions. I will introduce our online part of the system which hosting hundreds of models and walkthrough our model implementation and deploy process. There are many challenges to build an online computing service. It is hard to host hundreds of models which have different structure and algorithms and also keep the same SLA. Data changing, latency and new feature adoption make it difficult to maintain. And one research interest from our customer, how to support time-series clustering and classification. Even more, how to adopt real-time learning in risk detection area.

Searching for causality

Speaker: TBD, TLV

During the past decade, PayPal has tailored a transaction pipeline for risk assessments which contains tree main components or business units. In this talk we will walk through the business units and point challenges we are facing today at the first unit, known as Core. Core is responsible for feature engineering. We have generated tens of thousands features up till now and continuing to do so. These features are high level features build

on top of low level transaction elements (such as IP, sender country etc.). Each of these high lever features is engineered by small to medium scale machine learning techniques. The processes in Core incorporates intuition and our understanding of the world of fraud where a significant part of it is manual. The manual part is reasoning with the data and selecting the most relevant features within enormous features pool. Recently, we have started exploring causal feature selection methods such as Markov Blankets, though we are still facing different challenges
such as scale and sparsity.

Clustering very large graphs with constraints

Speaker: Ohad Raviv, TLV

Using our diverse data sources, we would like to detect real world entities, such as actual user, related users and workplaces. Performing such “Entity Resolution” on graph data is commonly referred to as community detection.

Many community detection algorithms have been proposed, but incorporating prior knowledge into the detection process remains an open and challenging problem. This is essentially a semi-supervised clustering problem, where the labels come from manual tagging using external domain expertise. There are several approaches to this problem, which we will discuss in this talk and consider future steps.