Machine Learning Based, Fraud Detection Algorithms

Brandon Sim, **R**yan Lee, **W**esley Chen (**BRW)**

Advisor: Pavlos Protopapas

Collaborator: Bank of America

Goal/Problem: To use machine learning algorithms to detect anomalous transactions and entities given data set with meta data and transaction list as a time series

To analyze and quantify various approaches for outlier detection in the given data set

Roadmap and Potential TImeline (tentative):

1. Problem Definition and Refinement: 9/25
2. Literature Search (while waiting for data) (10/2)
3. Data Exploration (10/9)
4. Cycles 1 to N
    a. Propose Methodology (10/16)
    b. Implementation
    c. Update Meeting with BoA

Data Set: Initial data provided by Bank of America, order of 10K users.

Possibility for scale-up

Literature Search:

- Nun/Protopapas: http://iopscience.iop.org.ezp-prod1.hul.harvard.edu/0004-637X/793/1/23/pdf/0004-637X_793_1_23.pdf
- Papadimitriou:
- Unsupervised Anomaly Detection: http://link.springer.com/chapter/10.1007/978-1-4615-0953-0_4

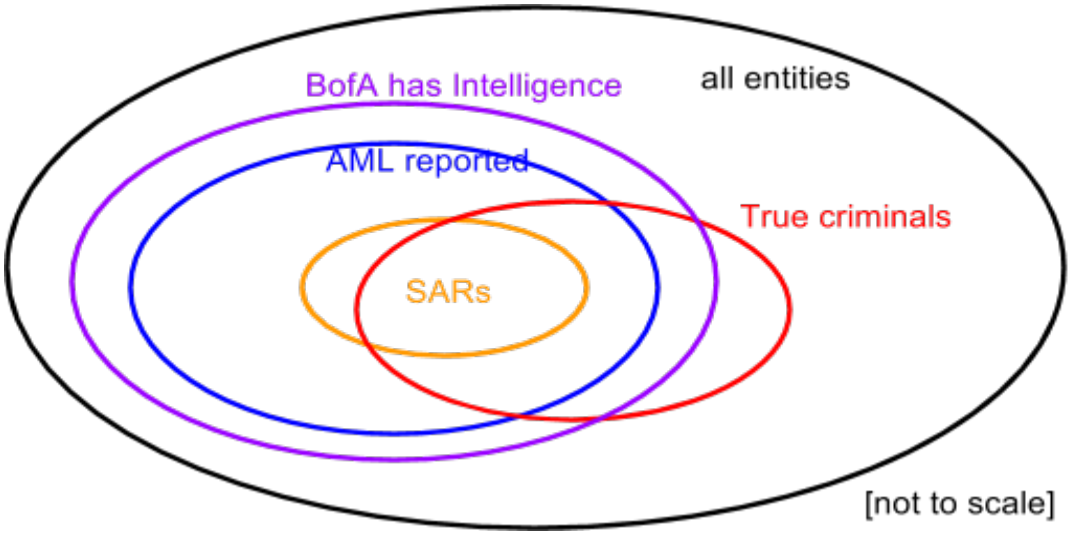
Nomenclature: [ADD]

- **SAR:**

Notes / Questions for BoA: *as of Sep 25th*

- What is "intelligence" – purely internal or combined with external (non-BoA caught by others)?
    - Does the SAR include reports on those not reported by BoA but by others?
- For the first project, how do we know where the current algorithm fails if there is no SAR to ever confirm uncaught criminals (type II errors)
    - Or is this project just to minimize type I error (detecting a false positive)
- Can data format be shown?
    - How does it break down transactions - by categorical payments?
    - Which variables are included? Time? Date? Location? Amount? Others?
    - Any missing or inconsistent entries? (entry errors? incomplete data?)

- Is the diagram below representing the data valid?



Challenges and Possible Solutions:

- No ground truth
    - SAR report is supposed subset of those already detected
    - SAR report is noisy within itself - perhaps just assume SAR report are 100% correct
    - Unknown type II error space - some criminals are never caught
- Nominal Labels
    - Consider using indicator variables at the cost of high-dimensionality
    - Scale qualitative labels on a scale (project to quantifiable space)
        - ex: profession - > salary
- Data (potential challenges as at this point unseen)
    - Detail of data set
    - Consistency of data set
    - Any missing data per given entity
    - Any missing entities?

Strategies:

Semi-Supervised Approaches
Feature Space Boundaries

Unsupervised Approaches
Feature Space Density
    Cluster-Based
    K-Nearest Neighbor
SVM

Time series analysis