10/2

Went over random forest code, LOCI code and Eskin paper
Get another index to toss in with LOCI

TODO
B:
Literature Search find papers, put in google drive
Update Brandon summary

R:
Bayesian Network implementation with random forest stuffs

W:
Implement Eskin paper approach
DARPA data

9/25

NDA was not accepted
Return offer to see if other things can be made

Potential for diff data set
Malware
Astro

Papadimitriou
New Metric Multigranularity deviation factor
alpha computes quickly
approximate version in linear time
Metric eval >3SD of outlier averaged over all metric calcs for the set
No arbitrary cutoff

alpha r is what matters

TODO
Wesley: Read another paper, write summary
Brandon: Explain how r is chosen, implement LOCI
Ryan: Explore random forests in scikitlearn, test on some dummy data

---

9/18

Problem 2 is favored by everyone for being more flexible, and more like what we had envisioned

List of Entities each with Features metadata and then transaction list

Detect abnormal entities relative to metadata
Find features in the list of transactions, sample (iid) and time series (non iid)
Detect abnormal transactions

Feature space (2 features as axis) find area of high density and find area of low density as outliers

No true positives

Challenges:
Missing data
Nominal classifiers (not quantified)
Incomplete ground truth
Labelling
Feature space density finding outlier (low density of points) areas in the feature space
Can also use label to reverse engineer when the classification is for when it's confused (requires complete boundaries which is not a density approach)
Be sure that classifications themselves can't select features

How to Handle Nominal Data?
Something to think about
0/1 indicator variables – would require many dimensions
Perhaps scale it in a reasonable way like average salaries for professions based on expectation

Active Learning Phase on Training Set

Reference Paper:
Nun/Protopapas
http://iopscience.iop.org.ezp-prod1.hul.harvard.edu/0004-637X/793/1/23/pdf/0004-637X_793_1_23.pdf
under related work chapter 2 – outlier detection in machine learning

12-14 week Roadmap:
Problem Definition
Literature Search (skipping data due to access lag)
Data Exploration
Propose Methodology

Implementation
Cycle 1
Second Propose Methodology
Second Implementation
Cycle 2
Rinse and Repeat

Assignments:
Read Nun/Protopapas and 2 others, no need to implement Pavlos'
Google Doc Update

List of Questions Summarized at the End:
What is “intelligence” – purely internal or external?
Is the data set for the first project a subset of the second project?
For the first project, how do we catch where your algorithm misses if we never have an SAR generated for those cases?