Project proposal

1. What data set are you planning to use and why it is interesting to you?

- I am planning on using this fraud detection dataset. This particular dataset is interesting to me because of the opportunity it presents to use a lot of the skills I learned this and last semester, and apply them in to classify fraud. I think this is important because fraud comes in many different ways, and I'd like to observe specific trends with it and be able to predict if a transaction is fraudulent, and be able to construct a relational graph as a result. This dataset is also very well put together, with minimal cleaning necessary, and provides a large amount of different features.

2. What is the problem you are solving or the question you are asking? Show that you have thought about it and share your insights.

- What specific features are key indicators of fraud? Do specific threat actors/fraudsters typically repeat or is it a one time offense? Can you construct a graph linking transactions to specific people/groups based on their actions?
- I want to be able to train a basic model and find the most important features necessary to classify fraud. This application is important because financial has a significant impact on both small businesses and large corporations alike, and being able to classify fraud will allow them to stop it, saving time and money.
- I think that there is a definite link in the fraudulent accounts, and being able to spot fraud (whether account id, location, device, etc) will be a good step forward.
- Fraud is typically collaborative, and being able to graph the data will allow me to detect patterns such asfraud rings, shared devices or collusion. This behavior could show up as high, centrally, tight knit communities and outlier structures (unusual clustering metrics)
- I can use BFS/Dikstra's clustering and centrality to find this out.

3. What are the steps/components needed to accomplish the project? Specify the milestones with approximate dates you will accomplish them and how you plan to test the individual components.

- 4/6: I want to be able to use a combination of classification techniques as well as machine learning (sklearn packages ported via Py03 or specific random forest/test_train_split crates).
- 4/13: Build BFS
- 4/13: Build Centrality function
- 4/20: Build Clustering function
- 4/20: Implement SKlearn - like packages for ML implementation.
- 5/1: Figure out how to graph data
- 5/1: Build tests
- 5/1: Finalize code, add readme, push all work to github.