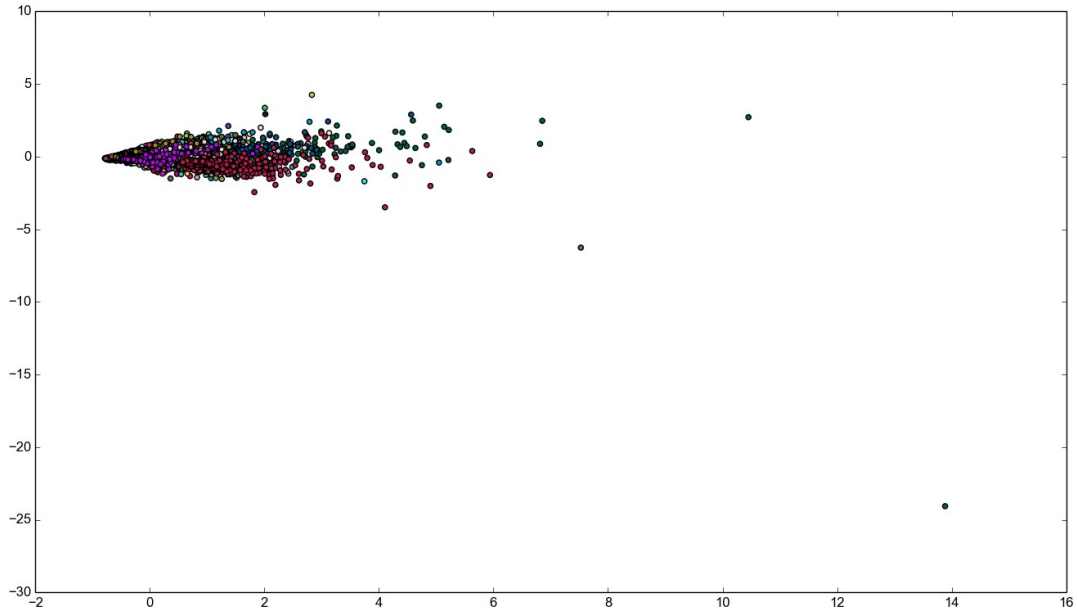
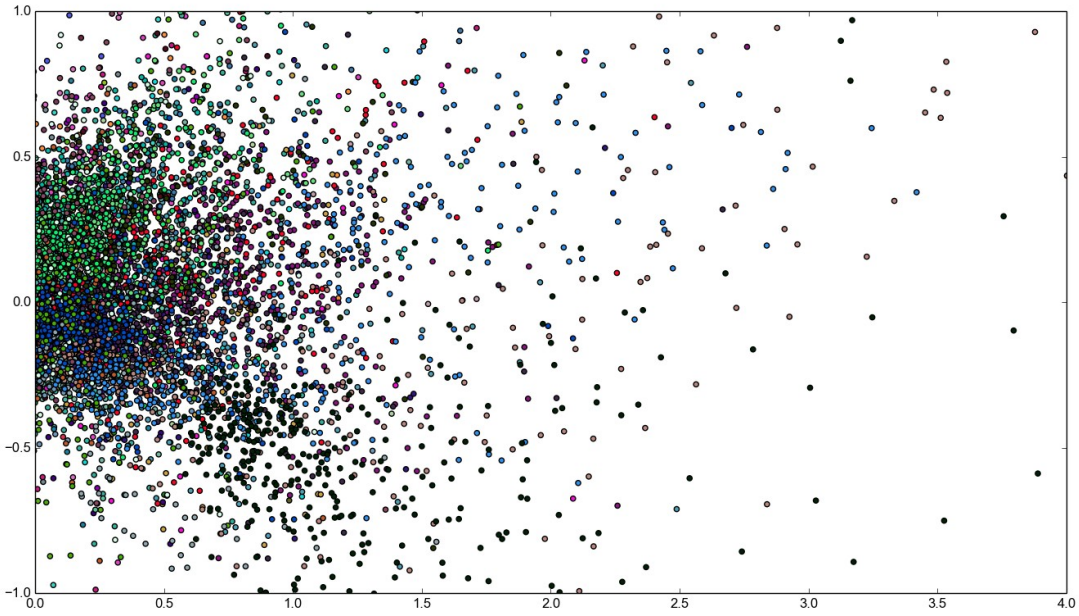
Problem 3 discussion

We initially decided to just run the problem data through a dimensionality reduction to visualize the data in a comprehensible way. We used PCA and took the top 2 dimensions for this.



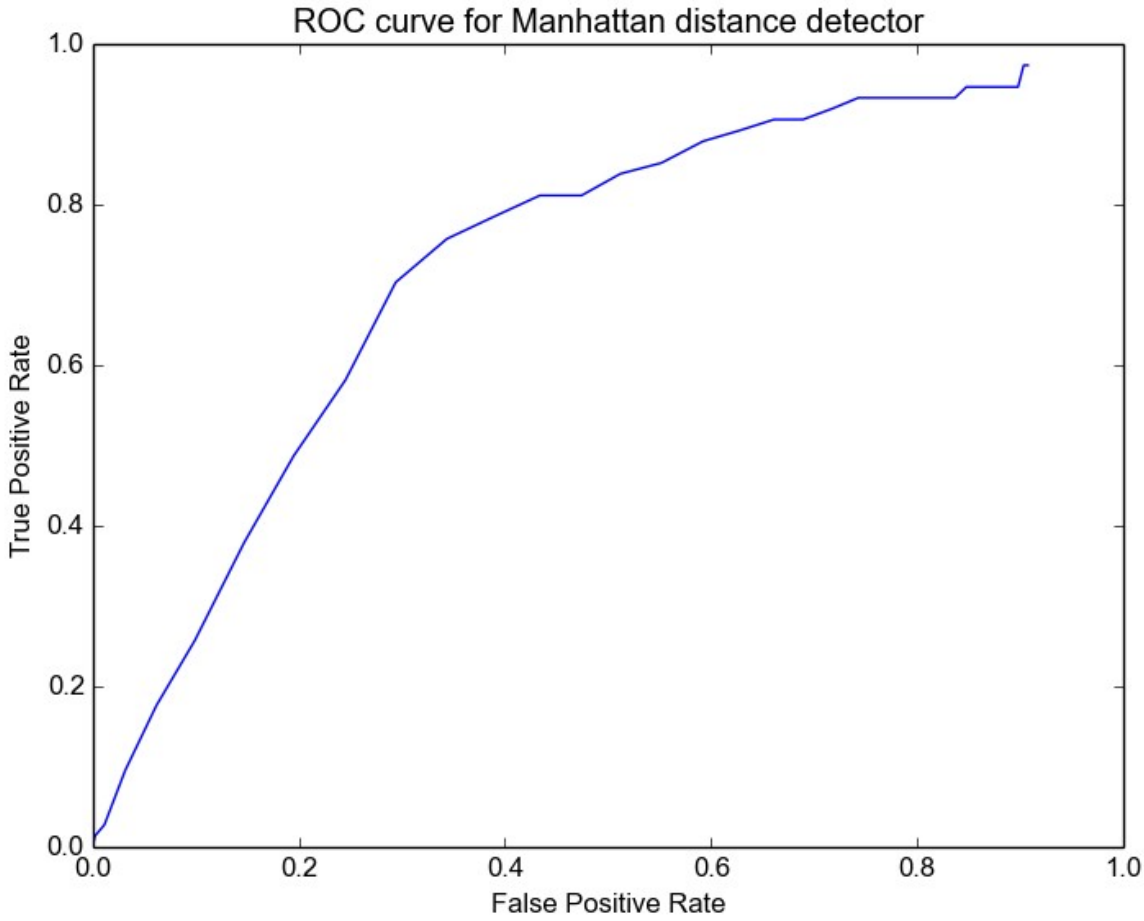
Scaled to remove outliers:



We can see that even in this reduced dimensional space it is possible to make out a few clusters. However, most of the clusters seem to overlap with each other.

Our initial approach was to take a relatively simple detector and train it with all the feature vectors to see how it would perform. For this we chose Manhattan distance with the mean and variance of the 1-norm of the training data used to determine the threshold value for each user. We noticed that we could reduce the size of our feature vectors by removing one of three features for each key (either Hold, Down-down, or Up-down), since any third feature can be determined by the other two. However, we chose not to remove redundant features as they would not actually affect the outcome of the classifier, since it would scale every vector and the threshold values by a constant factor.
Due to the observed overlap between the clusters, we didn't expect this to perform very well.

We see that our classifier has a consistently high true positive rate, but a moderately high false positive rate as well (where we define a true positive to be a user labeled as authentic when the user is indeed authentic, and a false positive to be a masquerader labeled as authentic). For example, the classifier would achieve a 88% true positive at roughly a 50% false positive rate.
The results were reported by dividing the training set into a 80:20 ratio to get a validation set and the results are based on the experiments performed on the validation set.



As a result we tried to use a detector that looked at the nearest neighbors based on the manhattan distance (since this was also observed to perform the best by the CMU paper).

As in the 1st case, we divided the training set into a training set and validation set. The experiment performed was effectively a nearest neighbor based multi-class classification on the validation set.
The algorithm reported an accuracy of 80.89% for multi-class classification on the validation set when compared against the sequences in the new training set.
This agrees with our observation from the pca analysis that the data points for a single user seem to be clustered together, however, the clusters might not have a well defined 'cuboidal' shape as inferred from the failure of the simple manhattan distance based detector.