Kayla Straub Thesis Outline

1. Introduction

   (a) Build up context for the report. Cite email statistics.

   (b) Describe the problem of email analysis and challenges associated with it (privacy concerns, accurate labels, etc.).

   (c) Briefly cover the process/solution I have implemented (data collection, developed features, machine learning algorithm).

   (d) My contributions: new dataset from raw material, combining features not used before, and the job-classification problem in general

   (e) This project is important because it shows the amount of information that can be gleaned from a single form of communication.

   (f) Touch on potential applications of this work.

2. Literature Review - very similar to paper

   (a) Show other applications of email analysis (mostly using Enron).

   (b) Cover the graph theory fundamental papers.

   (c) Discuss the most similar research studies and highlight how few have been performed as well as their lack of quantifiable results.

3. Data Collection - very similar to paper, but with more detail

(a) Section: Enron issues

Discuss the need for a modern dataset with known labels. Intimate knowledge of the environment that produced our data is a big advantage.

(b) Section: Data collection process

Describe the steps taken to collect, anonymize, and store the data properly.

(c) Section: Dataset statistics

Provide detailed descriptions/statistics of the dataset. Use Table 1 from the paper here.

4. Feature Analysis

Go into depth explaining all features used. Cover all math behind all metrics. This will all be very similar to the paper but in more detail.

(a) Section: Graph-based features

Use Figure 1 from the paper here.

(b) Section: Social-based features

Use Figures 2 and 3 from the paper here.

(c) Section: Feature selection (using mutual information anlaysis)

5. Algorithm Design

Again, this section will be very similar to the paper but with more detail.

(a) Discuss model selection. Use Figure 4 from the paper here. Go into the math behind how random forests work and what they are used for. Cover their advantages

and disadvantages.

(b) Describe and justify all choices made (max depth, splitting metric, etc. - selected via cross-validation). Use Table 2 from the paper here.

6. Testing and Results

Talk about all evaluations. What did and did not work. Describe results and what they mean.

(a) Section: Classification results - very similar to paper

Cover the classification results from the paper. Use Figure 5 from the paper here. Add that LOOCV performed worse, with only about 61.194% correct. Make a similar graph to Figure 5 for this result. Report the results of the feature analysis, and use the graph from Figure 6 in the paper here.

(b) Section: Demonstrate that behavior is constant in time

Compare email statistics for a person over time. Measure distance from the centroid for each person. Compare these statistics between classes. Make a chart comparing these distances.

(c) Section: Hierarchy analysis

Compared which directors and PM's employees interacted with the most, and compared with the official hierarchy. Use the statistics from the paper here.

7. Conclusions

(a) Summarize the results. We successfully categorized people using part of their

email accounts. This worked for first- and second-ring employees. We learned and proved that over time a person's email behavior is constant. We uncovered some of the official hierarchy, but were incorrect on other parts.

(b) Discuss the assumptions inherent to this analysis and potential sources of bias. For example, the fact that we are training and testing on the same people. For the basic classification analysis, we assumed that behavior was constant in time.

(c) Discuss the inferences that can be drawn from this work. Official hierarchies do not always reflect the day-to-day office relationships. Behavior patterns can be inferred from second-hand data.

(d) Explain applications: corporate espionage might give access to 1 email account - what could we uncover? These methods are not limited to email, and can be applied to other communication systems (call records, text message metadata, social networks, etc.)

8. Future Work

(a) Describe what we could do if we had a bigger dataset. How big would that dataset need to be?

(b) Discuss applying deep learning to this problem and the math behind how that would work.