

Kayla Straub Thesis Outline

1. Introduction

Build up context for the report. Describe the problem and challenges associated with it. Why is this important? Briefly cover the process/solution I have implemented (data collection, developed features, machine learning algorithm). Touch on potential applications of this work.

2. Literature Review

Cover background works in similar studies. Show other applications of email analysis (mostly using Enron). Also cover the graph theory fundamental papers. Finally, discuss the most similar research studies and highlight how few have been performed as well as their lack of quantifiable results.

3. Data Collection

Section: Enron issues

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

Section: Data collection process

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

Section: Dataset statistics

Provide detailed descriptions/statistics of the dataset.

4. Feature Analysis

Go into depth explaining all features used. Cover all math behind graph-based metrics.

Section: Graph-based features

Section: Social-based features

Section: Feature selection (using mutual information analysis)

5. Algorithm Design

Discuss model selection. Go into the math behind algorithms used (random forest for now). Describe and justify all choices made (max depth, splitting metric, etc. - selected via cross-validation)

6. Implementation

Cover the actual process by which everything was calculated and implemented. Talk about the software packages used and any special techniques.

7. Testing and Results

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

Section: Demonstrate that behavior is constant in time

Section: Classification results

Section: Hierarchy analysis

8. Conclusions

Summarize the results. Discuss the inferences that can be drawn from this work.

Section: Applications Corporate espionage and alternative communication networks

(call records, text message metadata, social networks, etc.).