1 Introduction

This project will explore the realm of personal data aggregation, analysis, and presentation. Specifically, I plan to develop a tool suite for mining iPhone SMS data and presenting that data in meaningful ways.

2 Background

The theme behind this project builds off of two larger industry trends. The first is the rise of data visualization. The past few years have seen an explosion in this realm, with sites like Visual.ly (a marketplace for info-graphics) and tools like D3 (for generating complex visualizations in JavaScript) gaining wide popularity. The second trend is personal data. Facebook is one of the best (and most discussed) sources of personal data, but there are other rich services emerging as well. Consider the success of FitBit and other “health” apps at tracking your weight loss, Mint at tracking your finances, and WakeMate at tracking your sleep cycles.

Key takeaway: We love data visualizations and we especially love data visualizations about ourselves.

3 Data Sources and Inspiration

3.1 Nicholas Feltron

Nicholas Feltron (now a designer at Facebook) has become a brand name in personal analytics over the past few years through the creation of his annual reports. Feltron’s reports are famous not just for their impeccable sense of design, but also for the uniquely personal approach they take to data aggregation, analysis, and reflection. If the data stream analysis from this project reveals aggregate information that is interesting or otherwise valuable, project deliverables will include a visualization mechanism in this style.
3.2 Wolfram Alpha / Facebook Graph

In 2012 Wolfram Alpha released a data analysis and visualization tool for the express purpose of exploring one’s own Facebook data. While many of the generated charts are gimmicky (neither computationally impressive nor personally revealing), there are nevertheless some very interesting visualizations and unique insights captured. A small sampling of the data presented is shown below.

3.3 Twitter

A myriad of services exist to analyze your twitter history, activity, and profile. These range from the most simple, such as favstar.fm which aggregates your popular posts over time, to the more ambitious such as tweetpsych.com which purports to psychoanalyze your tweets.

3.4 Gmail

There are also several outlets for analyzing your Gmail (or generic email) usage. Gmail itself releases monthly “insight” summaries, although the presented information is relatively limited in scope. There are also more complex services such as MailTrends which can download your message history and then present a more complete picture of your usage over time.
Figure 2: WA - A generated word-cloud from Facebook posts

Figure 3: WA - Visualization of one’s extended friend network, with social clustering
Figure 4: WA - Generated insights from graph analysis over the friend network

Figure 5: WA - Posts graphed by time of day and type of post
To date, no service offers a similarly popular data analysis tool for SMS data. Which makes this data source especially exciting to explore in my senior project. Although SMS data will present only 3 “columns” (name, message, time) to work with, there is much second-degree analysis that can be layered on top. For instance, any of the Wolfram Alpha analytics shown above (excluding the graph visualizations) could be replicated with SMS data. Further, the structure of SMS data lends itself well to a more complex StreamGraph generation.

4 StreamGraph

Based on empirical results, project deliverables may include creation of StreamGraph for SMS analysis (i.e. if anecdotal testing reveals StreamGraph anal-
ysis over SMS has utility). StreamGraph, as pictured below, is a novel graphing style used to display aggregate trends over time, often in an especially intuitive or even emotional manner. The generation of these graphs, however, is quite technically challenging. If pursued, this work could add a good deal of complexity to the overall project.

Figure 8: StreamGraph of a Last.fm user’s history

Figure 9: One hurdle to implementing a StreamGraph is an intelligent ordering of the individual data layers
5 Formal Proposal

The main project deliverable will be an open source tool-suite for generating personal SMS analytics. At its core, this entails a basic python platform for extracting texts from an iPhone, cleaning them for storage, and exposing an api for “modules” to plug in to for analysis. These individual modules will then each output to a standardized design template, and the python platform will produce a summary PDF of the aggregated data. I would build the platform first, and then add in individual modules as time and interest allow.

Modules could be as simple as “word clouds” or histograms of when texts are sent throughout the day. But modules could also do more complex tasks, such as assign “labels” to SMS contacts based off of content and time analysis, or suggest insights as to certain relationships and their trends.

The StreamGraph could also be implemented, showing how relationships (as measured by texting) grow, fluctuate, and fade over time. Such a module would be the most difficult but potentially also the most rewarding to construct.

6 Deliverables

1. SMS extraction tool (for iPhone, maybe Google Voice as well)
2. Standardized report tool, with output to PDF
3. Several analysis modules, of ranging complexity
4. Python source code for all of the above
5. Sample generated report
6. Final paper with discussion and analysis