DATASCI W205: Storing and Retrieving Data – Final Paper

Document Overload, Finding a Needle in a Haystack

Jason Blauvelt, Geoffrey Link, Megan Pera, Guangzhi ‘Frank’ Xie

**Motivation**

The key motivation for this project is the financial crisis of 2008 in which regulators did not have enough information to understand the breadth and depth of the crisis and its effect on financial institutions and the economy as a whole.

Data obfuscation is a key tactic of organizations under investigation or regulation. For example, Goldman Sachs Group Inc. was subpoenaed by the Financial Crisis Inquiry Commission (FCIC) after the FCIC accused the firm of engaging in a document dump to hinder a probe. Goldman Sachs sent more than a billion pages of documents, not all of which is the information that the FCIC requested. In essence, transmitting such a large amount of information in a time constrained investigation is a deliberate attempt to run out the clock such that no tangible insights can be obtained from the information before the investigation ends[[1]](#footnote-1).

The analysis of large volumes of data is a near impossibility for organizations that face both time and budget constraints. Hiring hundreds of $200 per hour junior lawyers to shuffle through a billion pages of documents is an option only a few organizations can afford; moreover, this process is subject to human error and fatigue.[[2]](#footnote-2)

To sum up, we propose to create a proof-of-concept implementation of a solution to this issue utilizing the tools we learned in W205.

**Data Sources**

For the purposes of this project, we chose to use the publicly available Enron Email Data from the Federal Energy Regulatory Commission (FERC) available on Amazon Web Services.[[3]](#footnote-3) Enron’s complex financial statements were confusing to shareholders and analysts; off-balance-sheet vehicles, complex financing structures and deals were so confusing that few people could understand them.[[4]](#footnote-4) Enron’s auditor, Arthur Andersen, was found guilty of illegally destroying documents relevant to the SEC investigation. Similar to financial institutions after the financial crisis of 2008, Enron elected to disclose minimal details on its use of special purpose entities; example of which include: JEDI, Chewco, Whitewing, and LJM.

1. Bloomberg, Goldman Subpoenaed After FCIC Says Firm Slowed Probe, 2010. bloomberg.com/news/articles/2010-06-07/goldman-sachs-documents-subpoenaed-by-u-s-financial-crisis-investigators [↑](#footnote-ref-1)
2. CNN, The fall and rise of lawyers, 2015. cnn.com/2015/05/22/opinions/barton-rise-and-fall-of-lawyers/ [↑](#footnote-ref-2)
3. Amazon Web Services, Enron Email Data, aws.amazon.com/datasets/enron-email-data/ [↑](#footnote-ref-3)
4. Wikipedia, Enron scandal, en.wikipedia.org/wiki/Enron\_scandal [↑](#footnote-ref-4)