Introduction

Introduction

Goal

The goal of this paper is to illustrate a combination of machine learning, natural language processing and graph analysis techniques applied to corporate email to identify potential witnesses in litigation.

Background

> In times of political turmoil, events can move from impossible to inevitable without even passing through improbable. Anatole Kalesky

Enron Corp. and its affiliates were engaged in energy-related businesses, as described in its Annual Report on Form 10-K for the year ended December 31, 2000

```
*    the transportation of natural gas through pipelines to
markets throughout the United States;

*    the generation, transmission and distribution of
electricity to markets in the northwestern United States;

*    the marketing of natural gas, electricity and other
commodities and related risk management and finance services
worldwide;

*    the development, construction and operation of power
plants, pipelines and other energy related assets worldwide;

*    the delivery and management of energy commodities and
capabilities to end-use retail customers in the industrial
and commercial business sectors; and

*    the development of an intelligent network platform to
provide bandwidth management services and the delivery of
high bandwidth communication applications.
```

As of December 31, 2000, Enron employed approximately 20,600 persons.

For the year ended December 31, 2000, it had operating revenues of $100,789 million, according to the same report.

On December 2, 2001, Enron filed for bankruptcy protection.

In less than a year, Enron underwent a complete reversal of fortune as its business strategies ran afoul of applicable regulations, among which were those of the Federal Energy Regulatory Commission (**FERC**).

FERC became aware of irregularities in the California wholesale electricity market prices. An orientation to the issues is provided by testimony before FERC, which provides a concise summary.[1]

[1] The short version, which I can relate as a former California regulatory official from personal knowledge, is that public electric utilities were losing a large share of industrial customers to self-generation. Many businesses found it cheaper to generate on-site than to pay tariff rates. Foreseeably, residential and business customers without the option to self-generate would come to bear

Following Enron's bankruptcy, FERC began an intense investigation, including the email records of 149 Enron employees. A preliminary staff report issued six months later.

Motivating Data

FERC obtained approximately 500,000 emails. Copies of these were acquired by Leslie Kaelbling of MIT and published by William W. Cohen of Carnegie Mellon University. It is one of the largest publicly available datasets of corporate email and is referred to as the Enron Corpus. The term *corpus* is used in natural language processing to denote a collection of related text.

At the time, electronic record examination (*ediscovery*) in litigation was in a primitive state. It was not uncommon, for example, for paper copies of email to be offered. These would typically be read by teams of freelance attorneys looking for keywords. Advanced technology included scanning with optical character recognition and some proprietary software options to organize emails and capture the status of review.

Much of the focus was directed to keyword searches, sometimes called the *smoking gun* approach. Brute force examination misses opportunities to understand the social networks that reflect how the organization operates, what their concerns are and the haphazard exposure of document reviewers inevitably poses the Elephant and the Blind Men Problem. To triage the corpus quickly and efficiently, it should first be be distilled and analyzed.

Analysis

Data acquisition

I obtained a copy of the 2009 version of the corpus. It contains copies of emails of a private nature that involve three users have who since requested to be redacted. I have removed those 27 emails.[2]

Conversion

Each email was a plaintext file[3] Each user had a directory tree similar to the one below.[4]

Although tedious, traversing the directory tree, parsing the emails and loading them into an SQL database, was accomplished with a combination of Python and Perl scripting and standard bash tools. I do not reproduce that process here as it has little bearing on the main topic of this paper.

Data structure

While the emails were not in native format, the plain text versions contained nine principal segments, as shown in the figure below

The following were extracted from my 2010 analysis for this paper:

---

the entire cost of unamortized utility fixed assets (termed *stranded costs*), and rates for retail, commercial and small industrial customers would increase. The adopted solution was to require the utilities to sell their generation plants and buy power on a new public market on a *day-ahead*, tomorrow's estimated demand, and an *hour-ahead* basis for unanticipated demand. Although much thought was devoted to the dangers that participants would game the system to sell or buy at discounts from market, insufficient consideration was given to multi-participant cooperation.

[2]Most of my work on data wrangling and preliminary took place in 2010 in Python, relying heavily on the NLTK and networkx packages.

[3]Most had been generated by Microsoft Outlook, but some older emails were produced in IBM Notes, which created some character encoding issues.

[4]This user had 10 directories with 3048 files (the directory tree has been pruned to omit spurious detail) containing 12,147 lines and 69,226 words.

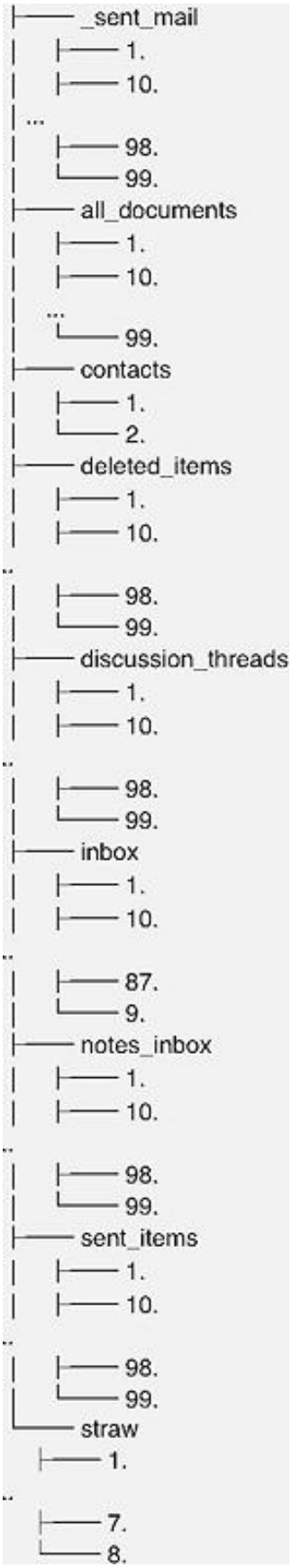

Figure 1: Typical user data tree

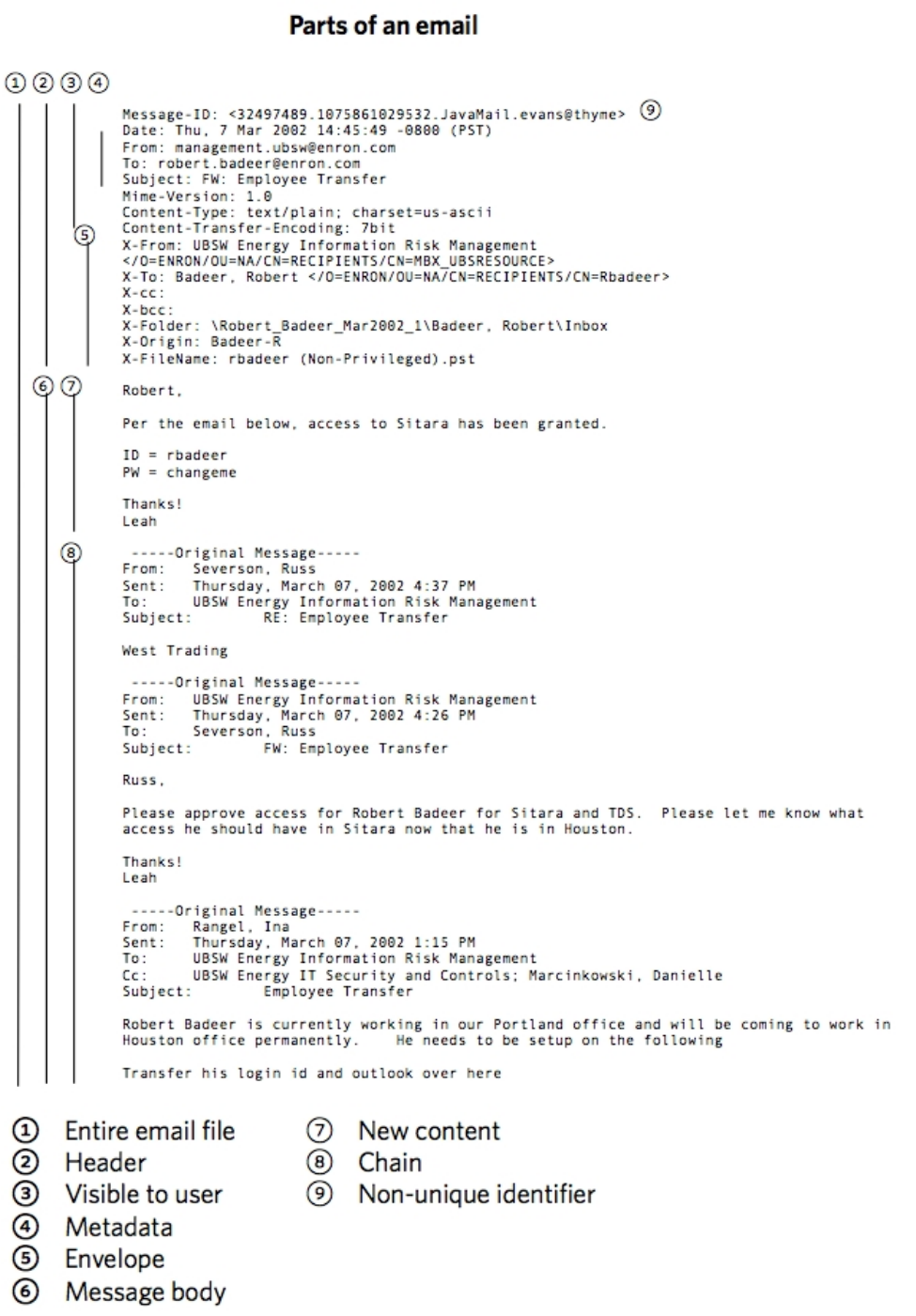

Figure 2: Structural analysis of an Enron email

Field	Type	Null	Key	Default	Extra
body lastword	mediumtext	YES	UNI	NULL	
hash sender	mediumtext	YES		NULL	
tos mid ccs	varchar(250)	YES		NULL	
date subj	varchar(250) text	YES		NULL	
tosctn ccsctn	varchar(250) text	YES		NULL	
source	datetime	YES		NULL	
	varchar(500)	YES		NULL	
	mediumint(9)	YES		NULL	
	mediumint(9)	YES		NULL	
	varchar(250)	YES		NULL	
		YES		NULL	
		YES		NULL	

- sender
- date
- subject
- recipient
- new content (*lastword*)

Deduplication

Using scripting tools, each text file extraction created a *payload* of the new content in the related email, capturing the text between the beginning metadata and the following metadata for email purposes. A `payload` hash, an md5 encoded message digest[5] was used in the initial analysis as a primary key to assure the uniqueness of each record. Approximately half of the corpus consisted of duplicates, such as the original message in the sender's sent file and one or more copies in the recipient's inbox, at a minimum. Multiple recipients and recipients who used email folders as a filing system were another source of duplicate messages. Applying this filter reduced the corpus to approximately 250,000 emails.

Text isolation

For natural language processing (**NLP**) purposes, treating the `payload` rather than the `message body` as the unit of analysis avoided an *echo chamber* effect of `chains` quoting and re-quoting the original message, multiplying the frequency of the words it contained.

Prioritization

Traditional analysis of emails was conducted on the principe that *something may be overlooked,* which delays the value of email in preliminary analysis. Prioritizing always leaves open the option of reviewing the set-asides later.

After deduplication, the first filter applied was to eliminate all email from external addresses that were not also recipients from internal addresses. Spam, newsletters and the like have low information potential. This filter reduced the remaining half of the corpus by half again, leaving approximately 125,000 emails.

A second filter for internal email was used to eliminate broadcast messages and high frequency administrative messages. Indicia of broadcast messages were large numbers of recipients, high frequency, paucity of return correspondence and keyword in context screening. Administrative messages to single recipients were

[5]In theory, it is possible that two non-identical sequences of bytes be encoded identically; the probability is low enough to make an md5 digest usable as a checksum verification, its purpose here.

identified by frequency, lack of return correspondence and high frequency words. Many of these were nagging emails concerning the lack of approval of expense reports, for example. This filter reduced the dataset to approximately 24,000.

The third filter limited the dataset to emails sent before Enron's December 2, 2001 bankruptcy. This filter reduced the email count to approximately 13,500, about 2.7% of the original total.[6] A few emails dated "1979-12-31" were reviewed and deleted.[7] The resulting dataset was named `g_enron` for its primary purpose, network graph analysis

Augmentation and transformation

Each unique Enron address in the reduced dataset was assigned a userid. The primary purpose was to facilitate social network analysis with node identifiers of uniform length; the second, to reduce analyst bias arising from gender stereotyping, frequency of exposure and similar subjective pattern seeking behaviors.

The next transformation was to create an additional field with a `corpus` object to facilitate natural language processing.

```
g_enron <- g_enron %>% mutate(edge_corp = map(payload, tm::VectorSource)
```

To achieve a computationally practicable dataset for initial social network analysis, emails were limited to single Enron sender to Enron single recipient, reducing the dataset further, to 9,615 emails. The resulting graph had low density, and the the dataset was further trimmed by restricting it to reciprocal correspondents, each of whom sent an email to the other, either as a reply or an original message, excluding emails by a user to herself. This further reduced the dataset, to 6,009 emails among 465 users. Finally, many emails were found to be blank or extremely short – fewer than 10 characters, resulting in a final count of 5,922 emails among 445 users.

[6] *Ninety percent of everything is crap.* Theodore Sturgeon's Revelation, made in a dominantly paper-based information environment. *See also* Pareto distributions.

[7] In the technology of the day, user desktop computers relied on an internal clock powered by a battery; when the battery died, the date reset to what the operating system, usually a variant of Windows, considered as the beginning of time.