The Enron Email Dataset
Database Schema and Brief Statistical Report[1]

Jitesh Shetty
University of Southern California
Los Angeles, CA
jshetty@usc.edu

Jafar Adibi
USC Information Sciences Institute
Marina del Rey, CA
adibi@isi.edu

Introduction

Email logs have been considered as a useful resource for research in fields like link analysis, social network analysis and textual analysis. Most of the experiments in these fields of research are performed on synthetic data due to lack of an adequate and real life benchmark. The Enron email dataset is a touchstone for such research. This dataset is very similar to the kind of the data collected for fraud detection and counter terrorism hence it is a perfect test bed for testing the effectiveness of techniques used for counter terrorism and fraud detection. In this report we describe the MySql database prepared for the dataset and also statistically analyze its appropriateness for research. We further derive a social network constituting of 151 employees from the email logs, by defining a social contact to be someone with whom an individual has exchanged a pre decided threshold number of emails.

Enron Dataset

The Enron email dataset was made public by the Federal Energy Regulatory Commission during its investigation. It had a lot of integrity problems. It was later collected and prepared by Melinda Gervasio at SRI for the CALO (A Cognitive Assistant that Learns and Organizes) project; most of the integrity problems in the dataset had been resolved. It contains all kind of emails personal and official. Some of the emails have been deleted as part of the redaction effort due to requests from affected employees. William Cohen from CMU has put up the dataset on the web for researchers (http://www-2.cs.cmu.edu/~enron/). This version of the dataset contains around 517,431 emails from 151 users distributed in 3500 folders. These messages don't include attachments.

The dataset contains the folder information for each of the 151 employees. Each message present in the folders contains the senders and the receiver email address, date and time, subject, body, text and some other email specific technical details.

We created a MySql database for the dataset to catalyze the statistical analysis of the data. Figure 1 shows the schema of the database. The Enron database contains four tables. The first table contains information of each of the 151 employee. The second table contains the information of the email message the sender, subject, text and other information. The third table contains the recipient's information. It contains the email address of the recipient and the type (To, CC, BCC) in which the message was sent to the recipient. The fourth table contains information of all those messages that have been referenced after being sent once, either as a forward or reply.

We cleaned the dataset by removing a large number of duplicate emails. Folders such as "discussion_threads", "all documents" were generated by the computer and were not user created. All the messages present in these folders were already present in some other user created folder or the inbox. The folder sent_mail also contained duplicate sent messages. All the sent messages present in this folder were already present in some other sent messages folders. We dropped all the messages present in such folders.

---

[1] This database contains a lot of private emails, while using this database please be considerate about the privacy of the people who were not involved in any of the actions which precipitated the investigation.

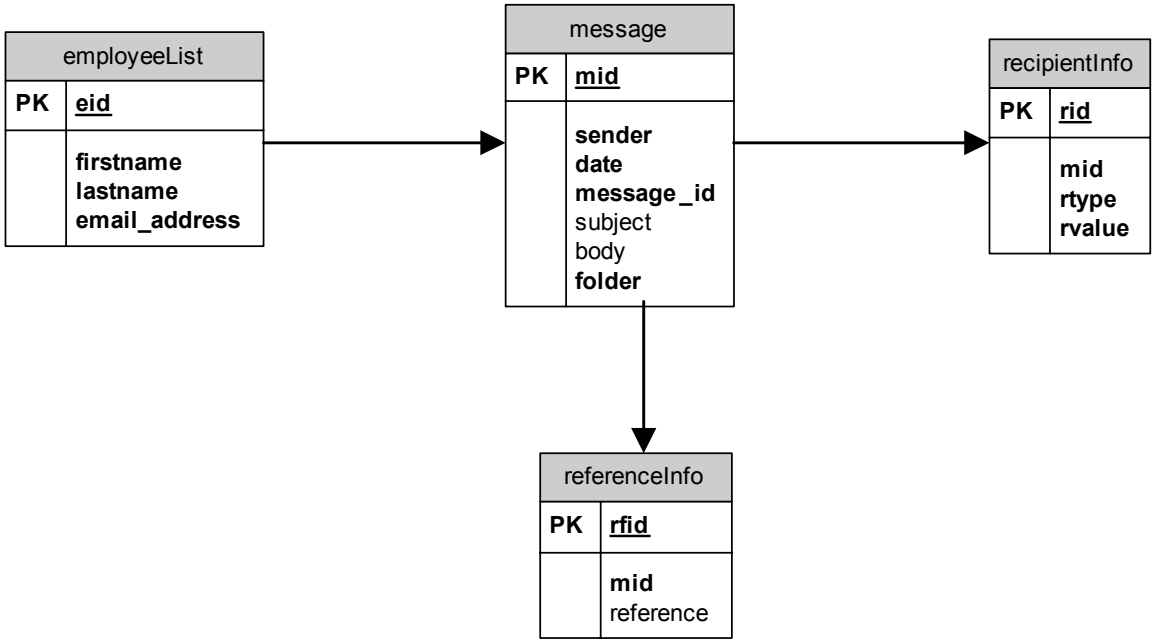

Figure 1: Enron Database Schema

There were a few messages which had junk data in them such as left over of past attachments, all such messages were dropped. There were a few messages which were completely blank, we dropped all such messages. There were certain messages which contained invalid email addresses they were changed to "no.address@enron.com". There were certain recipients whose addresses were not disclosed all such addresses were changed to a common format "undisclosed-recepients@enron.com". There were certain messages returned by the email system as an email transaction failure etc, all such messages were dropped. A large number of emails were sent to email groups, we had no information of the members of the groups, so the group name is retained as it is and the recipient information for a particular message sent to a group can be determined from the folder information. Our cleaned Enron email dataset contains 252,759 messages from 151 employees distributed in around 3000 user defined folders.

Description of the tables in the MySql database:

1. *The Employee List Entity*: The Employee List entity has a corresponding table in the physical schema; it maintains a list of the employees, their first name and their last name and their email address. This table has four attributes
    1. *EID*, this is the primary key of the table, it is an auto increment.
    2. *First Name*, this attribute has the first name of the employee.
    3. *Last Name*, this attribute contains the last name of the employee.
    4. *Employee Email*, this attribute contains the email address of the corresponding employee

MySql Description for the Table "Employeelist"

Field	***Type***	***Null***	***Key***	***Default***	***Extra***
eid	int(10) unsigned		Primary	Null	Auto Increment
firstname	varchar(31)	-	-	-	-
lastname	varchar(31)	-	-	-	-
Email_id	varchar(31)	-	-	-	-

2. *Message Entity*: The message entity has a corresponding table in the physical schema, it has the message of the email, the attributes present are as follows:

1. *MID*, this is the primary key of this table, it is auto incremented.
2. *Message ID*, this is message id present in each email.
3. *Folder*, this is folder information present in each email.
4. *Sender Email*, this contains the email address of the sender.
5. *Subject,* this is the subject of each email.
6. *Email Text*, this is the text of the email.
7. *Date***,** this is the date the email has been sent or received.

MySql Description for the Table "Message"

Field	***Type***	***Null***	***Key***	***Default***	***Extra***
mid	int(10) unsigned	-	Primary	Null	Auto Increment
sender	varchar(127)	-	-	-	-
date	datetime	Yes	-	Null	-
message_id	varchar(127)	Yes	-	Null	-
subject	text	Yes	-	Null	-
body	text	Yes	-	Null	-
folder	varchar(127)	-	-	-	-

3. *Recipient Information Entity*: This entity has a corresponding table in the physical schema; this stores the recipient's information. The attributes present in this table are as follows:

    1. *RID***,** this is the primary key of this table and it is auto incremented.
    2. *MID*, this is the foreign key from the Message table.
    3. *R-type*, this attribute stores information about how the message was sent to the recipient, as a TO, CC or BCC.
    4. *Recipient Email Value*, this attribute contains the email address of the recipient.

MySql Description for the Table "recipientInfo"

Field	***Type***	***Null***	***Key***	***Default***	***Extra***
rid	int(10) unsigned	Yes	Primary	Null	Auto Increment
mid	int(10) unsigned	Yes	-	Null	-
rtype	Enum('TO,'CC','BCC')	Yes	-	Null	-
rvalue	varchar(127)	Yes	-	Null	-

4. *Reference Text Entity*: This entity has a corresponding table in the physical schema. This table stores information about the emails forwarded or replied with the original email. This table contains the following attributes.

    1. *RFID*, this is the primary key and it is auto incremented.
    2. *MID*, this is the foreign key from the Message table.
    3. *Reference Text*, this is part of the text that is forwarded or replied.

MySql Description for the Table "referenceInfo"

Field	***Type***	***Null***	***Key***	***Default***	***Extra***
rfid	int(10) unsigned	Yes	Primary	Null	Auto Increment
mid	int(10) unsigned	Yes	-	Null	-
reference	text	Yes	-	Null	-

Statistical Analysis

We analyzed the statistics of the dataset to verify its appropriateness. Figure 2 shows the distribution of the messages per user. The x-axis represents the number of email messages in log scale. The y-axis represents the number of Enron employees in log scale. The graph clearly shows that the messages are not evenly distributed between the users. A small number of users have a large number of messages. However, there are employees distributed through out the y-axis which reflects that the dataset contains employees with all amount of email messages.

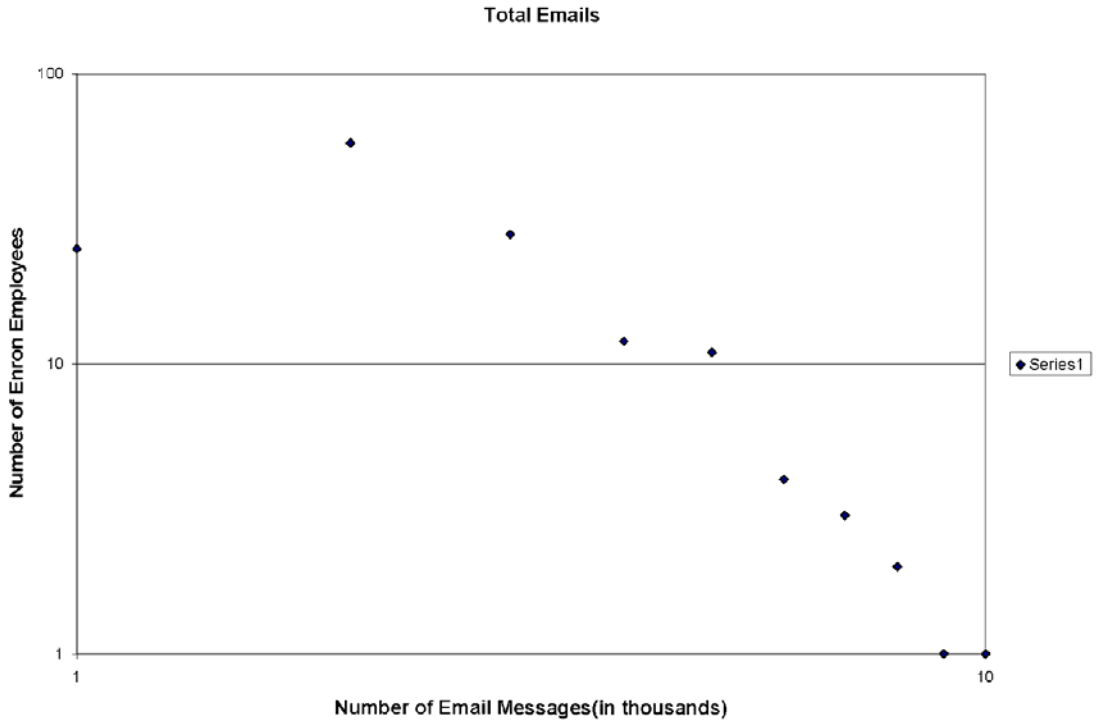

Figure 2: Distribution of emails per user

Most of the email messages present in the users folders are the email messages received by the user. The sent messages are very few in number as compared to the received messages.

Figure 3 shows the distribution of the sent emails. The x-axis represents the Enron employees. The y-axis represents the number of emails.

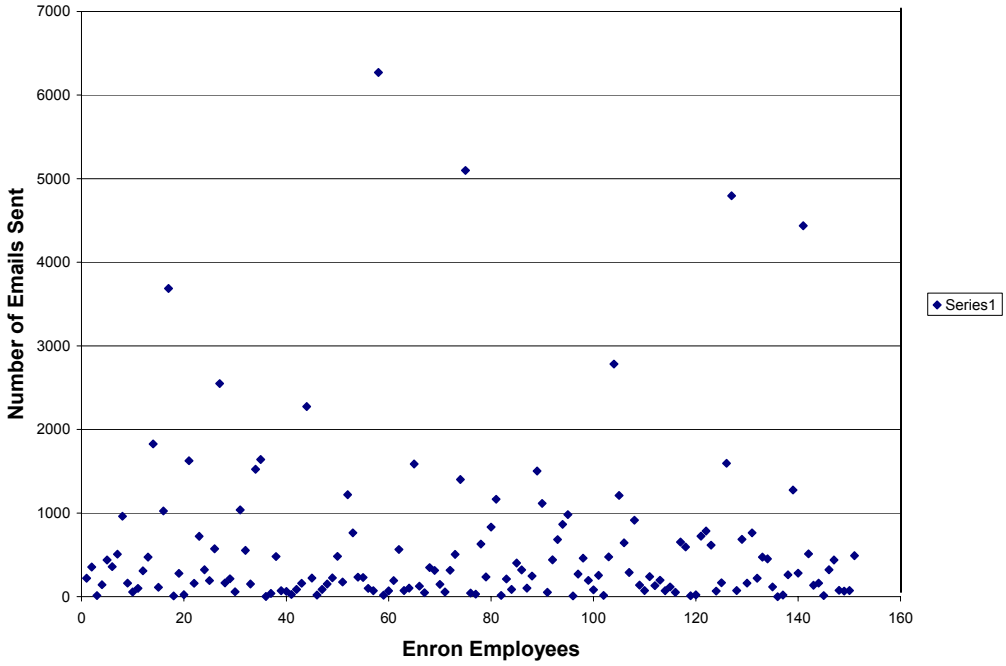

Figure 3: Distribution of Sent emails per user

Figure 4 again shows the distribution of the sent emails. The x-axis represents the number of email messages (in thousands) in log scale. The y-axis represents the number of Enron employees in log scale. The graph clearly shows that the messages are not evenly distributed between the users. A small number of users have sent a large number of messages.

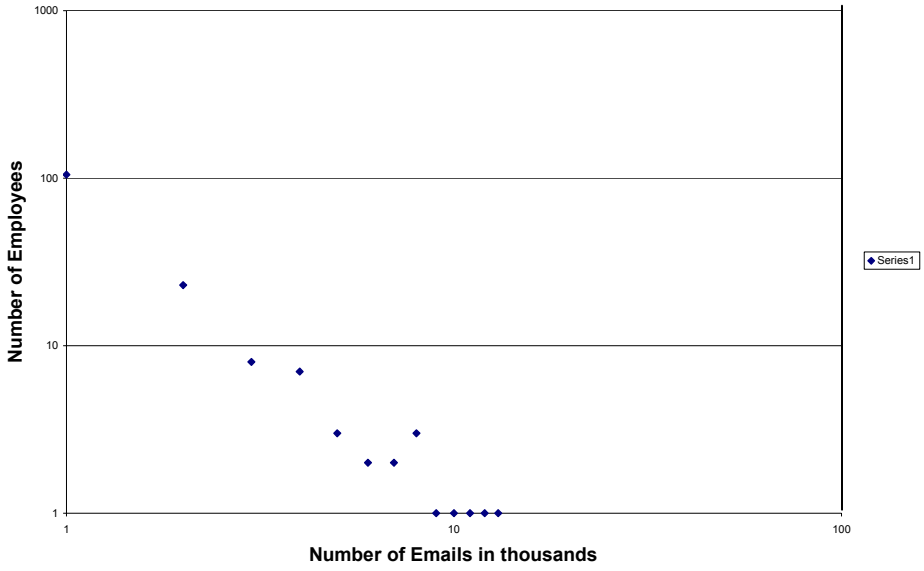

Figure 4: Distribution of Sent emails per user

Figure 5 shows the distribution of the emails over time. The figure clearly reflects that most of the emails have been sent and received in the year 2001. The x-axis represents the year in which the email has been sent or received and the y-axis shows the number of emails.

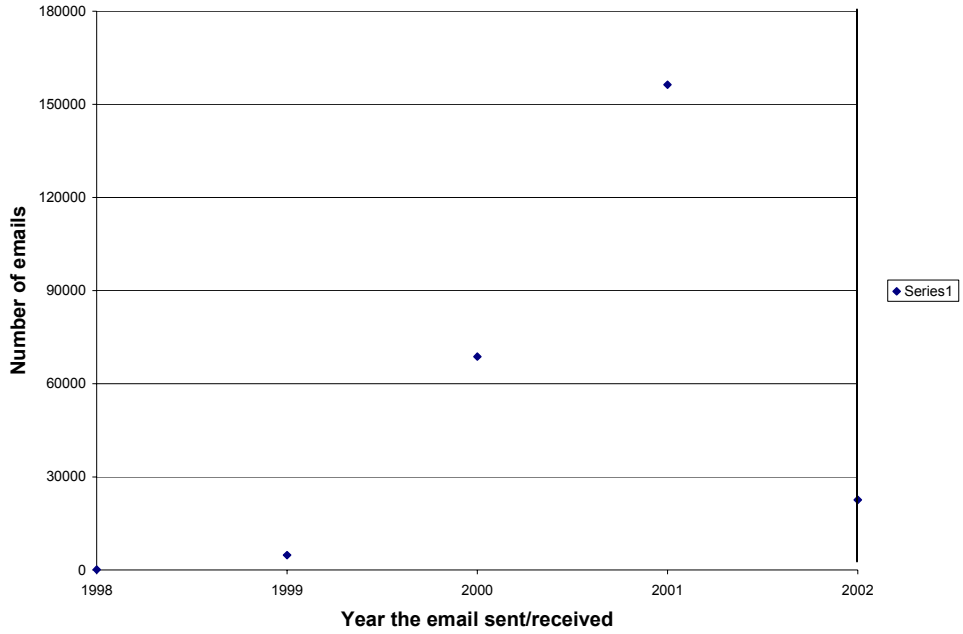

Figure 5: Distribution of emails over time (Year)

Figure 6 shows a more detailed distribution of the emails over time.

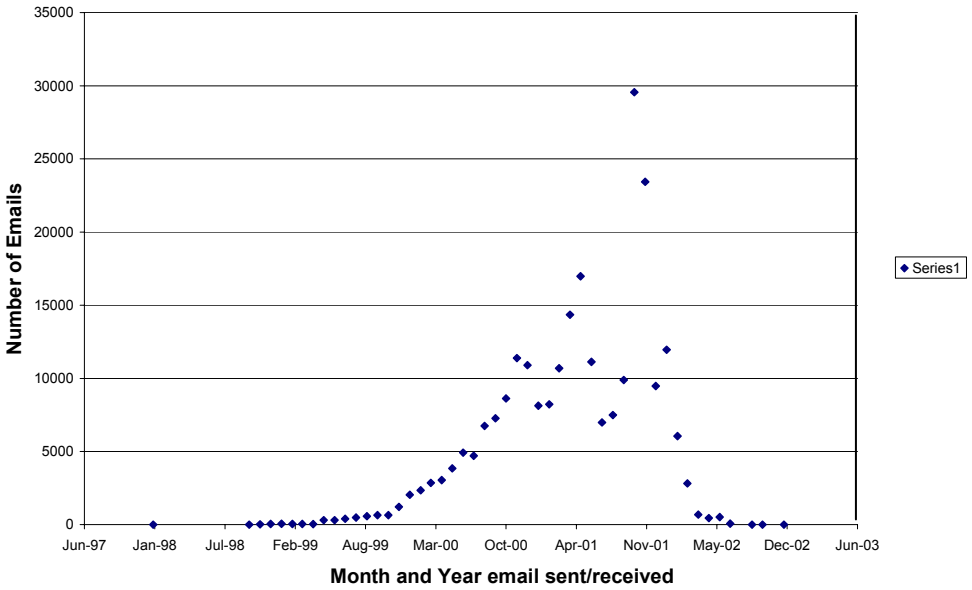

Figure 6: Distribution of emails over time in detail

Social Network

We derived a social network from the dataset. This network constitutes of 151 employees of Enron. The basic definition of a social contact between two employees is a pre defined threshold number of exchange of emails. The threshold we consider here is 5. We considered only bidirectional links, this means that we consider a contact only if both have sent emails to each other. This guarantees that there has been some exchange of information between the two employees of Enron and hence they are involved in some kind of conversation. We also considered the position of every employee in the organization hierarchy to manifest the flow of information in an organization. Figure 7 shows a graph of the social network as a Gower layout.

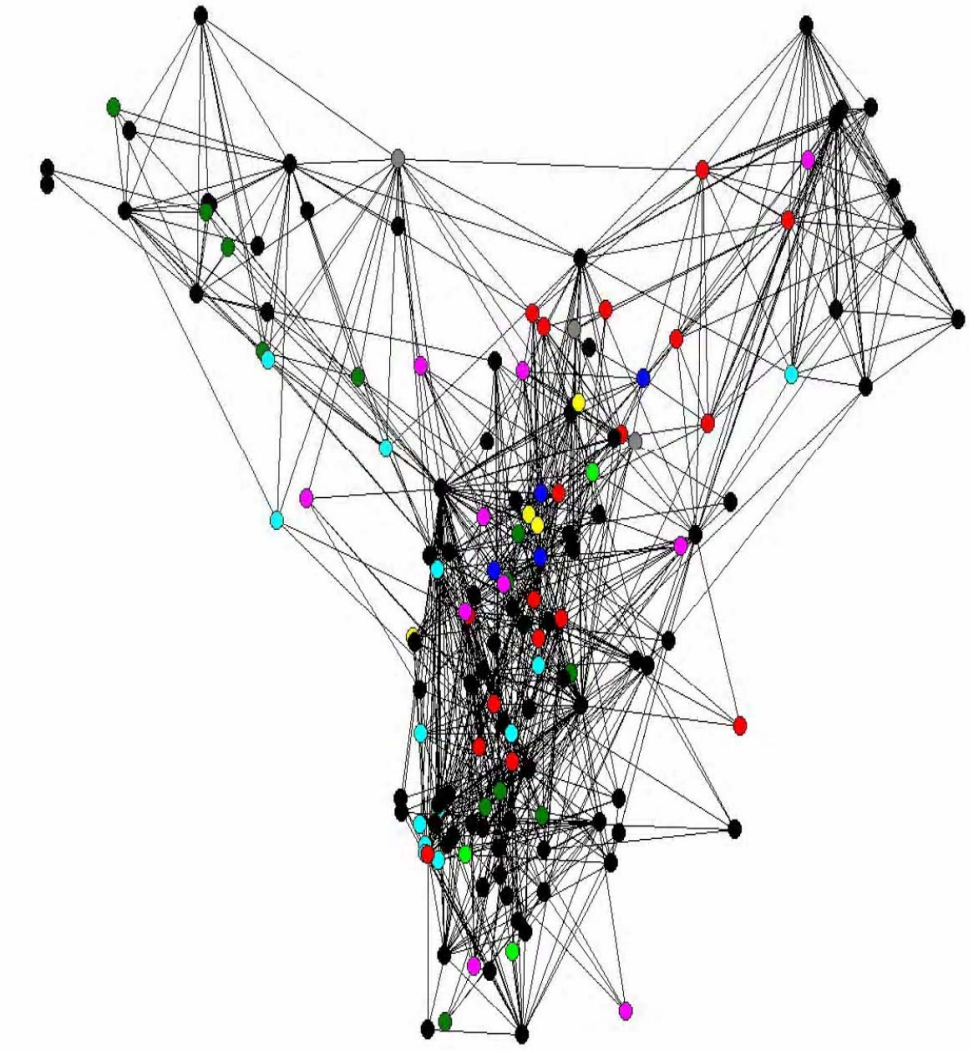

Figure 7: Network showing how the ex employees were connected

Red: Vice President, Blue: President, Black: Employee (non managerial), Grey: In House Lawyer, Pink: Manager, Dark Green: Trader, Light Green: Managing Director, Light Blue: Director, Yellow: CEO

References

1. The original dataset downloaded from William Cohen's web page (http://www-2.cs.cmu.edu/~enron/
2. Introducing the Enron Corpus. Bryan Klimt, Yiming Yang.
3. Used the document issued by the United States Bankruptcy Court, Southern district of Texas, Houston Division to find the status of the ex Enron employee.