Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset

Yingjie Zhou
Rensselaer Polytechnic Institute
zhouy5@rpi.edu

Mark Goldberg
Rensselaer Polytechnic Institute
goldberg@cs.rpi.edu

Malik Magdon-Ismail
Rensselaer Polytechnic Institute
magdon@cs.rpi.edu

William A. Wallace
Rensselaer Polytechnic Institute
wallaw@rpi.edu

Abstract

Archived organizational email datasets have been considered valuable data resources for various studies, such as spam detection, email classification, Social Network Analysis (SNA), and text mining. Similar to other forms of raw data, email data can be messy and needs to be cleaned before any analysis is conducted. However, few studies have presented investigation on the cleaning of archived organizational emails. This paper examines the properties of organizational emails and difficulties faced in the cleaning process. Cleaning strategies are then proposed to solve the identified problems. The strategies are applied to the Enron email dataset.

Contact:
Yingjie Zhou
Dept. of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute
Troy, NY 12180

Tel: 1-518-276-8457
Fax: 1-518-276-8227
Email: zhouy5@rpi.edu

Key Words: Organizational Email Cleaning, Enron Email Dataset

Support: This paper was supported in part by National Science Foundation Grant No. 0621303, Social Communication Networks for Early Warning in Disasters.

Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset

Yingjie Zhou, Mark Goldberg, Malik Magdon-Ismail, and William A. Wallace

Archived organizational email datasets have been considered valuable data resources for various studies, such as spam detection, email classification, Social Network Analysis (SNA), and text mining. Similar to other forms of raw data, email data can be messy and needs to be cleaned before any analysis is conducted. However, few studies have presented investigation on the cleaning of archived organizational emails. This paper examines the properties of organizational emails and difficulties faced in the cleaning process. Cleaning strategies are then proposed to solve the identified problems. The strategies are applied to the Enron email dataset.

To simplify the denotations in this paper, "lname", "fname", "mname" and "emailid" will be used to represent an employee's last name, first name, middle name, and organizational email ID respectively. Names having a middle name will be used in examples unless otherwise mentioned. All the code is written, compiled, and run in Perl [Wall, Christiansen & Orwant, 2000].

Properties of Organizational Email Data

Email, one of the important communication tools, is faster than conventional mail but not as interactive as telephone and online chatting. It has become an important way for people to exchange thoughts, opinions, and feelings. In general, an internet email message contains two parts: the header and the body. The header contains structured information, including sender, receiver(s), subject, and date. The body contains unstructured text, including the content of the email and sometimes a signature block at the end; often the message may have quotations ("forwarded" or "replied to" text) and attachment(s). Information such as who talks with whom at what time regarding what issues may be retrieved from the header and the body. Moreover, forwarded text gives the origin of the information, and the replied to text explains the intent of current email. The signature normally gives the contact information of the sender and may be used to identify the sender's position in the organization.

Email data has several favorable features for SNA research. *First,* emails are formatted, and the format is usually defined and followed. An email starts with the header, followed by a blank line, and then the body. The header usually includes several header fields, such as "From", "To", "Subject", and "Date". Each header field starts with a field name, followed by a ":", then a space or a tab, and then field value. *Second,* emails are easily collected and normally stored in a server. Assuming privacy of the emailers has been ensured, data collection and storage are relatively inexpensive. *Third,* emails are unobtrusive. They document the "real" past enabling researchers to trace past events. It is usually difficult to guarantee the unbiasedness and accuracy for other types of SNA data because of various factors affecting data collection. For example, the unbiasedness and accuracy of survey data depends on the design of the survey form, the understanding, and judgment of the clients. The employees' information, such as position and responsibilities, are usually available and very helpful in interpreting results from the SNA. Despite the favorable properties, email data has data cleaning problems like many other data types. Identifying these problems and proposing a thorough and systematic approach to solving them are the purposes of this paper.

Difficulties in Cleaning Organizational Emails

In general, organizational email datasets have three problems. *First,* multiple email addresses, names, or IDs exist for the same person. Even though the header name, such as "To" in an email, makes itself clear, that is, the value followed is the information about the receiver(s), the value could be in an email address form or name form. One person can be labeled with various name formats by other people. For example, one can be labeled as "lname, fname" by one colleague and "fname mname lname" by another. These optional names will not affect the actual delivery because the delivery address is supplied in the protocol, but the difficulty of mapping these names correctly is greatly increased. Moreover, one person may have multiple email addresses from domains in addition to his/her organizational one, such as yahoo, gmail, hotmail, etc. Even within the same organizational domain, people may have several email addresses. When the people who send and receive emails are of interests in SNA research, mislabeling a person may lead to confusing or even wrong conclusions. *Second,* duplicate emails exist. For example, when A sends an email to B, that email would be in the "Outbox" of A and "Inbox" of B simultaneously. If there are multiple recipients, each one of them would have a copy of that email. Duplicate emails should be removed if email or word frequency is being studied, otherwise the number of emails will be overestimated in a nonlinear way and the accuracy of results of SNA cannot be determined. *Third,* the content of the email is difficult to extract. The email content is normally mingled with the signature and the quotation. If the content of emails is an essential part in the research, the irrelevant parts should be removed before any text mining techniques are applied.

Our work mainly focuses on the first and second problems identified above. The third problem will not be addressed in this paper. One approach for solving it has been proposed using the Support Vector Machine (SVM) method [Tang, Li, Cao & Tang, 2005].

Strategies for Cleaning an Organizational Email Dataset

Strategies must be developed for cleaning the emails in an organization. The strategies proposed consist of four tasks: entities identification, formats extraction, formats generalization, and duplicate emails removal. Task 1, 2, and 3 are sequential, while Task 4 can be implemented independently.

The first task is to identify the communicating entities. We assume that the archival organization emails are organized by employee folders that may contain further subfolders, such as "Inbox", "Sent_Items", and "Deleted_Items", etc. By observing the "From" header field in the folder "Sent_Items" or similar folders that contain emails sent by the owner, his/her name and email ID normally can be determined. The study subjects of SNA are therefore defined after each folder has been mapped to an employee of the organization. However, the "From" header field itself is not sufficient to detect all the email addresses and optional names belonging to that employee. Therefore, the second task of the cleaning strategies is to collect all the email addresses and optional names for each employee by examining the "To", "Cc", and "Bcc" header fields in addition to the "From" field. By matching the employee's first name, last name and email ID in the header fields, we would collect many raw formats of email addresses and optional names. The raw formats should be converted to the extracted formats by removing the unnecessary characters. Since we assume that all the emails in one employee's folder are either sent from or to that employee, at least one of the fields should match that employee. If none of the fields in an email do, that email should be recorded for further investigation. In most cases that email is sent to a "list" that the employee is in. As a result, a list of email addresses and optional names is compiled for each employee when the second task is completed. The third task is then to generalize the formats of email addresses and optional names of each employee by taking the union operation on the extracted formats of all employees. Extra attention should be paid to unusual names and addresses. The purpose of generalization is to make the parsing procedure applicable to all employees.

The fourth task is to remove duplicate emails from the entire email dataset. Duplicate emails exist because both sender and receiver(s) may save a copy of the email, or occasionally one email is mistakenly sent more than once. Any two emails having the same recipients, day, and content are considered as duplicate emails in our approach. The day, not the time, is used as one of the constraints because one email could be sent more than once unintentionally.

The generalized formats from the third task are used in parsing the "From", "To", "Cc", and "Bcc" header fields to find the sender and recipient(s) for each unique email identified by the fourth task. Various statistics and social networks based on email frequencies can be computed and constructed. The cleaning strategies are general to organizational email datasets although each dataset may have its own problems. The Enron email dataset is used to test the effectiveness of cleaning strategies proposed in this paper.

Introduction to the Enron Email Dataset

In this section, a brief history of the Enron email dataset is introduced, followed by the organization and the format of these emails. The Enron email dataset is valuable because it is one of the very few collections of organizational emails that are publicly available. The emails of this period (1998.11 - 2002.6) record the dynamics of Enron, from glory to collapse.

Enron was the World's Leading Energy Company; it declared bankruptcy in December 2001, which was followed by numerous investigations. During the investigations, the original Enron email dataset, consisting of 92% of Enron's staff emails, i.e. 619,446 email messages in total, was posted to the web by the Federal Energy Regulatory Commission (FERC) in May of 2002 [Federal Energy Regulatory Commission, 2002]. Leslie Kaelbling at Massachusetts Institute of Technology (MIT) purchased the dataset, and found integrity problems with it. A group of researchers at SRI International worked on these problems for their Cognitive Assistant that Learns and Organizes (CALO) project, and the resulting dataset was sent to and posted by Professor William W. Cohen at Carnegie Mellon University (CMU) [Cohen, 2004]. This dataset is called the March 2, 2004 Version, which is widely accepted by many researchers. In this version, the attachments are excluded, and some messages have been deleted upon the request of Enron employees.

The resulting corpus contains 517,431 messages organized into 150 folders. The folder's name is given as the employee's last name, followed by a dash, followed by the initial letter of the employee's first name. For example, folder "allen-p" is named after Enron employee Phillip K. Allen. Presumably we guess that each folder matches one employee, but this conjecture is not correct, which will be discussed in detail in the Data Cleaning Experiment section. Each employee folder contains subfolders, such as "inbox", "sent", "_sent_mail", "discussion_threads",

"all_documents", "deleted_items", and subfolders created by the employee. A large number of duplicate emails exist in those folders.

An Enron email message contains the following header fields in order (the header field in parenthesis is optional): "Message-ID", "Date", "From", ("To"), "Subject", ("Cc"), "Mime-Version", "Content-Type", "Content-Transfer-Encoding", ("Bcc"), "X-From", "X-To", "X-cc", "X-bcc", "X-Folder", "X-Origin", and "X-FileName". The email content is separated with the headers by a blank line. The signature and quotation are continued if they exist. The header field names initiated with an "X-" means that the field values are from the original email message. The field values of "From", "To", "Cc" and "Bcc" are converted from those of "X-From", "X-To", "X-Cc" and "X-Bcc" correspondingly by the SRI researchers based on their rules. As we noted in the previous section, the field value could be in address form or name form; the SRI researchers recognized this problem and tried to convert into a uniform email address form. However, the conversion process is not satisfactory as shown in the Data Cleaning Experiment section.

An abridged Enron email message "allen-p\inbox\62" is shown in **Figure 1** as an illustration. Unimportant header fields are removed; the content of the email and quotation message is shortened as <EMAIL CONTENT> and <Quotation Message> to save space. In **Figure 1**, lines from1 to 6 are the header, the body is from line 7 to 10, and line 11 is the abridged quotation message.

```
1   Date: Fri, 2 Nov 2001 12:18:31 -0800 (PST)
2   From: renee.ratcliff@enron.com
3   To: k..allen@enron.com
4   Subject: RE:
5   X-From: Ratcliff, Renee </O=ENRON/OU=NA/CN=RECIPIENTS/CN=RRATCLI>
6   X-To: Allen, Phillip K. </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Pallen>

7   Phillip,
8   <EMAIL CONTENT>
9   Thanks,
10  Renee

11  <Quotation Message>
```

Figure 1: An Enron Email Message "allen-p\inbox\62"

Related Work on the Enron Email Data Cleaning

Most data cleaning methods, such as the detection of outliers and treatment of missing data, are designed for numerical data. Although several products are available for email cleaning, they are mainly concerned with the readability of individual emails. For instance, eCleaner 2000 and Text Money Pro are able to clean forwarded email messages by removing the '>' characters; they are also capable of removing line breaks, empty white space, and extra blank lines to improve readability [Boxer Software, 2004; Decker, 2000]. However, their capabilities do not satisfy the requirements for cleaning organizational archived emails. Few studies have done a thorough investigation on cleaning archived emails to our knowledge. As Carley and Skillicorn noted: "The sheer size and complexity of the dataset resulted in massive amounts of time being spent simply "cleaning" the data; e.g., eliminating copies of messages, identifying when the same person had multiple ids, and so on." [Carley & Skillicorn, 2005]. Despite the difficulties in cleaning organization emails, researchers have done part of the cleaning work for the Enron email dataset. Previous research efforts on duplicate emails removal for the Enron email dataset are discussed first, followed by the mapping problem.

The easiest way of removing duplicate emails is to remove folders since some folders contain a large number of duplicate emails [Klimt & Yang, 2004; Shetty & Adibi, 2004]. Although simply removing folders was fine for some research purposes, it was not adequate for the purpose of removing duplicate email messages. Andrés Corrada-Emmanuel removed the duplicate emails with a different approach. By calculating the MD5 digest of the body constrained by the same day, he concluded that the dataset contained 250,484 unique email messages belonging to 149 employees [Corrada-Emmanuel, 2004]. This approach of identifying duplicate emails is more rational than simply deleting folders. Therefore, the idea is adopted in our research by adding recipient(s) as another constraint.

Chapanond, Krishnamoorthy, and Yener specifically discussed the problem of mapping multiple email addresses and names to a person [Chapanond, Krishnamoorthy &Yener, 2005]. They extracted email addresses from "From" and "To" fields, which were converted from "X-From" and "X-To" by the SRI researchers. However, they found that one person could have multiple email addresses. For instance, Vince J Kaminski had email addresses: vince.kaminski@enron.com, vince.j.kaminski@enron.com, j..kaminski@enron.com, vince_j_kaminski@enron.com,

kaminski@enron.com, vincent.j.kaminski@enron.com, j'.'kaminski@enron.com, and j.kaminski@enron.com. They claimed that manual inspection was necessary to deal with the employees having the same last name or unexpected characters. It is not hard to notice that some of the email addresses, such as "j'.'kaminski@enron.com", are unusual. In addition, email addresses for employees in an organization are normally assigned based on clearly defined rules. It is suspicious that one employee could have so many distinct email addresses. We cannot help thinking that there is something wrong with the "From" and "To" fields. Indeed, we found that the conversion itself was flawed during our cleaning process. Although part of the cleaning work has been done in the previous research, a systematic way of cleaning the Enron email dataset should be performed with improvement and corrections.

Data Cleaning Experiment with the Enron Email Dataset

The proposed strategies guide the Enron email data cleaning experiment. Most of the employees involved in the Enron email dataset are senior managers and traders. One assumption made in most previous research work is that one folder corresponds to one employee. Priebe, Conroy, Marchete, and Park, however, found 184 employees in Enron email dataset [Priebe, Conroy, Marchete & Park, 2005], which was the number of unique addresses they obtained from the "From" header of emails in the "Sent" boxes after removing some addresses that were clearly not related to the 150 employees. Therefore, our first task was to identify who were the employees. We found 156 employees after removing the secretaries, assistants, and people or addresses that were irrelevant to the employees by searching the "From" field of every email in subfolder "_sent_mail", "sent" and "sent_items". Three cases of folder-employee matching exist in the Enron email dataset: 1 folder to 1 employee, 1 folder to 2 employees, and 2 folders to 1 employee.

Case I: One folder corresponds to one employee. For example, in subfolder "_sent_mail", "sent" and "sent_items" of user folder "allen-p" we found 8 emails sent by "Ina Rangel", 1 email by "Pam Butler", and 1500 emails by "Phillip K Allen". By further examining the content of the emails, Ina Rangel seemed to be an assistant of Allen K Phillip, but there was no clue to the identification of Pam Butler. In this case, we claim Phillip K. Allen is the only owner of folder "allen-p".

Case II: One folder corresponds to two employees. This case happens when two employees have same last name and same initial of first name. Sometimes two employees have the same first name but distinct middle names. For example, in folder "campbell-l", we found 5 emails sent by "Robert J. Baker", 303 emails by "Larry Campbell", and 359 emails by "Larry F Campbell". The five emails from "Robert J. Baker" can be ignored. However, Larry Campbell and Larry F Campbell are two different persons, which can be determined by the organizational email IDs (lcampbe versus lcampbel) and email content from "campbell-l\all_documents\952". In this case, we claim that both Larry Campbell and Larry F Campbell are the owners of the folder "campbell-l". The same situation applies to folder "dean-c", "gay-r", "hernandez-j", "hodge-j", "mcconnell-m", "scott-s", and "taylor-m".

Case III: Two folders correspond to one employee. Two such cases happen. In folders "panus-s" and "phanis-s", Stephanie Panus is the only sender, so we concluded that both folders belonged to her. Another two folders, "whalley-g" and "whalley-l", were found to belong to Greg Whalley.

We identified 156 Employees with their names and Enron email IDs from Enron email dataset. We then limited our scope to the emails among these 156 employees. The second cleaning task was to investigate the recipient information of all emails in each employee folder to find his/her email addresses and optional names. As we said, the "To", "Cc", and "Bcc" fields were converted by the SRI researchers from the original email fields "X-To", "X-cc", and "X-bcc". The intention of conversion is good: all the messy formats are converted to a format like string1.string2@enron.com. For example, "Rick Buy" was converted to rick.buy@enron.com. However, even if the conversion is correct, various email addresses have to be mapped to an employee [Chapanond, Krishnamoorthy & Yener, 2005]. Also, the rules for the conversion have flaws:

- The rules are aggressive; email addresses from other domains were converted to Enron domain when there were multiple recipients. For example, "Kilburn, Bobbi <Bobbi.Kilburn@vynpc.com>" in the "X-To" field was converted to "bobbi.kilburn@enron.com", and "Braintree Electric Light Dept. <kstone@beld.com>" was converted to "dept..braintree@enron.com" in message "baughman-d\inbox\218". The conversion is obviously wrong.
- The rules may lead to confusing and incorrect results even for Enron employees. For example, "Taylor, Mark E (Legal) </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Mtaylo1>" in "X-To" was converted to "legal <.taylor@enron.com>" in message "cash-m\sent_items\505", which doesn't make sense. Worse than that, two different names may be converted to the same address. For example, in message "davis-d\deleted_items\101", "Davis, Mark Dana </O=ENRON/OU=NA/CN=RECIPIENTS/CN=MDAVIS>" sent an email to "Davis, Dana </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Ddavis>", but the converted

results showed that "dana.davis@enron.com" sent an email to the same address "dana.davis@enron.com". Employees who have the same last name and same initial of first name may be confounded.

- Useful information is lost during the conversion. For example, in message "allen-p\sent_items\68", "Mike Grigsby <Mike Grigsby/HOU/ECT@ECT>" in "X-To" was converted to "mike.grigsby@enron.com"; the information that Mike Grigsby worked in the department of Enron Capital and Trade (ECT) in Houston (HOU) was lost.

The original header fields, therefore, were used in our cleaning procedure. Since "X-bcc" field was not provided in the Enron email dataset, it was ignored in the cleaning process. The "X-To" and "X-cc" fields of all emails in each employee folder were examined for collecting the email addresses and optional names by matching the last name, first name, and Enron email ID. We found that two formats dominated; one was "fname mname lname", and the other was "lname, fname mname. </o=enron/ou=na/cn=recipients/cn=emailid>". Other standard formats include emailid@enron.com and fname.mname.lname@enron.com. The raw formats contained various unnecessary characters associated with the real email addresses. The extracted formats were obtained by removing these characters. For example, in folder "allen-p", the raw formats found for Phillip K. Allen are shown in the left column of **Table 1**. The corresponding extracted formats are listed in the right column of **Table 1**.

Table 1: Raw Formats and Extracted Formats for Employee Phillip K. Allen

Raw Formats	Extracted Formats
phillip k allen	phillip k allen
allen, phillip k. </o=enron/ou=na/cn=recipients/cn=pallen>	allen, phillip k. </o=enron/ou=na/cn=recipients/cn=pallen>
pallen@enron.com "phillip allen" <pallen@enron.com> "pallen@enron.com" <pallen@enron.com> phillip <pallen@enron.com> phillip allen <pallen@enron.com> "allen, phillip k" <pallen@enron.com> <pallen@enron.com>	pallen@enron.com
pallen70@hotmail.com 'pallen70@hotmail.com' phillip allen - home (email) <pallen70@hotmail.com> "phillip allen" <pallen70@hotmail.com>@enron	pallen70@hotmail.com
pallen@hotmail.com	pallen@hotmail.com
phillip.k.allen@enron.com <phillip.k.allen@enron.com>	phillip.k.allen@enron.com
pallen@ect.enron.com phillip k allen <pallen@ect.enron.com>	pallen@ect.enron.com

As we have identified in the first task, one folder may have two users. Interestingly, we found that not only were we confused by their similar names, but so were Enron employees themselves. They kept forwarding misdirected emails to each other, and complained about the mess, such as "scott-s\all_documents\1328", "scott-s\sent_items\389", and "taylor-m\all_documents\1740". After we have finished the second cleaning task for all the 156 employees, 156 tables similar to Table 1 were created. In the third task, we compared these tables, and found that the extracted formats followed explicit rules although some employees might miss one format or another. The Union operator was taken to make a comprehensive list of all the formats. Missing formats might be found for individuals by comparing with the formats in the comprehensive list. For example, we found "lname, fname <something>" was also a common format. However, we failed to discover it for Phillip K. Allen from his folder "allen-p", instead we obtained format "allen, phillip </o=enron/ou=na/cn=recipients/cn=notesaddr/cn=ba4cd662-58db2db2-862564b8-5b412a>" from other user folders. Same argument held for format "phillip k allen/hou/ect". For those email addresses outside of the Enron domain, we had to deal with them individually. Most of the time, these email addresses were used to forward emails from or to himself/herself. Occasionally they were used to communicate with colleagues. In conclusion, the standard formats that were general to all Enron employees were summarized as: (1) fname mname lname; (2) lname, fname mname. <...=emailid>; (3) lname, fname <…>; (4) emailid@enron.com; (5) emailid@dept.enron.com; and (6) fname.mname.lname@enron.com. Other formats, which were only applicable to one employee or a small group of employees, were listed only for those employees. The general formats together with the individual formats of all employees were used for parsing the headers.

The fourth task is to remove the duplicate emails. We suggest that any two emails having the same recipients, day, and content be considered as duplicate emails. Calculating MD5 Digest for these three elements, 252,830

"unique" messages were found. This number is slightly less than 250,484 when only day and content were used [Corrada-Emmanuel, 2004]. This method is very sensitive to the emails with quoted messages. Therefore, the tools for cleaning the quotations could be used before removing the duplicate emails to achieve better results.

Upon completion of all four tasks, we have cleaned the emails among 156 Enron employees identified from the Enron email dataset distributed by William W. Cohen. By "cleaned", we mean that we have removed the duplicate emails, identified 156 Enron employees from 150 folders, mapped optional names and multiple email IDs for each individual, and finally picked out the emails among these 156 employees from the whole dataset. This subset of the emails can then be used for the SNA.

Conclusions and Future Work

Data cleaning strategies for archived organizational emails are proposed and applied to the Enron email dataset. These strategies should provide guidance when cleaning organizational emails. However, an organizational email dataset may have its own unique problems. In general, the strategies are practical and served well in cleaning the Enron emails. In the future, the strategies will be applied to other email datasets to test their robustness. Machine Learning techniques will be used to distinguish persons having similar or even the same names by learning their social contacts and email content. Finally, cleaning the content of the organizational emails is also part of our future research.

References

[Boxer Software, 2004] Boxer Software, 2004, "Ugly Email Messages: Beware the Text Monkey." Retrieved October 20, 2006, from http://www.boxersoftware.com/textmonkey.htm

[Carley & Skillicorn, 2005] Carley, Kathleen M. & David Skillicorn, 2005, "Special Issue on Analyzing Large Scale Networks: The Enron Corpus." *Computational & Mathematical Organization Theory,* 11(3), 179-181, 2005.

[Chapanond, Krishnamoorthy & Yener, 2005] Chapanond, Anurat, Mukkai S. Krishnamoorthy & Bülent Yener, 2005, "Graph Theoretic and Spectral Analysis of Enron Email Data." *Computational & Mathematical Organization Theory,* 11(3), 265-281, 2005.

[Cohen, 2004] Cohen, William W., 2004, "Enron Email Dataset." Retrieved March 12, 2005, from http://www.cs.cmu.edu/~enron/.

[Corrada-Emmanuel, 2004] Corrada-Emmanuel, Andrés, 2004, "Enron Email Dataset Research." Retrieved March 12, 2005, from http://ciir.cs.umass.edu/~corrada/enron/index.html.

[Decker, 2000] Decker, John, 2000, "eClean 2000." Retrieved October 20, 2006, from http://www.jd-software.com/eClean2000/index.html.

[Federal Energy Regulatory Commission, 2002] Federal Energy Regulatory Commission, 2002, "FERC: Information Released in Enron Investigation." Retrieved March 12, 2005, from http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.

[Klimt & Yang, 2004] Klimt, Bryan & Yiming Yang, 2004, "Introducing the Enron Corpus." *Conference on Email and Anti-Spam,* Mountain View, CA, 2004.

[Priebe, Conroy, Marchete & Park, 2005] Priebe, Carey E., John M. Conroy, David J. Marchette & Youngser Park, 2005, "Scan Statistics on Enron Graphs." *Computational & Mathematical Organization Theory,* 11(3), 229-247, 2005.

[Shetty & Adibi, 2004] Shetty, Jitesh & Jafar Adibi, 2004, "The Enron Email Dataset Database Schema and Brief Statistical Report." Technical report, Information Sciences Institute, 2004.

[Tang, Li, Cao & Tang, 2005] Tang, Jie, Hang Li, Yunbo Cao & Zhaohui Tang, 2005, "Email Data Cleaning." *In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Chicago, IL, August 21-24, 2005.

[Wall, Christiansen & Orwant, 2000] Wall, Larry, Tom Christiansen & Jon Orwant, "Programming Perl." 3rd Edition, O'Reilly, July 14, 2000.