Enron Case Study: Analysis of Email Behavior Using *EmailTime*

Minoo Erfani Joorabchi, Ji-Dong Yim, Mona Erfani Joorabchi, and Christopher D. Shaw

Simon Fraser University

Abstract

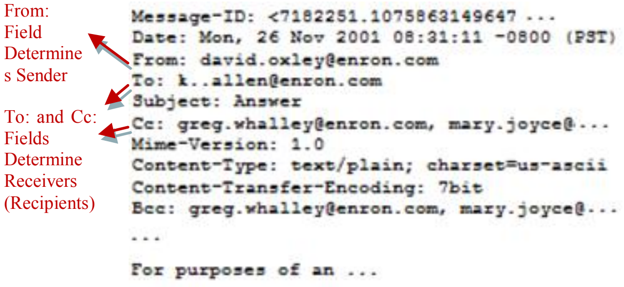
This paper presents a case study with *Enron* email dataset to explore the behaviors of email users within different organizational positions. We defined *email behavior* as the email activity level of people regarding a series of measured metrics e.g. sent and received emails, numbers of email addresses, etc. These metrics were calculated through *EmailTime*, a visual analysis tool of email correspondence over the course of time. Results showed specific patterns in the email datasets of different organizational positions.

Keywords: Email, Enron, Case Study, *EmailTime*, Visual Analysis.

Index Terms:

1 Introduction

Exploring *Enron* case study, we defined *email behavior* through measured metrics using *EmailTime* [1]. The metrics were namely, Sent emails, Received emails (as 'To' and 'Cc' fields, See Figure 1), numbers of the email addresses, numbers of the created folders and Size of the sent emails. We defined the *size of a Sent email* as the number of recipients in the 'To' and 'Cc' fields of that email.



```
Message-ID: <7182251.1075863149647 ...
Date: Mon, 26 Nov 2001 08:31:11 -0800 (PST)
From: david.oxley@enron.com
To: k..allen@enron.com
Subject: Answer
Cc: greg.whalley@enron.com, mary.joyce@...
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: greg.whalley@enron.com, mary.joyce@...
...

For purposes of an ...
```

Figure 1. An email processed by *EmailTime*. 'From', 'To' and 'Cc' fields were specified. We used these terms throughout the paper.

2 Case Study

2.1 Hypotheses

General research question is derived as whether there are specific patterns in the way people use email. More specifically, metrics are utilized to discover the impact of the organizational positions on the email behavior of people by providing detailed quantitative information about them.

2.2 Benchmark: Enron Email Dataset

Finding and working with a real world benchmark has been a challenge especially when the nature of the data is private. Email datasets of organizations are good examples of this private data area. *Enron* email archive is a unique large dataset which contains more than 200,000 emails [2]. The dataset used in this paper is selected from two years of time span between January 2000 and December 2001 as the email collaborations in this period of time look most realistic (data of many Enron people are public). Nevertheless, we still see that missing actors and emails in the original dataset mostly for unknown or sensitive reasons. We grouped the email users in the archive using seven categories identified in the public dataset [3] including 35 employees, 4 CEOs, 4 presidents, 21 vice presidents, 12 directors, 14 managers and 11 traders (See Figure 2). In the next section, we detail the differences of email behaviors within organizational positions.

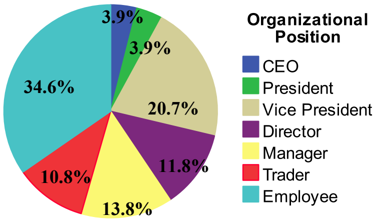


Figure 2. Composition of Enron organizational positions.

3 Results And Discussion

3.1 Analysis on the Activity Level

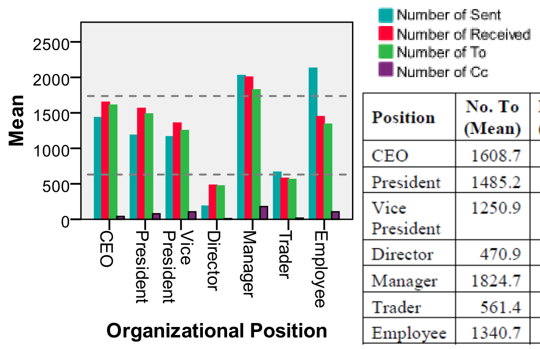
From comparing activity levels of organizational positions, some groups behaved similarly regarding number of Sent, Received, and exchanged emails (sum of Sent and Received). From Figure 3 we divided them into three categories – Modest, Moderate and Active (See Table 1). Figure 3-A graphically specifies the three categories with dashed lines. Results identified Managers and Employees were Active, Traders and Directors were Modest, and the rest were Moderate. Employees have the highest average number of Sent emails while Managers have the highest average number of Received emails.

Table 1. Classification of organizational positions.

No of Emails	Modest	Moderate	Active
Exchanged (Sent+Received)	[<1500] Directors, Traders	[1500-3500] CEOs, Presidents, Vice Presidents, Employees	[>3500] Managers
Sent	[<1000] Directors, Traders	[1000-1500] CEOs, Presidents, Vice Presidents	[>1500] Employees, Managers
Received	[<1000] Directors, Traders	[1000-2000] CEOs, President, Vice Presidents, Employees	[>2000] Managers

SIAT, Simon Fraser University, 102 Avenue, Surrey, Canada
Send correspondence to Minoo.E.J.
Minoo.E.J.: E-mail: mea18@sfu.ca
J.D.Y.: E-mail: jdyim@sfu.ca
Mona.E.J.: E-mail: mea16@sfu.ca
C.D.S.: E-mail: shaw@sfu.ca

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 29, Salt Lake City, Utah, USA
978-1-4244-9486-6/10/$26.00 ©2010 Crown



Position	No. To (Mean)	No. Cc (Mean)
CEO	1608.7	41.0
President	1485.2	78.0
Vice President	1250.9	105.5
Director	470.9	9.7
Manager	1824.7	179.5
Trader	561.4	16.8
Employee	1340.7	105.9

A. B.

Figure 3. Results of exchanged emails. A) Average number of Sent and Received (as Cc and To) emails for Enron organizational positions from 2000 to 2001. B) Average number of Received emails as To and Cc for each organizational positions.

3.2 Analysis on the Roles

Another approach to interpret Figure 3 is to compare numbers of Sent and Received emails to determine their role. Table 2 specifies the roles of different organizational positions.

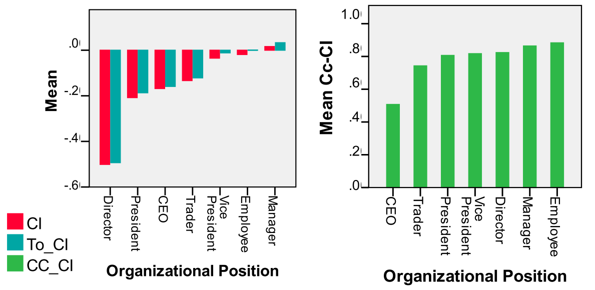
Table 2. Organizational positions' roles.

Relationship	Organizational Positions
#Sent < #Received	Presidents, Directors, CEOs and Vice Presidents
#Sent = #Received	Managers
#Sent > #Received	Employees and Traders

Email addresses hold different roles within different period of time. Gloor et al. defined Contribution Index (1) to specify the role of email addresses [4].

$$\frac{\text{emails sent – emails received (To+Cc)}}{\text{total of emails sent and received (To+Cc)}} \tag{1}$$

This index is near to –1 for the receivers and +1 for the senders. We expand this formula to *received as To* (To-CI) and *received as Cc* (Cc-CI) to see the impact of To and Cc separately.



A. B.

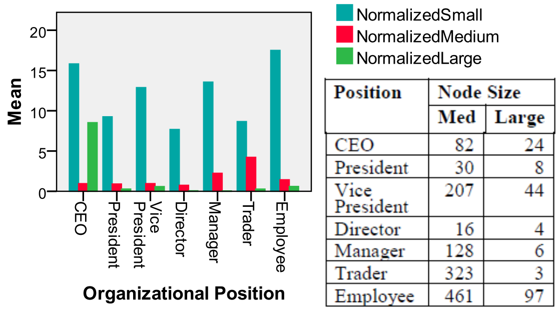
Figure 4. Contribution Index (CI) for organizational positions. A) Average of CI and To-CI. B) Average of Cc-CI.

Figure 4 shows that the average Contribution Index of administrative people tends to be lower than that of lower positional staff. CI and To-CI follow a same trend while Cc-CI has a different trend. It may be the result of the nature of their positions (e.g. employees ask and report whereas administrators are reported and make orders). Figure 4-A shows CIs of Employees, Traders and Managers are near zero, which means that they had same amount of send and received emails in average. Then there is a jump to Director, President and Vice President where the three groups behaved as weak receivers. Finally, there is a jump to CEO group where the executive officers behaved as strong receivers.

On the other hand, Figure 4-B shows Cc-CI for organizational positions where most of them except CEOs were near +1. This means that they had received a little number of emails as Cc.

3.3 Analysis on the Size of Sent Emails

We divided the Sent emails into four categories based on the number of recipients; Single nodes (including only one recipient), Small nodes (2-9), Medium nodes (10-29) and Large nodes (30 and up). We then calculated the normalized of Small, Medium and Large to graph them with the organizational positions (See Figure 5). Comparing the ordinary and normalized graph we figured out the normalized graph contribute more in terms of presenting the habits of different positions in sending emails to group of people. It showed a huge difference between CEOs and other groups in sending Large size emails. Traders and then Managers sent Medium size emails more than any other groups.



Position	Node Size	
	Med	Large
CEO	82	24
President	30	8
Vice President	207	44
Director	16	4
Manager	128	6
Trader	323	3
Employee	461	97

A. B.

Figure 5. Results from size of Sent email. A) Average of normalized number of Small, Medium and Large sent emails for organizational positions. B) Average number of Medium and Large emails sent by different organizational positions.

4 Conclusion and Discussion

We presented a case study to analyze the activity level, role and size of sent nodes by Enron people between January 2000 and December 2001. We found some groups behaved similarly. According to the results, since no relationship between the numbers of created folders and organizational positions was found in our dataset, we believe this is a subjective issue. Moreover 80% of cases had the numbers of email addresses (with Enron domain) within the range of 2 to 6.

We performed ANOVA (using SPSS) on the Enron dataset, where the dependent variables were numbers of exchanged, Sent, received as To and Cc emails, numbers of email addresses, CI, To-CI, Cc-CI, numbers of Small, Medium and Large size for sent emails. The independent variables were the seven organizational positions. Results presented no significant differences. There was only a significant difference for number of Email Address between CEOs and other groups, $F(6,94) = 2.67$, $p < .05$.

References

[1] M. Erfani Joorabchi, J. D. Yim, C. D. Shaw. EmailTime: visualization of the temporal email. GHC, October 2010.

[2] B. Klimt and Y. Yang. Introducing the Enron corpus. *First Conference on Email and Anti-Spam* (CEAS), Mountain View, CA. October 2004.

[3] W. W. Cohen, CALD, CMU. October 2004, from http://www-2.cs.cmu.edu/~enron/

[4] P. Gloor, S. Niepel, Y. Li. Identifying potential suspects by temporal link analysis. MIT CCS working paper, January 2006.