Analyzing the Enron mails

Peter Heringer
Matrikelnummer 6109174
peter.heringer@student.uni-tuebingen.de

Felix Seidel
Matrikelnummer 5969276
felix.seidel@student.uni-tuebingen.de

Abstract

The Enron mail corpus provides real world corporate mail communication data. The corpus contains mails by Enron personell during the phase in which Enron declared bankruptcy. We extract the core working hours of the Enron personell from this dataset. The dataset indicates that Enron performed a migration of their mail systems shortly before the scandal became public, which introduces an interesting shift in the mails' time information. However, we show that there is no significant change in core working hours over time. [1]

1 Introduction

After the US energy cooperation Enron declared bankruptcy in 2001 [2], law enforcement agencies released enormous amounts of e-mail communication between Enron employees to the public. This raw database export was cleaned up as the *Enron Mail Corpus* by Klimt and Yang [5]. Today, this dataset contains over 500M e-mails from the mailboxes of Enron employees and can be accessed via https://www.cs.cmu.edu/~enron/. This dataset is used in various research projects, e.g. for e-mail classification research by Klimt and Yang [4] or graph analysis by Chapanond et al. [1]. The e-mails contained in this dataset span over multiple years, and interestingly, also contain e-mail conversations throughout 2000 and 2001. As the Enron scandal enrolled during 2001, this period is interesting in order to check if Enron employees worked differently, i.e. worked more or less.

2 Analyzing the time data

For our work we want to investigate whether there was any significant change in the amount that employees worked during the time they declared bankruptcy. This of course is no proof of a direct causal link between these two events, but makes one at least more plausible.

To analyze this hypothesis, one needs to infer the probable working time from the times the e-mails were sent. Thus, the first task is to get these timestamps. As the files contained the headers of the e-mails, the timestamps can be read from the `Date` header [6]. As we only care for the working hours of Enron employees (and not of those outside people who's mails just happen to be included in the Enron Mail Corpus), we only consider the e-mails that are actually sent by `@enron.com` addresses. Mails from Enron employees that are sent *to* Enron employees can be included multiple times in the Mail Corpus. The `Date` header is usually set by the sender [3]. This can be the sender's client or the e-mail server, but never the receiver. Thus, we consider all e-mails with the same sender, subject and `Date` header to be the same.

[1] The git-repository for this project can be found at https://github.com/gnampfelix/enron-data-literacy.

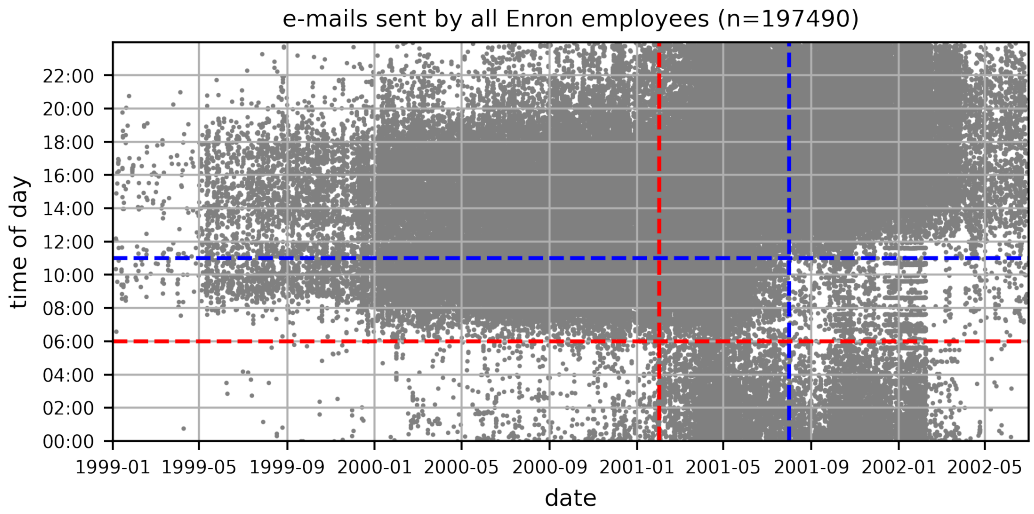

Figure 1: All mails in the Enron Mail Corpus that are sent by `@enron`.com addresses. The vertical red dashed line marks the beginning of the migration period with the usual start of the core working hours indicated by the horizontal red dashed line. The vertical blue dashed line marks the assumed end of the migration period with the new usual start of the core working hours indicated by the horizontal blue dashed line.

In Figure 1 all mails that are left by removing such duplicates are drawn. During 1999, the amount of e-mails is small and it increases drastically by January 2001. During this period, the majority of e-mail communication occurs between 06:00 and 20:00 and night times are clearly reflected in the amount of e-mail traffic. Starting February 2001, the silence in the night disappears and the data is becoming more noisy. From August 2001, we start to see a separation between working hours and non-working hours again, but the time frame appears to be shifted +5 hours in comparison to the pre-February period.

The reason for this shift can be found in the content of the e-mails: Starting in February 2001, Enron performed a migration from their Lotus Notes e-mail infrastructure to a Microsoft Outlook infrastructure. Each employee received a couple of e-mails from `team.outlook@enron.com`, informing them of their individual migration date. On migration day, the e-mails of the last 30 days of the employee were moved to the new infrastructure automatically. Each employee was responsible to manually migrate older e-mails that they need. When looking at the mailbox of individual employees, one can find duplicates that seem to be created during this period. Interestingly, the duplicates are the exact same e-mails, but have a `Date` header that is shifted by exactly five, six, ten or twelve hours. We can only speculate what the precise reason for this shift is. One possible explanation is a misconfiguration of the different e-mail databases from which the dataset was exported.

This is why the dataset as parsed for Figure 1 is not suitable to infer the core working hours directly, especially in the migration period: Duplicate mails with shifted time information that might be even duplicated and shifted multiple times more, due to them being present in multiple mailboxes that were migrated on different days, make the results virtually unusable.

To reduce this noise, we limit our e-mail parsing to the folders `sent_items`, `_sent_mail`, `_sent` and `sent`. By doing this, we naturally remove many duplicates, as only one mailbox can be the sender mailbox of a particular e-mail. Furthermore, we exclude sender addresses that are not represented with a mailbox in the dataset and would therefore only provide few data points. The remaining duplicates are removed by considering e-mails with the same sender, subject, minute and second as duplicate. This respects the time being shifted in some of the duplicates by exactly five, six, ten or twelve hours.

The result of this is visualized in Figure 2. By limiting the parsing to those four folders and by removing some more duplicates, we are able to reduce the noise drastically.

When using this de-duplicated dataset and plotting boxplots (Figure 3) for each employee, one still can see big differences between the individual employees. This might have different reasons like

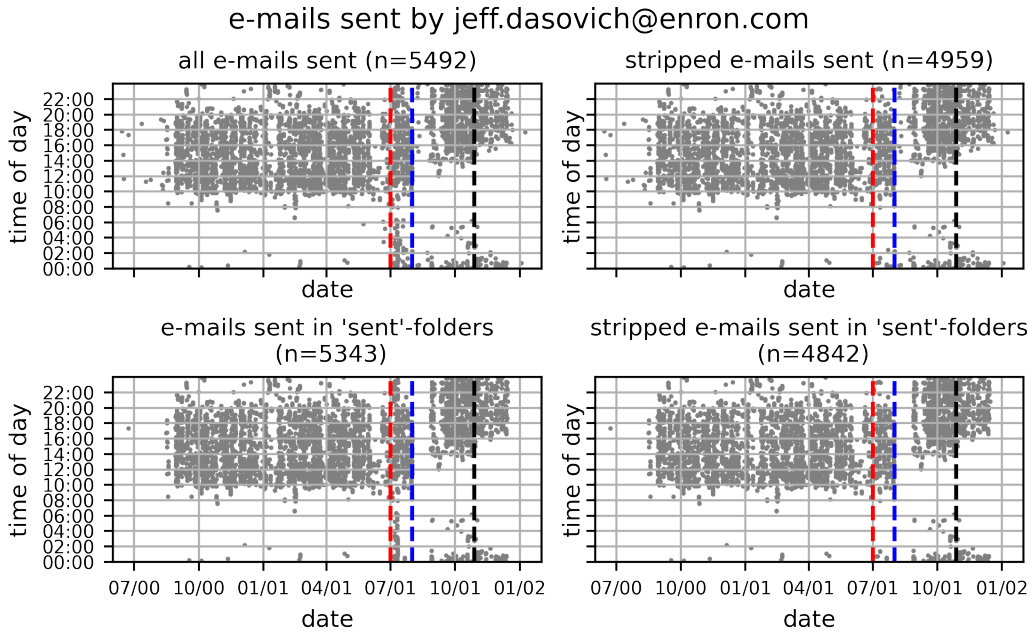

Figure 2: All mails in the Enron Mail Corpus that are sent by `jeff.dasovich@enron.com`. In all four plots, the red dashed line marks the begin of Jeff's e-mail migration period, the blue dashed line the end. The black dashed line marks the end of the 2001 daylight saving time in the USA. The upper row shows all e-mails in the mail corpus where the sender is `jeff.dasovich@enron.com`, the lower row only those e-mails by `jeff.dasovich@enron.com` that are in a `.*sent.*`-folder. The left column considers e-mails with the same subject, sender, time and date duplicate. The right column considers e-mails with the same subject, sender, minute and second duplicate.

Table 1: The average working hours per month. The first half of 2000 acts as a baseline against which the following months are checked for significant divergence.

month	H1 00	06/01	07/01	08/01	09/01	10/01	11/01	12/01
avg. hours	05:00	05:45	05:16	05:43	04:48	05:00	05:12	04:30
p-Value	-	0.524	0.530	0.522	0.503	0.525	0.521	0.520

employees having different core working hours, flexible time management, working only before or after the e-mail migration or working in different timezones. The boxplot uses the 25th and 75th quantile for the box limits, which is reasonable to us. We will use those quantiles to calculate the core working hours for each employee in each month. For each month, we will exclude employees who wrote less than 10 e-mails in that month, as for those, the working hours cannot be inferred.

We test whether we can reject H_0, that the amount of hours worked by the Enron staff has not changed significantly over time. We assume the first half of 2000 to be our baseline, as there are only few data points before that date in the Enron Mail Corpus. We re-sampled that baseline 10000 times for each following month with a sample size equal to the amount of data points for that month and checked whether the average amount of hours worked is more extreme on both sides than the average amount of hours worked inferred from the data of that month. An excerpt of this data is depicted in Table 1. We found no significant divergence from the baseline for any month in the dataset.

3 Discussion

We've made some assumptions for our analysis that we want to address briefly. Firstly, we assumed that we can infer the core working hours from the mail data. Every individual person might start

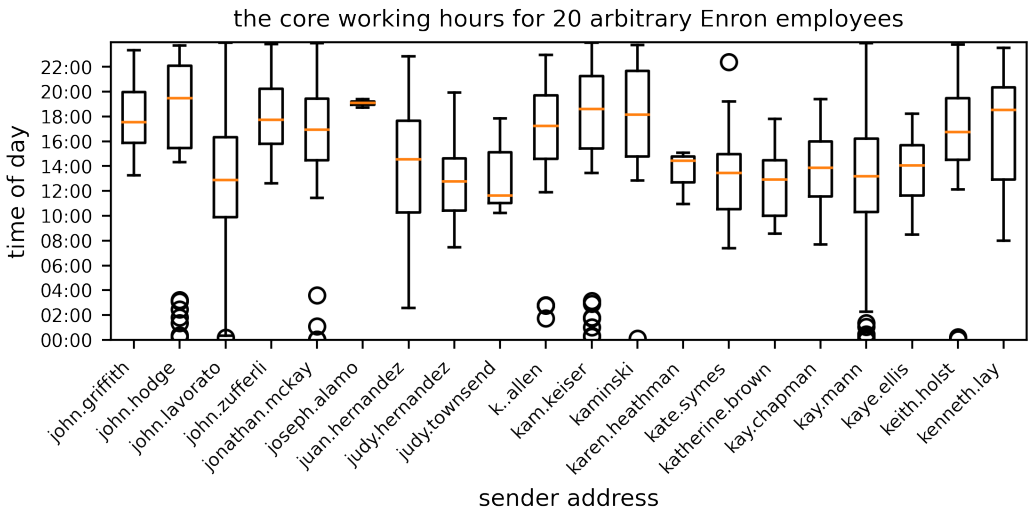

Figure 3: The mail communication for 20 arbitrary Enron employees, the box limits are the 25th and the 75th quantile.

or end their working day differently, e.g. by participating in daily meetings (no e-mail data at all) or by answering to e-mails of the last day (very much e-mail data). In general, we assumed that each working day has a well-defined start and end and did not consider more flexible arrangements (e.g. to watch over the kids during an extended lunch break). Also, there might be differences in the working hours that occur periodically (e.g. such as working less on Friday).

Secondly, the amount of e-mails sent by individuals depends in parts on the amount of e-mails they receive, i.e. an answer B to an e-mail A can naturally only be written after the initial e-mail A was received. We acknowledge this by looking at the average working time for all employees in each month in terms of the 25th and the 75th quantile.

Finally, we base our analysis on the assumption, that the time and date reconstruction from the original database export was performed consistently according to the appropriate RFC 5322 [6] and RFC 2821 [3], which might not have been the case.

References

[1] Anurat Chapanond, Mukkai S. Krishnamoorthy, and Bülent Yener. Graph theoretic and spectral analysis of enron email data. *Computational and Mathematical Organization Theory*, 11(3): 265–281, oct 2005. doi: 10.1007/s10588-005-5381-4.

[2] Paul M. Healy and Krishna G. Palepu. The fall of enron. *Journal of Economic Perspectives*, 17 (2):3–26, June 2003. doi: 10.1257/089533003765888403. URL https://www.aeaweb.org/articles?id=10.1257/089533003765888403.

[3] Dr. John C. Klensin. Simple Mail Transfer Protocol. RFC 2821, April 2001. URL https://www.rfc-editor.org/info/rfc2821.

[4] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

[5] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, 2004.

[6] Pete Resnick. Internet Message Format. RFC 5322, October 2008. URL https://www.rfc-editor.org/info/rfc5322.