Chapter 10
Engagingness and Responsiveness Behavior Models on the Enron Email Network and Its Application to Email Reply Order Prediction

Byung-Won On, Ee-Peng Lim, Jing Jiang, and Loo-Nin Teow

Abstract In email networks, user behaviors affect the way emails are sent and replied. While knowing these user behaviors can help to create more intelligent email services, there has not been much research into mining these behaviors. In this paper, we investigate user engagingness and responsiveness as two interaction behaviors that give us useful insights into how users email one another. Engaging users are those who can effectively solicit responses from other users. Responsive users are those who are willing to respond to other users. By modeling such behaviors, we are able to mine them and to identify engaging or responsive users. This paper proposes four types of models to quantify engagingness and responsiveness of users. These behaviors can be used as features in email reply order prediction, which predicts the email reply order given an email pair. Our experiments show that engagingness and responsiveness behavior features are more useful than other non-behavioral features in building a classifier for the email reply order prediction task. When combining behavior and non-behavior features, our classifier is also shown to predict the email reply order with good accuracy. This work was extended from the earlier conference paper that appeared in [9].

B.-W. On (✉)
Advanced Institutes of Convergence Technology, Seoul National University, 864-1 Iui-dong, Yeongtong-gu, Suwon-si Gyeonggi-do 443-270, Korea
e-mail: on.byung.won@gmail.com

E.-P. Lim · J. Jiang
School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902, Singapore
e-mail: eplim@smu.edu.sg; jingjiang@smu.edu.sg

L.-N. Teow
DSO National Laboratories, 20 Science Park Drive, Singapore 118230, Singapore
e-mail: tloonin@dso.org.sg

T. Özyer et al. (eds.), *The Influence of Technology on Social Network Analysis and Mining*, Lecture Notes in Social Networks 6, DOI 10.1007/978-3-7091-1346-2_10,

10.1 Introduction

10.1.1 Motivation

Electronic mail (also known as email) despite its long history has remained to be the most popular communication tool today. Unlike other newer communication tools such as weblog, twitter, messenger, etc., email has been widely adopted in the corporate world and often seamlessly integrated with business applications. As users email one another within and outside the corporate boundaries, they form different kinds of email networks. Within each network, users demonstrate behaviors that also affect how emails are sent and replied.

In this paper, we study user interaction behaviors in email networks and how they are relevant to predicting future email activities. An email network is essentially a directed graph with nodes and links representing users and messages from users to other users respectively. Each email is assigned a timestamp and has other attributes including sender, recipients, subject heading, and email content. We focus on two user interaction behaviors that are closely related to how users respond to one another in email networks, namely **engagingness** and **responsiveness**.

We define *engagingness* behavior as the ability of an user to solicit responses from other users, and *responsiveness* behavior as the willingness of an user to respond to other users. A user at the low (or high) extreme of engagingness behavior are known as to be disengaging (or engaging). Similarly, a user can range from unresponsive to highly responsive. As suggested by their definitions, user engagingness and responsiveness have direct or indirect implications on the way emails are sent and responded, and the strength of relationships users may have with other users in the networks. Nevertheless, these implications have not been well studied. The use of interaction behaviors to enhance email functions has been largely unexplored.

This paper therefore aims to provide a fresh approach towards modeling the engagingness and responsiveness behaviors in email networks. These models are quantitative and assign to each user an engagingness score and a responsiveness score. The scores are within the [0,1] such that 0 and 1 represent the lowest and highest scores respectively. With the scores, we can rank all users by engagingness or responsiveness. Moreover, we derive new features from these behavior scores and use them in an example email activity prediction, i.e., email reply order prediction.

The engagingness and responsiveness behavior models can be very useful in several applications. In the context of business organizations, they help to identify engaging and responsive users who may be good candidates for management roles, and to weed out lethargic users who are neither engaging and responsive making them the bottleneck in the organization. For informal social email networks, engaging and responsive users could be the high network potential candidates for viral marketing applications. Engaging users may solicit more responses for viral messages while responsive users may act fast on these messages. By selecting

these users to spread viral messages to targeted user segments by word-of-mouth, marketing objectives can be achieved more effectively.

In this paper, we specifically introduce the **email reply order prediction** task as an application, and show that engagingness and responsiveness behavior models contribute significantly to prediction accuracy. Email reply order prediction refers to deciding which of a pair of emails received by the same user will be replied first. This prediction task effectively helps an email recipient to prioritize his replies to emails. For example, if e_1 and e_2 are two emails sent to user u_k who plans to reply both. The outcome of prediction can either be e_1 replied before e_2 or vice versa. The ability to predict reply order of emails has several useful benefits, including helping users to prioritize emails to be replied, and to estimate the amount of time emails get replied. Here, our main purpose is to use the task to evaluate the utility of engagingness and responsiveness behavior models.

10.1.2 Research Objectives

This paper proposes to model engagingness and responsiveness behaviors quantitatively. In order to develop these quantitative behavior models, we first preprocess the emails so as to remove noises from the data and to construct the reply and forward relationships among emails. From the email relationships, we also derive email threads which are hierarchies of emails connected by reply and forward relationships. We then systematically develop a taxonomy of engagingness and responsiveness models using the reply relationships and email threads. These models are applied to the Enron email dataset, a publicly available dataset consisting of 517,431 emails from 151 ex-Enron employees. The email reply order prediction task is addressed as a classification problem. Our approach derives a set of features for a email pair based on the emails' metadata as well as engaging and responsive behaviors of their senders. As we evaluate the performance of the learnt prediction models, we would like to identify the interplay between behavior features and prediction accuracy. Our approach does not depend on email content or domain knowledge which are sometime not available and time costly to process. Given that there are only two possible order outcomes, we expect any method should have an accuracy of at least 50 %. In order for email reply order prediction to be useful, a much higher prediction accuracy is required without relying on content analysis.

Both behavior modeling and email reply order prediction are novel problems in email networks. Research on engagingness and responsiveness behaviors is a branch of social network analysis that studies node properties in a network. Unlike traditional social network analysis which focuses on node and network statistics based on static information (e.g., centralities, network diameter) of social networks, behavior analysis is conducted on networks with users dynamically interacting with one another.

In the following, we summarize the important research contributions of this paper.

- We define four of models for engagingness and responsiveness behaviors prevalent in email networks. They are (a) email based, (b) email thread based, (c) email sequence based, and (d) social cognitive model categories. For each model category, one can define different behavior models based on different email attributes. To the best of our knowledge, this is the first time engagingness and responsiveness behavior models are studied systematically.
- We apply our proposed behavior models on the Enron email network, analyze and compare the proposed behavior models. We conduct data preprocessing on the email data and establish links between emails and their replies. In our empirical study, we found engagingness and responsiveness are distinct from each other. Most engagingness (responsiveness) models of users are shown to be consistent with each other.
- We introduce email reply order prediction as a novel task that uses engagingness, responsiveness and other email features as input features. An SVM classifier is then learnt from the features of training email pairs and applied to test email pairs. According to our experimental results, the accuracy of our SVM classifier is about 77 % which is better than random guess (50 %). This indicates that user behaviors are useful in the prediction task.

Unlike most previous research on behavior analysis in email networks which focuses on mainly direct statistics of emails such as recipient list size, rate of emails from receiver to sender, and email size to characterize an email user [4, 13], our modeling of engagingness and responsiveness behaviors relies mainly on email reply and forward relationships not available directly in the email data. Previous research on email prediction tasks include the prediction of (a) social hierarchy of email users [12], (b) topics of emails [7], and (c) viral emails[13]. Email reply order prediction is thus a new task to be investigated. Although engagingness and responsiveness behaviors and reply order prediction task are defined in the context of email networks, our proposed approaches and results are also applicable to other form of information exchange networks such as messaging and blog networks.

10.1.3 Enron Email Dataset

Throughout this research, we use Enron email dataset in our empirical study of real data. This dataset is so far the only known publicly available email data with messages assigned with specific senders and recipients [6]. This dataset provides 517,431 emails for 151 Enron employees. Each email message has a unique message ID and contains header information such as the date and time when the message was sent, sender, recipients (To and Cc lists), subject and body in plain text format. We performed two data preprocessing steps on the email data, namely *duplicate elimination* and *email relationship identification*.

Duplicate elimination. As noticed by the previous studies on this corpus, there are many duplicate emails in either different folders of the same user (e.g., in computer generated folders such as all_documents, discussion_threads) or folders of different users (e.g., a message in sender's sent_mail folder often appears in some recipient's inbox or other folder). Message_IDs cannot be used to identify duplicate emails as such emails also have unique message IDs. We therefore use a strategy similar to [7] by computing the MD5 sum on email fields: Date/Time, Sender, To, Cc, Subject and Body. This will assign the same MD5 value (128 bit integer) to all duplicate emails that exactly match on these fields. After duplicate elimination, the dataset contains 257,044 unique emails.

Email relationship identification. To identify reply and forward relationships between emails, we first group all emails of each matching subject after ignoring the Reply and Forward prefixes (e.g., RE, FW, FWD, etc.) and order them by time. Each reply (or forward) email e_i in the group is then assigned a reply relationship (or forward relationship) with the most recent earlier email e_j such that e_i's sender is one of the e_j's recipients and that $t(e_i) - t(e_j) \leq 90$ days where $t(e)$ denotes the send time of e. With this approach,[1] we found 34,008 email relationship of which 27,730 and 6,278 are reply and forward relationships respectively. When a set of emails form a connected component by reply and forward relationships, we call it an email thread. From the email relationships identified, we derive 18,593 threads that connect 52,601 emails (about 20 % of all unique emails).

To evaluate our email relationship identification approach, we first compute precision that measures how many links are correct among the links detected by our method. For this, we selected a random sample of 100 link relationships from the total of 34,008 links. For every pair, we manually verified whether or not an email is sent and the other is the correct reply email. Our manual evaluation showed a precision of 91 %. To compute recall, we randomly selected 30 subject groups, each of which contains about five or ten emails. For each subject group, we manually created threads by connecting emails to their follow-up responses. This sample includes about 120 emails with 79 reply links and 21 forward links. For each link of two emails, we manually examined if the link is correctly found by our method. We found 79 % correct links which are actually present in the Enron dataset. This suggests that our identified relationships are quite accurate. In our subsequent experiments, we therefore use these identified email relationships.

10.1.4 Paper Outline

The remainder of this paper is organized as follows: In Sect. 10.2, we present engagingness and responsiveness behavior models. Subsequently, we discuss a

[1] We discuss the detailed description of this algorithm in Appendix.

challenging problem of predicting email reply order based on our behavior models in Sect. 10.3. The proposed models are evaluated and compared in Sect. 10.4 using a set of experiments on the Enron dataset. In particular, Sect. 10.4.3 describes experiments that evaluate the performance of email reply order prediction using different classifiers trained with different features including those based on email behaviors. In Sect. 10.5, we briefly introduce works related to our behavioral analysis problem. Finally, we offer our concluding remarks in Sect. 10.6.

10.2 Engagingness and Responsiveness Behavior Models

In this section, we describe our proposed behavior models for user engagingness and responsiveness. All the models assume that emails have been preprocessed as described in Sect. 10.1.3. We divide our models into the following categories:

- **Email based models**: These models consider emails as the basic data units for measuring user behaviors. Email attributes such as sender, recipient list, date, etc., are used.
- **Email thread based models**: These models consider email threads as the basic data units for measuring user behaviors. The models therefore use attributes of email thread to quantify behaviors.
- **Email sequence based models**: These models examine the sequence of emails received and replied by each user and derive the user behaviors from the gaps between emails received and their replies.
- **Social cognitive models**: These models consider social perception of user behaviors within the email network and measure behaviors accordingly.

Figure 10.1 shows the taxonomy of behavior models in the above categories to be further defined in the following sections. Each model (M) consists of a pair of engaging (E^M) and responsive (R^M) score formulas defined based on some principles. The E^M and R^M score values are in [0,1] range with 0 and 1 representing the lowest and highest values respectively. Table 10.1 shows a list of symbols and their meanings that we use in this paper.

10.2.1 Email Based Models

Email Count Model (EC)

The email count model is defined based on the principle that an engaging user should have most of his emails replied, while a responsive user should have most of his received emails replied. The engagingness and responsiveness formulas are thus defined by:

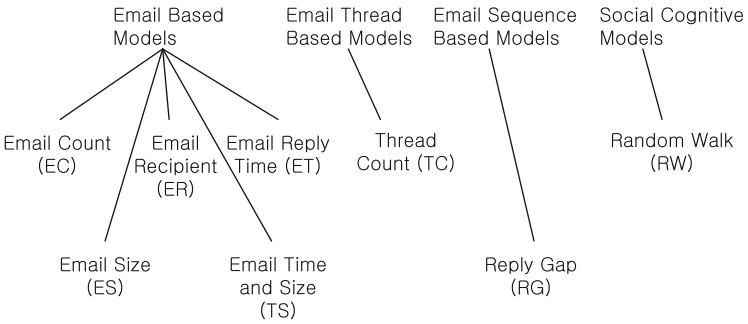

Fig. 10.1 Taxonomy of models

Table 10.1 Notations

$S(u_i)$	Emails sent by user u_i
$R(u_i)$	Emails received by u_i
$RB(u_i)$	Email replies sent by u_i
$RT(u_i)$	Emails replying to u_i's earlier emails
$SZ(e)$	Size of e added by the email sender excluding the forwarded content
$TH(u_i)$	Threads started by an email sent by u_i
$r(e)$	Reply to email e
$Sdr(e)$	Sender of email e
$Rcp(e)$	Recipients (in both To and Cc lists) of email e
$t(e)$	Sent time of email e
$E(u_i \rightarrow u_j)$	Emails from u_i to u_j
$E(u_i \leftrightarrow u_j)$	Emails between u_i and u_j
$rt(u_i \rightarrow u_j)$	Average response time from u_i to u_j
$rt(u_i \leftrightarrow u_j)$	Average response time between u_i and u_j
$RE(u_i \rightarrow u_j)$	Reply emails from u_i to u_j
$RE(u_i \leftrightarrow u_j)$	Reply emails between u_i and u_j

$$E^{EC}(u_i) = \frac{|RT(u_i)|}{|S(u_i)|} \tag{10.1}$$

$$R^{EC}(u_i) = \frac{|RB(u_i)|}{|R(u_i)|} \tag{10.2}$$

For users with empty $S(u_i)$ (or $R(u_i)$), $E^{EC}(u_i)$ (or $R^{EC}(u_i)$) is assigned a zero value.

Email Recipient Model (ER)

The intuition of this model is that an email with many recipients is likely to expect very few replies. Hence, an engaging user is one who gets replies from many recipients of his emails while an disengaging user receives very few or no reply

when his emails are sent to many recipients. On the other hand, a responsive user is one who replies emails regardless of the number of recipients in the emails. A non-responsive user is one who does not reply even if the emails are directed to him only. The engagingness and responsiveness formulas are thus defined by:

$$E^{ER}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{|\{u_j \in Rcp(e) \wedge r(e) \in RB(u_j)\}|}{|Rcp(e)|} \tag{10.3}$$

$$R^{ER}(u_i) = \frac{1}{|R(u_i)|} \sum_{\substack{e \in RB(u_i) \ s.t. \\ \exists u_j, \exists e'' \in S(u_j), r(e'')=e}} \frac{|Rcp(e)|}{MaxRcpCnt} \tag{10.4}$$

where *MaxRcpCnt* (=291) denotes the largest recipient count among all Enron emails.

Email Reply Time Model (ET)

The reply time of an email can be an indicator of user engagingness and responsiveness. The email reply time model adopts the principle that engaging users receive the reply emails sooner than non-engaging users, while responsive users reply to the received emails quicker than non-responsive users.

Given an email e' which is a reply of email e, $e' = r(e)$, the *reply time* of e', $Rpt(e') = t(e') - t(e)$. The z-normalized reply time $\hat{Rpt}(e')$ is defined by $\frac{Rpt(e') - \overline{Rpt}}{\sigma_{Rpt}}$ where $\overline{Rpt}$ and σ_{Rpt} are the mean and standard deviation of reply time respectively. Now, we define the engagingness and responsiveness of ET model as:

$$E^{ET}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{1}{|Rcp(e)|} \sum_{\substack{u_j \in Rcp(e), \\ \exists e' \in RB(u_j), e'=r(e)}} (1 - f(\hat{Rpt}(e'))) \tag{10.5}$$

$$R^{ET}(u_i) = \frac{1}{|R(u_i)|} \sum_{e' \in RB(u_i), e \in R(u_i), r(e)=e'} (1 - f(\hat{Rpt}(e'))) \tag{10.6}$$

where

$$f(x) = \frac{1}{1 + e^{-x}} \tag{10.7}$$

The function $f()$ is designed to convert the normalized reply time to the range [0,1] with 0 and 1 representing extreme slow and extreme fast reply times respectively.

Email Size Model (ES)

The email size model is analogous to the email reply time model except that we take the content size of emails into account rather than the reply time of emails.

The principle behind this model is based on the size of reply email roughly representing the amount of a recipient's effort. For instance, let us assume $SZ(e) = k$ and e' is a reply email of e. If $SZ(e') > k$, the engagingness score of the sender of e will be high. The amount of content in a reply email can be used to measure the amount of eagerness of the user sending the reply email. Let $\hat{SZ}(e)$ be the z-normalized $SZ(e)$. We then develop the engagingness and responsiveness measures based on email size as

$$E^{ES}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{1}{|Rcp(e)|} \sum_{\substack{u_j \in Rcp(e), \\ \exists e' \in RB(u_j), e' = r(e)}} f(\hat{SZ}(e')) \tag{10.8}$$

$$R^{ES}(u_i) = \frac{1}{|R(u_i)|} \sum_{e' \in RB(u_i), e \in R(u_i), r(e) = e'} f(\hat{SZ}(e')) \tag{10.9}$$

Email Time and Size Model (TS)

This model combines both email reply time and size into a hybrid model as

$$E^{TS}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{1}{|Rcp(e)|} \sum_{\substack{u_j \in Rcp(e), \\ \exists e' \in RB(u_j), e' = r(e)}} (1 - f(\hat{Rpt}(e'))) f(\hat{SZ}(e')) \tag{10.10}$$

$$R^{TS}(u_i) = \frac{1}{|R(u_i)|} \sum_{e' \in RB(u_i), e \in R(u_i), r(e) = e'} (1 - f(\hat{Rpt}(e'))) f(\hat{SZ}(e')) \tag{10.11}$$

Examples

To illustrate the behavior models in Sect. 10.2.1, suppose a simple email network as shown in Fig. 10.2. In Fig. 10.2a, u_i is a sender, while u_1, u_2, and u_3 are recipients. u_i sends email e_1 to u_1 and u_2, and then another email e_1' is replied by u_1. However, u_2 does not respond to u_i. In the email network, the engagingness score of the user u_i is calculated as $E^{EC}(u_i) = \frac{3}{5} = 0.6$ and $E^{ER}(u_i) = \frac{\{\frac{1}{2} + \frac{2}{3}\}}{5} = 0.23$. In Fig. 10.2b, u_2, u_4, and u_i are recipients, whereas u_1 and u_3 are senders. u_3 sends email e_2 to u_2 and u_i. While u_2 does not reply to u_3, u_i replies to u_3 and to u_2 as Cc. In the figure, the responsiveness score of the user u_i is measured as $R^{EC}(u_i) = \frac{2}{2} = 1$ and $R^{ER}(u_i) = \frac{\{\frac{1}{4} + \frac{2}{4}\}}{2} = 0.38$, where we assume $MaxRcpCnt = 4$. In particular,

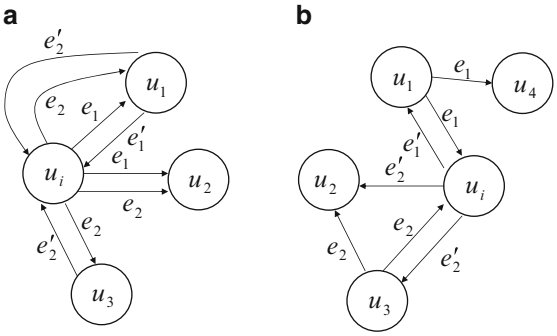

Fig. 10.2 An example for email based models. (**a**) Engagingness of u_i. (**b**) Responsiveness of u_i

Table 10.2 Distribution of reply times in the Enron email dataset

Reply time up to	1 h	5 h	10 h	1 day	5 days	10 days	90 days
# of email pairs	2,100	5,029	5,854	8,405	11,569	13,810	19,167

we order emails by the number of recipients in the ascending order, and then assign to *MaxRcpCnt* the number of recipients in an email at the 90 percentile. Recall that the email count model is a *macro* approach, while the email recipient model is a *micro* approach.

In the email reply time model, $E^{ET}(u_i) = \frac{\frac{1}{2}\cdot(x_1+\frac{1}{\infty})+\frac{2}{3}\cdot(x_2+\frac{1}{\infty}+x_3)}{2}$, where we compute x_i as follows:

- $x_1 = 1 - f(Rpt(t(e'_1(u_1,\{u_i\})) - t(e_1(u_i,\{u_1\})) = 5\,\text{s})) = 1 - 0.29 = 0.71$
- $x_2 = 1 - f(Rpt(t(e'_2(u_1,\{u_i\})) - t(e_2(u_i,\{u_1\})) = 10\,\text{s})) = 1 - 0.45 = 0.55$
- $x_3 = 1 - f(Rpt(t(e'_2(u_3,\{u_i\})) - t(e_2(u_i,\{u_3\})) = 20\,\text{s})) = 1 - 0.75 = 0.25$

where $e_v(u_x,\{U_y\})$ denotes email e_v sent by u_x to recipients U_y and e'_v denotes the reply of email e_v. To compute function $f(\hat{Rpt})$, we transform Rpt to z-scores. For instance, the z-score of $Rpt(t(e'_1(u_1,\{u_i\})) - t(e_1(u_i,\{u_1\})) = 5\,\text{s}) = \frac{5\,\text{s}-\bar{x}}{\sigma} = \frac{5-11.67}{7.64} = -0.87$, where $\bar{x}$ and σ denote the mean and standard deviation of reply times. According to our observation on reply times of Enron emails (see Table 10.2), the mean of reply times is much larger than the median. This indicates there are many outliers of reply times, and further most z scores can be negative. Thus we remove extreme reply times prior to computing z-scores. Then, $f(-0.87) = \frac{1}{1+e^{-(-0.87)}} = 0.29$. In particular, the term $\frac{1}{\infty}$ in $E^{ET}(u_i)$ indicates that u_i sends e_1 and e_2 to u_2 but u_2 does not reply to u_i. As a result, $E^{ET}(u_i) = \frac{\frac{1}{2}\cdot(0.71+0)+\frac{2}{3}\cdot(0.55+0+0.25)}{2} = 0.45$. The responsiveness of u_i is calculated in the same manner. In addition, the email size model computes engagingness scores in the same manner except that the length of email content is considered instead of the reply time of emails.

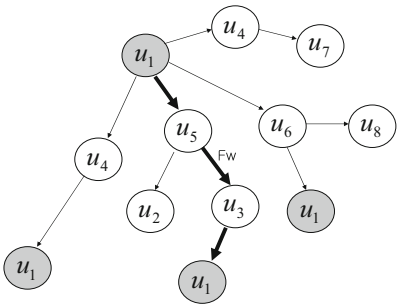

Fig. 10.3 An email thread example

Table 10.3 Distribution of emails per thread in the Enron email dataset

# emails	2	3	4	5	6	≥7	Total
# threads	11,302	3,925	1,614	732	404	616	18,593

10.2.2 Email Thread Based Models

Here, we define the **thread count model (TC)** as an email thread based model. In the email count model, engagingness is measured by emails sent by a sender and sent emails directly replied by some recipient(s). However, direct reply is not the only type of response to an email. Email may be indirectly replied in email threads due to forwarded emails. For example, as illustrated in Fig. 10.3, user u_1 advertises a job position by sending an email to professor u_5 who subsequently forwards it to his student u_3. If u_3 replies to u_1, we say that the original email is replied indirectly in an email thread.

Email thread is defined by a tree of emails connected by reply and forward relationships. Table 10.3 shows the distribution of threads by the number of emails per thread. As we can notice, the distribution follows Zipf's law. Majority of threads (11,302) contain only two emails. There are 3,925 threads that include three emails. The largest thread contains 37 emails.

Based on email threads, the thread count model includes indirect replies to emails forwarded between users using the principle: the user is highly engaging if he receives many of his emails replied directly or indirectly by recipients, and is highly responsive if he replies or forwards most emails earlier received. In the following, the engagingness and responsiveness of a user u_i are defined as:

$$E^{TC}(u_i) = \frac{1}{|S(u_i)|}|\{e \in S(u_i)|\exists t \in TH(u_i), \exists e', e \underset{t}{\twoheadrightarrow} e' \wedge u_i \in Rcp(e')\}| \tag{10.12}$$

$$R^{TC}(u_i) = \frac{1}{|R(u_i)|}|\{e \in R(u_i)|\exists u_j, e', t \in TH(u_j), e \underset{t}{\twoheadrightarrow} e' \wedge u_j \in Rcp(e')\}| \tag{10.13}$$

where $e \twoheadrightarrow_t e'$ returns TRUE when e is directly or indirectly connected to e' in the thread t, and FALSE otherwise.

10.2.3 Email Sequence Based Models

Email sequence refers to the sequence of emails sent and received by a user ordered by time. To derive engagingness and responsiveness from email sequences, we consider the principle that an engaging user is expected to have his sent emails replied soon after they are received by the email recipients, and a responsive user replies soon after they receive emails. As users may not always stay online, the time taken to reply an email may vary very much. Instead, we consider the number of emails received later than an email e but are replied before e by a user as a proxy of how soon e is replied.

The above principle is thus used to develop the **reply gap model (RG)**. Let seq_i denote the email sequence of user u_i. When an email received by u_i is replied before other email(s) received earlier, the reply of the former is known as an *out-of-order reply*. Formally, for an email e received by u_i, we define the *number of emails received* and *number of out-of-order replies* between e and its reply e' in seq_i, denoted by $n_r(u_i, e)$ and $n_{\overline{o}}(u_i, e)$ respectively, as

$$n_r(u_i, e) = \begin{cases} \text{\# emails received between } e \text{ and } e' \text{ in } seq_i, & \text{if } \exists e' \in RB(u_i), \\ & r(e) = e' \\ -1, & \text{otherwise} \end{cases} \tag{10.14}$$

$$n_{\overline{o}}(u_i, e) = \begin{cases} \text{\# emails received between } e \text{ and } e' \text{ in } seq_i \\ \text{and have been replied,} \\ -1, & \text{otherwise} \end{cases} \quad \begin{array}{l} \text{if } \exists e' \in RB(u_i), \\ r(e) = e' \end{array} \tag{10.15}$$

The -1 value is assigned to n_r and $n_{\overline{o}}$ when e is not replied at all. The user engagingness and responsiveness of the RG model are thus defined as:

$$E^{RG}(u_i) = \frac{\sum_{e \in S(u_i)} \left(\frac{1}{|Rcp(e)|} \sum_{u_j \in Rcp(e)} \left(1 - \frac{n_{\overline{o}}(u_j, e)}{n_r(u_j, e)}\right)\right)}{|S(u_i)|} \tag{10.16}$$

$$R^{RG}(u_i) = \frac{\sum_{e \in R(u_i)} \left(1 - \frac{n_{\overline{o}}(u_i, e)}{n_r(u_i, e)}\right)}{|R(u_i)|} \tag{10.17}$$

For example, let $seq_i = \{e_1, e_2, e_3, e_4, e'_1, e'_4, e'_2\}$ be the email sequence of user u_i where $e'_k = r(e_k)$'s. Note that $\frac{n_{\overline{o}}(u_i, e_1)}{n_r(u_i, e_1)}$, $\frac{n_{\overline{o}}(u_i, e_2)}{n_r(u_i, e_2)}$, $\frac{n_{\overline{o}}(u_i, e_3)}{n_r(u_i, e_3)}$, and $\frac{n_{\overline{o}}(u_i, e_4)}{n_r(u_i, e_4)}$ are $\frac{0}{3}$, $\frac{1}{2}$, $\frac{-1}{-1}$, and 0 respectively. Hence, $R^{RG}(u_i) = \frac{\{(1-\frac{0}{3})+(1-\frac{1}{2})+(1-\frac{-1}{-1})+(1-0)\}}{4} = 0.625$. The engagingness of u_i can be computed in the same manner.

10.2.4 Social Cognitive Models

A social cognitive model is based on *social cognitive theory* which suggests that people learn by watching what others do [8]. Such kind of models thus measure a user's engagingness and responsiveness behaviors by observing what the other users react to emails sent from the user and observe the email interaction among one another. In this paper, we introduce a **random walk (RW)** social cognitive model.

For engagingness, each user u_k perceives a user u_i to be more engaging than another user u_j if more emails from u_i are replied ahead of emails from u_j based on the emails in the mailbox of u_k. For instance, suppose that u_k has an email sequence $seq_k = \langle e_1(u_1, \{u_k\}), e_2(u_2, \{u_k\}), e'_2(u_k, \{u_2\}), e'_1(u_k, \{u_1\})\rangle$, where $e_v(u_x, \{U_y\})$ denotes email e_v sent by u_x to recipients U_y and e'_v denotes the reply of email e_v. u_k receives e_1 before e_2 but the reply e'_1 comes after e'_2. This indicates that u_k considers u_2 more important than u_1. Furthermore, u_2 is more engaging than u_1 from u_k's standpoint. Based on the above observation, we say that u_k observes the engagingness superiority of u_2 over u_1.

Similarly for responsiveness, u_k perceives a user u_1 to be more responsive than another user u_2 if u_k observes reply emails from u_1 earlier than u_2 for the same emails sent to both u_1 and u_2 which can be from u_k or other users.

Formally, we represent an **engagingness weighted directed graph** $G^E = \langle U, L^E\rangle$ as follows:

- U represents the set of all users.
- L consists of directed edges. When in the mailbox of some u_k, u_i has x_k emails replied ahead of emails from u_j, we represent this by a directed edge $u_j \rightarrow u_i$.
- The weight of $u_j \rightarrow u_i$, $weight(u_j \rightarrow u_i)$, is the sum of x_k's for all u_k's. The larger is $weight(u_j \rightarrow u_i)$, the more users observe that u_i is more engaging than u_j.

In a similar manner, we can define a **responsiveness weighted directed graph** $G^R = \langle U, L^R\rangle$.

The engagingness (responsiveness) weighted directed graph will be further processed to derive the degree of engagingness (responsiveness) of users. Each directed graph so far captures the perceived relative difference between users in engagingness (responsiveness). It however does not immediately assign engagingness (responsiveness) scores to the users. We therefore propose to perform random walk on the engagingness (responsiveness) graph so as to determine the user engagingness (responsiveness) values as the stationary probabilities of visiting them.

The random walk process on the engagingness graph to obtain the engagingness of users denoted by $E^{RW}(u_k)$'s consists of the following steps:

1. Determine the largest node aggregated edge weight, $MaxWeight = Max_{u_j}\{\sum_{u_i} weight(u_j \rightarrow u_i)\}$

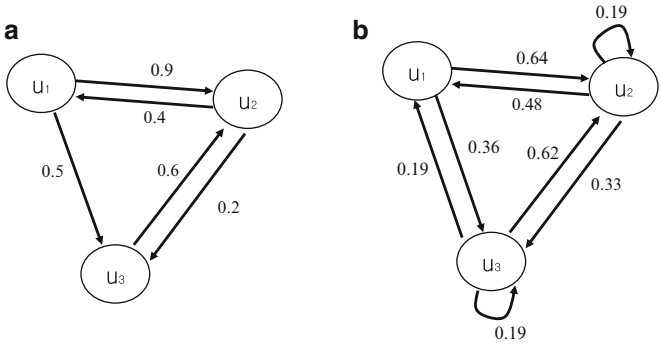

Fig. 10.4 Social cognitive model. (**a**) Engagingness weighted directed graph G^E. (**b**) Engagingness graph for random walks

2. For each user u_j,
   (a) $sum_j = 0$
   (b) For each edge $u_j \rightarrow u_i$,
      (i) Assign a transition probability to $u_j \rightarrow u_i$ as $p(u_j, u_i) = \frac{weight(u_j \rightarrow u_i)}{MaxWeight}$
      (ii) $sum_j = sum_j + p(u_j, u_i)$
   (c) Assign to the remaining weights to all users.
   Create an edge $u_j \rightarrow u_t$ for all u_t with $p(u_j, u_t) = \frac{1-sum_j}{|U|}$ if $u_j \rightarrow u_t$ does not exist;
   Increment $p(u_j, u_t)$ by $\frac{1-sum_j}{|U|}$ otherwise
3. For each user u_i, initialize $E^{RW}_{new}(u_i)$ randomly
4. Repeat the following steps:
   (a) For each u_i, $E^{RW}(u_i) = E^{RW}_{new}(u_i)$
   (b) For each u_i, $E^{RW}_{new}(u_i) = \sum_{u_j \rightarrow u_i} p(u_j, u_i) \cdot E^{RW}(u_j)$
5. Until $|E^{RW}(u_i) - E^{RW}_{new}(u_i)| \leq \epsilon$[2] for all u_i's

To illustrate the above algorithm, consider examples in Fig. 10.4. u_2 is more engaging than u_1 by $weight(u_1 \rightarrow u_2) = 0.9$. On the other hand, u_1 is less engaging than u_2 by $weight(u_2 \rightarrow u_1) = 0.4$. In Fig. 10.4a, the total engagingness weight of u_1 to all nodes u_2 and u_3 in G^E is $weight(u_i) = weight(u_1 \rightarrow u_2) + weight(u_1 \rightarrow u_3) = 1.4$. Similarly, the engagingness weight of u_2 and u_3 are 0.6 and 0.6 respectively. Then, the weight value of each edge is normalized by the maximum weight value, $MaxW = weight(u_1)$. For example, $weight(u_2 \rightarrow u_3) =$

[2]In our experiment, we used $\epsilon = .0000001$ and numbers of iterations required to compute E^{RW} and R^{RW} are 8 and 12 respectively.

$\frac{weight(u_2 \to u_3)}{MaxW} = \frac{0.2}{1.4}$. For nodes with total weight <1, the unused weight will be used to create edges with equal weights to all the nodes. For example u_2, it has unused weight of $\frac{\{MaxW - weight(u_2)\}}{weight(u_1)} = \frac{\{1.4-0.6\}}{1.4}$. As a result of the new edges for the unused weight, $weight(u_2 \to u_3) = \frac{0.2}{1.4} + \frac{\{1.4-0.6\}}{1.4} \cdot \frac{1}{3} = 0.33$. In this process, the engagingness graph is row-stochastic because its rows are nonnegative and the sum of each row is 1. This stochastic matrix can be viewed as a transition matrix associated to a family of Markov chains, where each entry (u_i,u_j) represents the probability of a transition from state u_i to state u_j.

10.3 Email Reply Order Prediction

We now consider the email reply order prediction which has the following setup. Given a pair of emails (e_i, e_j) sent to the same user (u) from users u_i and u_j respectively, we want to determine the order in which the two emails will be replied. Here, we assume that both e_i and e_j require some replies and u_i and u_j are not the same person. The outcome of prediction is either e_i or e_j first.

Our proposed method is to train a Support Vector Machine (SVM) classifier using labeled email pairs, and to apply the trained classifier on unseen email pairs. For each email pair, we can derive features directly from the emails themselves and their senders including the previous emails they have sent and received. There are three types of features used, namely: (a) *comparative email features* ($\mathbb{E}$), (b) *comparative interaction features* ($\mathbb{I}$) and (c) *comparative behavior features* ($\mathbb{B}$).

Table 10.4 lists the email features used in our classifier. For each email feature f_k, we derive a corresponding comparative feature f_k^c of an email pair (e_i, e_j) by

$$(e_i, e_j).f_k^c = e_i.f_k - e_j.f_k.$$

For email send time $t(e)$ feature, we further convert the positive and negative comparative feature values to 1 and -1 respectively. Interaction features refer to set of features derived from the sender of the email to the common recipient u_r as shown in Table 10.5. In the following sections, we will discuss the non-behavior features in more depth. The behavior features refer to the eight E^M and eight R^M behavior scores of email senders. The comparative interaction and behavior features are defined similar to that of email features.

For instance, we formulate our email reply order prediction as a binary classification problem. Each email pair is assigned a class label such that

$$\begin{cases} \text{Class} = -1 \text{ if } t(e_i) - t(e_j) < 0 \\ \text{Class} = 1 \quad \text{if } t(e_i) - t(e_j) > 0 \end{cases}.$$

The class label stands for u's preference to reply e_i before or after e_j with the assumption that e_i and e_j are received and replied by u. Now we suppose that $E^{ET}(u_i) = 0.8$ and $E^{ET}(u_j) = 0.4$. If we consider E^{ET} to be a feature (f_1),

Table 10.4 Email features $\mathbb{E}$

No	Description	No	Description
1	$t(e)$	9	$\lvert S(Sdr(e))\rvert$
2	$size(e)$	10	$\lvert R(Sdr(e))\rvert$
3	$size(r(e))$ (assuming we can determine the reply)	11	Avg. $\lvert S(Sdr(e))\rvert$ per day
		12	Avg. $\lvert R(Sdr(e))\rvert$ per day
4	$size(e) + size(r(e))$	13	$\frac{\lvert RB(Sdr(e))\rvert}{\lvert S(Sdr(e))\rvert}$
5	$Rcp(e)$	14	$\frac{\lvert RT(Sdr(e))\rvert}{\lvert R(Sdr(e))\rvert}$
6	$indegee(Sdr(e))$ (# users sending emails to $Sdr(e)$)	15	$\frac{\lvert RT(Sdr(e))\rvert}{\lvert S(Sdr(e))\rvert}$
		16	$\frac{\lvert RB(Sdr(e))\rvert}{\lvert R(Sdr(e))\rvert}$
7	$outdegee(Sdr(e))$ (# users receiving emails from $Sdr(e)$)	17	Avg response time for emails in $RT(Sdr(e))$
8	$indegree(Sdr(e)) + outdegree(Sdr(e))$	18	Avg response time for emails in $RB(Sdr(e))$

Table 10.5 Interaction features $\mathbb{I}$

No	Description	No	Description
19	$\lvert E(Sdr(e) \rightarrow u_r)\rvert$	27	$\frac{\lvert RE(Sdr(e) \leftrightarrow u_r)\rvert}{\lvert E(u_r \leftrightarrow Sdr(e))\rvert}$
20	$\lvert E(u_r \rightarrow (Sdr(e))\rvert$	28	$rt((Sdr(e) \rightarrow u_r)$
21	$\lvert E((Sdr(e) \leftrightarrow u_r)\rvert$	29	$rt(u_r \rightarrow (Sdr(e))$
22	$\lvert RE((Sdr(e) \rightarrow u_r)\rvert$	30	# threads involving $(Sdr(e)$, u_j as senders/recipients
23	$\lvert RE(u_r \rightarrow (Sdr(e))\rvert$		
24	$\lvert RE((Sdr(e) \leftrightarrow u_r)\rvert$	31	# threads involving $(Sdr(e)$, u_r as senders
25	$\frac{\lvert RE((Sdr(e) \rightarrow u_r)\rvert}{\lvert E(u_r \rightarrow (Sdr(e))\rvert}$		
26	$\frac{\lvert RE(u_r \rightarrow (Sdr(e))\rvert}{\lvert E((Sdr(e) \rightarrow u_r)\rvert}$		

the comparative feature of f_1 is $E^{ET}(u_i) - E^{ET}(u_j) = 0.8–0.4 = 0.4$. Furthermore, if we suppose $t(e_i) - t(e_j) < 0$, the feature vector used in SVM can be represented as $\{-1\ f_1{:}0.4 \ldots\}$.

10.3.1 Non-behavior Features

10.3.1.1 Email Features $\mathbb{E}$

As shown in Table 10.4, Feature No. 1 represents the order in which emails e_i and e_j are received. For simplicity, let us denote by f_1 the feature value. $f_1 = -1$ if e_i arrived before e_j. $f_1 = 1$ if e_i arrived after e_j. $f_1 = 0$ if e_i and e_j arrived at the same time. Feature No. 2–4 measures the amount of effort required by a replier in terms of reading the content of a received email and writing the content of a reply email. The size of an email e represents the reading effort, while the size of the reply email $r(e)$ stands for the writing effort. Feature No. 5 counts the number of recipients in an email e, based on the fact that an email sent to many recipients is unlikely to be

replied. Feature No. 6–8 measure indegree, outdegree, and total degree, respectively. Given a sender $Sdr(e)$ in an email e, the indegree of the sender is the number of users who send emails to the sender. The outdegree of the sender is the number of neighbors who receive emails from the sender. The total degree of the sender is the total number of users who exchange emails with the sender. Feature No. 9 and 10 are the total number of emails sent or received by a user. On the other hand, Feature No. 11 and 12 are the average number of emails sent or received by a user per day. Feature No. 13 and 14 estimate the proportion of reply emails in a user's sent and received emails. On the other hand, Feature No. 15 and 16 compute the proportion of emails that a user replies or receives a reply for. Feature No. 17 and 18 represent the average response time for the reply emails sent or received by a user.

10.3.1.2 Interaction Features $\mathbb{I}$

Recall the framework of our email reply order prediction task, where u receives the emails from u_i and u_j, and then u will reply to either e_i or e_j first. Feature No. 19 counts the number of emails from u_i to u. We expect that u is likely to reply to e_i earlier than e_j if u_i usually sends more emails to u than u_j does. Similarly, Feature No. 20 counts the number of emails from u to u_i. Feature No. 21 counts the total number of emails exchanged between u_i and u. Feature No. 22 counts the number of reply emails exchanged between u and u_i and the total number of emails from u_i to u. Feature No. 23 counts the number of reply emails from u to u_i. Similarly, Feature No. 24 counts the total number of reply emails exchanged between u and u_i. Feature No. 25 estimates the proportion of emails replied. Feature No. 26 computes the proportion of emails replied out of the emails sent by u to u_i. We expect that u is likely to quickly reply to u_i who also responds to most of the emails received from u. Feature No. 27 measures the ratio of the total number of replies by the total number of emails exchanged between u and u_i. Feature No. 28 computes the average response time over all reply emails from u_i to u. Feature No. 29 also computes the average response time from u to u_i. Feature No. 30 counts the number of threads shared between u_i and u. It is because users who are involved in many threads are likely to be co-workers. Feature No. 31 counts the threads in which u and u_i actively participate. We define active participants to users who send at least one email in a thread.

10.4 Empirical Study

10.4.1 Set-Up

Dataset

For our task, we used the Enron email dataset that is publicly available at http://www.cs.cmu.edu/∼enron. This dataset provides 517,431 emails for 151 Enron employees. Each email message contains a unique message_ID, header

information such as the date and time when the message was sent/received, sender, recipients (To and Cc lists), subject and body in plain text format.

Using the email thread assembly algorithm (please see Appendix), we created a link database that stores a pair of emails linked via Reply or Forward relationships. The database consists of 34,008 links which includes 27,730 Reply links and 6,278 Forward links. These binary links make up a total of 18,593 threads that connect 52,601 emails.

Data Characteristics

We have conducted some analysis on the preprocessed email dataset to derive some statistics of Enron employees using and replying/forwarding emails. The interesting findings obtained include:

1. 52.6 K emails are involved in some threads.
2. Large majority (>90 %) of 18.5 K threads are short with two email messages each.
3. Large majority of threads last for at most 1 day.
4. Large majority of emails are replied within a day.
5. User response time is correlated with number of emails received, number of users he emails to, and number of users emailing him.

Evaluation Metric

To validate the effectiveness of our proposed models, note that we are not able to perform a direct evaluation on our behavior models because the ground truth is absent in the Enron dataset. Instead, we indirectly evaluate them, comparing the four types of behavior models on Enron dataset. To compare the ranked user lists produced by two models, we utilize the **Kendall τ distance measure**. In each ranked list, first and last ranked users represent the most and least engaging (or responsive) users respectively. Formally, we denote the rank of a user u_i in a ranked list L_k by $l_k(u_i)$. The Kendall τ distance between two ranked lists L_1 and L_2 is defined as $\frac{K(L_1,L_2)}{\frac{1}{2}n(n-1)}$ such that $K(L_1, L_2) = |(u_i, u_j) : u_i < u_j, (l_1(u_i) < l_1(u_j) \wedge l_2(u_i) > l_2(u_j)) \vee (l_1(u_i) > l_1(u_j) \wedge l_2(u_i) < l_2(u_j))|$. Note that Kendall τ distance is 0 if $l_1 = l_2$ for all users, and 1 if there is no correlation between l_1 and l_2 [3,5].

10.4.2 Analyzing Behavioral Models

Correlation Between Engagingness and Responsiveness

We first show the correlation between engagingness and responsiveness in our proposed models. Table 10.6 illustrates the Kendall τ distance between two lists,

Table 10.6 Kendall τ distance between engagingness and responsiveness

Behavior model (M)	$\tau(E^M, R^M)$
M = EC	0.46
M = ER	0.52
M = ET	0.49
M = ES	0.47
M = TS	0.48
M = TC	0.46
M = RG	0.5
M = RW	0.11

where the Enron employees in one list are ordered by engagingness scores and the same employees in the other list are ordered by responsiveness scores in each model. By definition, if the Kendall τ distance is 0, the two lists stand for perfect match, while there is no correlation between the two lists in case of the Kendall τ distance is 1. Interestingly, our proposed models show that most τ distances range in between 0.4 and 0.5. These results indicate that engaging employees are not necessarily the same as responsive employees in the Enron email data.

Correlation Between Different Models

Tables 10.7 and 10.8 show the correlations of pairs of different models by engagingness and responsiveness respectively. For instance, we calculate the Kendall τ distance between two lists, where employees in one list are ordered by E^{EC} and the same employees in the other list are ordered by E^{ER}. The Kendall τ distance between E^{EC} and E^{ER} is 0.14 as shown in Table 10.7. In particular, our proposed models are more correlated by responsiveness rather than by engagingness. The email based models such as ER, ET, ES, and TS are highly correlated in both engagingness and responsiveness. On the other hand, the social cognitive approach is not highly correlated with the other models. For example, the Kendall τ distances between RW and the other models are 0.26 on average, while the distances between other models are considerably small. According to the results in Tables 10.7 and 10.8, the social cognitive approach shows low correlation with the other models. For example the Kendall τ distance between E^{ES} and E^{RW} is 0.24 and the Kendall τ distance between R^{RG} and R^{RW} is 0.27. In the social cognitive approach, each user u_k perceives a user u_i to be more engaging than another user u_j if more emails from u_i are replied ahead of emails from u_j based on the emails in the mailbox of u_k. Our further investigation reveals that most emails tend to be replied by the last-in-first-out principle. While some users may reply emails in the same order as they arrive (follow the first-in-first-out), most users exhibit a strong *recency bias* towards more recently received emails that appear higher in the inbox. Indeed, there are a few emails from u_i which are replied ahead of emails from u_j based on the emails in the mailbox of u_k. For instance, let us present that Sean Crandall (u_k), Fran Chang (u_i), and Alan Comnes (u_j) in the Enron dataset. u_i has a set of replied emails $\{e'_{126}, e'_{127}, e'_{15,126}, e'_{15,129}, e'_{15,456}, e'_{15,457}, e'_{15,458}, e'_{15,459}, e'_{27,518}\}$, where subscripts stand for

Table 10.7 Kendall τ distance between two models by engagingness

	E^{ER}	E^{ET}	E^{ES}	E^{TS}	E^{TC}	E^{RG}	E^{RW}
E^{EC}	0.14	0.16	0.18	0.2	0.01	0.18	0.22
E^{ER}		0.12	0.14	0.16	0.14	0.15	0.24
E^{ET}			0.19	0.13	0.16	0.15	0.22
E^{ES}				0.09	0.18	0.19	0.24
E^{TS}					0.19	0.18	0.24
E^{TC}						0.18	0.22
E^{RG}							0.24

Table 10.8 Kendall τ distance between two models by responsiveness

	R^{ER}	R^{ET}	R^{ES}	R^{TS}	R^{TC}	R^{RG}	R^{RW}
R^{EC}	0.06	0.03	0.03	0.04	0.01	0.03	0.26
R^{ER}		0.07	0.07	0.07	0.06	0.08	0.25
R^{ET}			0.05	0.03	0.03	0.03	0.26
R^{ES}				0.03	0.03	0.05	0.26
R^{TS}					0.05	0.05	0.26
R^{TC}						0.03	0.26
R^{RG}							0.27

email ID. Similarly, u_j has a set of replied emails $\{e'_{400}, e'_{3,065}, e'_{9,321}, e'_{12,248}, e'_{17,495}, e'_{19,143}, e'_{19,144}, e'_{19,672}\}$. Then, u_k has a sequence of emails ordered by time $\{e_{9,321}, e_{19,675}, e'_{19,672}, e_{15,126}, e_{15,129}, e'_{15,129}, e_{126}, e'_{126}, e_{127}, e'_{127}, e_{19,144}, e'_{19,144}, e_{19,143}, e'_{19,143}, e_{400}, e_{15,495}, e_{3,065}, e'_{3,065}, e_{27,518}, e'_{27,518}, e_{15,457}, e_{15,456}, e_{15,458}, e_{12,248}, e_{17,495}\}$. Some emails such as $e_{15,129}$ and e_{126} are replied right after the email arrives at a recipient. On the other hand, for each replied email of Alan Comnes (e.g., $e_{3,065}$), there is no emails from Fran Chang that comes before an email from Alan Comnes and replies after the reply to Alan Comnes.

Interestingly, the email thread based model shows similar result to that of the email count model regardless of engagingness or responsiveness. This is because there are fewer number of forwarded emails among 151 Enron employees. For instance, from our email thread assembly, we obtained 7,291 email threads, each of which has more than or equal to three emails. In addition, we observed that an email sent by a sender is forwarded by recipients, and the sender finally receives reply emails by not the recipients but some other users. The total number of such emails is 313 among 151 Enron employees. For E^{TC}, only one thread contains eight forwarding emails, but most threads include at most one or two forwarding emails. Such small number of forwarding emails causes TC to be similar to EC.

Most Engaging and Responsive Users

Table 10.9 shows the top five engaging users and top five responsive users after averaging the ranks of our proposed models. The table shows that the two sets of

Table 10.9 Top-five users by engagingness and responsiveness. Note that we derive the overall engagingness and responsiveness of each user by averaging the engagingness and responsiveness of different models

	Engagingness		Responsiveness	
Rank	Enron employee	Position	Enron employee	Position
1	Ryan Slinger	Trader	John Lavorato	CEO
2	Larry Campbell	N/A	Monika Causholli	Employee
3	Joe Quenet	Trader	Jeff Dasovich	Employee
4	Mike Swerzbin	Trader	Kate Symes	Employee
5	Jeff King	Manager	Kay Mann	Employee

top users are different, consistent with our earlier results. It is interesting to note that most engaging users are traders. Other than CEO John Lavorato, the top responsive users are general employees. Interestingly, there exists no common actors between the two top-five employee lists by engagingness and responsiveness. In other words, there is no both high engaging and responsive actor among 151 Enron employees. This result is consistent with that in Table 10.6.

Role Analysis in Correlation

Figure 10.5 shows the scatter plot of engagingness and responsiveness scores of Enron employees with different roles. In the figures, we just present that 93 Enron employees among them are 3 chief executive officers, 9 directors, 35 employees, 3 house lawyers, 8 managers, 2 managing directors, 4 presidents, 12 traders, and 17 vice presidents. Since the job positions of the remaining employees from 151 Enron users are not known, we exclude them from Fig. 10.5. In particular, we show the engagingness and responsiveness scores of Enron employees with different roles in terms of the email reply time model. Note that most employees, managers, and traders tend to have higher engagingness scores than the other employees. In other words, employees, managers, and traders can effectively solicit responses from other actors. In contrast, vice presidents show wide range of engagingness scores. Unlike engagingness, we observed that responsiveness is not correlated to particular job appointments. Rather than actor roles, responsiveness is more related by actors' individual personality. Some actors are willing to respond to other actors, while other actors are not.

10.4.3 Email Reply Order Prediction

The goal of this experiment is to evaluate the performance of our proposed classification approach to predict email reply order. We also want to examine the usefulness of engagingness and responsiveness behaviors in prediction task.

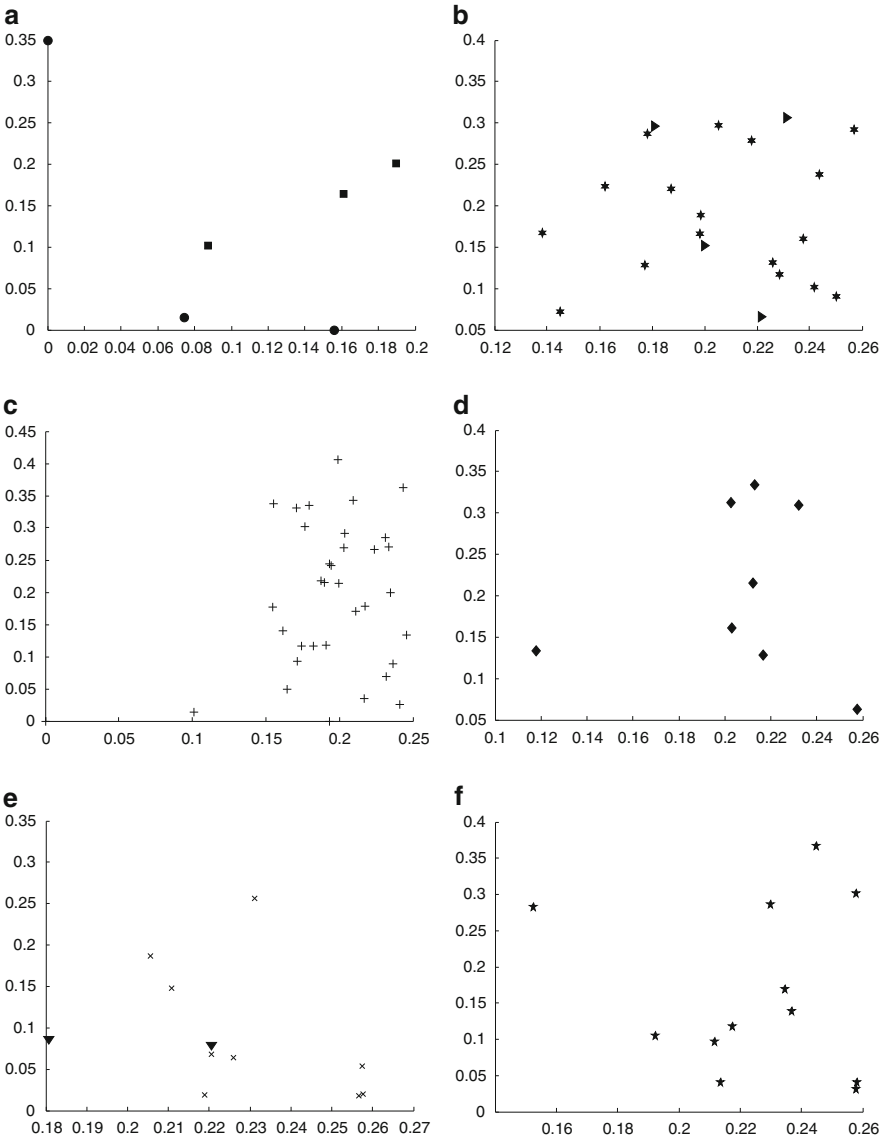

Fig. 10.5 Actor role in the email reply time model (X-axis and Y-axis denote E^{ET} and R^{ET}, respectively). Since the other models show similar results to the email reply time model, we omit those of the other models. (**a**) CEO and house lawyer. (**b**) President and vice president. (**c**) Employee. (**d**) Manager. (**e**) Director and managing director. (**f**) Trader

Table 10.10 Results of email reply order prediction

Features used in SVM	Average accuracy (%)
$SVM_{\mathbb{E}+\mathbb{I}}$	76.68
$SVM_{\mathbb{U}}$	77.3
$SVM_{\mathbb{B}}$	67.25
$SVM'_{\mathbb{E}+\mathbb{I}}$	65.33
$SVM'_{\mathbb{U}}$	69.31

There are five SVM classifiers trained, namely: (a) using comparative email and interengaging features (denoted by $SVM_{\mathbb{E}+\mathbb{I}}$); (b) using comparative behavior features only (denoted by $SVM_{\mathbb{B}}$), (c) using all features (denoted by $SVM_{\mathbb{U}}$), (d) using comparative email and interengaging features except $t(e)$[3] (denoted by $SVM'_{\mathbb{E}+\mathbb{I}}$), and (e) using all features except $t(e)$ (denoted by $SVM'_{\mathbb{U}}$). Classifiers (d) and (e) are included as earlier study [4] has shown that email replies often follow the last-in-first-out principle. $SVM'_{\mathbb{E}+\mathbb{I}}$ and $SVM'_{\mathbb{U}}$ allow us to find out if we can predict without knowing the email time information. From the 27,730 email reply relationships, we extracted a total of 19,167 email pairs for the prediction task. The emails in each pair have replies that comes after the two emails are received by the same user. For each email pair, we computed feature values based on only email data occurred before the pair. In addition, we used complement email pairs in training. The complement of an email pair (e_i, e_j) with class label c is another email pair (e_j, e_i) with class label $\bar{c}$. Fivefolds cross validation was used to measure the average accuracy of the classifiers over the fivefolds. The accuracy measure is defined by $\frac{\#\ correctly\ classified\ pairs}{\#\ email\ pairs}$.

Table 10.10 illustrates the results of all the five SVM classifiers. $SVM_{\mathbb{U}}$ produces the highest accuracy of 77.04 % due to the use of all available features. By excluding the email arrival order feature, the accuracy (of $SVM'_{\mathbb{U}}$) reduces to 68.97 %. This performance is reasonably good given that random prediction gives an accuracy of 50 %. The above results show that email arrival order feature is an important feature in the prediction task. We however notice that behavior features contribute to prediction accuracy especially when the email arrival order feature is not available.[4]

Table 10.11 depicts the top ten features for the $SVM_{\mathbb{U}}$ classifier. The table shows that engagingness based on the email reply time model *ET* is the most discriminative feature. This suggests that engagingness and responsiveness are useful in predicting email reply order.

[3] See Table 10.4.

[4] Recently, [4] reported that email replies often follow the last-in-first-out principle.

Table 10.11 Top-ten features for $\text{SVM}'_{\mathbb{U}}$

Rank	Feature	Weight
1	$E^{ET}(Sdr(e_i)) - E^{ET}(Sdr(e_j))$	0.66
2	$R^{RG}(Sdr(e_i)) - R^{RG}(Sdr(e_j))$	0.59
3	$Indegree(Sdr(e_i)) - Indegree(Sdr(e_j))$	0.55
4	$E^{RW}(Sdr(e_i)) - E^{RW}(Sdr(e_j))$	0.54
5	Engaging threads xy	0.48
6	$R^{ES}(Sdr(e_i)) - R^{ES}(Sdr(e_j))$	0.37
7	$E^{ER}(Sdr(e_i)) - E^{ER}(Sdr(e_j))$	0.36
8	$E^{TC}(Sdr(e_i)) - E^{TC}(Sdr(e_j))$	0.32
9	emails y2x	0.27
10	$size(r(e_i)) - size(r(e_j))$	0.26

10.5 Related Work

We first review related work on engagingness and responsiveness behavior modeling. Engagingness and responsiveness behaviors have not been well studied in the past. There is one work on responsiveness [1] (even though it is not sufficient) but no work on engagingness. In [1], responsiveness behavior of a user (in the context of Enron email data set) was defined as the average deviation in response time of user from the other users. Users with positive deviations are known to be lethargic and those with negative deviations are responsive.

Since we are using the Enron dataset, we also review other research on the data set comparing their works with ours. These works can be divided into:

- **Knowledge extraction:** Rowe et al. present an automatic method for extracting social hierarchy data from user communication behavior on the Enron dataset [12]. Such communication patterns are captured by computing the social score based on a set of features: number of emails, average response time, number of cliques, degree centrality, clustering coefficient, mean of shortest path length from a specific vertex to all vertices in the graph, betweenness centrality, and Hubs-and-Authorities importance. Then, by performing behavior analysis and determining the communication patterns, their method ranks main users of an organization, groups similarly ranked and connected users to reproduce the organizational structure, and understand relationship strengths among users. Pathak et al. investigate socio-cognitive networks based on email communication in an organization [11]. Socio-cognitive network analysis involves understanding who knows who knows who in a social network. For analysis, the authors propose a model using probability distributions for communication probabilities, in which a Bayesian inference technique is used for updating the probabilities.
- **Email thread detection:** To exploit parent-child relationships from email messages, grouping messages together based on which messages are replies to which others, Yeh and Harnly propose email thread detection using undocumented

header information from the Microsoft Exchange Protocol and string similarity metrics [14]. Then, their method recovers missing messages from email threads.

- **Email label prediction:** Karagiannis and Vojnovic study various parameters including the email size, the number of recipients per email, role of the sender and recipient in the organization, information load on the user, etc., and their effect on reply probability and response time [4]. While their results shed some interesting insights into how these parameters affect users' replying behavior, further research is required to actually implement a learning model that can automatically prioritize emails based on these findings. Interestingly, through our experimental analysis, we found that email replies often follow the last-in-first-out principle which has been reported by Karagiannis and Vojnovic [4]. The study of [15] builds a supervised classifier to automatically label emails with priority levels on the scale of 1–5. Their model primarily focuses on graph-based metrics such as node degree, centrality, clique count, etc. derived from the underlying social networks of users. McCallum et al. present the author-recipient-topic model which learns topic distributions based on the direction sensitive messages sent between users [7]. In particular, this model works based on Latent Dirichlet Allocation and the author-topic model in which distribution over topics is conditioned distinctly on both the sender and recipient according to the relationships between users. Unlike our models, the authors have explored Enron dataset mostly from a Natural Language Processing (NLP) perspective. Recently, B. On et al. conducted preliminary study of behavior models on mobile social networks [10].
- **Email interaction prediction:** To predict whether emails need replies, Dredze et al. present a logistic regression model with a variety of features e.g., dates and times, salutations, questions, and header fields of emails [2].

10.6 Conclusion

In a nutshell, we study user engagingness and responsiveness behaviors in an email network. We have developed four types of behavior models based on different characterization principles. Using the Enron dataset, we evaluate these models. We also apply the models to email reply order prediction task.

The work is a significant step beyond the usual node and network statistics to derive node behavior measures for a given network. While our results are promising, there are still much room for further research. We will develop new behavior models based on probability and email content. We also plan to conduct a more comprehensive study of engagingness and responsiveness behaviors on a much larger and complete information exchange dataset (e.g., twitters, blogs, SMS, etc.). This will remove some shortcomings of the existing Enron dataset which does not have complete emails of each user. We will also expand our work to apply the behaviors to other interesting email prediction tasks.

Appendix

Email Thread Assembly The algorithm to identify reply and forward relationships among emails is as follows:

- Step 1: Group all emails with matching subjects after ignoring the prefixing Reply (RE, Re) and Forward (FW, Fw, FWD, Fwd) tags from the subject field. Emails with blank subject or subject matching *"no subject"* are ignored.
- Step 2: Sort emails in each subject group by date and time. We use the Perl module Time: Local to convert the message date and time into an integer number that indicates the number of seconds since the system epoch (Midnight, January 1, 1970 GMT).
- Step 3: For each email e_1 whose subject starts with one of the Reply or Forward tags (Re, RE, Fw, FW, Fwd, FWD)
  - Step 3a: Scan all the previous emails in its group
  - Step 3b: Find the most recent email e_2 such that the sender of e_1 is one of the recipients of e_2
  - Step 3c: If the subject of e_1 begins with a Reply tag, also check that the sender of e_2 is one of the recipients of e_1
- Step 4: Compute the time difference $t(e_1) - t(e_2)$
- Step 5: If $t(e_1) - t(e_2) \leq \xi$, add link $e_1 \rightarrow e_2$ to indicate that e_1 is a reply or forwarded email of e_2

Here, the parameter ξ specifies a time-window between emails e_1 and e_2 to consider it as a valid thread link. In our experiments, we set $\xi = 90$ days (3 months) and discard pairs that have a time difference larger than that.

References

1. Deepak, P., Garg, D., Varshney, V.: Analysis of Enron email threads and quantification of employee responsiveness. In: Workshop on Text Mining and Link Analysis (TextLink), Hyderabad (2007)
2. Dredze, M., Blitzer, J., Pereira, F.: Reply expectation prediction for email management. In: 2nd Conference on Email and Anti-Spam (CEAS), Stanford University (2005)
3. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top K lists. SIAM J. Discret. Math. **17**, 134–160 (2003)
4. Karagiannis, T., Vojnovic, M.: Behavioral profiles for advanced email features. In: 18th International World Wide Web Conference (WWW), Raleigh (2009)
5. Kendall, M.: Rank Correlation Methods. Charles Griffin and Company Limited, London (1948)
6. Klimt, B., Yang, Y.: The Enron corpus: a new dataset for email classification research. In: 15th European Conference on Machine Learning (ECML), Pisa (2004)
7. McCallum, A., Wang, X., Coorada-Emmanuel, A.: Topic and role discovery in social networks with experiments on Enron and academic email. J. Artif. Intell. Res. **30**, 249–272 (2007)
8. Miller, N.E., Dollard, J.: Social Learning and Imitation. Yale University Press, New Haven (1941)

9. On, B.-W., Lim, E.-P., Jiang, J., Purandare, A., Teow, L.: Mining interaction behaviors for email reply order prediction. In: 2nd International Conference on Social Network Analysis and Mining (ASONAM), Odense, (2010)
10. On, B.-W., Lim, E.-P., Jiang, J., Chua, F., Nguyen, V., Teow, L.: Messaging behavior modeling in mobile social networks. In: Symposium on Social Intelligence and Networking (SIN) in conjunction with 2nd IEEE International Conference on Social Computing (SocialCom), Minneapolis (2010)
11. Pathak, M., Mane, S., Srivastava, J.: Who thinks who knows who? socio-cognitive analysis of an email network. In: 6th IEEE International Conference on Data Mining (ICDM), Hong Kong (2006)
12. Rowe, R., Creamer, G., Hershkop, S., Stolfo, S.: Automated social hierarchy detection through email network analysis. In: 9th WEBKDD and 1st SNA-KDD Workshop, San Jose (2007)
13. Stolfo, S., Hershkop, S., Hu, C., Nimeskern, O., Wang, K.: Behavior-based modeling and its application to email analysis. ACM Trans. Internet Technol. **16**(2), 187–221 (2006)
14. Yeh, J., Harnly, A.: Email thread reassembly using similarity matching. In: 3rd Conference on Email and Anti-Spam (CEAS), Mountain View (2006)
15. Yoo, S., Yang, Y., Lin, F., Moon, I.: Mining social networks for personalized email prioritization. In: 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Paris (2009)