Behavioral Profiles for Advanced Email Features

Thomas Karagiannis and Milan Vojnović
Microsoft Research Cambridge, UK
{thomkar,milanv}@microsoft.com

ABSTRACT

We examine the behavioral patterns of email usage in a large-scale enterprise over a three-month period. In particular, we focus on two main questions: (Q1) what do replies depend on? and (Q2) what is the gain of augmenting contacts through the friends of friends from the email social graph? For Q1, we identify and evaluate the significance of several factors that affect the reply probability and the email response time. We find that all factors of our considered set are significant, provide their relative ordering, and identify the recipient list size, and the intensity of email communication between the correspondents as the dominant factors. We highlight various novel threshold behaviors and provide support for existing hypotheses such as that of the least-effort reply. For Q2, we find that the number of new contacts extracted from the friends-of-friends relationships amounts to a large number, but which is still a limited portion of the total enterprise size. We believe that our results provide significant insights towards informed design of advanced email features, including those of social-networking type.

Categories & Subject Descriptors: H.4.3 [Communications Applications]: Electronic mail
General Terms: Design, Measurement, Human Factors
Keywords: Reply time, reply probability, email profiles.

1. INTRODUCTION

The proliferation of social computing services has recently sparked an interest for enhancing the email service with features common in social network applications [5]. Such integration, however, requires a comprehensive understanding of the intrinsic usage characteristics of the email service, and in particular, of the user behavior and how this translates to information flows through email networks. This understanding is thus important in order to (i) guide the design of new features for the service overall but also at the email client side, and (ii) to potentially leverage the social network induced by email communications for other online services.

This paper is a step towards this direction. Specifically, we present results of an extensive study of "behavioral" profiles that characterize users, their actions in processing emails, and properties of the exchanged emails. By behavioral we refer to characteristics that describe interactions of correspondents and properties that reveal specific human behaviors. Our goal is to investigate to what extent these profiles may inform the design of advanced email features such as email prioritization, recommendation and filtering mechanisms, or expert finding and people search by exploiting the email social graph.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Our study is based on the analysis of emails exchanged in a large-scale, multinational corporation with more than 100,000 employees spread across several countries. The measurement trace covers the email communications of all employees for a period of three months amounting to roughly 315 Million sent emails (Section 2.1). To examine user relationships and how these affect the overall information flow, we also leverage side information such as the global enterprise organizational structure.

Overall, our study is predominantly driven by exploring design possibilities focusing on two main questions that we feel are of particular relevance to enterprise environments:

Q1: What do replies depend on? Can we predict the email response time (Section 3.1), and which emails will be replied (Section 3.2)?

and

Q2: What is the gain if contact lists are augmented with contacts derived from "friends-of-friends" relationships of the email social graph (Section 4)?

With regards to **Q1**, we determined that the email *reply probability* depends on a set of factors related both to the correspondents (i.e., sender and receiver), but also to their interactions. Of these factors, the recipient list size appears to be the most important, followed by the rate of emails between the receiver and the sender, the organizational distance between the correspondents, and the elapsed time since the last email activity of the receiver. Through analysis of variance we show that while conditioning on one of these factors provides moderate prediction gains, combining factors provides significant gains with respect to predicting replied emails.

Replies constitute only a small portion of sent emails, surprisingly, as little as 3% on average. Furthermore, we note the significance of recency and in general the effect of the receive time, and time-of-day periodicity on the reply probability and the time it takes for a user to reply to a received email. These findings suggest that from a sender point of view, it does matter *when* an email is sent, not only to receive a prompt reply, but also to receive a reply in the first place. The significance of the recency factor confirms claims

based on survey results [10] that users tend to largely prioritize emails with respect to recency. However, it is surprising to observe the rather dramatic decrease of the probability of reply with the inactivity period of the receiver over the timescale of one hour. Finally, we identify a "congestion collapse" type of dependence of the probability of reply on both the rate of emails from the receiver to sender and from the sender to receiver. The probability of reply tends to increase with sender-to-receiver email volume up to a threshold; however, after this threshold, there is a qualitative change with a tendency to smaller values.

With regards to **Q2**, we highlight that the number of new contacts discovered through friends-of-friends types of relationships in the email graph exhibits diminishing returns for users with large contact lists. However, we found that typical contact lists comprise tens of contacts, facilitating discovery of order thousands of new contacts through friends-of-friends. This number appears significant and implies that advanced search and filtering mechanisms would need to complement such a feature to allow the user to take advantage of the augmented contact list. On the other hand, contacts discovered through friends-of-friends are at the same time limited, when we take into consideration that a contact list of order thousand corresponds to 1% of the total enterprise graph. This suggests that services such as expert finding or people search will have to examine contacts that are further than two hops away.

We believe that our findings provide insightful observations for application designers with respect to human interactions and behavior as seen through the email service. To the best of our knowledge, our work is among the first to study in large-scale the importance of such behavioral patterns.

2. COMMUNICATION PROFILES

We first examine the variety of communication profiles across users. As large-scale enterprises are characterized by employee heterogeneity, we explore to what extent email communications reflect this diversity. In particular, in the following subsections we describe our datasets, and examine properties such as the information load imposed on users, and the number of correspondents per user.

2.1 Datasets

The results presented in this paper are based on logs from Microsoft Exchange servers that cover email communications for all employees of a large multinational corporation for 3 months. This global enterprise consists of over 100,000 employees spread across 100 countries and 6 continents. During the three month period we observe a total of 315 Million emails. Each log entry specifies the sender and recipients per email, the subject, the sent timestamp, the size of the email in bytes, and other information such as the exchange servers involved, email ids etc. Hence, this information allows us to examine both internal (within the enterprise) and external communications for each enterprise user over time. By the information presented in the email subject, we can additionally separate emails that constitute replied (RE) or forwarded (FW) emails. Note that SPAM emails are rare in our datasets, as SPAM filters operate before the Exchange Servers that provided the logs for this study; typically only a few (less than ten) SPAM emails are observed per user per month. With regards to mailing lists, emails to such aliases were expanded to include all recipients of each such email. We do not exclude these emails from the analysis, since we are interested in the overall information load a user experiences. Examining whether the various properties of email replies differ between mailing list items and other emails is left for future work.

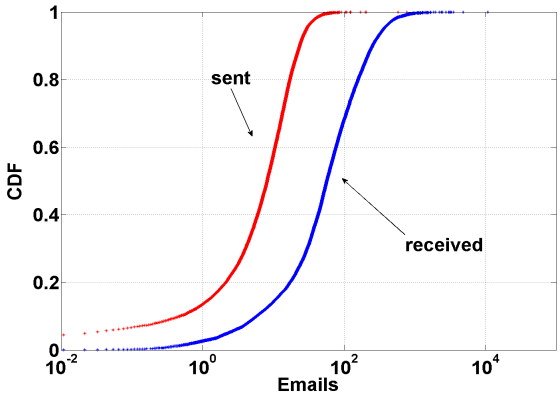

Figure 1: CDFs of sent and received emails per user per day. The ratio of receive-to-sent emails is approximately 7.

We further relate information flow and user behavior to the organizational structure of the enterprise. To this end, our *org-structure* dataset provides us with information regarding the names and email aliases of all employees, their physical location and distribution to buildings and offices across countries, and their organizational title. Using this information we are able to extract the organizational tree of the enterprise (henceforth referred to as "org-tree") and identify "report-to" relationships for each user (i.e., identify each user's manager and direct reports). We use the term "root-distance" to denote the length of the shortest-path of an employee to the root of the organizational tree. Similarly, we use the term "level-distance" between two employees to refer to the difference between the length of the shortest-paths (i.e., root-distances) of each of the two employees to the root of the organizational tree (CEO of the enterprise).

2.2 Basic Properties

In order to examine the email information load generated by and imposed on users, we consider the email volume per user for the three months of our study. We find that the median number of emails sent per user per day is approximately 8 with 10% and 90% quantiles of 0.5 and 24 emails per day respectively. For the received emails the corresponding numbers amount to 57, 6, and 252 (median and quantiles respectively). We find that one-third of users receive less than 34 emails per day, another one-third of users receive between 34 and 95 emails per day, and the remaining third more than 95 emails per day. In a survey conducted by Neustadter et al [10], 29% of participants reported to receive less than 50 emails per day (low-volume), 36% reported to receive in between 50 and 100 emails per day (medium-volume), and the remaining 34% of users reported to receive more than 100 users a day (high-volume). In contrast, we find that almost half of users qualify as low volume users – 45.25% of users receive less than 50 emails per day. We also find a larger portion of high-volume users than medium-volume – 31% vs. 23.75%.

Further analyzing the sent emails per day, we find that roughly 2 out of the sent emails per day correspond to replies, while forwarded emails amount to just 0.2 emails per day. We will extensively discuss replied emails in Section 3. Fig. 1 displays the corresponding Cumulative Distribution Functions (CDF) for the number of sent and received emails per user per day. The mean numbers of received and sent emails per day per user are 107 and 11 respectively. Overall, the median ratio of received-to-sent emails is approximately 7, which we have found to be consistent both with the mean recipient list size, and the minimum organizational group size when considering leaf nodes in the organizational tree (i.e., groups of users reporting to the same manager).

The number of correspondents per user is another basic property of email communications. This property also relates to the "notion" of degree when forming the "social" graph that results from such communications. Considering only interactions across enterprise users, the median out-degree (i.e., reflecting correspondents of sent emails) is roughly 190 for a period of three months with 10% and 90% quantiles of 13 and 2260. Similarly, the in-degree median is 291, with quantiles of 8 and 1960. High degrees here indicate participation in email distribution lists. As previous studies have also noted [13], applying thresholds on the formation of edges of such an email social graph (i.e., removing edges based on the number of emails sent across two correspondents) can filter out transient or rare communications that do not reveal "true" communication relationships between users. For example, applying a threshold of 15 emails (i.e., removing edges that correspond roughly to less than one email per week) results in median in-degree of 34 (quantiles of 5 and 183) and out-degree of 21 (quantiles of 3 and 108). In Section 4, we will further explore the potential gains of taking advantage of this enterprise social graph to expand user contact lists.

To investigate the effect of "strong-ties" [13] in the email graph on the generated email flow, we characterize the *top-k* correspondents per user based on the frequency of email communications (edge weights). This property essentially describes the "favorite" correspondents per user. Additionally, this metric informs the design of both network applications and devices, e.g., by examining the possibility of displaying the favorite correspondents of a user in the limited screen of a mobile device.

Fig. 2 highlights the portion of user-initiated email communications (i.e., sent emails) covered by the top-k correspondents of a particular user. Fig. 2 shows that roughly half of the users' conversations are covered through the set of the top-6 correspondents (2 and 15 for the 25% and 75% quantiles respectively) for half of the user population. Increasing this number to top-10 correspondents covers more than 60% of emails sent. These observations provide important clues for the dimensioning of contact list sizes. Achieving large hit rates, e.g., larger than 50%, requires contact lists of order 10 users. As previously mentioned, small contact lists are especially attractive for mobile devices.

Information flow vs. org levels. As previously discussed, the small number of correspondents per user is consistent with the average group size for the leaves of the organizational tree. It is thus natural to examine whether the imposed organizational structure influences email communications between employees, and how this influence manifests in the email network.

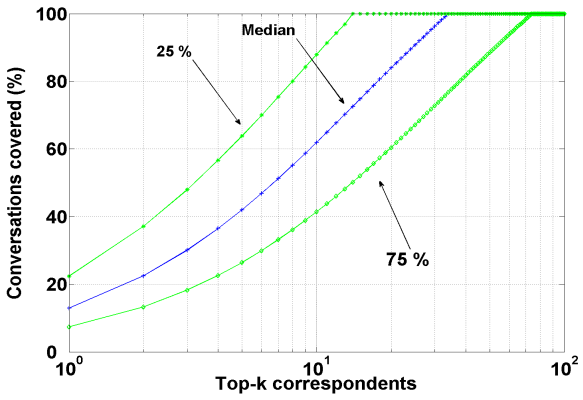

Figure 2: Email sends covered by top-k correspondents. Top-5 correspondents cover more than 40% sent emails for half of the user population.

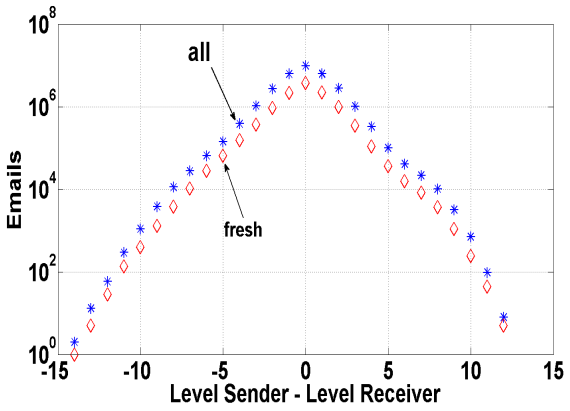

Figure 3: Email flow vs. level-distance between the sender and the receiver. The flow is symmetric and decreases exponentially with the level-distance.

We observe that the rate of email flow decreases exponentially with the level-distance between the correspondents. Fig 3 displays the number of emails sent versus the level-distance (note that levels increase as we move away for the root). The flow is symmetrical with respect to the organizational levels of the correspondents. To examine whether this symmetry is the result of responses to original emails, we condition on "fresh" emails only, where we exclude all replies. Even after removing replies, the symmetry is still present in the flow communications. Further, the peak of the plot suggests high inter-level communication. This is confirmed by examining intra-level and inter-level communications, where only a minor asymmetry exists with users of larger root-distances directing more flow upwards, compared to the flow users closer to the root send downwards.

3. WHAT DO REPLIES DEPEND ON?

We now turn our attention to replied emails. We examine factors that may determine user's action with respect to a received email, and specifically, here, replies to emails. In particular, we study the effect of several metrics on the (i) *reply time*, i.e., how much time it takes a user to reply to a received email, and ii) the *reply probability*, i.e., the probabil-

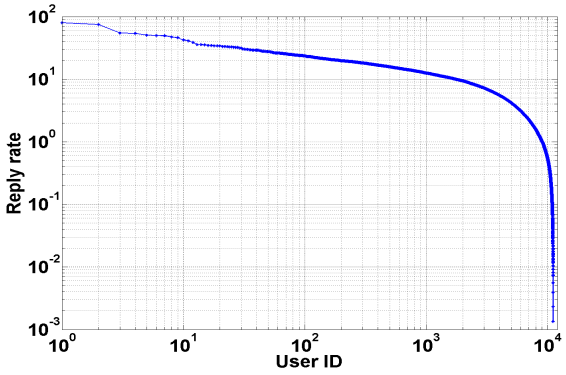

Figure 4: Reply rate per user. 25% of users reply to less than 1% of their received emails, with only 15% of users replying to more than 10%.

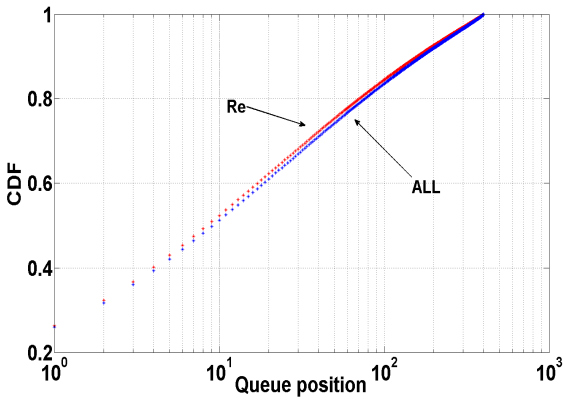

Figure 5: CDF of the queue position of replied emails. "ALL" refers additionally to forwarded emails.

ity that a received email will be replied. Since the content of the emails is not available, our analysis will focus mostly on factors that characterize correspondents, their status within the enterprise, and their interactions over time. Determining such metrics that affect email replies may inform the design of email prioritization or filtering mechanisms, simplifying the email processing for enterprise users. Our analysis in this section will focus on a random sample of 12,000 users (roughly 10% of all users) for the period of three months. [1]

3.1 Reply Time

As discussed in Section 2.2, only a small portion of received emails are actually replied. In this section, we characterize reply time for all emails that were replied. By reply time, we refer to the time difference between the receive time of an email, and the time of the corresponding reply. We examine the following factors: (i) The email triage referring to prioritization strategies employed by users in handling of the emails, i.e., the order in which emails are processed. For our purposes, processing of an email refers to an email being replied or forwarded as we cannot observe other actions (e.g., deletions) from our data; (ii) The time required to process an email – for example, comprising the time to read the email and prepare the reply; (iii) The user idle time, e.g., the time of day effects; and, finally, (iv) under other factors, we consider the reply time versus the in-flow rate of emails for the receiver.

Before examining individual factors, we first consider the rate of email replies across users. In Fig. 4, we display the reply rate per user in decreasing order of the rate. Note that the median reply rate is roughly 3% with only 15% of users showing a reply rate larger than 10%, and 25% of users having a reply rate less than 1%. Small reply rates could also reflect the effect of distribution lists or announcement messages that are broadcasted and rarely replied and are common in enterprises.

3.1.1 Email Triage Strategies

Email processing strategies employed by users determine the order by which emails are processed or replied, and thus can provide hints with regard to the expected reply time. Understanding email processing strategies is also important for the design of personalized email client applications.

To infer reply strategies, we apply the following procedure. We form a queue per each distinct user, storing received emails. Most recent items (emails) are placed at the first position (head) of the queue. Items are departing from the queue if we observe a reply ("RE") email sent from the specific user with a subject that matches existing items in the queue (excluding the "RE" characters – also excluding automatic "Out-of-office" replies), and the sender of the original email is included in the reply recipient list. Thus, items may depart from the queue at any order. Due to the size of the dataset and memory limitations, we had to restrict the queue size to 400 emails. When the queue is full, the oldest item in the queue is removed (thus creating a small bias in the results for emails that may have been processed after being removed from the queue).

We consider the recency ranks (queue positions) of replied emails – Fig. 5. The figure presents the histogram of queue positions for replied items and indicates high bias of processed items towards recently received emails. The figure also presents a similar curve ("ALL") for all processed items in the queue (i.e., after applying a similar procedure for forwarded emails). In particular, the median recency rank of an email reply is only about 10 with the 90% quantile being at roughly 200 emails. This recency bias could be the result of several reasons. For example, users may bias their email views so that recent emails appear higher in the list of received emails; additionally, most users may process emails in a timely manner so typically emails get processed while they are still high ranked with respect to their recency.

Reply time depends significantly on the recency rank of the email (Fig. 6). The figure indicates that the median email reply time is about 1 day for the emails of recency rank up to 150. Overall, Fig. 6 implies that email reply times vary rather widely, with reply times exploding as email items are pushed towards the tail of the queue.

In summary, we found indications that email triage strategies employed by the users prioritize recently received emails. This is in line with previously reported results of surveys [10].

The previous discussion implies that recency is a significant factor on the processing order of emails. Being in an enterprise environment, however, it would be perhaps nat-

[1]This sampling was necessary for data processing purposes due to the magnitude of our traces. The results presented here are qualitatively similar when taking multiple samples, and samples of various sizes from our logs.

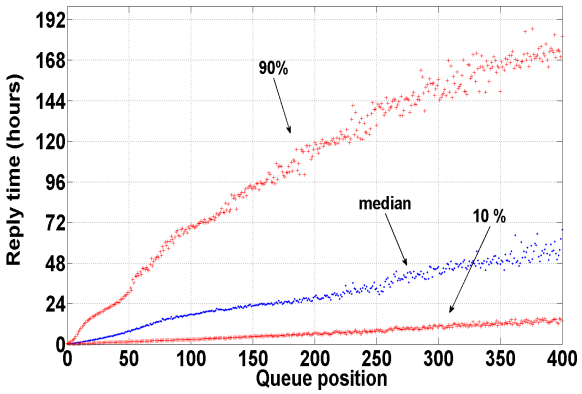

Figure 6: Queue position vs. reply time of replied emails. Reply times significantly increase as the emails are pushed to the tail of the queue.

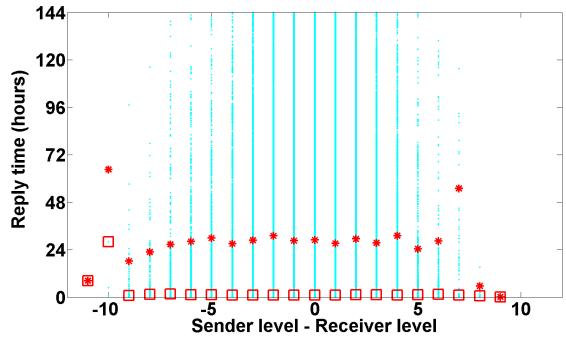

Figure 7: Email reply time versus level-distance of correspondents. Squares denote the median values while the stars ones the 90% percentile.

ural to assume that users would reply faster to emails originating from individuals with higher organizational status (e.g., an employee responding to a manager higher up in the organizational hierarchy). We find that this assumption is not supported by our data.

To test this hypothesis, we evaluate whether correlation exists between reply time and the level-distance between the sender and the receiver. Fig. 7 presents a scatter plot of the level-distance versus the reply time in hours, along with median, and the 10% and 90% quantile values. We observe that the median email reply time (denoted by the stars in Fig. 7) does not significantly depend on the level-distance between the email correspondents. In fact, the range of email reply times appears to tend to be larger, the smaller the absolute level-distance between the email correspondents is.

3.1.2 Email Processing Time

The second factor that may impact reply time relates to the time required to process a received email. This processing time may relate to the time required to (i) read the email, and (ii) prepare its reply. In our case, the email processing time might be reflected by the size of the received email.

Fig. 8 presents this relationship of email size and reply time as seen in our data. Clearly, the figure presents a strong dependency, with the reply time increasing as the email size becomes larger. In particular, emails larger than 1 Mbyte, most likely containing attachments, take more than 6 hours to reply to, while on the other end, small emails are replied

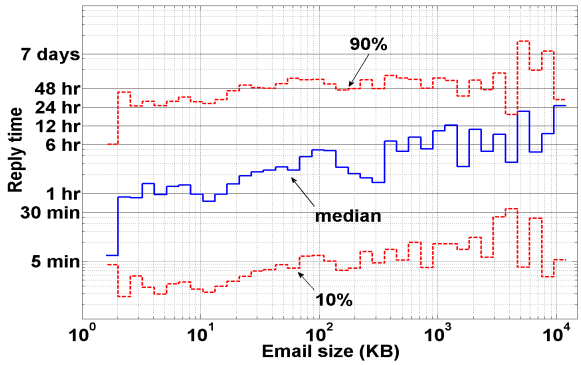

Figure 8: Reply time versus email size. Larger emails result to higher reply times.

within an hour. This confirms the suggested hypothesis [10] that users would tend to follow the *least-effort principle* in processing of the emails.

3.1.3 Time-of-Day

While we expect the processing time to exhibit some diurnal and weekly periodicities, it is not clear what exact shape such periodicities would assume and what values would hold during work hours and weekend days, in particular. Indeed, Fig. 9–left shows a strong weekly periodicity with reply times being considerably larger for emails sent at the end of a work week or weekend days. Over weekdays, the email reply times vary naturally over the range from less than 1 hour to order half a day. Perhaps more interestingly, we note that during work hours, the median email reply time appears rather concentrated around a value smaller than 1 hour. This suggests that employees may use the email service for instant-messenger-like interactions where delays of a few minutes are acceptable (e.g., we identified several such cases with emails related to eliciting participation to a group lunch). This is further confirmed by Fig. 9–right, which presents the CDF of the reply time for all replied emails. Fig. 9–right highlights that the reply time for 20% of replied emails is less than 5 minutes. Overall, combining the time-of-day observations with the recency processing bias implies that emails have a higher probability of being replied to at the beginning or during the working hours.

3.1.4 Other Factors

The size of the queue relates to the information load imposed on the user, where large queues might lead to "information overload" effects and increased reply times. We study the relationship between the overall received emails and the reply time. Interestingly, Fig 10 highlights that the larger the receive rate the smaller the reply time for users that receive a large number of emails per day (for example, more than 30 emails per day). There are at least two possible explanations for such an effect. First, email is a significant portion of the work activity for some users requiring timely replies. Indeed, examining the roles of users with high receive rates and small reply times reveals that a significant portion corresponds to management roles. Second, this effect could be attributed to users engaging in conversations through email, forming a sequence of received and replied emails within short time intervals.

Summarizing the discussion in this section, we find that reply time depends strongly on time of the day effects, as

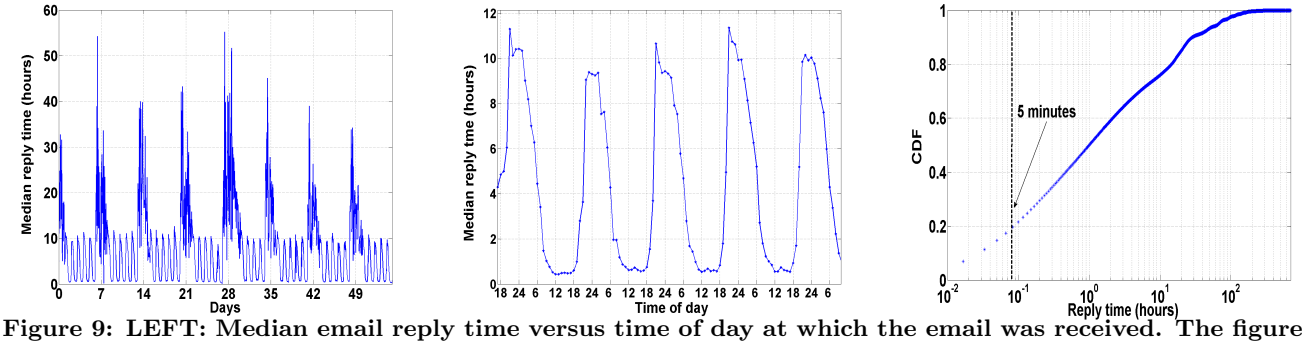

Figure 9: LEFT: Median email reply time versus time of day at which the email was received. The figure on the middle is a zoomed version of the one on the left. RIGHT: CDF of email reply time. The reply time for 20% of replied emails is less than 5 minutes.

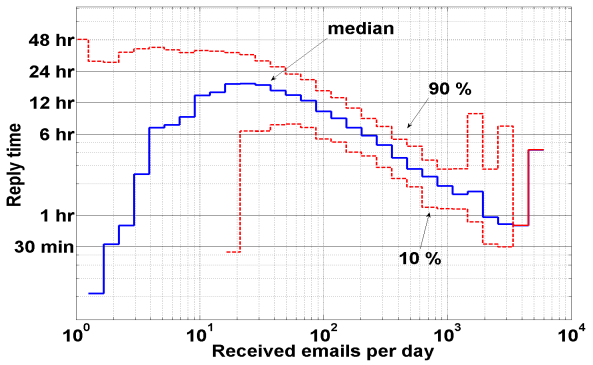

Figure 10: Reply time vs. the information load as seen by the number of received emails per day. For large receive rates, reply times decrease.

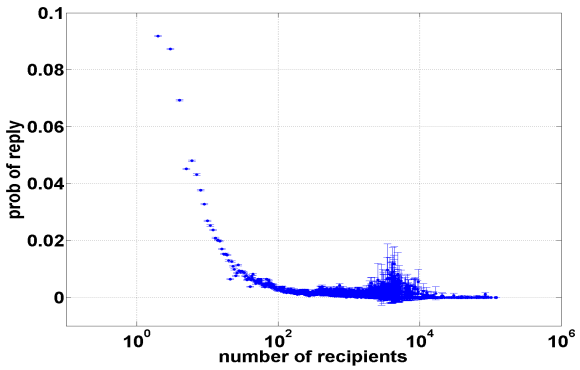

Figure 11: Probability of reply conditional on the number of recipients.

well as the recency of the email received. Reply times increase with the size but appear to decrease with the information load for large receive rates. Finally, despite examining an enterprise environment, org structure does not appear to directly impact the reply time.

3.2 Reply Probability

In this section, we analyze the influence of several factors on the probability of reply to an email. Our analysis aims at identifying which factors can best predict the probability of reply to an email. To this end, we first perform a visual exploration of the probability of reply conditional on each of the factors. This study already suggests qualitative dependencies. We further conduct quantitative analysis of variance to evaluate the quality of predictions of individual factors and their combinations. This factorial analysis is motivated by the possibility of a "reply-to" recommendation functionality that would assist users in the email triage process. In particular, we consider the following factors:

1. Recipient list size;
2. The number of emails sent from the receiver to the sender;
3. Org distance between receiver and sender measured as the maximum hop distance until the first common ancestor node in the org tree; we refer to this distance metric as "ancestor-distance".
4. Time elapsed since the last observed email activity of the receiver;
5. The number of emails sent from the sender to the receiver;
6. Org level of the sender (root-distance);
7. Email size in Bytes.

In addition, we also considered other factors including the identities of senders and receivers. While the above is a non-exhaustive list of factors that may influence the probability of a reply, we believe the set captures a wide-range of factors. For example, factor 6 profiles the sender; factor 4 profiles the receiver; the remaining factors profile a sender-receiver pair (i.e., edge in the email graph). Additionally, some factors capture the status of correspondents (e.g., org level) and the level of interaction (e.g., the rate of emails sent between the correspondents).

In the following paragraphs, we analyze factors 1 to 7 and present results that suggest that all factors are significant in the descending order of appearance in the above list.

3.2.1 Recipient List Size

The conditional probability of reply tends to decrease with the number of recipients of an email. Fig. 11 presents this conditional probability showing the points with the 95%-confidence intervals less than or equal to 0.01 (i.e., small variance). The probability of reply rapidly decreases with the number of recipients; for example, for emails with 3 recipients the probability drops to about a half compared to emails with 1 recipient; further, for 20 recipients, we have an order of magnitude decrease. We also observe some noticeable spikes for order 1000 recipients which likely correspond to replies to email lists.

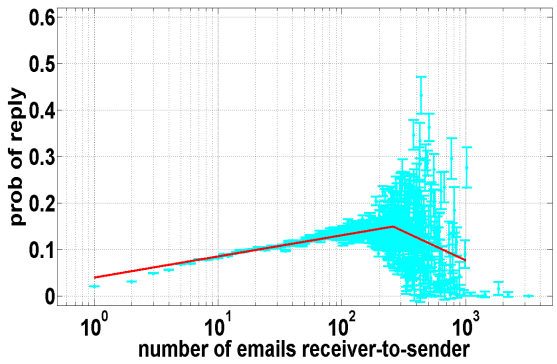

Figure 12: Probability of reply conditional on the number of emails from receiver to sender.

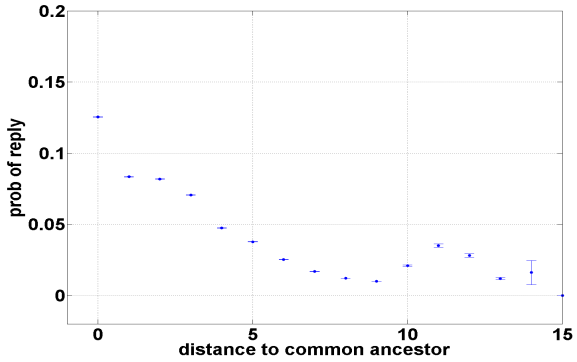

Figure 13: Probability of reply conditional on the ancestor-distance between sender and receiver.

3.2.2 Rate of Emails from Receiver to Sender

The probability of the email reply conditioned on the rate of email flow from the receiver to sender exhibits the following *threshold* behavior (see Fig. 12). The probability of email reply increases with *diminishing returns* (in particular, the growth is *logarithmic*) with the rate of the email flow from the receiver to sender up to a threshold. Beyond this threshold, there is a qualitative change of the dependence, with a deviation from the linear dependence in the logarithmic scale and a tendency towards smaller values.

We estimate the value of the threshold by a linear regression model where the relation between the email rate and the probability of reply is assumed to be a concatenation of two linear segments intersecting at a threshold value. This defines a family of models indexed by the threshold value. We choose the model with the maximum likelihood estimate of the threshold [3] which results in an estimate of about 2.8 emails per day, or 4.3 emails per work day; the 95%-confidence interval ranges from 2.3 to 3.5 emails per day.

The above analysis suggests a "congestion collapse" type of dependence where the reply probability increases up to a threshold and decreases beyond. It would be natural to argue that the value of the threshold could be interpreted as the capacity of information processing by users.

3.2.3 Ancestor-Distance

We found evidence of the following *homophily* bias. The probability of reply tends to decrease with the ancestor-distance between the correspondents. For "report-to" relationships (i.e., one of the correspondents reporting to the other), we set the ancestor-distance to 0. For example, two employees are at ancestor-distance 1 if they report to the same manager. Fig. 13 reveals roughly a probability of 13% to respond to a sender that is at ancestor-distance 0. The probability is about 8% for ancestor-distances 2 and 3 and assumes smaller values for large ancestor-distances.

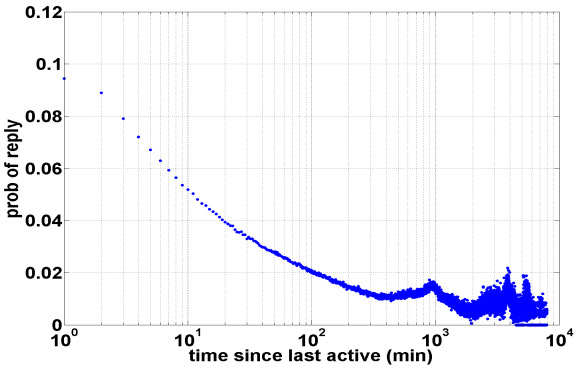

Figure 14: Probability of reply conditional on the time elapsed since the receiver was last active.

3.2.4 Time Since Receiver's Last Email Activity

We found the following *activity bias*. The probability of reply decreases with the time elapsed since the last observed email activity of the receiver (Fig. 14). Note that our observation of user's email activity is limited to email send events. We would expect that an email would be replied more likely if received while the user is active (i.e., processing email). Due to the observations of small median response time in Section 3.1.3, we expect that the estimated probability of reply time for emails received shortly after an email was sent would be larger because of the potential instant-messenger like conversations among employees through email. Indeed, we found that the probability of email reply to an email received shortly after an email was sent is roughly 0.1 and drops to about half if received 10 minutes after an email was sent. The probability of email reply drops for an *order of magnitude* if received about 6 hours after an email was sent. This appears aligned with the diurnal periodicity.

While it is intuitive to expect that the probability of email reply would decrease with the duration of email inactivity of the receiver due to the accumulated email processing load, it is surprising to find the observed magnitude of decrease over the timescale of 1 hour. These results suggest that users prioritize processing emails with respect to recency and is consistent with the recency bias observed in Section 3.1. The activity bias suggests that from an email sender viewpoint, it does matter *when* an email is sent, not only to expect a prompt reply, but also in order to elicit a reply in the first place!

3.2.5 Rate of Emails from Sender to Receiver

For this factor we observe the same qualitative properties as for the receiver-to-sender activity (Section 3.2.2). Thus, we only discuss the differences here. We find that, overall, the conditional probability is of smaller magnitude compared to the receiver-to-sender interaction (Fig. 15). The maximum likelihood threshold is smaller with 1 email per day or about 1.7 emails per work day with the 95%-confidence interval of 0.7 to 1.6 emails per day.

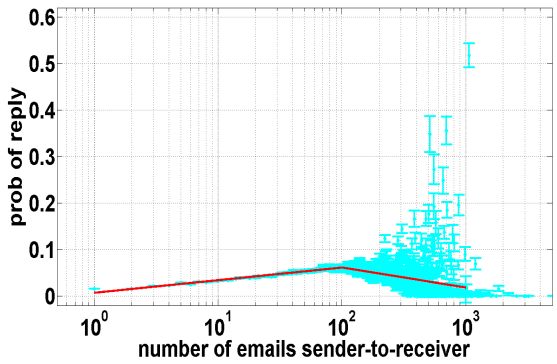

Figure 15: Probability of reply conditional on the number of emails from sender to receiver.

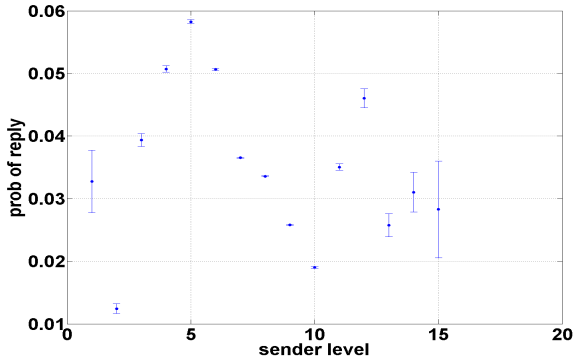

Figure 16: Probability of reply conditional on the sender level.

The observed threshold phenomena suggests existence of a threshold on the rate of emails sent to a correspondent beyond which there is an adverse effect on eliciting a reply.

3.2.6 Org Level of Sender

By visual inspection of Fig. 16, we observe that the dependence of the probability of reply on the sender level does not present any monotonic patterns with a notable spike centered around at sender level 5. We have also examined the dependence on the identity of the sender. Fig. 17 demonstrates that there is a dependence on the sender identity, indicating large diversity across user profiles.

3.2.7 Email Size

Fig. 18 shows the estimated probability of reply conditional on the email size. As in the previous figures, the reported points are the ones with 95%-confidence intervals less than or equal to 0.01. We observe that the probability of reply grows from about 0.005 to 0.3 over the interval from 1 to 10 KB and then remains concentrated at around 0.3 over the interval from 10 to 100 KB. We also observe some spikes at around 500 KB with values roughly at 0.07.

3.3 Analysis of Variance

To examine the significance of these factors more rigorously and obtain an estimate of error should these factors were used to predict a reply, we conducted one-way analysis of variance (ANOVA) on a random sample of 0.5M emails. The analysis was performed for discrete values of factors ob-

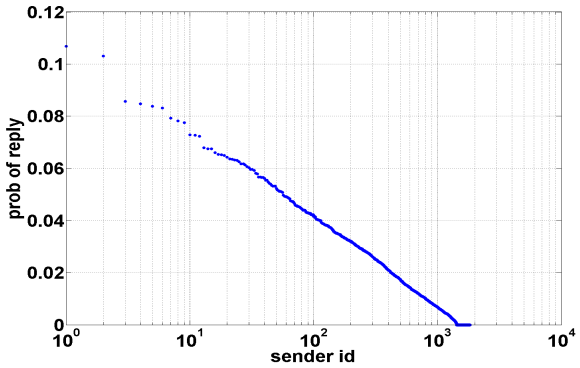

Figure 17: Probability of reply conditional on the sender identity.

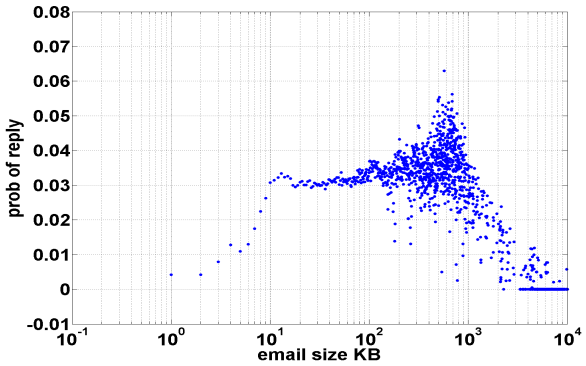

Figure 18: Probability of reply conditional on the email size.

tained by log binning and rounding of the original values to limit the number of values for factors such as email size.

For factors 1 to 7, ANOVA analysis resulted in p-values equal to zero and F values ranging from 275 to 4325 indicating that all factors are important. The orders of factors in decreasing F value and increasing mean square value are the same and follow the discussion order over the previous sections from the most to the least significant according to the test. The mean square values span the range of values from 0.0307 to 0.0349. As a point or reference, we consider the unconditional probability of reply that amounts to 0.0344 which corresponds to the mean prediction error of about 0.0332. Hence, conditioning on factor 1 (recipient list size), we obtain a moderate gain of about 5%.

We also ran a two-way analysis of variance over all distinct pairs of factors for random samples of 20,000 emails obtaining consistent results. In summary, based on the p-values, we observed that all the factors are significant, with the exception of the sender level combined with the recipient list size or the ancestor distance. The four factor pairs with the smallest mean square error all contain the recipient list size factor, which, recall, was already found to be most significant by one-way analysis of variance. The smallest mean square error was obtained by combining the recipient list size with the email size, equal to 0.287 corresponding to 21.60% prediction gain.

In summary, we found that individual factors yield moderate gain with respect to the prediction error and that this can be significantly improved by conditioning on more than one factor. We have attempted further running ANOVA

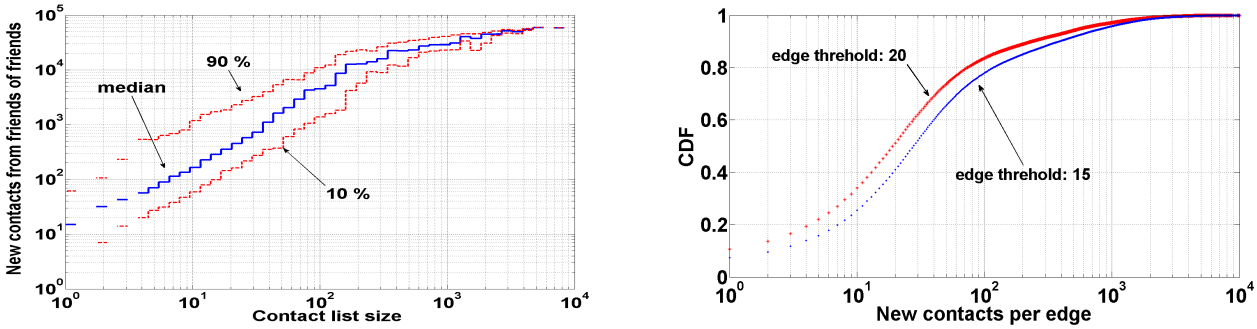

Figure 19: LEFT: New contacts discovered through friends-of-friends relationships versus the original contact list size. The number of new contacts grows linearly up to roughly contact list sizes of order 100. RIGHT: CDFs of new contacts discovered per local contact for two edge thresholds in the email graph. About 100 new contacts can be discovered through roughly 80% of the local contacts.

combinations of several factors, for example combinations of three or more. However due to the explosion of the combinations of values for the various factors, processing was not feasible with non-trivial sample sizes (more than a few tens per combinations of values).

4. AUGMENTING CONTACT LISTS THROUGH FRIENDS-OF-FRIENDS

Most contemporary email client applications are limited to extracting information from "local" emails, either sent or received by the the specific user. Using this local knowledge, these clients then support features for processing emails and other business-related tasks through email. Use of global information, for example a company's address book, is sometimes present but is rather employed for a limited set of tasks, such as meeting requests or sharing of basic information (e.g., business titles).

We believe that email clients and the email service in general could greatly benefit by information extracted from several email users and aggregated through collaborative filtering techniques. Such a service could support advanced features, present in most of today's social networks, such as expert finder, people search, etc. Thus, integration of social-networking features such as discovery of new contacts or information sharing among friends appears attractive.

Motivated by these potentials, we explore here the extent to which the graph induced by email communications can support discovery of new contacts. Specifically, we are interested in the two-hop relationships in the email graph, namely the friends-of-friends (or contacts-of-contacts) relationships. To this end, we consider two different viewpoints. First, we examine how much a local contact list of a specific user can be augmented by new contacts discovered from this user's one-hop neighbors. Second, we consider the discovery of new contacts per contact, i.e., the augmentation of the contact list conditioning on selected users. This viewpoint is similarly of interest in scenarios of search or filtering friends-of-friends for selected contacts.

Fig. 19-left presents the number of new contacts a local contact list could be augmented with by adding friends-of-friends versus the size of the original contact list per user. For this analysis, we consider thresholding of 15 emails per edge as described in Section 2.2 to identify true contact relationships. We observe that the number of new contacts increases with diminishing returns with the size of the contact list. This is expected; if the graph was a balanced tree with branching factor b, then the number of new contacts would be equal to b^2, which is equal to the number of second-hop neighbors. Otherwise, the growth would be slower due to the existence of *triads* in the email graph, i.e., loops of length three hops.

Indeed, examining Fig 19-left, we note that the median number of new contacts follows a linear growth conditional on the contact list size of up to order 100 consistent roughly with a tree-like structure. For a contact list size of order 100, the number of new contacts from friends-of-friends is about two orders of magnitude larger compared to the size of the local contact list. Considering that a typical user in our data has a contact list of size 60, this allows a discovery of about 2,000 new contacts according to Fig. 19. While this number appears rather large, it corresponds to only an 1/100 portion of the enterprise's employees. Nevertheless, handling 2,000 new contacts would require efficient search and filtering features to assist the user.

We turn now our attention to the new contacts discovered per contact in the contact list. Fig. 19-right shows the empirical cumulative distribution function for the number of new contacts discovered per contact for two edge thresholds of 15 and 20. In both cases, we observe qualitatively the same behavior. The median number of contacts discovered is approximately 20 and about 80% of the users's first-hop contacts can offer less than 100 new contacts. This suggests that the number of new contacts discovered from a local contact may not necessitate sophisticated processing tools.

5. RELATED WORK

While there has been various studies that consider email graphs, analysis has thus far concentrated on the graph structures and their evolution over time.

In particular, Adamić and Adar [1] studied the performance of greedy forwarding algorithms and provided empirical performance results for the email communication graph of a moderate-size enterprise that employed about 400 employees. In particular, an edge of the graph was formed between two individuals if and only if they exchanged at least 6 emails over a course of 3 months (1/2 emails per week, on average). They found that the greedy forwarding that biases to next-hop with the largest degree performs poorly and argued that this is because of the exponential decay of the degree distribution. In contrast, the greedy scheme that

biases forwarding to a node with the smallest organizational distance to the destination node was found to perform well. Our work differs both on the focus, and also the scale of our study. Our analysis concentrates on examining the potential of augmenting the email service with social-networking type of features. The scale of our study is considerably larger as we consider a global, large-scale enterprise with more than 100,000 employees.

Additionally, Kossinets and Watts, as well as Eckmann et al. examined the structure of email graphs and its evolution over time [8, 9, 7]. These studies are concerned more with the temporal evolution of the relationships in email graphs; instead, our work focuses on profiling the nodes (email correspondents) and their interactions.

Part of our work on the understanding the flow of emails and its relation to the underlying organizational structure and user profiles is related to sociological literature such as that of Allen and Cohen [2]. Therein, authors considered interactions as they happen in research laboratories. The main factors that determine information flow are identified as (a) organizational structure and (b) through "technological gatekeepers". Another related work is that of Sproull and Kiesler [14] that, in particular, suggested that email promotes status equalization within the medium. While from our data we cannot test this hypothesis, our results suggest symmetry in email flow between correspondents at different organizational levels.

Email processing strategies have been examined through user studies such as that of Venoila et al [17], Tyler et al[15], or Neustadter et al [11]. These works consider factors such as the status of the correspondent or the email importance. Our work is instead based on measurable data of user actions, and considers a broader set of factors in a larger-scale measurement of user behavior regarding the email service.

Finally, we point to the line of work on the general problem of ranking expertise and interest that, in particular, considered graphs induced by email communications, e.g., [12, 4, 6, 18]. While we characterize user profiles across various dimensions, our focus is not on expertize ranking. Other related work is that of Tyler et al [16], which considers clustering of an email graph in communities.

6. CONCLUSION

We presented analysis of behavioral profiles observed in usage of emails in enterprise scenarios. We focused on understanding which factors influence email replies, and assessed their significance. We observed that only a small portion of email-send events corresponds to replies. We demonstrated that the prediction of email reply can already be improved by simple models that condition on values of a few factors, most notably the recipient list size and the email size. We also focused on evaluating the augmentation of contact lists with new contacts taken from the friends-of-friends relationships in the email social graph. This would be of interest for an advanced email service, complemented with social-networking features.

Future work may study other prediction models for email replies. In particular, of interest would be online learning models, that would account for possible temporal correlations in the email behavior of individual users. Complementing such analysis with results of a user study would enable detailed observations of email processing strategies, and how user interface designs impact human behavior.

7. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] T. J. Allen and S. I. Cohen. Information Flow in Research and Development Laboratories. *Administrative Science Quarterly*, 14(1):12–19, 1969.

[3] J.-Y. L. Boudec. Performance Evaluation Lecture Notes. `http://ica1www.epfl.ch/perfeval/lectureNotes.htm`.

[4] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise Identification using Email Communication. In *Proc. of CIKM 2003*, pages 528–531, 2003.

[5] K. J. Delaney and V. Vara. Will Social Features Make Email Sexy Again? *The Wall Street Journal, Oct. 2007.*

[6] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for E-mail expertise analysis. In *Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 42–48, San Diego, California, 2003.

[7] J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. In *Proc. Natl. Acad. Sci. 101:14333-14337*, 2004.

[8] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *ACM SIGKDD*, pages 435–443, New York, NY, USA, 2008. ACM.

[9] G. Kossinets and D. Watts. Empirical analysis of an evolving social network. In *Science, 311:88Ű90*, 2006.

[10] C. Neustaedter, A. J. B. Brush, and M. A. Smith. Beyond From and Received: Exploring the Dynamics of Email Triage. In *Proc. of ACM CHI, 1977-1980, 2005.*

[11] C. Neustaedter, A. J. B. Brush, M. A. Smith, and D. Fisher. The Social Network and Relationship Finder: Social Sorting for Email Triage. In *Fifth Conference on Email and Anti-Spam, CEAS*, 2005.

[12] M. F. Schwartz and D. C. M. Wood. Discovering Shared Interests Among People Using Graph Analysis of Global Electronic Mail Traffic. *ACM Communications*, 36(8):78–89, 1993.

[13] X. Shi, L. Adamic, and M. Strauss. Network of Strong Ties. *Pysica A*, 378(1):33–47, 2007.

[14] L. Sproull and S. Kiesler. Reducing Social Context Cues: Electronic Email in Organizational Communications. *Management Science*, 32(11):1492–1512, 1986.

[15] J. R. Tyler and J. C. Tang. When can i expect an email response? a study of rhythms in email usage. In *Proc. of ECSCW, 239-258, 2003.*

[16] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as spectroscopy: automated discovery of community structure within organization. *Communities and technologies*, pages 81–96, 2003.

[17] G. D. Venoila, L. Dabbish, J. J. Cadiz, and A. Gupta. Supporting Email Workflow. Technical Report MSR-TR-2001-88, Microsoft Research, 2001.

[18] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise Networks in Online Communities. In *Proc. of WWW 2007*, Banff, Alberta, Canada, 2007.