Email Information Flow in Large-Scale Enterprises

Thomas Karagiannis and Milan Vojnovic
Microsoft Research
{thomkar,milanv}@microsoft.com

May 2008

Technical Report
MSR-TR-2008-76

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Abstract–We present analysis results of email communications in a large-scale enterprise network. Our study first focuses on understanding the social graph induced by email communications between individual users. Specifically, we examine how email communication flows are correlated with user profiles, the organization structure, and how outside information penetrates the enterprise. We then concentrate on understanding the information processing load imposed to users and the strategies applied by users in email triage. To the best of our knowledge, this is the first measurement study of email communications of a global enterprise network comprising email data from over 100,000 employees spread across multiple continents. Our analysis results inform the design of network applications that takes into account typical user behaviour in social interactions and solitary information processing. Our large-scale dataset further allows us to examine the validity of several hypotheses suggested by the social network theory.

1. INTRODUCTION

Despite the widespread use of the email, surprisingly, little is known about the properties of the email service and the email social graphs formed in large-scale enterprises. Enterprise social networks exhibit several peculiarities when compared to other popular social networks, as email communications follow from the day to day work collaborations, and the information flow is intimately related to the underlying organizational structure and the existence of "information gateway" user personas, for example.

Recently, and influenced by the explosion of social computing services, there has been a lot of discussion on enhancing the email service with new social features, and on the other hand, integrating the email service with social networks and other services [4]. Such integration however requires a comprehensive understanding of the intrinsic usage characteristics of the email service, and in particular of how information flows through email networks. This understanding is thus important in order to (a) to guide the design of new features at the service overall but also at the email client side, and (b) to potentially leverage the social network induced by email communications for other online services.

In this paper, we examine the usage of the email service and the information flow in a large-scale, multinational corporation with more than 100,000 users spread across several continents. Our analysis is based on a measurement trace that covers all email communications within the enterprise over a five month period, and amounts to Billions of email exchanges across enterprise users (Section 2). 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 analysis provides a series of characterization results that could guide an informed design for the future enhanced email service. Specifically, our contributions are summarized as follows:

We characterize the overall email workload focusing on the global email volumes across the enterprise and the email size distribution (Section 4). This characterization is of relevance for both server capacity dimensioning, and the client design. We found that the typical email size is rather small, with a median of just 25KB; yet, it varies over a wide range.

We provide several characterization results of email service from the viewpoint of individual users (Section 5). We first asses the information load generated by and imposed on users by the email service.More than a quarter of the replied emails in our trace correspond to email replies of the most-recently received emails; additionally, more than half of all replies correspond to replying to the 10 most-recently received emails. These results indicate that typically users are rather efficient in handling the email processing load. These observations are further confirmed by the email response times versus the time of day, where median response times are consistently less than 1 hour throughout working hours. Further, we show that the distribution of email inter-send time by a user approximately follows a power law decay up to order half a day and is exponentially bounded beyond. Interestingly, similar observations were also found to characterize other aspects of human activity [6, 20].

We then focus on characterizing the email social graph (Section 6). The enterprise email graph is of interest for the informed design of social networking features. Our analysis presents results that go beyond previous work [1] (see Section 3) both with respect to the properties presented, but also to the scale of the email network. Our findings provide clues about the searchability of the social email graph and indicate that the graph is rather robust to removal of edges of infrequent conversants. Note that this is not an intuitive result as some graphs heavily depend on weak ties for global graph connectivity. We further characterize the effective number of correspondents for typical users with a significant number of email communications. We find that, typically, the users' 10 most favorite correspondents account for more than half of the emails sent by this user.

Finally, we examined how information flow relates to the organizational structure of the enterprise. We found that email exchanges are symmetric between users at different organizational levels. "Information relay" usertypes appear to be infrequent in the email enterprise

network, with 85% of users forwarding less than one external email per day. This finding suggests that external information cannot easily penetrate the email social graph. External emails typically reach small groups, of order 10 users, with only some cases reaching a larger coverage of internal users. We observe significant email flow within organizational levels that diminishes towards higher and lower levels. We believe that this is partly due to different numbers of users per level and the underlying process by which relationships are built between users. We further examine what the volume of email flow across users reveals about their organizational relationships. We find that the volume itself does not appear as an efficient indicator for a direct link in the organization graph between two users. However, this estimator improves when conditioning on the email recipient list size.

To the best of our knowledge, our analysis is the first large-scale characterization of the email service and its associated social graph across a global enterprise. While the characterization is based on a particular dataset which describes the environment of a specific corporation, we believe that most of the properties described in this paper will hold in most email enterprise networks.

2. DATA AND METHODOLOGY

The results presented in this paper are based on email logs from Microsoft Exchange servers that cover all email communications across employees of a large multinational corporation. This global enterprise consists of over 100,000 employess spread across 100 countries and 6 continents. The roles of the employees vary from developers, and administrative stuff to researchers and business units. Overall, the diversity of our data facilitates a comprehensive study of email flows at a large scale.

Specifically, the analysis throughout the paper is based on the following datasets:

Exchange logs: The logs contain approximately 705M entries, one per email item, across a period of roughly 5 months and amount to a size of roughly 145G compressed. Collection started on Sunday, 27 May 2007, and ended on Wednesday, the 31^{st} of October, 2007. Each log entry specifies the sender and recipients per email, the subject of the email, timestamp, the size of the email in bytes, and other information such as the exhange servers involved, email ids etc. As internal emails would appear in different exchange servers within the Enterprise, all logs were preprocessed so that duplicate entries were removed based on email-ids associated to each email item, and which was consistent across all exhange servers for the same email. After the preprocessing and cleaning phase, our logs contain around 1.2 Billion unique email exhanges (between a sender and a receiver [1] when counting only the number of internal correspondents.

Org-structure: A significant portion of our analysis examines the flow information with respect to the organizational structure of the enterprise. 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 organizationl title. Further, we can extract the organizational tree and identify "report-to" relationships for each user (i.e., identify each user's manager and direct reports). Each node of the tree represents an employee, with the parent node being the users' manager and the nodes' children reflecting the users' direct reports. Throughout the paper, we will use the notion of organizational distance between two employees. This distance is defined as the difference between the distances of each of the two employees to the root of the organizational tree. As such, larger levels denote greater distance from the root of the tree.

Due to the size of the dataset and the complexity in processing the large graph of the email social network, we will focus our analysis on ten days at the end of September 2007, unless otherwise specified in the paper. During these ten days, we have obseved roughly 22 Million log entries, with 31 Million unique internal email exchanges, and 75 Million receipts of email across users (externally or internally).

3. RELATED WORK

In this section we discuss related prior work and discuss how the present work differs from it.

In [1], Adamic and Adar studied performance of greedy forwarding algorithms and provided empirical performance results for 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 3 months (1/2 emails per week, on average). Their data suggested the hypothesis that the distribution of the node degree decreases exponentially. They further observed that the probability of an edge between two nodes decreases exponentially with the organizational distance.[2] They found that the greedy forwarding that biases to next-hop with the largest degree performs poorly and argued that this because of the exponential decay of the degree distribution. In contrast, the greedy scheme that biases forwarding to a

[1]Note that as often recipient lists contain more than one email recipients this number is significantly larger than the number of log entries.

[2]Note that their definition of organizational distance differs from ours – they define it as the length of the shortest path connecting two nodes in the organizational tree, while we define it as the difference of the organizational levels.

node with smallest organizational distance to the destination node was found to perform well. It was argued that this is because of matching hypotheses to a model proposed in [8]. Our work differs in that we consider a global, large-scale enterprise network and we consider a broader set of questions than focusing only on the graph robustness to edge removal. In the later particular scope a novelty of the present paper is in that we also consider directional graph connecting users.

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". The notion of technological gatekeepers refers to individuals who forward the information from an external source. This "two-hop" information flow is well recognized in the context of influence spread (e.g. a person referred to a product through mass-media such as TV and this person then refers to a friend) – see the book by Katz and Lazarsfeld [7]. In the present paper, we analyze the information flow and its relation to organizational structure as well user profiles such as relaying information from external sources that may be seen as a proxy for technological gatekeeper behavior. Another related work is that of Sproull and Kiesler [16] 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.

In the present paper, we examine how email information is forwarded through the network. This aspect has connections with the information spread, e.g. [9, 10] though note that one important difference is that in the referral systems the underlying user information goal or incentive may be entirely different – e.g. in our system the goal may not be aligned with that of promoting a product or making some personal material gain but could be information sharing reason or some form of social capital transaction (Lin [11]).

There have been some reported work on how users process emails. Early user studies such as that of Venoila et al [19] identified factors that users may use for email processing prioritization that include status of the correspondent and whether the message was directly addressed to the user. Neustadter et al [12] reported results of a user study that suggest that users triage emails either in single or multiple passes and handle "unimportant" emails first because they can quickly delete or file them. Neither of this paper does not appear based on large scale measurement of user behavior in using the service.

Shi et al [15] considered whether the information can be transmitted widely and rapidly in strong-tie social networks. Specifically, a tie was defined as a closed triad (two nodes connected directly and indirectly through a third node). The graphs were that of a student campus online community and instant messenger network. They found that removing weak ties shrinks the giant component only gradually and suggested that this might have been because users belong to different and overlapping communities of interest. The connection with our work is in that we examine the robustness of graph connectivity to edge removal, in particular to those of infrequent conversants.

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. [14, 3, 5, 21]. While we characterize users profiles across various dimensions, in this paper, our focus is not on expertize ranking. Other related work is that of Tyler et al [18], which considers clustering of an email graph in communities.

4. EMAILS

We first characterize basic properties of the email service as observed from the enterprise network as a whole. Specifically, in this section we focus on the total volume of emails over time for the entire dataset, and the distribution of email sizes. These global characteristics are of general interest for system dimensioning purposes both at the server and the client side.

4.1 Email Service Workload

To characterize the overall workload, we examine the volume of email traffic (in number of email items) over the 5 months covered by our dataset. Fig. 1 displays this volume for the whole enterprise. Additionally, we condition on internal emails, that is, emails originating from and destined to employees of the enterprise, and further for internal senders and receivers in the US. Fig. 1(top) presents the number of email items seen (irrespective of the number of senders or receivers per email) and highlights a traffic workload of 4 million emails per day. The time-series appears periodic, with weekly periodicities, and traffic peaking usually from Tuesday to Thursday. The two small dips that appear in the time-series are due to our measurement collection problems at a few of the servers for these periods.

While examining email items provides some intuition for the overall workload, email service load also depends on the number of recipients per email (i.e., how many times a sent email item needs to be replicated). This perspective is presented in Fig. 1(bottom), which examines the aggregate number of emails per sender-receiver pair per day for internal emails. As often recipient

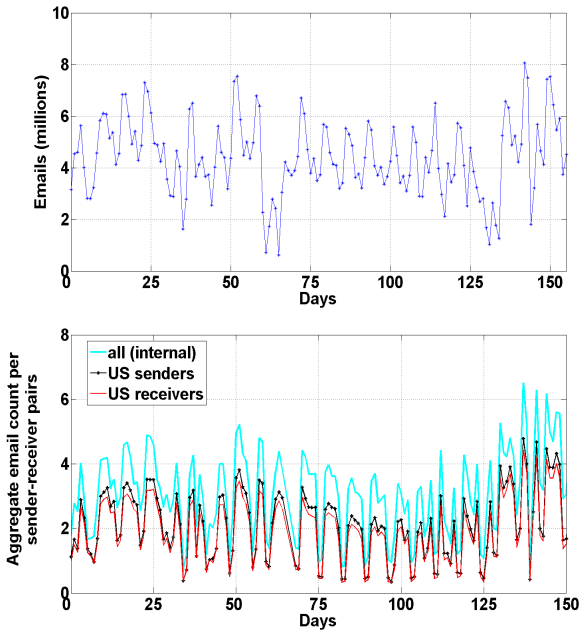

Figure 1: Number of emails per day. Total email items (top) and aggregate count across emails per sender-receiver pairs for US internal employees.

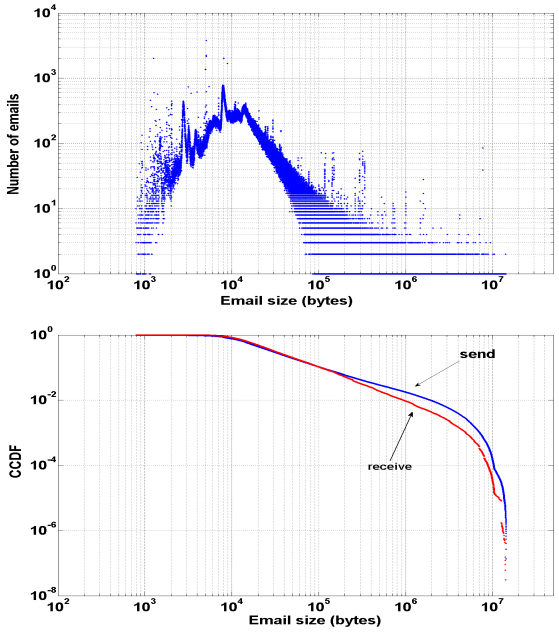

Figure 2: Email size in bytes: (Top) histogram and (Bottom) CCDFs for send and received emails.

lists contain more than one correspondent, this volume would be higher compared to the number of sent email items. We observe that weekly periodicities are evident when only users from US are considered. Seasonal trends are also observed as the volume increases towards the last months of the dataset (September-October). The mean workload is roughly 7.4 million emails per day. Examining separately sent and received emails, we observe that the workload for internal emails originating in the US is roughly 5 million (4.5 million for the receiver case). While these results provide further initial observations with respect to the email processing load imposed on users, we will examine this load in more detail in Section 5.

We next consider the distribution of email sizes in Fig. 2. The top figure presents the histogram of email sizes for our sample of 10 days (see section 2), and the bottom the Complementary Cumulative Distribution Function (CCDF). The histogram of the email size indicates multiple modes with peaks around 3KB and 8KB, median of 20KB and the 90% qunatile of the distribution being 1MB. The CCDF of the email size exhibits an approximate power law decay over a wide range, spanning email sizes of the order 10KB to 2MB. Beyond 2MB, the CCDF is exponentially bounded with a cut-off at 14MB. Overall, the results indicate high diversity of email sizes. However, the majority of emails appear to be textual conversations as denoted by the small median.

The characterization of email sizes is informative when considering the email transfer latency, especially in cases where the network capacity is a limiting resource, e.g. mobile device scenarios. Small email sizes are attractive as synchronization of mobile devices with the server inbox can allow for emails to be transfered in their entirety, as opposed to partial downloads with completion only if user makes an explicit request.

5. USERS

Having examined the email service workload as a whole, in this section, we characterize email from the user viewpoint. In particular, we (a) characterize the email processing load imposed to users, (b) infer how effective user prioritization strategies for email processing are, and (c) how this prioritization depends on factors such as the recency of a received email and the organizational status of the email sender.

5.1 Information Load

What is the typical information load generated by and imposed to users from email? To address this question we consider the number of emails sent and received per user. Fig. 3 presents the corresponding distributions, showing the histograms of all emails (left) and the histogram of internal only emails (right). Additionally, Fig. 4 presents the CCDF for the internal email communications (left) and the scatter plot of sent to received emails per user (right).

Specifically, we find that the median number of emails sent per user per day is approximately 3 with a 90% quantile of about 8 emails per day. These numbers im-

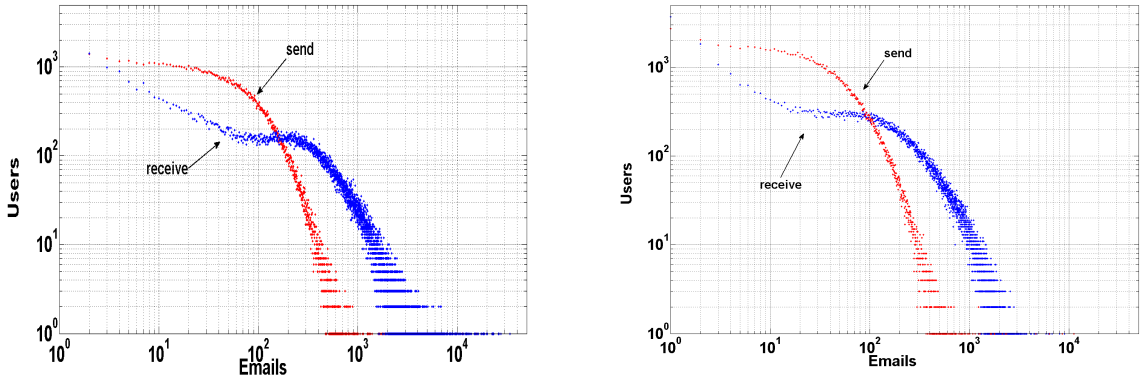

Figure 3: **Number of emails sent or received per user: Histograms of all (left) and internal-only (right) email communications.**

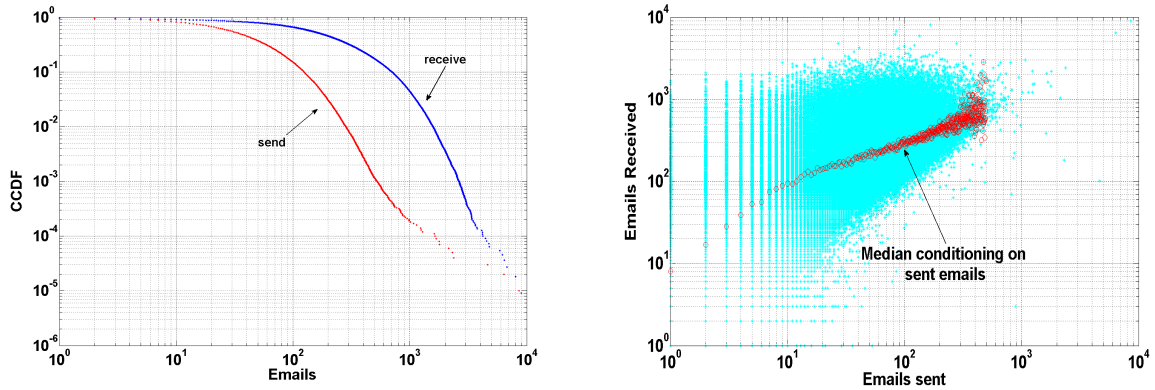

Figure 4: **Emails sent or received per user: CCDF for enterprise-internal email transfers (left). Scatter plot of sent vs. received emails, and the median when conditioning on emails sent (right).**

ply that most enterprise users do not generate significant information load.[3] We observed that the CCDF of the emails sent decays exponentially up to roughly 10 emails per day.

Looking at the receiver side, we observe an asymmetry when comparing the distributions. The median number of received emails per user is about 15 emails per day with a 90% quantile of about 45 emails per day. The ratio of received-to-sent emails is approximately 5, which is 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). We will discuss how information flows within the email network in more detail in Section 6.

Overall, these observations suggest that qualitatively users are rather diverse in the amount of their processed emails. Looking at the scatter plot, and conditioning on the sent emails, shows that sending an email will result in disproportionate number of received emails (e.g., the median number of received emails is 100 for users that sent 10 emails) up to roughly 200 emails sent, where points start to converge in the diagonal. Finally, some user profiles are also visible with users of "sender-type" and "receiver-type" for points significantly away for the diagonal.

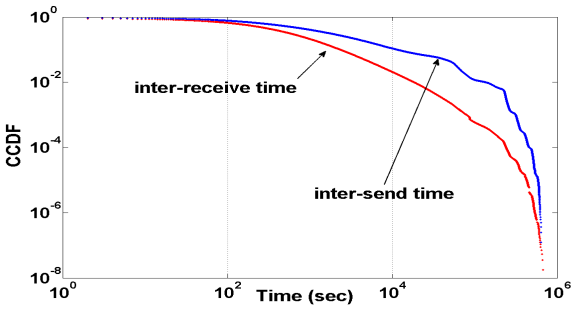

Figure 5: **CCDF of email inter-send and inter-receive times.**

[3]Enterprise users may sent or receive emails from non-enterprise email accounts, so our metrics here present a lower-bound on the generated load. However, these number do reflect work-related information load from email.

5.2 User's Email Activity in Time

We now assess user activity with respect to email processing. To this end, we examine the distributions of email inter-send and inter-receive time (i.e., time passed between two successive send and receive events per user respectively). Fig. 5 shows the respective CCDFs. We observe that the median email inter-send and inter-receive times are about 10 min and 5 min, respectively. The distribution of inter-receive time is concentrated to smaller values than the distribution of inter-send time, which is a natural ordering as sent emails per user originate from a single individual while those received originate in general from several individuals.

The inter-send time distribution is of particular interest since it reflects human activity. In particular, the small value of median email inter-send time and the substantially larger mean email inter-send time (2 hours) suggest that inter-send time distribution consists of a large number of short duration samples and a few long-duration ones (consistent with working hours). This is further reflected in the shape of CCDF of email inter-send time, which exhibits a slow decay up to order half a day and decreases faster beyond. Qualitatively similar distributions have been observed to characterize various aspects of human activity [6].

5.3 How Do Users Process Emails?

Efficient design of network applications presupposes knowledge of user behavior as for example temporal user behavior shown through the inter-send time distribution presented in the previous section. Another important factor relates to user email processing strategies. Here, we examine what is the order by which emails are processed and how this order depends on factors such as the organizational structure and the email originator.

To identify processing strategies, we form a queue per user, where emails are stored once received from the user. Most recent items are appended at the first position in the queue. Items (emails) are departing from the queue if we observe a reply ("RE") or a forward ("FW") email sent from the user with a subject that matches existing items in the queue (excluding the "RE" or "FW" characters – also excluding automatic "Out Of Office" replies). Thus, items may depart from the queue at any order. Due to the size of the dataset and memory processing limitations, we 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). Note also that user processing here ignores all user actions besides replies or forwards since we cannot observe email deletions or reads from our dataset. However, we believe that the processing strategies observed below should also hold for these cases.

Recency of the responded emails We first consider the recency ranks of replied (or replied and forwarded) emails – Fig. 6 (left). The figure presents the histogram of queue positions for processed items and indicates high bias of processed items towards recently received emails. In particular, the median recency rank of an email reply is only about 2 with the 90% quantile at roughly 6 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 process emails in a timely manner so typically emails get processed while there are still high ranked with respect to their recency.

We further examine the portion of replied emails for a given list of recency ordered emails presented to the user. Fig. 6–middle shows that more than 25% of email replies are served while assuming recency rank 1. A list of 10 emails would cover more than 60% of responses while a list of 30 emails would cover more than 80% responses. Looking at the response time versus email recency rank (Fig. 6–right) shows that the median email response time is about 1 day for the emails of recency rank 50. This number may be a bias of typical desktop screen sizes and MS Outlook list containing about 50 emails in the received rank view, in which case the result might imply higher reply probability for items in the current view pane. Overall, Fig. 6–right indicates that email response times vary rather widely.

Processing time vs. time-of-day. While we expect processing time to exhibit some diurnal and weekly periodicity, it is not clear what the exact shape such periodicities would assume and what values would hold during work hours and weekend days, in particular. Indeed, Fig. 7 shows a strong weekly periodicity with response times being considerably larger for emails sent at the end of a working week or weekend days. Over week days, the email response 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 response time appears rather concentrated around a value smaller than 1 hour. 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.

Finally, we examine the aggregate processing time distribution, as well as the ones for RE and FW emails (Fig. 8). The histogram displays peaks at around 1 and 10 minutes, while the median distribution is roughly over an hour. Interestingly, the median reply time is around 50 minutes, while the median forward time over 2 hours, which suggest that users are keen to reply, but less so when it comes to forwarding information. For-

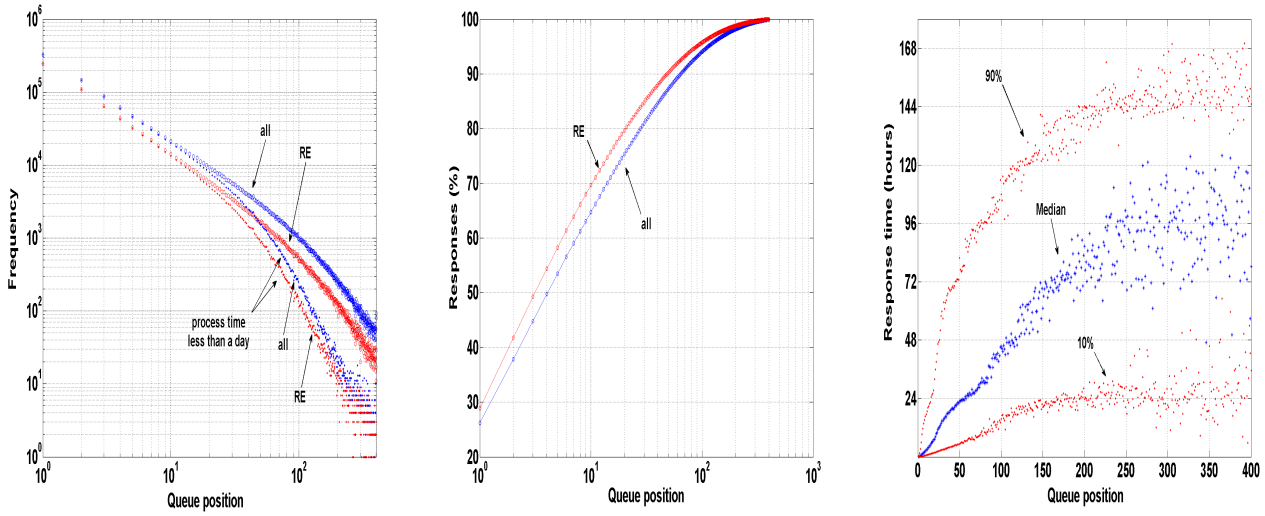

Figure 6: **Recency of email responses and response time: (Left) Histogram of recency ranks of email responses, (Middle) same as in the left plot but CDF, (Right) response time versus email recency rank.**

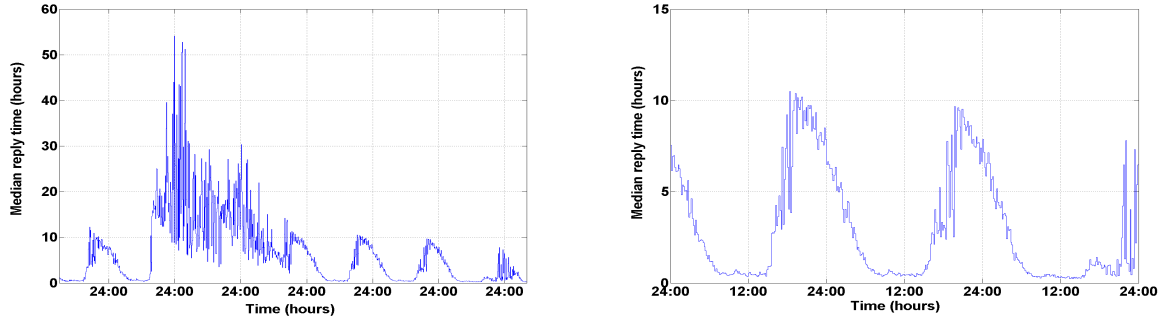

Figure 7: **Median email response time versus time of day at which the original email was received. (Right) is a zoomed version of (Left).**

warding time is an important property especially for information dissemination purposes in the email network which we extensively study in section 6. Our results suggest that forwarding does not appear to be an immediate action (very low forward times in the figure are due to automatic forwarding enabled by some users). A last note here is that the email service does not appear to be favored for interactive communication as only 5% (10%) of replies are within 1 (2) minutes of the original email (roughly 60% of the replies are within the first two hours).

5.3.1 High Org Status Reply First?

It is rather natural to assume that users would respond 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 distance between the sender and the receiver in the organizational structure. Fig. 9 presents a scatter plot of this distance vs. the response time in hours. We observe that the median email response time (denoted by the stars in Fig. 9) does not significantly depend on the organizational distance between the email correspondents. In fact, the range of email response times appears to tend to be larger, the smaller the absolute organizational distance between the email correspondents are.

6. SOCIAL GRAPH

We now turn our attention to the graph and information flow properties of the email communication network. Besides typical graph metrics such as the node degrees and connectivity, in this section, we examine (a) how information flows within the enterprise network, (b) how externally generated information pen-

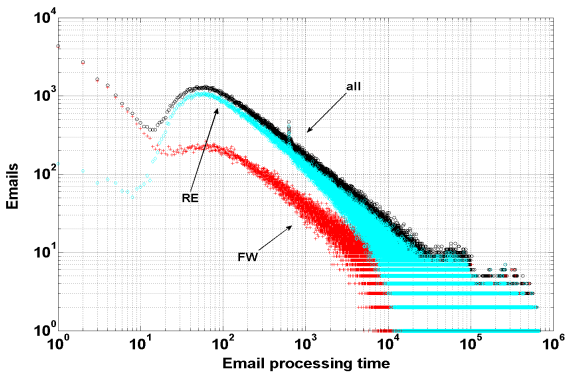

Figure 8: Histogram of email processing time.

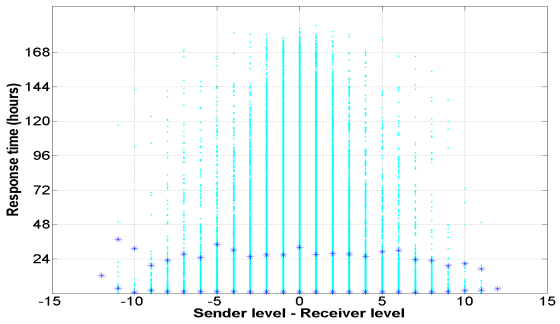

Figure 9: Email response versus org distance of correspondents. Upper stars denote the median values while the lower ones the 90% percentile.

etrates the enterprise and (c) whether the organization structure has an effect on the observed structure.

Throughout the section, we refer to the graph that is formed when considering users as nodes, and email communications between two users as the corresponding edges. In various cases, we condition the formation of an edge based on its "volume", i.e., the number of emails exchanged by two users. We refer to this volume as the "weight" of the edge. While most previous studies have concentrated on studying undirected graphs, in this section we consider both directed and undirected graphs, with the direction of the edge specifying the "sender-to-recipient" relationship. As information flow is not symmetric in the majority of the cases (e.g., Fig. 3), we believe that examination of the properties of the directed graphs is as important as those of the directed ones.

6.1 Searchability and Favorite Correspondents

One of the properties that define the graph structure is the degree of the graph nodes. For the email social network, the degree of a node translates to the number of correspondents of a specific user. Henceforth, indegree will denote the number of senders for all emails received by a particular user. Similarly, outdegree will denote the number of recipients for all messages of a specific user. In such a social graph, the degree distribution crucially affects the *searchability* of information and the discovery of expertise in the network, and as such informs the design of effective search strategies.

Our data suggests that *the distributions of indegree and outdegree appear qualitatively different.* Fig. 10 presents the distributions by applying various thresholds on the edge weight; that is, if the weight of an edge is less than the threshold, the specific edge is removed from the graph. The rational behind this thresholding scheme is to separate "well-connected" users that communicate frequently, with occasional or isolated communications that often occur within enterprises (e.g., announcements). We find that the shape of the distributions does not appear to be affected by the edge weight. Additionally, we also plot the degrees for the undirected case (Fig. 10).

The outdegree distribution exhibits power law over a wide range of contacts (10s to 1000s) for various thresholds that define an edge connecting two email correspondents (1 to 10 emails over 10 days). The median degree is 25 for a threshold of 1, and just 2 for a threshold of 10 (1 email per day). The 90% quantiles are 200 and 15 respectively. On the other hand the indegree displays an exponential decay, with median numbers of correspondents equal to 50 (threshold 1), and 2 (threshold 10). The 90% quantiles are 300 and 15. Overall, we observe that for larger weights, the number of edges decreases significantly, suggesting only a few correspondents for the majority of users.

The difference in the shape of the two distributions implies that examining only undirected graphs [1] may obfuscate significant properties of the graph structure. For example, while searchability on power-law graphs would bias queries towards high-degree nodes, such strategies may fail in exponentially degree distributions as the one observed for the indegree distribution. Our data suggests that depending on the view, different strategies may be more effective. The directed graph degree distribution appears to be a mix of the two distributions, with a small power-law range (especially for larger thresholds) that is similar to the one present in the outdegree distribution for the directed case.

6.2 Connectivity Robustness

Does a small degree imply a disconnected graph for the email social network? Intuitively, the organization structure imposed by the enterprise should ensure a connected email network, as managers communicate often with their reports and vice versa. However, it is of interest to examine how the removal of weak ties [15] (i.e., infrequent correspondents) affects the robustness of the overall graph.

To this extent, we examine the largest strongly connected component (SCC) of the social email graph by

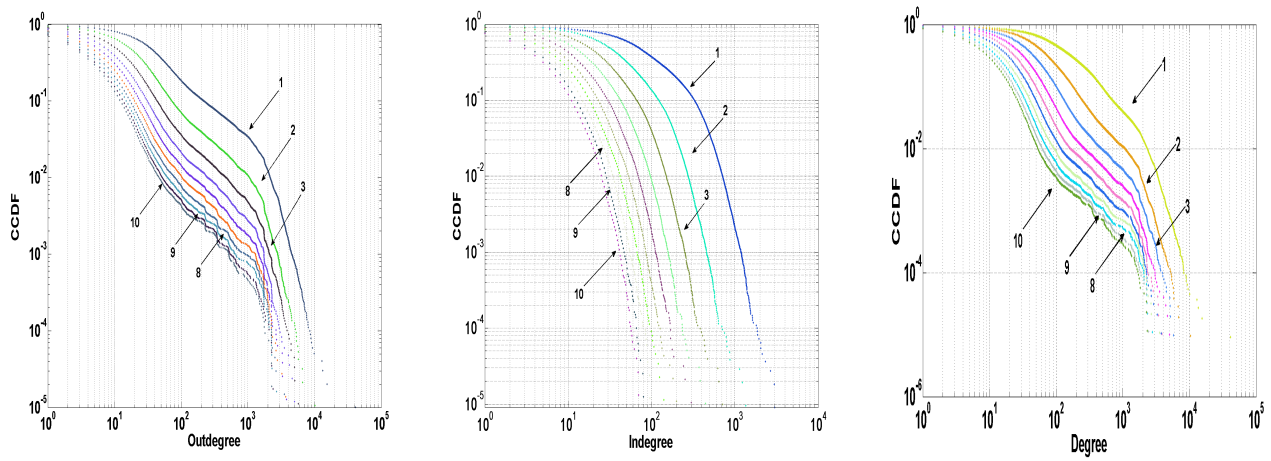

Figure 10: Outdegree and indegree for the directed graph, and degree for the undirected graph, for various thresholds of edge weight for the email social network.

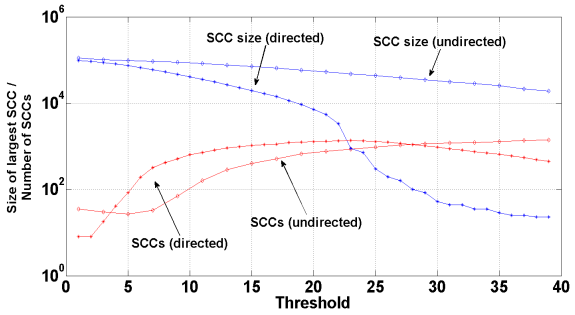

Figure 11: The size of the giant component and its relationship to the degree of the nodes, as edges are removed based on their weight. The number of strongly connected components saturate at around 1000.

conditioning on the edge weight. The SCC refers to a subset S of a (un)directed graph, such that any node in S is reachable from any other node in S. We computed the SCC with Tarjan's algorithm [17] and Nuutila's modifications [13].

Fig. 11 presents the size of the SCC for the directed and the undirected case as we increase the threshold (*x-axis*), which defines the minimum weight so that an edge is considered. Further, the figure also highlights the number of strongly connected components (excluding isolated nodes) as the threshold increases and the network becomes disconnected.

We observe that removing weak ties does not break the global connectivity of the email network. This observation suggests that a design that encourages users of an enterprise to connect to a few frequent correspondents (e.g., through the design of contact lists) would not result in global dis-connectivity. In particular, defining an edge between two users that exchanged 1 to 2 emails over 10 days, decreases the size of the giant connected components for about an order of magnitude. At the same, time, the number of connected components exhibits diminishing growth and saturates at around 1000 components. At this stage, on the average the size of the strongly connected components roughly matches the number of the minimal organizational groups in the enterprise (formed by examining the number of leaf nodes reporting to the same manager in the organizational tree). This correlation implies that the organizational structure does allow part of the network to stay connected even with nodes of small degrees.

For the undirected case, the shrinkage of the giant SCC is as expected slower, since the weight of the edge refers to communications in any direction. Finally, we observe that after roughly 2.5 emails per day, the number of SCCs diminishes for the directed case, indicating that most nodes in the graph are isolated at this point (degree 0).

6.3 Who are Emails Sent To?

The previous sections imply a small number of correspondents per users. Here we examine who are these small number of correspondents and their characteristics. Questions of interest in this section are, (a) how many people do users typically target with an email, and (b) what is the size of the recipient list of an email (e.g., how many copies of a query are propagated in the network typically?).

Recipients per email. We first consider the number of recipients per email (Fig. 12,left) across all emails, plotted against the number of distinct senders for this recipient list size. We observe that 75% of emails have just one recipient, with 95% of all emails having less than 6. Conditioning on internal senders, these percentages drop to 66% and 92% respectively. These re-

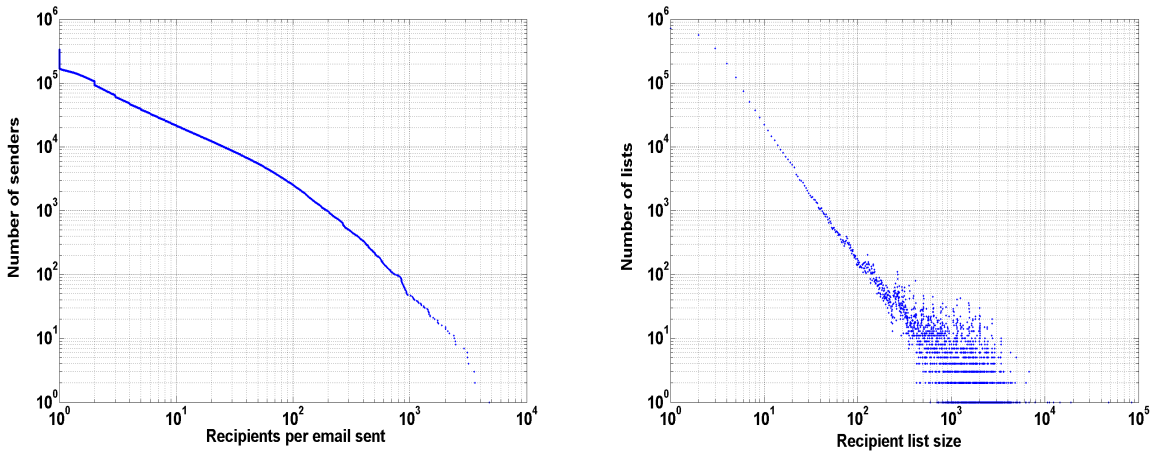

Figure 12: Histogram of the number of recipients per email sent (left) and the histogram of sizes of distinct email recipient lists (right).

sults suggests that users target a small number of recipients in the vast majority of their communication. Assuming that some of the emails also seek information, it is interesting to note that users are not inclined to use mass-emailing, but rather target small groups even within a single organization. These small group sizes are also consistent with the overall organizational structure where "report-to" group sizes are of size 6 on the average, and may suggest that users direct communications and queries to their peers. This hypothesis is further examined in section 6.5.

Distinct recipient lists. We want to characterize email recipient lists. This was already done in Fig. 12(left) for the recipient list size, but note that the histogram therein depends on the number of emails sent to a recipient list. We now provide information about the frequency of recipient list sizes over the set of all distinct recipient lists. Note that this measure depends only whether an email was sent to a recipient list or not, but otherwise not on the volume of the emails sent to the recipient list. In other words, this measure does not depend on recipient list "popularity" in terms of the emails sent, which may change over time. Fig. 12(right) displays the number of the number of recipient lists for given recipient list size. A large diversity is observed for small list sizes as expected, as most conversations are limited to small groups of users. Larger lists point towards email distribution lists of various sizes. This diversity of recipients suggests a need for personalized services. For example, recommendation engines that would provide potential recipient suggestions based on the content of the email may not achieve high hit ratios without user profiling.

Correspondents per User. So far we have seen that distinct recipients lists and correspondents are limited

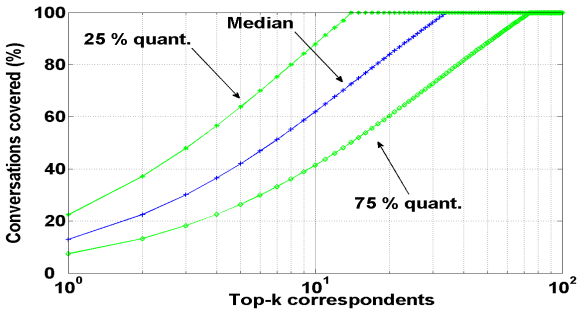

Figure 13: Percentage of user conversations covered by top-k correspondents. A list of top-5 correspondents covers more than 40% sent emails for half of the user population.

per user. However, another important property is the frequency of communications across these correspondents. Examining the "favorite" correspondents per user is of interest, as not only such edges provide the strong ties in the social network [15], but also inform the design of both network applications and devices. For example, could a mobile device display the favorite correspondents of a user in its limited screen without requiring scrolling actions from the user (typical mobile devices today fit 5-10 contacts)?

Fig. 13 exactly looks at user communications covered by the top-k correspondents of the user. For each user, we identify the percentage of the total emails sent to each of the users' recipients, and we then examine the median the quantiles across users.

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 50% of

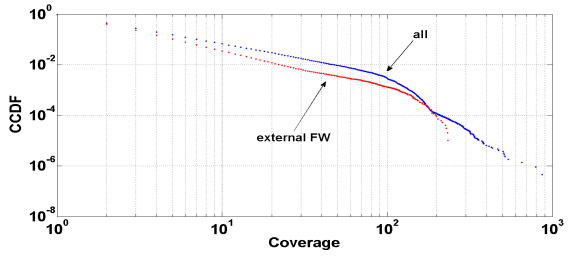

Figure 14: CCDF of the number of distinct users reached by an email.

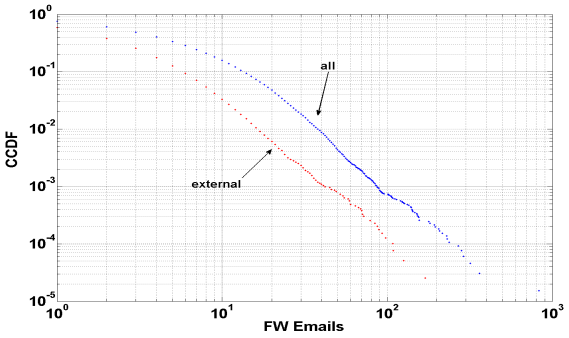

Figure 15: CCDF of the number of email forwards.

the users. Additionaly, 90% of the conversations are covered by the top-25 correspondents. Overall, a list of top-5 correspondents covers more than 40% of the emails sent for half of the user population. Increasing this number to top-10 correspondents covers more than 60% of emails sent for half of the user population. These observations provide important clues with regards to the dimensioning the 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.

6.4 Information Reach

While email as a service is targeted more towards direct communication between two correspondents, email forwarding may result in significant propagation of information within an email network, and possibly viral patterns. Here, we examine to what extent email conversations propagate within the enterprise and the overall information reach. Concretely, we wish to identify the number of users that are reached by the same email, conditioning on its subject. This reach or coverage may reflect discussions across large groups of users or indeed propagation of information through forwarding. While not directly addressed by this analysis, such information dissemination relates to questions such as, how many "hops" does a query traverse before being answered, or how far an interesting news piece propagates to?

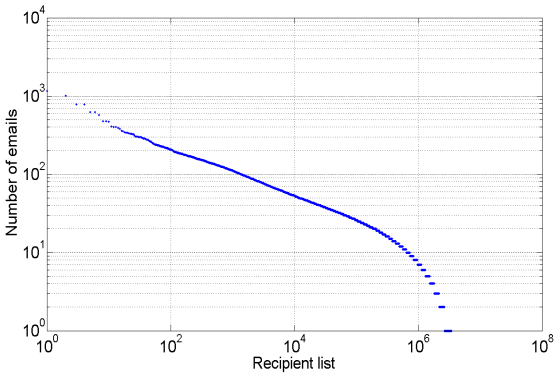

Figure 16: Histogram of email recipient lists.

Email coverage. We define coverage as the number of users that have been reached by the same email (subject). Note that since different groups can participate in discussions that feature the same subject, we require that the set of users covered by each subject must be connected (i.e., the graph formed between the corresponding users for the specific subject is connected); otherwise, if there are several disjoint groups discussing the same subject, we regard them as separate conversation groups. Of particular interest in this analysis are emails that originate outside the enterprise, but then are forwarded internally by the first internal recipient. This is a measure of how external information may penetrate the enterprise.

Fig. 14 examines both cases of coverage; the overall coverage, and the one generated by external forwards. The figure presents the CCDF of the number of users covered in each case for distinct conversations. Note that, we limit the analysis to emails with less than 200 recipients in the recipient lists in order to exclude mass emailing, and large distribution lists. We apply this threshold since we are interested in viral type of information propagation and user conversations.

Overall, 95% of emails reach less than 13 recipients, with a corresponding number of 8 for the externally generated emails. These findings highlight limited viral propagation overall. Further, we find that there is a strong correlation of the coverage with the number of initial emails sent by the original forwarder of the email. This observation further supports the previous indication of limited information propagation through email. Possible reasons contributing to these results are the various discussion tolls (e.g., forums, distribution lists) where information reaches faster and in a more direct manner all the interested parties.

External-information relays. In general, forwarding behavior is not common in the enterprise. Fig 15 shows that 85% of users forward less than 1 email per day

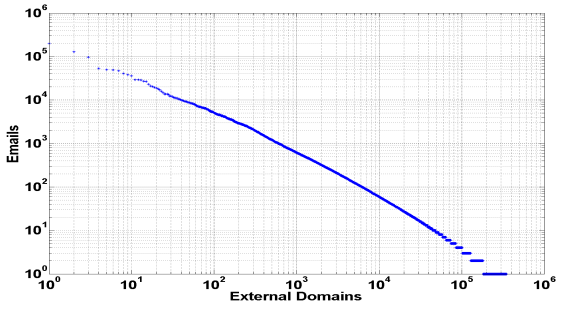

Figure 17: Histogram of external email domains.

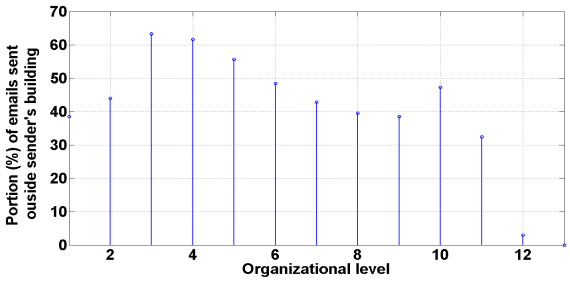

Figure 18: Inter-site (building) email transfers.

(96% for external emails) which suggests that only a small number of people act as information relays. This limited forwarding behavior indicates that it is not typical for external information to penetrate the enterprise email network. A possible explanation for this observation is that users may favor personal email accounts for non-work related information. Looking at the distribution of the number of external incoming emails per external domain where the account of the sender belongs to, indeed reveals that the three top domains are email service providers. However, in general, external emails are fairly diverse (Fig. 17) with respect to their origin domain, with the distribution exhibiting a power-law behavior.

Information propagation and physical locality. Typically, most co-workers and peers are collocated in a building within an enterprise. Coupling this physical co-location with the small number of correspondents per user, one would expect that most emails are between pairs of users that are physically located in the same building. To examine this hypothesis, we correlated email communications with physical user location as seen by the building where user offices are located. Interestingly, we found that roughly 70% of internal email communications are cross-building. Further, Fig 18 examines how locality of communication depends on the organizational level, with level numbers indicating the distance from the root. The figure presents the median value for each level, and shows that the higher the level of the user, the less locality email communications

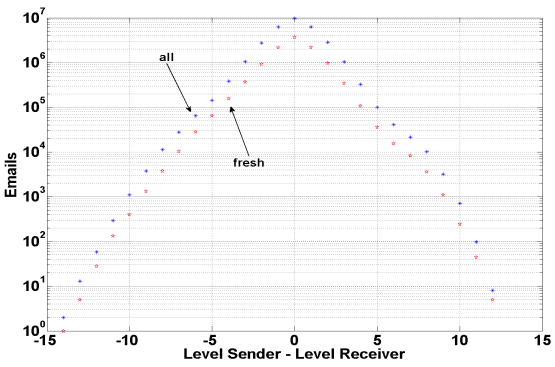

Figure 19: Email flow (log-scale) vs. level distance between the sender and the receiver for each email communication. The flow is symmetric and decreases exponentially with the distance.

feature. In the following section, we look at how the structure of the enterprise influences communications.

6.5 Email Flows and Org Structure

As discussed in the previous section, the small number of correspondents per user is consistent with the average group size for the leafs 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.

Information flow vs. org levels. We observe that the rate of email flow decreases exponentially with the organizational distance between the correspondents. Fig 19 displays the number of emails sent versus the distance in levels (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.

Fig. 19 however does not reveal how intra-level communications depend on the actual level, since this relationship is abstracted in the distance metric (*x-axis*). To examine possible dependencies, we plot the portions of intra-level and inter-level communications versus the level in Fig. 20. The figure on the left shows high intra-level communications for middle levels, while the one on the right highlights that upward (towards the root) email flow grows with the level. Note that both figures are partly explained by the organizational structure where the majority of employees are at the middle levels. However, the figure on the right does show some asymmetry where higher levels tend to direct more flow

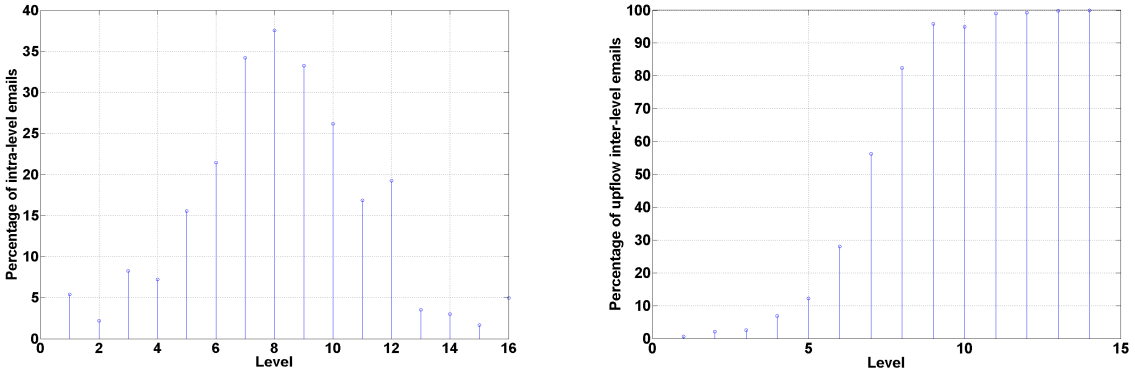

Figure 20: (Left) Portion of intra-level email transfers versus org level. (Right) Portion of inter-level flow to upper org levels.

Type/Rec. list size	1	2	3	4	5	6	7	8	9	≥ 10
Rep− >Mng	0.49	0.36	0.25	0.16	0.11	0.08	0.06	0.06	0.04	0.049
Mng− >Rep	0.3	0.23	0.17	0.12	0.08	0.08	0.06	0.07	0.6	0.03
Peer− >Peer	13.65	9.54	6.41	4.25	3.20	2.41	2.01	1.63	1.34	13.65
Other	85.56	89.87	93.17	95.48	96.79	97.43	97.87	98.24	98.55	85.56

Table 1: Frequency of user-pair types.

upwards, compared to the flow lower levels send downwards.

Org relationship vs. frequency of email communications. Does user behavior depend on the relationship between the email correspondents? Intuitively, most communications should reflect the organizational relationship between the correspondents, for example one's peers, manager or direct reports.

To examine this hypothesis, we first look at the portion of email communications that occur between users who are likely close collaborators according to the organizational structure. To this extent, we consider user pairs that are either in a peer relationship (i.e., report to the same manager), or have a report-to-manager, and vice versa relationship. Table 6.5 that only a small portion of email exchanges occur within these given relationship types. This observation holds irrespective of the size of the recipient list.

The above observations indicate that he portion of email exchanges is not an effective indicator of the distance between two correspondents within the enterprise. However, extracting such relationships from email communication is of interest for the identification of relevant contacts for employees that are close in the organization structure. To this end, we further examined the relationship between "favorite" correspondents; for each user, we identified the employee that exchanged the most emails with, and then examined their organizational relationship. Table 6.5 presents these results in a similar format as Table 6.5. Interestingly, conditioning on large recipients lists, we find larger portions of peer-to-peer relationships. This observation may be explained as a consequence of peers collaborating and addressing emails to the group of their collaborators.

In general however, frequency of emails does not directly relate to the organizational structure.

7. CONCLUSION

We presented characterisation results on email information flows in a large-scale enterprise. The results span characterization of email workload, user information load, user information processing, and social graph induced by email communications. We believe that these results enhance our understanding of the email service usage in corporate environments and inform the design of novel application features.

We believe that this is only a first step towards understanding the knowledge dissemination in enterprise environments through the email lense. Future work may broaden the analysis to cover other interesting aspects of the information flow and enterprise social graphs, and extend the analysis to corporations from other than software area.

Acknowledgements

We are grateful to the Microsoft Exchange team for providing us with data and valuable assistance. In particular, we would like to thank Georgia Huggins and Dave LeClair.

8. REFERENCES

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

Type/Rec. list size	1	2	3	4	5	6	7	8	9	≥ 10
Rep− >Mng	0.46	0.52	0.57	0.55	0.50	0.49	0.55	0.57	0.47	0.18
Mng− >Rep	0.40	0.40	0.44	0.47	0.45	0.55	0.46	0.58	0.57	0.22
Peer− >Peer	12.56	14.19	14.90	15.91	16.79	19.40	19.24	18.67	18.92	3.71
Other	86.58	84.89	84.10	83.08	82.26	79.56	79.75	80.18	80.04	95.90

Table 2: Type of the user's most frequent correspondent.

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

[3] 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.

[4] K. J. Delaney and V. Vara. Will Social Features Make Email Sexy Again? *The Wall Street Journal, October 18*, 2007.

[5] 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.

[6] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter Contact Times between Mobile Devices. In *Proc. of ACM Mobicom 2007*, pages 183–194, Montreal, Canada, 2007.

[7] E. Katz and P. F. Lazarsfeld. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers, 2006.

[8] J. Kleinberg. Small-world pheonmena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS) 14*, 2001.

[9] J. Leskovec, L. A. Adamic, and J. Kleinberg. The Dynamics of Viral Marketing. In *Proc. of ACM EC 2006*, pages 228–237, 2006.

[10] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of Influence in a Recommendation Network. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[11] N. Lin. *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press, 2001.

[12] 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*, 2005.

[13] E. Nuutila and E. Soisalon-Soinen. On finding the strongly connected components in a directed graph. In *Information Processing Letters 49(1): 9-14*, 1994.

[14] 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.

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

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

[17] R. Tarjan. Depth-first search and linear graph algorithms. In *SIAM Journal of Computing 1(2):146-160*, 1972.

[18] 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.

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

[20] M. Vojnovic. On mobile user behaviour patterns. In *Proc. of IEEE IZS 2008*, Zurich, Switzerland, 2008.

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