Analyzing Gossip in Workplace Email

Tanushree Mitra and Eric Gilbert
School of Interactive Computing & GVU Center
Georgia Institute of Technology

We spend a significant part of our lives chatting about other people. In other words, we all gossip. Although sometimes a contentious topic, various researchers have shown gossip to be fundamental to social life—from small groups to large, formal organizations. Adopting the Enron email dataset and natural language techniques, we present the first study of gossip in a large CMC corpus. We find that workplace gossip is common at all levels of the organizational hierarchy, with people most likely to gossip with their peers and that it is more likely for an email to contain gossip if targeted to a smaller audience.. Also, gossip appears as often in personal exchanges as it does in formal business communication. Exploring the sentiment of gossip, we observe that gossip is in fact quite often negative. Our study provides empirical evidence of an under-researched yet important societal phenomenon and provides empirical insights by testing existing gossip theories originating from anthropology, on a real world large email dataset.

1. INTRODUCTION

One of the most pervasive societal activities involves chatting about other people. Anthropologists call conversations like these *gossip:* the absence of a third party from the conversation [Besnier 1989; Hannerz 1967]. Despite some negative social connotations, gossip is fundamental to healthy societies—from small groups to large, formal organizations [Feinberg et al. 2012]. Simply put, we use it to trade social information, information we may find very useful in the future. In fact, Dunbar [1994] goes so far as to suggest that language itself developed so we could gossip about one another. Despite being identified as one of the most important societal and cultural phenomenon [Gluckman 1963], gossip is largely under-researched.

This article presents an exploratory study of gossip in a large corpus of computer-mediated communication (CMC). Using the Enron email dataset of 517,431 messages, we look to answer the following research questions. In systems characterized by power and hierarchy—like workplaces—what role does hierarchy play in shaping how people gossip? Going further, can we infer someone's corporate rank from their gossip behavior? Is gossip limited to personal email exchanges, or does it leak into more formal business communication?

Though sometimes overlooked in an always-changing internet, email was the internet's first widespread social medium [Henderson and Myer 1977]. Email affords conversations among both small and large groups. Networks of contacts form over time, like Twitter. Unlike Twitter, however, 92% of online adults use email [Purcell. 2011]. Madden and Jones [2008] recently reported a sharp increase in the number of adults who "constantly" check their work email, a figure that has almost certainly risen as smartphones find their way into more and more pockets. In other words, it may be fair to call email the world's most successful and pervasive type of social media.

With this as a backdrop, we turn to natural language methods—specifically, Named Entity Recognition—to identify gossip in the Enron corpus. We find it present at all levels of the corporate hierarchy. We demonstrate hierarchical signatures of gossip, showing specific pathways for the transmission of gossip via email. People belonging to certain ranks are the major sources of these messages, while other ranks silently receive it. Yet others do both. We find that people gossip most with their peers, indicating their tendency to gossip within their own group, the ones belonging to the same rank. Interestingly, people have a greater likelihood to send gossip messages to smaller audiences: a fact demonstrated by deriving a power law relation between the frequency of gossip email and the number of recipients on an email.

After exploring gossip as framed by hierarchical structure, we take a closer look at the content of gossip messages. Using sentiment analysis, we search for emotional signals in gossip. We explore the sentiment associated with gossip email, finding that gossip is in fact quite often negative: 2.7 times more frequent than positive gossip.

2. QUANTIFYING GOSSIP IN WORKPLACE EMAIL

Our research is based on four complimentary datasets:

Enron email corpus: This dataset has 517,431 email messages[1], sent by 151 people between 1997 to 2002 [Klimt and Yang 2004; Shetty and Adibi 2004].

Enron job titles dataset: Researchers at USC[2] and John Hopkins gathered the status of 132 employees within Enron and generated a job title dataset for them. Figure 1(a) shows the job titles assigned to employees [Shetty and Adibi 2004].

Ranks of job titles: We referred to Gilbert's work [2012] to match each job title with a numeric rank relative to its position in the organizational hierarchy. CEOs and Presidents have greatest power in an organization and reside at the top of the hierarchy. They are assigned rank 6, while employees are at the lowest level have rank 0. Figure 1(a) depicts the relationships.

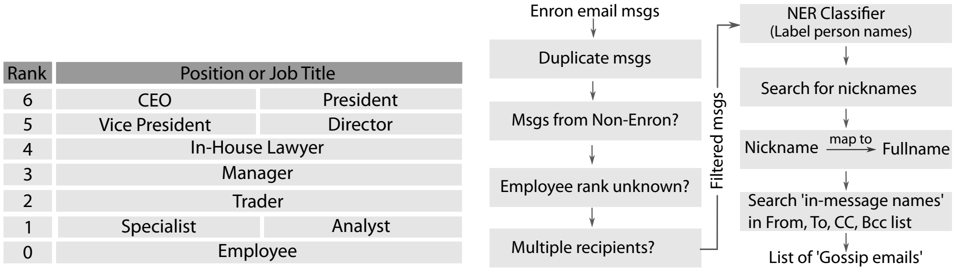
Personal vs. business email: Jabbari et al. [2006] manually annotated a subset of the CMU Enron email dataset, labeling 11,220 messages as "Business" and 3,598 as "Personal." We use this dataset for analyzeing gossip in personal and business email.

2.1 Unit of Analysis: Gossip email messages

Our computational method for isolating email messages containing gossip is outlined in Figure 1(b). Using the Enron corpus, we first remove all duplicate messages; i.e., we keep only the sender's copy of the message and discard any copies shared by other recipients. We also filter out those messages where the sender is a non-Enron employee or if his rank

[1] `http://www.cs.cmu.edu/~enron`

[2] `http://www.isi.edu/~adibi/Enron/Enron.htm`



Rank	Position or Job Title	
6	CEO	President
5	Vice President	Director
4	In-House Lawyer	
3	Manager	
2	Trader	
1	Specialist	Analyst
0	Employee	

(a)

(b)

Fig. 1: (a). Relative ranks of job titles. Figure has been reproduced from earlier work. (b). Steps for identifying gossip email from a list of Enron email messages.

is unknown. Next, we scan the body of each email to check if the sender has mentioned a name of a person and has not included him in the recipient list. Email messages satisfying this criteria are termed *gossip email messages*. They form our unit of analysis. We used the Stanford Named Entity Recognition (NER) classifier [Finkel et al.] to label words in the email body as person names. It is a common practice to shorten a person's full name (e.g., Abe for Abraham). NER also labels these nicknames as person names. In order to find the corresponding full names, we borrowed a nickname lookup file from an open source project hosted by the "Web Science and Digital Libraries Research Group" of Old Dominion University[3]. We check if any of the labeled person names are present in the nickname database. We then map any matched nicknames to its corresponding full name. For each email message we call the list of all these full names "in-message names." Next, we check to see if all the in-message names found in the earlier step are present in the recipient list. If not, then the missing names are the ones about whom the sender gossiped in the email. These gossip email messages comprise our corpus.

3. SOCIAL FACTORS UNDERLYING GOSSIP

Does position of an employee influence the amount of gossip email he sends and the audience of his gossip messages? In other words, who gossips more, bosses or employees?

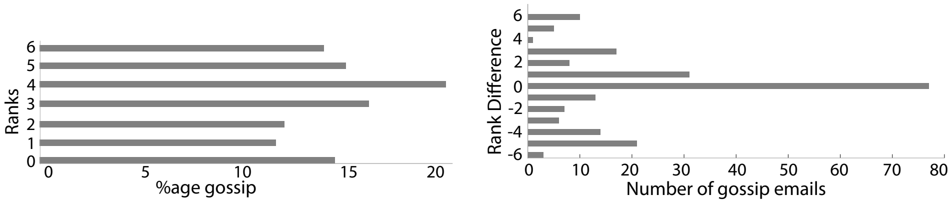
3.1 Who starts the gossip?

We first study the percentages of gossip email originating from each rank. Figure 2(a) shows these proportions. We see that gossip is a common phenomenon among every rank.

3.2 Where does gossip go?

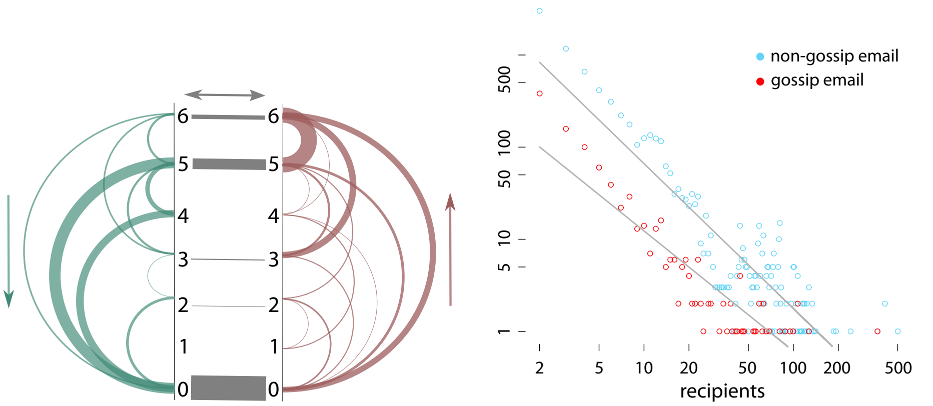
Now, let's consider the opposite question: who receives the most gossip email? We restricted our analysis to a dataset of single recipients because multiple recipients may belong to different ranks and such a mixture of ranks might be confusing for conclusions about the audience of gossip email and the flow of gossip across ranks. This resulted in a rather small dataset of 845 email messages. Next, we define "rank difference" as the rank of the recipient

[3] `http://bit.ly/m7tYcC`



(a) (b)

Fig. 2: (a). Gossip proportion varying with hierarchical rank. (b). Rank difference versus number of gossip emails



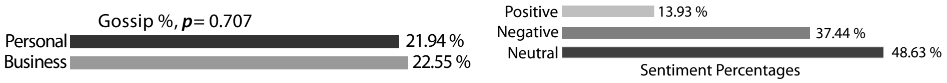
(a) (b)

Fig. 3: (a). Flow of gossip across ranks. (↑) denotes that gossip email flow up the hierarchy, while (↓) denotes downward flow. (↔) denotes that gossip stays within the same organizational rank. (b). Log-log plot of the number of recipients in the *To* list of an email versus frequency of such an email, shows that the power law relation holds true for both 'gossip email' and overall email traffic.

minus the rank of the sender. We see a huge peak at rank difference 0 (see Figure 2(b)). This implies that people mostly gossip with their peers (i.e., other employees belonging to the same rank). The next highest peak is at rank difference 1, implying that there is heavy flow of gossip messages one level up the hierarchy. There is also significant flow of gossip email four levels down the hierarchy, corresponding to the rank difference of -4.

What's behind these peaks? We produced Figure 3(a) to answer this question. Each arc in the figure corresponds to the flow of gossip email between any two ranks. Right side corresponds to flow up (↑) the hierarchy, while left corresponds to downward flow (↓). The thickness of the arc is proportional to the amount of gossip email sent, which gives us a sense of the major contributors in the flow of gossip.

One interesting thing to note from Figure 3(a) are the distinct "gossip sinks" and "gossip sources" present in either direction. "Gossip sources" correspond to the ranks which are the major contributors in generating gossip email, while "gossip sinks" correspond to the ranks which receive most of the gossip. Ranks 6 and 0 are the "gossip sinks" up and down the hierarchy respectively. Ranks 5, 3 and 0, on the other hand, are the major "gossip sources" for gossip flowing up the hierarchy, while the same is true for 5 and 4 for downward flow.



(a) (b)

Fig. 4: (a). Testing the effect of personal and business email on the proportion of gossip. (b). Percentage negative, positive and neutral emotions in the gossip email.

This clearly indicates that employees at the lowest level play a prime role in circulating gossip throughout the hierarchy.

3.3 Gossip as a function of audience size

We also explore the relationship between the number of gossip messages and the number of recipients on that email. Our test bed for this analysis consists of all email that had more than one recipient in the *To* list. We were left with a dataset of 16,500 email messages. We searched for gossip email in this corpus and noted the count of its corresponding recipients. Both regular email and gossip email roughly follow a power-law function (see Figure 3(b)). Letting y be frequency and x be the number of recipients on the *To* list, we can model the following relationship: $y \propto x^{-a}$. The exponent of the fitted line is a = 1.304 for gossip traffic and $a = 1.573$ for overall email traffic of the Enron corpus. These exponents demonstrate that sending email to a small set of people is more frequent and it is more common to see gossip in messages targeted to a smaller audience.

3.4 Gossip in personal vs. business email

For this study, we searched for gossip email in the Jabbari et al.'s [2006] manually annotated 3,598 "Personal" and 11,220 "Business" email dataset. We find that the proportion of gossip is independent of whether the email relates to personal matters or business ones, a seemingly counterintuitive result, $\chi^2(1, N = 1618) = 0.1413$, $p = 0.707$ (see Figure 4(a)).

3.5 Sentiment of gossip email messages

To analyze sentiments of gossip email, we first extracted the message body from each of 7,206 gossip messages and converted the text to lowercase. We then used the Natural Language Text Processing API provided by `text-processing.com`[4] to perform sentiment analysis. We find that a significant portion of the gossip text has neutral tone. However negative sentiment is predominantly higher compared to positive (see Figure 4(b)).

4. DISCUSSION AND FUTURE WORK

Our study reveals some important characteristics of organizational gossip. First, gossip is present in both personal and business email and across all sections of the hierarchy, which demonstrates its all-pervasive nature in organizations. Next, we showed that the hierarchical position of an employee affects his gossip behavior, both in terms of his frequency of gossip and the audience with whom he gossips. Our results indicate that people are most likely to

[4] `http://bit.ly/yGMudg`

gossip with their peers. These findings are in line with Gluckman [1963]'s theory: gossip maintains a group's unity and establishes its boundary. Building off this finding, perhaps using gossip as an indicator, an organization could build self-aware applications to spot peer groups and groups which have diverse interests.

We also show that organizational gossip is a social process. Some people are actively involved in generating gossip messages ("gossip source"), while others are silent readers of the messages ("gossip sink"'), and there are some who play both roles. Acting as a conduit of information, identifying gossip sources and sinks may help an organization locate its information hot spots.

We have studied gossip behavior in organizational email only. It would be interesting to perform the study in other communication media such as instant messaging or Facebook. While instant messaging is a purely dyadic private communication channel, where the third party has no knowledge about the interaction, gossip on a Facebook wall has every chance to be noticed by the third party. It would be interesting to see if and when gossip percolates from one social circle to another and what triggers this process. What would happen if, in group-level gossip, multiple people present conflicting facts? How would the listeners react to such conflicting information? Will it be detrimental to the unity of the group? Are gossipers eager to confirm the information from multiple social interconnections? Does the reputation of the gossiper (gossips too much or too little; gossips about positive things) determine his trustworthiness? Does the type of information (entertaining, concerning) determine their willingness to confirm? The process of confirmation might in turn cause the information to flow to other social connections. It would be interesting to see if and when the information gets garbled while cascading to different levels. More work needs to be done to explore these deep questions.

5. ACKNOWLEDGEMENTS

We would like to thank our entire comp.social group at Georgia Tech for their valuable feedback on early versions of this work.

REFERENCES

BESNIER, N. 1989. Information withholding as a manipulative and collusive strategy in nukulaelae. *Language in Society 18*, 315–341.

DUNBAR, R. 1994. *Grooming, Gossip and the Evolution of Language.* London, Faber and Faber.

EGGINS, S.AND SLADE, D. 1997. *Analyzing casual conversation.* London, Cassell.

FEINBERG, M., WILLER, R., STELLAR, J., AND KELTNER, D. 2012. The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*.

FINKEL, J. R., GRENAGER, T., AND MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc.* ACL '05.

FOSTER, E. K. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology 8,* 2, 78–99.

GILBERT, E. 2012. Phrases that signal workplace hierarchy. In *Proc.* CSCW '12. ACM, New York, NY, USA, 1037–1046.

GILBERT, E. AND KARAHALIOS, K. 2009. Predicting tie strength with social media. In *Proc.* CHI '09. ACM, New York, NY, USA, 211–220.

GLUCKMAN, M. 1963. Papers in honor of melville j. herskovits: Gossip and scandal. *Current Anthropology 4,* 3, 307–316.

GOFFMAN, E. 1959. *The Presentation of Self in Everyday Life*. New York, Doubleday.

GRANOVETTER, M. S. 1973. The strength of weak ties. *American Journal of Sociology 78,* 6, 1360–1380.

GROUP, N. W. 2001. The internet society. rfc 2822 - internet message format.

HANNERZ, U. 1967. Gossip, networks and culture in a black american ghetto*. *Ethnos 32,* 1–4, 35–60.

HENDERSON, JR., D. A. AND MYER, T. H. 1977. Issues in message technology. In *Proc.* SIGCOMM '77. ACM, New York, NY, USA, 6.1–6.9.

JABBARI, S., ALLISON, B., GUTHRIE, D., AND GUTHRIE, L. 2006. Towards the orwellian nightmare: separation of business and personal emails. In *Proc.* COLING/ACL '06. Association for Computational Linguistics, Morristown, NJ, USA, 407–411.

JELVEH, Z. AND RUSSELL., K. 2006. The rise and fall of enron. *The New York Times*.

KLIMT, B. AND YANG, Y. 2004. Introducing the Enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*.

KURLAND, N. B. AND PELLED, L. H. 2000. Passing the word: Toward a model of gossip and power in the workplace. *The Academy of Management Review 25,* 2, pp. 428–438.

LEVIN, J. AND ARLUKE, A. 1985. An exploratory analysis of sex differences in gossip. *Sex Roles 12*, 281–286.

MADDEN, M. AND JONES., S. 2008. Networked workers. Tech. rep., Pew Internet and American Life Project.

PAINE, R. 1967. What is gossip about? an alternative hypothesis. *Man 2,* 2, 278–285.

PALUS, S., BRÓDKA, P., AND KAZIENKO, P. 2010a. How to analyze company using social network? In *WSKS (1)*. 159–164.

PALUS, S., BRÓDKA, P., AND KAZIENKO, P. 2010b. How to analyze company using social network? In *WSKS (1)*. 159–164.

PENNEBAKER, J. W., FRANCIS, M. E., AND BOOTH, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Word Journal Of The International Linguistic Association*.

PETERSON, K., HOHENSEE, M., AND XIA, F. 2011. Email formality in the workplace: a case study on the enron corpus. In *Proc.* LSM '11. Association for Computational Linguistics, Stroudsburg, PA, USA, 86–95.

PURCELL., K. 2011. Search and email still top the list of most popular online activities. Tech. rep., Pew Internet and American Life Project.

ROSNOW, R. L. 1977. Gossip and marketplace psychology. *Journal of Communication 27,* 1, 158–163.

ROY, D. F. 1959. Banana time: Job satisfaction and informal interaction. *Human Organization 18,* 04, 158–168.

SHETTY, J. AND ADIBI, J. 2004. The enron email dataset database schema and brief statistical report.

STIRLING, R. B. 1956. Some psychological mechanisms operative in gossip. *Social Forces 34,* 3, 262–267.

VIEGAS, F. B., WATTENBERG, M., VAN HAM, F., KRISS, J., AND MCKEON, M. 2007. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics 13*, 1121–1128.

WATTENBERG, M. AND VIÉGAS, F. 2008. The Word Tree, an Interactive Visual Concordance. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*.

WILSON, P. J. 1974. Filcher of good names: An enquiry into anrhropology and gossip. *Man 9,* 1, 93–102.