IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Novel Approach for Detecting Stock Price Movements

Asif G Sayyad[1*], Nilesh R Wankhade[2]

[1,2] Department of Computer Engg, SAPKAL College, Nashik, India

Corresponding Author: asifsayyad444@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.191196 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

Abstract — Grounded on communication theories, we propose to use a data-mining algorithm to detect communication patterns within a company to determine if such patterns may reveal the performance of the company. Specifically, we would like to find out whether or not there exist any association relationships between the frequency of e-mail exchange of the key employees in a company and the performance of the company as reected in its stock prices. If such relationships do exist, we would also like to know whether or not the companys stock price could be accurately predicted based on the detected relationships. To detect the association relationships, a data-mining algorithm is proposed here to mine e-mail communication records and historical stock prices so that based on the detected relationship, rules that can predict changes in stock prices can be constructed. Using the data-mining algorithm and a set of publicly available Enron e-mail corpus and Enrons stock prices recorded during the same period, we discovered the existence of interesting, statistically signi_cant, association relationships in the data. In addition, we also discovered that these relationships can predict stock price movements with an average accuracy of around 80 percent. Given the increasing popularity of social networks, the mining of interesting communication patterns could provide insights into the development of many useful applications in many areas.

Keywords — Corporate communication, data mining, organizational, performance, stock prediction

I. INTRODUCTION

Recent research reveals the existence of interesting communication patterns [1] among different participants of different social network platforms. These patterns have been shown to be useful in predicting product sales [2] and stock prices [3]. Compared to a social network, which can be considered as representing connections among people in the public, a corporate network connects only employees in a big corporation. While participants of a social network can xpress opinions on an issues of interest, members of a corporate communication network are expected to mainly talk about company-specific business. If human communication patterns sales or stock performance, one may wonder if such patterns also exist among members in corporate communication network to allow the same to be done. Unlike social networks, in a corporate communication network, e-mails have long been used as a tool for inter organizational and intra organizational information exchange. In the same way, a social network platform is able to capture participants' behavior and their opinions about various issues and events. Thus, we argue that a corporate communication network in the form of an e-mail ecosystem also contains insightful information, such as organizational stability and obustness [4], about a company's development.
We believe our argument is in line with corporate communication theory [5], which suggests that "employee communication can mean the success or failure of any major change program" resulting from a merger, acquisition, new venture new process improvement approach, or other management issues.

We chose to look into the case of Enron. The reason why Enron is chosen is that an Enron e-mail corpus has been made publicly available. By mining the data set to determine if the communication patterns discovered have any association with Enron's stock prices recorded during the same period, we managed to confirm that interesting, tatistically significant, association relationships do exist in Enron's corporate communication network. In addition, we also discovered that the detected relationships could predict stock price movements with relatively high levels of accuracy. Such result confirm our belief that corporate communication has identifiable patterns and such patterns can reveal meaningful information about corporate performance. Given the increasing popularity of social networks, the mining of interesting communication patterns ould provide insights .

In other words, employee communication can serve a critical "business function that drives performance and contributes to a company's financial success" [6]. Based on these broad corporate communication theories, we hypothesize that every company has its own communication approach with identifiable patterns. We believe that these communication patterns can reflect how a company manages major corporate activities (such as mergers, acquisitions, new ventures, new process improvement approaches, going concerns, or bankruptcy) that may subsequently affect the company's performance in the stock market. In this paper, we propose that a company's performance, in terms of its stock price movement, can be predicted by internal communication patterns.

To obtain early warning signals, we believe that it is important for patterns in corporate communication networks to be detected earlier for the prediction of significant stock price movement to avoid possible adversities that a company may face in the stock market so that stakeholders' interests can be protected as much as possible. Despite the potential importance of such knowledge about corporate communication, little work has been done in this important direction. It is for this reason that we are proposing in this paper to make use of a computational approach to determine if patterns detected in a corporate communication network are related to corporate performance. There has been some effort to use computational methods to mine large-scale social media data for public sentiment [3] toward the stock market. What we propose here is similar except that, instead of public opinion in popular social network, we are investigating the existence of communication patterns in a corporate network and we propose to make use of data-mining techniques for this purpose. Specifically, we determine whether or not there exist any association relationships between the frequency of e-mail exchange of the key employees in a company and the performance of the company as reflected in its stock prices. To the best of our knowledge, there has not been any attempt to investigate possible linkages between corporate communications data and a firm's share price. If such relationships do exist, we would also like to know whether or not the company's stock price could be accurately predicted based on the detected relationships.

II. Related Work

A. Enron and Its E-Mail Corpus

As the largest collection of authentic e-mail messages that is publicly available, Enron's e-mail corpus has provided researchers with various research opportunities. Via the ocial network analysis (SNA) technique, the dynamics of the structural and properties of the organizational communication can be explored and classified by organizational levels (such as positions and ranks) [24]. Communication among employees was found across roles and ranks during the crisis. Meanwhile, the originally disconnected employees also engaged in mutual communication. In addition to identifying the key e-mail authors in Enron's e-mail communication network, McCallum et al. [25] established the role–author–recipient–topic model, a Bayesian network for SNA that focuses on discussion topics. It combines a directionalized connectivity graph with the clustering of words to form topics from probabilistic language modeling. Data mining is another research focus related to Enron's e-mail corpus. Via data-mining algorithms, the community structure in the Enron e-mail corpus was detected and visualized in a graph [26]. Similarly, Carvalho and Cohen [27] also formulated an algorithm to identify employees at Enron working together on a similar topic or project, and automatically recommended e-mail recipients based on a large multiclass multilabel classification. In a continuous effort, Pathak et al. [28] mapped the organizational networks with a high degree of detail and accuracy to analyze the proximity between actors' perceptions about such organizational networks and the divergence of an actor's misperceptions about organizational network from reality.Furthermore, researchers also organized the publicly available e-mails from Enron by establishing new databases. This stream of research includes automated classification of e-mail messages into user-specific folders and information extraction from chronologically ordered e-mail [29]. These efforts resulted in an advanced version to visualize the content-actor network that covers 123 000 unique e-mail addresses [30].

B. Communication and Organizational Performance

There have been some attempts by researchers to empirically investigate the impacts of a wide range of communication variables (such as communication patterns, feedback, culture, and specific behaviors) on performance-related variables (such as communication satisfaction, job satisfaction, individual performance and productivity and organizational performance, and productivity and change) [9]. For instance, the productivity of an individual in a corporation was found to be explainable by certain communication patterns that cover supervisory communication, subordinate communication, personal feedback, organizational integration, media quality, corporate information, communication climate, and coworker communication correlation [10]. More specifically, the downward form of communication was found to exert the greatest influence on productivity while co-workers' communication rendered the least impact. These findings are in line with other empirical experiments [11], [12] that produce results indicating that communication flow is one of the key determinants of corporate performance. In some specific domains, research [13], [15] seems to indicate that communication patterns can have significant impact on a corporation's research and development (R&D) performance. In marketing services and research domain, to top management, has been consistently shown to provide valuable information to an organization about performance of products and services [16]. In addition to internal communication among employees, there have been some suggestions that there exists a direct association between public communication strategy and share prices. For instance, corporate communication with key stakeholders has been shown in [19] to enhance potential payoffs, while stock appreciation and corporate performance have been found to be uncorrelated if communication efforts (with outsiders) are faulty. Such findings echoed studies on the strategic role of investor relations in which communication of corporate information is crucial for the perception and evaluation of the financial community including analysts, investors, and potential investors [20], [21].

Wenjing Duan et.al. examines the persuasive e_ect and awareness e_ect of online user reviews on movies' daily box o_ce performance. In contrast to earlier studies that take online user reviews as an exogenous factor. we consider eviews both inuencing and inuenced by movie sales. Nevertheless, we find that box office sales are significantly inuenced by the volume of online posting, Our analysis is, by necessity, restricted to online users who choose to post reviews and post them on YM. Thus, our estimates are conditioned on such a user population. While such a restriction does not bias the panel estimation results, they should be interpreted as applying to a
self-selected set of online users[11].

Enireddy. Vamsidhar et.al. used the back propagation neural network model for predicting the rainfall based on humidity, dew point and pressure in the country INDIA. Two-Third of the data was used for training and One-third for testing. For rainfall prediction. Artificial Neural Network was applied and the rainfall was predicted in India. According to the results backpropagation neural network were acceptably accurate and can be used for predicting the rainfall. So by using this method for prediction we can find the amount of rainfall in the region by using the attributes like humidity, dew point and pressure[12].

GREG MILLER et.al. study that Twitter is a great way to promote a product, keep up with far- ung friends and colleagues, connect with others who share your passion. Their findings paint a portrait of humanities mood swings. Positive emotion runs high in the morning, declines throughout the day, and rebounds in the evening. The same pattern occurs on the weekends, suggesting its not just work bringing people down, Golder notes. People are happier overall on weekends, but the morning peak in good vibes is delayed by a couple of hours, suggesting they sleep in. There was this intriguing paradox where for most of the 20th century we seemed to know more about exploding stars at the edge of the galaxy and the proteome of yeast than we knew about how large human social groups function, says Jon Kleinberg, a computer scientist at Cornell. But the digital detritus of 21st century life online may change all that, Kleinberg says: Interesting things happen when you can take what was once invisible and make it visible [13].

Carol Hargreaves et.al. construct a framework that enables us to make class predictions about industrial stock performances. In order to have a systemized approach for the selection of stocks and a high likelihood of the performance of the stock price increasing, several analytical techniques are applied. A trading strategy is also designed and the performance of the stocks evaluated. Future researchers may include more methods for finding the best model for predicting stock prices. We used stocks from the Industrial Sector however it would be interesting to expand our study to see whether our stock selection and trading strategy will work in other sectors. Another interesting idea we have in mind, is to build an application tool that applies our stock selection and trading strategy. Finally, our trading strategy has a 10 percent upper limitation on the portfolio and an exit strategy on a loss of 5 percent or more. Other trading strategies may be investigated [14].

Abhijit A. Sawant and P. M. Chawan et.al. describes about di_erent data mining techniques used in _nancial data analysis. Financial data analysis is used in many financial institutes for accurate analysis of consumer data to find defaulter and valid customer. For this diffierent data mining techniques can be used. The information thus obtained can be used for Decision making Boosting has already increased the efficiency of decision trees. The assessment of risk will enable banks to increase profit and can result in reduction of interest rate [15].

III. METHODOLOGY

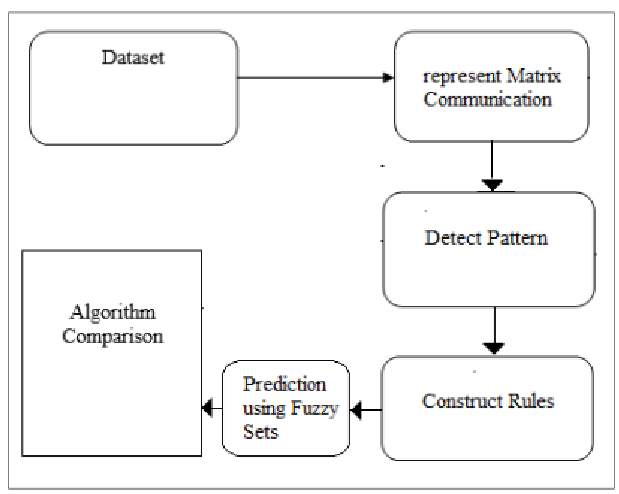

Figure 1. System Architecture

Using the social communication network among employees, the proposed algorithm can discover the relationship between communication and the movements (ups and downs) of stock price. We assume that the values of movements of stock price are temporally related to the values of frequency of communication in the previous time point. The task of our proposed algorithm is to uncover these temporal relationships and predict the movements of stock price using communication network. The proposed algorithm includes the following steps:

- discretizing the weight matrices and the value of movement.
- discovering patterns for describing the relationship between Communication frequency and stock price;
- constructing predication rules based on patterns;
- predicting the movement of stock price using prediction rules.

Discretizing Communication Matrix and Stock Price:
The value of edge dij between two different nodes for representing communication frequency between two nodes is numerous. In order to reduce and simplify the original data,numerous values are always replaced by a small number of interval labels, which leads to a concise easy-to-use, knowledge-level representation of mining results. The existing and mature methods on unsupervised discretization are primarily equal frequency discretization and equal width discretization. The equal width method is typically used in every statistic program to produce regular histograms. However, equal width discretization can hardly handle the situation if outliers exist in the data set. Equal frequency can overcome the limitations of the equal width discretization by dividing the domain in intervals with the same distribution of data points.

In a relative sense, 10 e-mails per day might be already well above the mean level of the communication frequency for a front line employee. We standardize the communication level according to each person/nodes mean of his/her communication frequency, and all continuous data are transformed into discrete data, such as dr1,dr2,dr3. Three levels are applied for Enron case which represent no communication, weak relationship, and strong relationship.

Detecting Patterns:
After discretization, we detect the prediction rules using an inductive method. It is important to know that how the value of stock price movement is dependent on the preceding e-mail communications between employees for accurate predictions. That means, if the value of ups and downs (i.e., the movement of stock price) is dsk that is always preceded at one time point earlier by the relationship between central node vc and the j th node vj on value drk , we can conclude that dsk is dependent on drk with a time point lag of one. However, if it is observed that drk is never followed by dsk at one time point later, we can also conclude that dsk is dependent on drk in a negative sense with a time point lag of one.That is, whenever a relationship between vc and vj is observed to have the characteristic dsk , the relationship that is located at one time point later in the stream will not possess the value of ups and downs in drk . When we predict the situations of price increasing and decreasing for a stock, we need to consider all the links between vc and vj.It is obvious that the association between communication frequency and stock price is interesting and worth scholarly and practical attentions. To discover all such kinds of associations patterns, we make use of residual analysis as described below.

Constructing Predication Rules Based on the Detected Patterns:
Based on the value of adjusted residual, we can determine whether stock price is significantly associated with communication frequency. If it is the case, it can be utilized to construct a characteristic description for the changes (i.e., ups and downs) of stock price. Thus, we need to ensure that they are utilized in the construction of prediction rules. The weight of evidence for or against a certain prediction of the attribute values of future objects by using the same information measure can be assessed quantitatively records stream. The weight of evidence measures the amount of positive or negative evidencethat is provided by the value of relationship between central node and destination node supporting or refusing the stock price being observed together. Hence, vjk that provides positive evidence supporting stock at one time point later in the graph stream. The value vil and vjk is the characteristics of weight for R(Nc, NJ) in the preceding time point.

Estimate Stock Price Movement Using the Prediction Rules:
After constructing a set of prediction rules, the weights of evidence are assigned for each pattern. When combining all the weights provided by the patterns that support the observation, the total weight (TW) of evidence is used for prediction by computing a total weight.The total weight of evidence for stock to be highly or lowly expressed is computed and the value for stock is determined by the one ith the greatest total weight of evidence. Finally, the same operations are iterated when other communications between other node pairs are used. Finally, a part of the records are selected as training data to construct prediction rules and obtain the value of TW to describe the correlation between communication network and the change of stock price. The remaining records are selected as the testing data.After preprocessing and discretizing the stock price and communications between central node and other nodes, the proposed algorithm needs to iterate in order to calculate the adjusted residual to measure the correlations between communication frequency and stock price in the next time point. In order to calculate the dependence measure between

one pair of communication frequency and the stock price, all the states of stock price and communication frequency need to be traversed. Hence, the run-time complexity is O(srp) for this step. For the Enron case, we discrete the communication into three levels and stock price into two levels, meaning the complexity should be O(6p). The number of iterations is n because the total number of nodes is n in the community network. Hence, the run-time complexity is srO(np). We then compare the running-time complexity with the traditional classication algorithm Decision Tree as detailed as follows: 1) the data size is p since p is the total time points and 2) the feature number is n since there are n nodes in the communication network. The complexity should be O(n2p) for the classication. When sr is equal to n, the complexity of the proposed algorithm is the same with Decision Tree. However, the srn especially when the network is large for the big data case.

IV. Results and Discussion

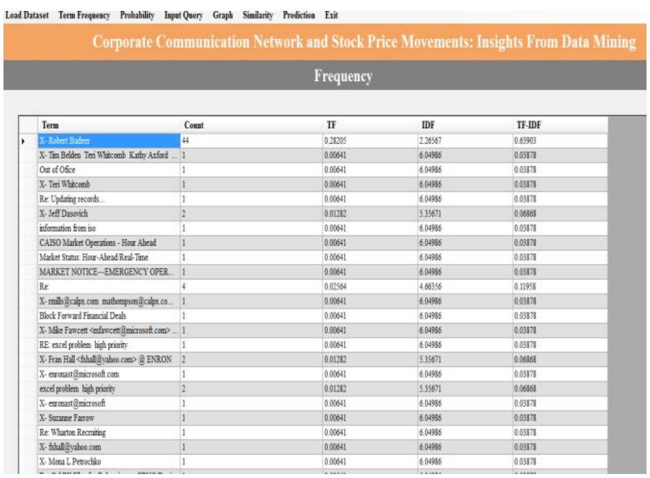

Figure 2. Frequency Data Set

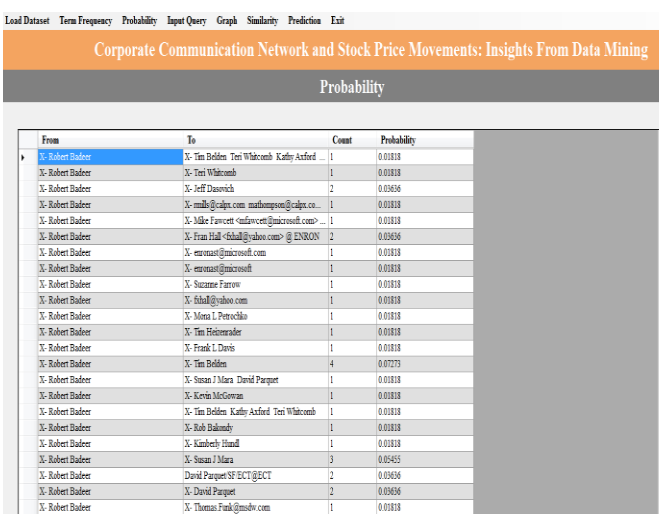

Figure 3. Calculation of Probability

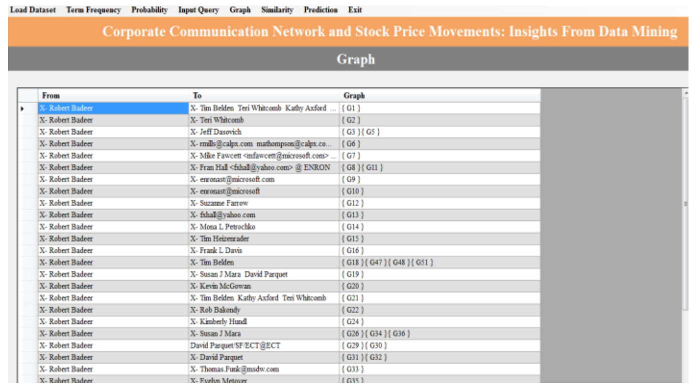

Figure 4. Graph Generation

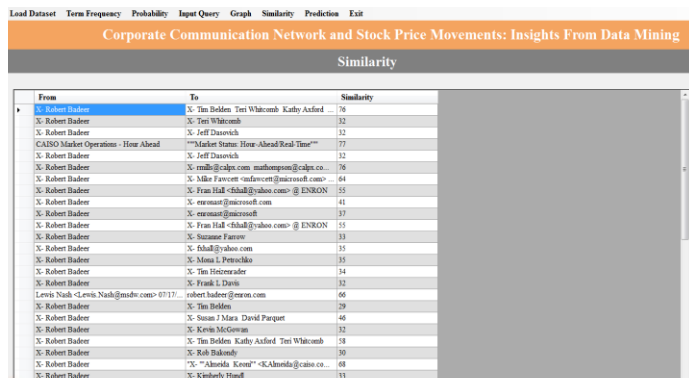

Figure 5. Similarity between two user

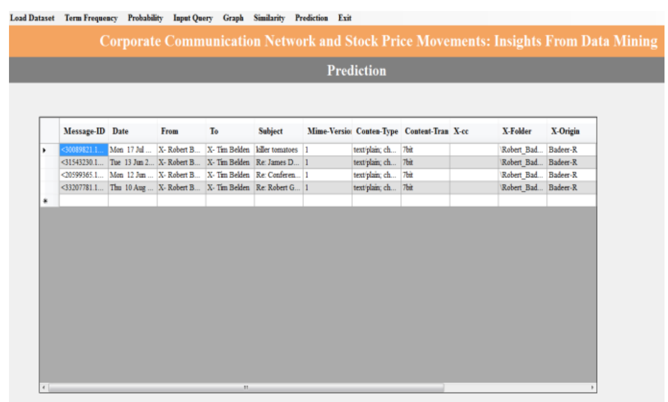

Figure 6 . Final Results

V. Conclusion and Future Scope

The The findings and theoretical implications from this project are two fold. On one hand, we can captured the communications among nodes in Enrons major corporate communication network and identified employees communication patterns. This project demonstrates that a corporate e-mail ecosystem contains meaningful information about employees communication patterns. Even if we only focus on the communication frequency, a company (Enron in our case) has identifiable patterns of e-mail exchange. Such identifiable patterns can reveal important information about major corporate activities and organizational stability that may subsequently influence the focal companys performance in the stock market. Therefore, cooperate communication patterns can serve as a good proxy to predict a companys stock performance. Our experimental results demonstrated the existence of dependence between e-mail communication network and stock price for Enron. This project extended the existing communication theories to capture the patterns of corporate communication and the focal companys stock performance. Will apply algorithms in the areas of information science, management, and finance such as for the investigation of the communication and organizational performance. look forward to more studies that build upon our proposed algorithms and extend our findings from Enron to general corporate research.

ACKNOWLEDGMENT

The It is my immense pleasure to work on this project Novel Approach For Detecting Stock Price Movements. It is only the blessing of my divine master which has prompted and mentally equipped me to undergo the study of this seminar. I would like to thank Dr. S. B. Bagal, Principal, Late G. N. Sapkal College of Engineering for giving me such an opportunity to develop practical knowledge about subject. I am also thankful to Dr. N. R. Wankhade, Head of Computer Engineering Department for his valuable encouragement at every phase of my seminar work and competition. I oer my sincere thanks to my guide Dr. N. R. Wankhade, who very aectionately encourages me to work on the subject and gave his valuable guidance time to time. While preparing this Dissertation Report am very much thankful to him. I am also grateful to entire sta_ of Computer Engineering Department for their kind co-operation which helped me in successful completion of seminar.

REFERENCES

[1] W. Duan, B. Gu, and A. B. Whinston, *Do online reviews matter An empirical investigation of panel data*, Decision Support Syst., vol. 45, no. 4, pp. 10071016, 2008.

[2] E. Vamsidhar, K. V. S. R. P. Varma, P. S. Rao, and R. *Satapati, Prediction of rainfall using backpropagation neural network model*, Int. J. Comput. Sci. Eng., vol. 2, no. 4, pp. 11191121, 2010.

[3] G. Miller, *Social scientists wade into the tweet stream, Science*, vol. 333, no. 6051, pp. 18141815, 2011.

[4] C. Hargreaves and Y. Hao, *Prediction of stock performance using analytical techniques,* J. Emerg. Technol. Web Intell., vol. 5, no. 2, pp. 136142, 2013.

[5] Sawant and P. M. Chawan, *Study of Data Mining Techniques Used for Financial Data Analysis,* Int. J. Eng. Sci. Innov. Technol., vol. 2, no. 3, pp. 503509, 2013.

[6] S. C. Nicholis and D. J. T. Sumpter, *A dynamical approach to stock market uctuations,* Int. J. Bifurcation Chaos, vol. 21, no. 12, pp. 35573564, 2011.

[7] M. Gahirwal and M. Vijayalakshmi, *Inter time series sales forecasting*, Int. J. Adv. Stud. Comput., Sci. Eng., vol. 2, no. 1, pp. 5566, 2013.

[8] [8] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, *Can blog communication dynamics be correlated with stock market activity? in Proc*. 19th ACM Conf. Hypertext Hypermedia, 2008, pp. 5560.

[9] R. Dash, R. L. Paramguru, and R. Dash, *Comparative nalysis of supervised and unsupervised discretization techniques,* Int. J. Adv. Sci. Technol., vol. 2, no. 3, pp.2937, 2011.

[10] J. Diesner, T. L. Frantz, and K. M. Carley, *Communication networks from the Enron email corpus Its always about the people. Enron is no di_erent, School Comput*. Sci., Inst. Softw. Res., Carnegie MellonUniv., Pittsburgh, PA, USA, Tech. Rep., vol. 11, no. 3, 2006. [Online]. Available: http://repository.cmu.edu/isr/46.