Which Group Do You Belong To? Sentiment-Based PageRank to Measure Formal and Informal Influence of Nodes in Networks

Lan Jiang(✉), Ly Dinh, Rezvaneh Rezapour, and Jana Diesner

University of Illinois at Urbana-Champaign, Champaign, USA
lanj3@illinois.edu

Abstract. Organizational networks are often hierarchical by nature as individuals take on roles or functions at various job levels. Prior studies have used either text-level (e.g., sentiment, affect) or structural-level features (e.g., PageRank, various centrality metrics) to identify influential nodes in networks. In this study, we use a combination of these two levels of information to develop a novel ranking method that combines sentiment analysis and PageRank to infer node-level influence in a real-world organizational network. We detect sentiment scores for all actor pairs based on the content of their email-based communication, and calculate their influence index using an enhanced PageRank method. Finally, we group individual nodes into distinct clusters according to their influence index. Compared to established network metrics designed or used to infer formal and informal influence and ground truth data on job levels, our metric achieves the highest accuracy for inferring formal influence (60.7%) and second highest for inferring informal influence (69.0%). Our approach shows that combining text-level and structural-level information is effective for identifying the job level of nodes in an organizational network.

Keywords: PageRank · Formal influence · Informal influence · Organizational networks · Sentiment analysis

1 Introduction

In this paper, we develop a new measure called sentiment-based PageRank that combines the structural information provided by the PageRank metric with text-based information via sentiment analysis as edge attribute. We test our measure by grouping nodes into three levels and compare them to the pre-defined groups of formal and informal roles in an organizational network. Prior studies from the field of social network analysis have provided evidence for differences in individual's power and influence in formal versus informal organizational structures [17,21]. In this context, formal structure is often defined by the organizational

R. M. Benito et al. (Eds.): COMPLEX NETWORKS 2020, SCI 944, pp. 623–636, 2021.
https://doi.org/10.1007/978-3-030-65351-4_50

chart that depicts job levels and titles or roles. Informal structure emerges from the social relationships among individuals. Scholars have examined the relationship between these two types of structures and power, e.g., by using network analytical metrics to determine influential nodes in formal and informal structures [8,21,30]. One widely used measure to characterize the influence of nodes is PageRank [24], which considers the node's immediate neighbors and the extent to which the neighbors are also connected. Originally, the PageRank algorithm [24] was used for rating web pages based on the assumption that a page has a high rank if the sum of ranks of its in-degree (edge(s) pointing to it) is high. Since then, PageRank has also been used to evaluate the influence of online users in social networks [12,32]. The column-stochastic matrix that serves as the input to PageRank is often constructed based on the presence/absence [24] or frequency of interactions in social networks [12]. Given that many real-world networks not only involve structural information, but also information exchange between individuals in the form of text data, there is a need to consider both, structural and textual data, to better comprehend node-level influence [5,6].

Natural language processing (NLP) methods have been broadly used in prior studies to analyze text-based information dissemination, flow, and exchange in social networks. Scholars have extracted text-level features from emails, such as words and phrases [3], linguistic coordination [22], and dialog structure [25], to infer people's levels of influence in an organization. Recognizing the applicability of sentiment analysis in analyzing linguistic cues that signal different types of influence, and the usefulness of PageRank to detect structurally central individuals, we extend these lines of literature by integrating sentiment analysis as an edge attribute into the PageRank measure to gain a more comprehensive understanding of individuals' formal and informal influence in signed and directed organizational networks [9].

In this paper, we use the Enron email data [26] to create the network (step A in Fig. 1) and analyze the communication data and patterns between 84 employees across the outlined three different levels of the job titles [23]). We enhance the traditional PageRank calculation by leveraging the learning rate for optimization [29] (step B in Fig. 1), and further infer the nodes' job level by employing

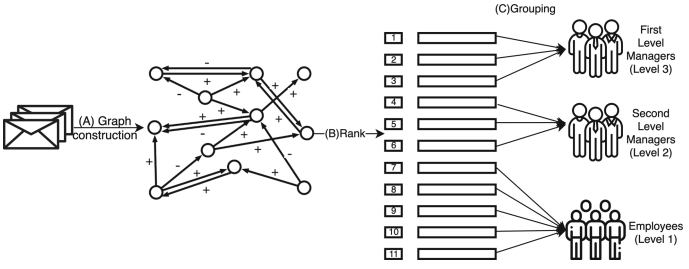


Fig. 1. Experimental pipeline: (A) We construct a network using sentiment scores of emails, (B) use PageRank to calculate an influence index per node, and (C) use three clustering methods to group nodes into three levels of influence.

unsupervised clustering methods to group the ranked nodes into three clusters, which represent first-level management, second-level management, and employees (step C in Fig. 1). Our experimental results show that using sentiment as edge attribute for calculating PageRank can be helpful in determining key groups of nodes with high formal and informal influence in this organizational communication network. Our method achieves 60.7% and 69.0% accuracy in grouping employees with respect to their formal and informal roles, respectively. Our paper makes the following contributions: first, we show how sentiment of emails can be used as a text-level indicator of the relational dynamics between pairs of nodes who communicate in organizational settings. We further propose a novel and advanced PageRank model that combines PageRank with sentiment scores to leverage both structural and language-based features to ultimately measure formal and informal influence in social networks.

2 Related Work

2.1 Formal and Informal Influence in Social Networks

Inferring formal and informal organizational structures is a research topic in a number of disciplines, including computer science [1], business [21], and organizational science [23]. A prominent example of early attempts to capture the formal and informal structure of organizational networks is Krackhardt's set of graph-theoretic dimensions [16,17], which captures the extent to which a network structure resembles a typical hierarchy (an 'outtree' graph). Another group of studies explored informal organizational structures using network-analytic characteristics, such as centralization metrics [8,23] and community detection [4].

2.2 Constructing Networks from Text Data

Prior studies have combined network-analytic concepts with NLP methods to automatically detect influential nodes in organizational networks. McCallum, Xang, and Corrada-Emmanuel [20] have captured topic distributions in email exchanges between pairs of employees at Enron using Latent Dirichlet Allocation (LDA). They found that employees who share similar topics often have similar positions in the hierarchy. Gilbert [10] presented an n-gram based Support Vector Machine classifier with psychometric properties such as "affect" and "certainty" to predict whether an email was sent *upwards* or *downwards* in the (formal) organizational hierarchy. Sentiment of expressions and words has been used as an additional feature to infer individuals' importance in social networks [3,31]: Bramsen and colleagues [3] as well as Tchokni and colleagues [31] found features such as sentiment, emotion, and emoticons (for social media data) to be relevant for predicting people's positions in networks.

Building upon the outlined prior literature, in this paper, we use sentiment scores per nodes to calculate the node's *influence index*.

2.3 Personalized PageRank

Xing and Ghorbani [34] introduced the weighted PageRank to calculate the aggregated importance of web pages. They assigned a large weight to more important pages to capture their prestige more accurately. Bollen and colleagues [2] used the weighted PageRank to measure the prestige of journals, where they assigned a high weight to journals with more citations. PageRank has also been used to identify important users in social networks. Heidemann and colleagues [12] weighted each user's PageRank score based on their friendship graph and online activities, such as wall posts and messages on Facebook.

Based on the intuitions discussed in this section and the introduction, we examine the relationship between sentiment and influence by applying an enhanced PageRank algorithm that uses sentiment of an email as an edge attribute in order to infer formal and informal influence.

3 Notations and Definitions

Table 1 lists the symbols used in the paper. We denote a directed signed graph as $G = (V, E, w, \sigma)$, where V and E are the sets of nodes and directed edges, respectively, and w is the weight function of the edges. Signed digraph G contains $|V| = m$ nodes and $|E| = n$ directed edges. σ is the sign function $\sigma : E \rightarrow \{-1, +1\}$. Given a set of text data, we construct edges $E = \{e_1, e_2, \ldots, e_n\}$ if there is communication (email sent or received) between $v_1, v_2, \ldots, v_m$, where $v_1, v_2, \ldots v_m \in V$. We extract $R(F, T, S, P)$ according to E. For each tuple $t \in R$, $t[F]$ is the source node of the document t, $t[T]$ is the target node of the document t, and $t[S]$ is the sentiment score of document t. We formalize the task addressed in this paper as follows:

Table 1. Symbols and definition

Symbol	Definition
$G = (V, E, w, \sigma)$	Graph constructed by $R(F, T, S)$
$E = \{e_1, e_2, \ldots, e_n\}$	Edges in the network
w	Weight of edges in the network after aggregation
W_{pos_i}	The weight of positive sentiment graph's PageRank score for node i
W_{neg_i}	The weight of negative sentiment graph's PageRank score for node i
$R(F, T, S)$	Tuple extracted from original text dataset
$t[F]$ or $t[T]$ or $t[S]$	The value of tth tuple
$V = \{v_1, v_2, \ldots, v_m\}$	Nodes in the network
$G_{output}(V, E, w, \sigma)$	Output of each node after weighted PageRank with learning rate

4 Proposed Methodology

4.1 Constructing Directed Signed Graph

In order to construct a graph, we build an edge $e_t \in E$ by extracting $t[F]$, $t[T]$ as the "source" and "target" nodes. We add a directed edge from v_1 to v_2 if there is an email from v_1 to v_2. We operationalize the weight of an edge w according to the sentiment score $t[S]$ extracted from each email, and define the signs σ of an edge as the sentiment polarity of each email. To deal with multiple edges between the same "source" and "target" nodes, we average all email sentiment scores from a given pair of nodes. E are the edges formed with averaged sentiment score S. We use S to determine the σ of an edge by using the following rule: σ is positive if S is higher than 0, and σ is negative if S is lower than 0. We discard edges if S is 0 (neutral). We also discard self-loop edges from E and isolated nodes from V since they cannot affect our results. The remaining graph G is used to calculate the influence index for each node.

4.2 Ranking Algorithm

We leverage the enhanced PageRank to calculate the influence index of a node in our network with respect to the quality of the neighboring nodes pointing to the focal node. We faced two challenges when implementing this method:

Challenge 1. PageRank requires a column-stochastic matrix to be a non-negative matrix. However, sentiment scores can be negative.

Challenge 2. Source nodes that have no incoming links make column-stochastic matrix not invertible, which affects PageRank calculation, in general.

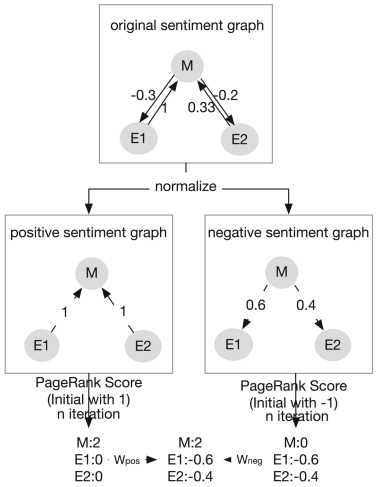


Fig. 2. An example for calculating sentiment-based PageRank

Table 2. Three job levels in Enron organization

Job level	Job title(s) included	Counts
Employees (Level 1)	Admin. Assistant	39
	Specialist	
	In-house Lawyer	
	Senior Specialist	
	Cash Analyst	
	Analyst	
Second level managers (Level 2)	Director	38
	Manager	
	Managing Director	
	Vice President	
First level managers (Level 3)	President	7
	CEO	

To address Challenge 1, we split the original graph G into two subgraphs $G_p = (V, E, w, \sigma = 1)$ and $G_n = (V, E, w, \sigma = -1)$. The positive versus negative sentiment graph G_p and G_n contain nodes connected by edges with positive versus negative sentiment scores, respectively. We initially considered updating positive and negative PageRank scores at the same time. However, this approach may potentially bias the final PageRank score because we have a disproportionate amount of positive (n = 5,943) versus negative edges (n = 1,563).

We calculate the column-stochastic matrix for each graph using sentiment scores. Additionally, we rescale the weight of outgoing links for each node as follows. The column-stochastic matrix represents the probability of an individual having a lower or higher status than the individuals pointing to them. Thus, we want outgoing links to add up to 1, as described in Eq. (1).

$$w_{v_j v_i} = \frac{w_{v_j v_i}}{\sum_{j=1}^{N} w_{v_j v_i}} \tag{1}$$

To tackle Challenge 2, we need to consider that the traditional PageRank algorithm relies on an invertible column-stochastic matrix [15]. When a column-stochastic matrix is not invertible, the PageRank score of source nodes would converge immediately (i.e., becoming 0 after 1 iteration). Traditionally, the update rule of PageRank scores is as follows (Eq. 2):

$$PR_{v_i} = \frac{1-\lambda}{n} + \lambda \sum_{j=1}^{N} w_{v_j v_i} * PR_{v_j} \tag{2}$$

where PR_{v_i} indicates *influence index* of node v_i. The fraction $\frac{1-\lambda}{n}$ represents the minimal amount of power assigned to each node. To address Challenge 1, we introduce a learning rate to prevent PageRank scores from converging too fast. Learning rates [29] are widely used in machine learning to obtain parameters in a neural network that minimize a loss function during an iterative procedure. Specifically, each parameter is updated with a small proportion of a partial derivative of the error. We use this proportion as our learning rate. Thus, we propose to update a small proportion of PageRank scores in each iteration until the scores converge. The updated rule of our approach is described in Eq. (3):

$$PR_{v_i} = (1 - lr) * PR'_{v_i} + \sum_{j=1}^{N} lr * w_{v_j v_i} * PR'_{v_j} \tag{3}$$

where lr is the learning rate, and PR'_{v_i} represents the PageRank score of node v_i in the last iteration. After calculating the PageRank score for each node within the positive and the negative sentiment graph, we add up the scores with respect to their weights, defined by the number of nodes' incoming degrees (Eq. 4), per graph. This gives us the final PageRank score for each node $G_{output}(V, E, w, \sigma)_{v_i}$ (Eq. 5).

$$W_{pos_i} = \frac{N_{pos_{v_i}}}{N_{pos_{v_i}} + N_{neg_{v_i}}} \tag{4}$$

where N_{pos_i} and N_{neg_i} are the number of incoming degree of node v_i in the positive and negative sentiment graph.

$$G_{output}(V, E, w, \sigma)_{v_i} = W_{pos_{v_i}} * PR(G_p)_{v_i} + W_{neg_{v_i}} * PR(G_n)_{v_i} \quad (5)$$

We optimize our method with $lr = 0.1$. For the positive sentiment graph, we set the initial value to 1, and for the negative sentiment graph, we set it to -1. Figure 2 shows an example of our method for calculating sentiment-based PageRank. M represents a manager, E_1 represent employee 1, and E_2 represent employee 2. The result of our enhanced PageRank with learning rate can be further used for distinguishing communities in the network, as explained next.

4.3 Clustering Methods

We use three unsupervised clustering methods: K-means [19], Gaussian Mixture Modelling (GMM) [28], and Hierarchical agglomerative clustering (HAC) [14], to place nodes with similar influence indices into the same groups. This step enables us to observe whether nodes with a similar *influence indices* actually belong to the same group; thus enabling us to validate our method by comparing the results against given ground truth data (refer to 5.3 for details).

5 Experiments

5.1 Dataset Description

Enron Corporate Email. The Enron email data is a large-scale, over-time dataset from a U.S. based energy company that filed for bankruptcy in 2001. In this paper, we use the latest version of the dataset,[1] which consists of 517,401 emails from the inboxes of about 150 employees. For each email we use the email addresses of the senders and receivers, and the related email bodies. We disambiguate the names and emails in the dataset since some individuals used more than one email addresses to communicate with others within and outside the corporation [8].

5.2 Sentiment Analysis for Graph Construction

We construct a signed graph based on the sentiment scores from each email, which we identify by using a given subjectivity lexicon [33]. We first domain adapted the lexicon by (1) extracting the top 2000 words with the highest TF-IDF score from the emails, (2) labeling each word as positive, negative, or neutral with respect to the context and its Part of Speech (POS) if they were not in the lexicon, (3) verified the labels of the words if they were in the lexicon, and finally (4) pruned the list by adding (removing) words that we found appropriate (redundant) for the context of our study.

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

After domain adapting the lexicon, we parsed each email, tokenized the sentences, and POS tagged the words using spaCy [13]. We counted the words in each sentence if the word and its POS matched an entry and its POS in our domain adapted lexicon. We then counted the aggregated number of positive, negative and neutral tokens per sentence, and tagged each sentence with the majority polarity class. We also performed negation detection for each sentence using the NLTK package [18], and flipped the polarity of the score to the opposite one if the sentence was found negated. Finally, we aggregated the sentiment scores of all sentences per email, and normalized the score by the number of sentences per email. The range of our sentiment scores is -5 to $+8$.

5.3 Formal and Informal Groupings

Enron's Formal Organizational Structure. The Enron organizational chart [26] includes names of employers, corresponding job titles, and job levels in the Enron organization. The job titles include employee, trader, manager, director, managing director, vice president, president, CEO, and N/A. We removed individuals with N/A position. We established our ground truth data with 84 individuals with a known job title and job category, using a combination of hierarchical roles labeled in [10,23]. While both datasets propose similar hierarchical structures, there were differences for how "Vice President" (level 1 in [23] versus level 2 in [10]) and "In-house Lawyer" (level 3 in [23] versus level 2 in [10]) are labeled. We categorized "In-house Lawyer" into level 2 because this position did not hold any managing responsibility. We categorized "Vice President" into level 2 because this position in Enron entailed managing a specific department/branch.

Enron's Informal Organizational Structure. We construct the informal organizational network using the frequency of email exchanges between employees. In addition, each node is given a degree centrality score based on the number of emails they sent and received. This approach allows us to capture individuals' (regardless of their formal role) prominence in the Enron social network with respect to their emailing frequencies. Agarwal and colleagues [1] constructed an undirected weighted network of Enron's informal structure and compared it with the formal structure. As an extension, we consider the directionality of edges to construct the informal network.

5.4 Clustering Based on PageRank vs. Formal and Informal Groupings

We evaluate the utility of our method in reflecting Enron's formal structure by comparing (1) the ground truth data against the clustering result based on our enhanced PageRank method, and (2) clustering results based on established network metrics, i.e., degree, betweenness, closeness, eigenvector. We also compare clustering results based on our enhanced PageRank method to the clustering

results based on degree centrality to evaluate the utility of our method in reflecting informal influence. We utilize normalized mutual information (NMI, [35]), adjusted rand index (ARI, [27]), and accuracy to evaluate our method. All three evaluation measurements range from 0 to 1, with 1 being the best result.

6 Results

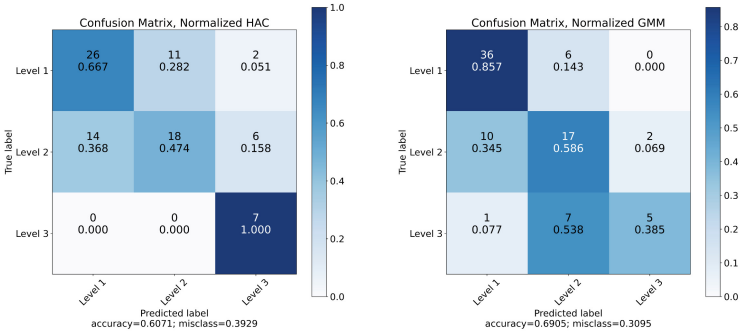
6.1 Clustering Results vs. Enron's Formal Groupings

Table 3 presents the accuracy, ARI, and NMI scores of our clustering results based on the enhanced PageRank method, and comparisons to clustering results based on alternative metrics outlined above for the task of inferring Enron's formal structure. Our method achieves the highest accuracy out of all clustering methods. However, accuracy might not be the best suited measure to evaluate clustering in our case due to an imbalance in our dataset, where 45% of individuals belong to level 1 cluster (as shown in Table 2). To mitigate this issue, we evaluate the clustering outputs with two other measures, namely ARI and NMI. Clustering results based on our enhanced PageRank method outperforms clustering results based on other network metrics in a majority of cases, except for NMI of K-means clustering, and ARI and NMI for GMM, where eigenvector centrality has the highest clustering result. Our enhanced PageRank result also outperforms frequency-based PageRank for all cases. These findings show that using sentiment as an edge label for PageRank calculation may be more suitable than using frequency of communication for PageRank calculation when inferring formal structure.

We further examine the clustering results of our method using confusion matrices. Based on Table 3, our method achieves the highest results using the HAC clustering method. As shown in Fig. 3a), our method achieves an overall clustering performance of 60.71%. There are 66.7% of correctly-identified

Table 3. Accuracy, ARI, and NMI of grouping results

Formal structure									
Method	K-means			GMM			HAC		
	Accuracy	ARI	NMI	Accuracy	ARI	NMI	Accuracy	ARI	NMI
Frequency-PageRank	0.429	0.022	0.018	0.440	−0.001	0.005	0.429	0.014	0.023
Degree centrality	0.488	0.090	0.063	0.476	0.092	0.064	0.500	0.096	0.073
Betweenness centrality	0.512	0.042	0.036	0.524	0.054	0.051	0.524	0.064	0.051
Closeness centrality	0.357	0.137	0.117	0.369	0.113	0.100	0.369	0.151	0.128
Eigenvector centrality	0.560	0.159	**0.145**	0.571	**0.170**	**0.150**	0.548	0.142	0.093
Our enhanced PageRank	**0.595**	**0.169**	0.123	**0.607**	0.164	0.131	**0.607**	**0.213**	**0.142**
Informal structure									
Betweenness centrality	0.607	0.341	0.440	0.643	**0.382**	**0.454**	0.595	0.289	0.402
Closeness centrality	0.381	0.278	0.311	0.536	0.359	0.370	0.417	0.259	0.341
Eigenvector centrality	**0.726**	**0.367**	**0.396**	**0.702**	0.359	0.370	**0.762**	**0.408**	**0.453**
Our enhanced PageRank	0.679	0.319	0.302	0.690	0.312	0.270	0.655	0.228	0.244



(a) Formal structure (b) Informal structure

Fig. 3. Confusion matrix for GMM, with clustering performance measured by NMI. The x-axis is the predicted label, and y-axis is the ground truth label.

instances for level 1, 47.4% for level 2, and 100.0% for level 3. A small proportion of instances that are actually level 2 were misclassified as level 1 (28.2%). There is also a small number of instances that are actually level 2 but were misclassified as level 3 (15.8%). This shows some difficulty of our model to distinguish second-level managers from other categories.

The mean *influence index* is 0.461 (SD = 0.506) for level 1 cluster, 0.728 (SD = 0.581) for level 2 cluster, and 2.006 (SD = 0.711) for level 3 cluster. Conducting a one-way analysis of variance (ANOVA), we found statistically significant differences between the means of the three groups in the formal hierarchy, $F(2,81) = 23.026$, $p < 0.001$. Post-hoc comparisons using a Tukey HSD test shows that all three groups differ significantly from each other at $p < 0.001$ level.

6.2 Clustering Results vs. Enron's Informal Groupings

Table 3 presents the results of using our enhanced PageRank method to infer the informal structure of Enron's email network. We also compare the results of our method to those obtained with established network metrics, with the exception of degree centrality because this metric is used to construct the informal network. We observe that eigenvector centrality performs the best in most cases. Exceptions are ARI and NMI using GMM clustering, where betweenness centrality performs best (ARI = 0.382, NMI = 0.454). Our method performs second best to eigenvector centrality in terms of accuracy for K-means (0.679), GMM (0.690), and HAC (0.655). However, our method does not perform well when compared to ARI and NMI. The range of scores for informal influence is 0.0 to 0.5, while the range for PageRank scores is wider (−0.5 to 3.2). The range of scores for eigenvector centrality (0 to 0.2) is similarity to range of the scores for informal influence, which may explain the high performance across different clustering methods. We examine the clustering performance of our method

with respect to the informal structure using confusion matrices. As presented in Fig. 3b, our method achieves the highest performance using GMM clustering (accuracy = 0.691). There are 85.7% correctly identified instances for level 1, 58.6% for level 2, and 38.5% for level 3. In terms of misclassifications, we observe a notable proportion of level 3 nodes classified as level 2 (53.8%).

7 Discussion

Our method takes both sentiment of text data (emails) and the structure of a focal node's neighborhood (PageRank) into consideration. For inferring formal influence, HAC and K-means clustering based on our enhanced PageRank method yields the highest performance for ARI and accuracy. GMM clustering based on eigenvector centrality yields the highest performance for ARI and NMI. This finding shows that our method can detect nodes that are influential in Enron's formal structure, along with eigenvector centrality. As these two metrics both take into account the relative influence of a node based on its ego-network, our findings illustrate that the reliable measurement of influence should move beyond counting first-degree connections. Alternatively, we should consider the number of links that a node's connections have and the quality of these connections. For the informal communication structure, we observe that nodes with higher PageRank scores have high informal influence scores (measured by frequency-based degree centrality).

Another notable finding is that sentiment as an edge label in networks constructed from text data is more effective than using frequency of communication as an edge label. This might indicate that sentiment of emails reveals more information about an individual's orientation towards another individual than counts of how many emails they exchanged. Our findings also show that the average sentiment score of emails received and sent per node are distinctive from one another, and the extent to which they differ from one another may be related to a node's formal position. For instance, nodes in level 3 have an average sentiment score of 0.127 for all incoming emails, and a score of −0.038 for their outgoing emails. In contrast to that, nodes in level 1 have an average sentiment score of −0.008 for their incoming emails, and 0.193 for their outgoing emails. Finally, nodes in level 2 have similar average sentiment score for both sent and received emails, 0.071 and 0.040, respectively.

8 Conclusions and Future Work

In this paper, we have introduced an enhanced PageRank score that considers sentiment information as extracted from text data to calculate a node's position of power and influence in formal and informal organizational networks. Using these information, we grouped the individuals into three groups (level 1, 2, and 3) based on their PageRank scores and compared our results with three levels of job categories. Our findings show that our method can reliably reflect Enron's informal structure, where nodes with higher PageRank scores often have higher

degree centrality scores. For the formal structure, we find a number of cases where nodes with higher PageRank scores are not in the higher job levels.

Our method has several limitations. First, there may be other linguistic features (other than sentiment) that capture hierarchical information in email communications. Therefore, we aim to include features such as word phrases [10] and language use [3] in hope to improve clustering performance. Second, we hope to incorporate a temporal dimension into our analysis, which would take into account the evolving events surrounding the Enron crisis. Based on Diesner and Evans [7], sentiment profiles vary across time periods during the Enron crisis. We want to examine whether and why sentiment information is more indicative of hierarchy in any particular time period(s). In future work, we also hope to incorporate both positive and negative edges simultaneously in one graph to calculate PageRank, as demonstrated in [11].

References

1. Agarwal, A., Omuya, A., Harnly, A., Rambow, O.: A comprehensive gold standard for the Enron organizational hierarchy. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 161–165. Association for Computational Linguistics (2012)
2. Bollen, J., Rodriquez, M.A., Van de Sompel, H.: Journal status. Scientometrics **69**(3), 669–687 (2006)
3. Bramsen, P., Escobar-Molano, M., Patel, A., Alonso, R.: Extracting social power relationships from natural language. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 773–782. Association for Computational Linguistics (2011)
4. Clauset, A., Moore, C., Newman, M.E.: Structural inference of hierarchies in networks. In: ICML Workshop on Statistical Network Analysis, pp. 1–13. Springer (2006)
5. Diesner, J., Carley, K.M.: Extraktion relationaler daten aus texten. In: Handbuch Netzwerkforschung, pp. 507–521. Springer (2010)
6. Diesner, J., Carley, K.M.: A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In: Proceedings of the 2010 IEEE 2nd International Conference on Social Computing, pp. 687–692. IEEE (2010)
7. Diesner, J., Evans, C.S.: Little bad concerns: using sentiment analysis to assess structural balance in communication networks. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 342–348 (2015)
8. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the Enron email corpus: "it's always about the people. Enron is no different". Comput. Math. Organ. Theor. **11**(3), 201–228 (2005)
9. Dinh, L., Rezapour, R., Jiang, L., Diesner, J.: Structural balance in signed digraphs: considering transitivity to measure balance in graphs constructed by using different link signing methods. arXiv preprint arXiv:2006.02565 (2020)
10. Gilbert, E.: Phrases that signal workplace hierarchy. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 1037–1046 (2012)

11. He, X., Du, H., Feldman, M.W., Li, G.: Information diffusion in signed networks. PLOS ONE **14**(10), e0224177 (2019)
12. Heidemann, J., Klier, M., Probst, F.: Identifying key users in online social networks: a PageRank based approach. In: Proceedings of the International Conference on Information Systems (2010)
13. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1373–1378 (2015)
14. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
15. Knight, P.A.: The Sinkhorn-Knopp algorithm: convergence and applications. SIAM J. Matrix Anal. Appl. **30**(1), 261–275 (2008)
16. Krackhardt, D.: Assessing the political landscape: structure, cognition, and power in organizations. Adm. Sci. Q., 342–369 (1990)
17. Krackhardt, D.: Graph theoretical dimensions of informal organizations. In: Computational Organization Theory, pp. 89–111. Lawrence Erlbaum Associates Inc. (1994)
18. Loper, E., Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the ACL on Interactive Poster and Demonstration Sessions, pp. 31–44. Association for Computational Linguistics (2002)
19. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, pp. 281–297 (1967)
20. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on Enron and academic email. J. Artif. Intell. Res. **30**, 249–272 (2007)
21. Michalski, R., Palus, S., Kazienko, P.: Matching organizational structure and social network extracted from email communication. In: Proceedings of the International Conference on Business Information Systems, pp. 197–206. Springer (2011)
22. Ngoc, P.T., Yoo, M.: The lexicon-based sentiment analysis for fan page ranking in Facebook. In: Proceedings of the International Conference on Information Networking 2014, ICOIN 2014, pp. 444–448. IEEE (2014)
23. Nurek, M., Michalski, R.: Combining machine learning and social network analysis to reveal the organizational structures. Appl. Sci. **10**(5), 1699 (2020). https://doi.org/10.3390/app10051699
24. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
25. Prabhakaran, V., Rambow, O.: Predicting power relations between participants in written dialog from a single thread. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 339–344 (2014)
26. Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y.: Scan statistics on Enron graphs. Comput. Math. Organ. Theor. **11**(3), 229–247 (2005)
27. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
28. Reynolds, D.A.: Gaussian mixture models. In: Encyclopedia of Biometrics, vol. 741 (2009)
29. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat., 400–407 (1951)

30. Rowe, R., Creamer, G., Hershkop, S., Stolfo, S.J.: Automated social hierarchy detection through email network analysis. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 109–117 (2007)
31. Tchokni, S.E., Séaghdha, D.O., Quercia, D.: Emoticons and phrases: status symbols in social media. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (2014)
32. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp. 261–270 (2010)
33. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, pp. 486–497. Springer (2005)
34. Xing, W., Ghorbani, A.: Weighted PageRank algorithm. In: 2004 Proceedings of the 2nd Annual Conference on Communication Networks and Services Research, pp. 305–314. IEEE (2004)
35. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical report, TR 01–40, Department of Computer Science, University of Minnesota (2001)