What matters more: how you speak, or whom you know? E-Mail Communication Patterns of Top Performers

Qi Wen, Peter A. Gloor, Andrea F. Colladon, Praful Tickoo, Tushar Joshi

pgloor@mit.edu

Abstract: In the information economy, individuals' digital communication strategies are closely associated with their work performance. This study combines social network and semantic analysis to develop a method to identify top performers based on email communication. By reviewing existing literature, we identified the indicators that quantify email communication into measurable dimensions. To empirically examine the predictive power of the proposed indicators, we collected a 2 million email archive of the 578 executives in an international service company. Panel regression was employed to derive interpretable association between email indicators and top performance. The results suggest that top performers tend to assume central network positions and have high responsiveness to emails. In email contents, top performers use more positive and complex language, with low emotionality, but rich in influential words that are more likely reused by coworkers. To better explore the predictive power of the email indicators, we employed AdaBoost machine learning models, which achieved 83.56% accuracy in identifying top performers. With cluster analysis we further find three categories of top performers, "networkers" with central network positions, "influencers" with influential ideas, and "positivists" with positive sentiments. The findings suggest that top performers have distinctive email communication patterns, laying the foundation for grounding email communication competence in theory. The proposed email analysis method also provides a tool to evaluate the different types of individual communication styles.

Keywords: top performer; email communication; social network analysis; semantic analysis.

1. Introduction

Effective communication becomes increasingly indispensable to achieve high work performance in an age of hyper-specialization, as there is a need for intensive information sharing to integrate multidisciplinary expertise (Malone et al., 2011). The rapid development of electronic communication technology enables people to continuously keep in contact, free from the restrictions of time and space (Byron and Baldridge, 2007; Butts et al., 2015). In order to facilitate effective communication, many firms are making huge investments in advanced information systems, however, without fully understanding the effects of communication patterns on performance (Erhardt et al., 2016). Academic research is especially lacking on how email communication patterns are associated with work performance (Mesmer-Magnus and DeChurch, 2009; Sosa et al., 2015).

Email communication has long been reported to occupy a significant proportion of knowledge workers' time (Sarbaugh-Thompson and Feldman, 1998; Fragale et al., 2012; Mazmanian et al., 2013). It provides a ubiquitous information sharing channel, in which people are available even when they are physically absent (Mazmanian et al., 2006). Due to its large information carrying capacity, emails have become a dominant means of communication in large organizations (Sarbaugh-Thompson, 1998). Although emails have high potential to reflect organizational communication patterns and behaviors, they have only been used in a limited number of past studies (Fragale et al., 2012). This is primarily attributable to two difficulties. Firstly, emails miss some important behavioral cues that are integral to face-to-face communication, such as speech tone and facial expressions (Kruger et al., 2005). However, some researchers argued that people tend to compensate for this shortcoming with additional cues in emails, such as the use of capitalized words and emoticons (Sarbaugh-Thompson and Feldman, 1998; Byron and Baldridge, 2007). Secondly, email data are inherently unstructured, making it hard to quantify constructs and test research hypotheses (Fragale et al., 2012). However, recent developments in social network analysis (SNA) and text mining methods enable large-scale unstructured data analyses, which are promising for identifying communication behaviors from emails (George et al., 2014; Sharaff and Nagwani, 2016).

Despite the critiques and arguments, it is clear that email data can act as a rich information source, from which meaningful signals of communication behaviors can be extracted (Gloor et al., 2017). Some pioneering studies have already identified several behavioral indicators of email communication (Butts et al., 2015; Fragale et al., 2012; Gloor et al., 2017), however, in a relatively fragmented manner, with each of them focusing on a few particular indicators. This study aims to further these research efforts and systematically investigate how the rich information in email communication can be operationalized into quantifiable indicators that are predictive of individuals' work performance. Thus, this study addresses the following research question: How can the information in email communication be aggregated to identify top performers in organizations?

Based on the email archive of 578 executives working in a global software services company, we combined regression and machine learning models to examine how top performers can be identified with email indicators. The findings reveal the most predictive email indicators and the various email communication patterns of top performers. By addressing the research question, this study provides contributions in two areas. First, it develops a set of email communication indicators and demonstrates their predictive power in identifying top performers. The identified email communication indicators of top performers lay foundation for operationalizing communication competence in the context of email communication from a social network perspective. Second, the identified influential email indicators provide a practical tool to map different individuals' contributions to organization communication. This facilitates the reflection on individuals' communication patterns and in turn generates valuable implications for improving email communication efficiency.

The rest of the study are organized as follows. The second section reviews the related literature on communication and work performance, and identifies the research gap to be addressed. The third section describes the data collection and analysis procedures, followed by the empirical analysis results in the fourth section. The fifth section discusses the theoretical and practical implications of the findings. Finally, the sixth section draws the conclusions.

2 Literature review

2.1 Email communication and individual performance

As a basic information channel for modern organizations, emails can reflect not only communication behaviors but also various other behaviors beyond communication (Byron and Baldridge, 2007; Fragale et al., 2012; Erhardt et al. 2016). Correspondingly, two streams of research implied that email communication contains meaningful signals that are predictive of individuals' performance. The first contains abundant research efforts to explore how team- and individual-level communication influences performance. Through the theoretical lens of the conduit model of communication, team communication is instrumental to effective teamwork (Cornelissen et al., 2015). It serves the purpose of diffusing task-relevant information (He et al., 2016), addressing coordination problems (Brandts et al., 2015) and developing a shared understanding of team states (Wasiak et al., 2011; Shore et al., 2015). Many researchers further argued that communication not only reflects but also in turn shapes team states and even organization institutional settings (Cornelissen et al., 2015; Ocasio et al., 2015), which have profound influence on performance. Following these theoretical arguments, empirical findings substantiated the positive relationship between email communication and team performance (see Marlow et al., 2017 for a meta-analysis). In this light, effective email communication is indispensable for becoming top performers in teamwork.

At the individual-level, communication competence is one of the most frequently cited enablers of superior work performance (Brass et al., 2004; Cross and Cummings, 2004). With increasing work virtualization, many studies propose the notion of individuals' "virtual competence" (and similarly "virtual intelligence" in Makarius and Larson, 2017), of which email communication is an important ingredient. The number of communication relationships is closely related to individuals' ability to access information and resources (Sarker et al., 2011; Lomi et al., 2014). The diversity of communication relationships is advantageous for individuals when performing tasks that require multidisciplinary knowledge (Aral and Van Alstyne, 2007). Lower communication frequency was found to be associated with less voice in team decisions (Gajendran and Joshi, 2012), more feelings of isolation (Golden et al.,

2008) and lower work performance (Chan and Lai, 2017).

On the other hand, some researchers pointed to the undesirable effects of frequent email communication that emails may act as a potential source of distraction and work stress (Kushlev and Dunn, 2015). Dealing with emails and recovering from the interruption caused by emails were reported to consume a considerable proportion of knowledge workers' time (Jackson et al., 2003). The pressure of responding to a number of emails can easily increase individuals' work stress (Barley et al., 2011). These effects reduce individuals' work performance and contribute to the productivity puzzle of "being able to do more work but not to do work more productively" in information society (Mano and Mesch, 2010). Many research efforts have been devoted to empirically examine the effects of email communication on work performance, but the findings are mixed and the effect of email communication is still controversial.

Despite the debates on how email communication influences performance, existing studies suggest that the quality, diversity and frequency of communication are influential to work performance, and hence imply a large potential of emails in predicting performance (Jarvenpaa et al., 2004; Aral et al., 2012). As pointed out by Barley et al. (2011), the conflicts in empirical findings are partly attributable to the fact that most existing research measured email communication with questionnaire instruments instead of real-world email data. This is echoed by Aral et al. (2012), who utilized email data as an objective measure of communication that contributes to overcoming the bias in surveys based on participants' memory of their communication networks. In this light, we expect that email communication is predictive of work performance and empirically examine the causal relationship using real-world email data.

The second stream of closely related research focuses on the behavioral cues embodied in email communication. The large information carrying capacity of email makes it a rich source of individual-level interactions (Mazmanian et al., 2006; Fragale et al., 2012; Mazmanian et al., 2013; Sharaff and Nagwani, 2016). According to social attribution theory, people attribute communication relationships to be not only *instrumental* to fulfill organizational roles and obligations, but also *expressive*, for example, to voice feelings and gain satisfaction (Barry and Crant, 2000; Barry and Fulmer, 2004). Therefore, email can reflect various individual behaviors, and in fact, some pioneering research already identified multiple kinds of behaviors from email data. For example, Fragale et al. (2012) utilized email to measure deference behaviors and argued that email data are advantageous for studying behaviors that are easily distorted by researchers' interventions. The studies by Mazmanian et al. (2006), Mazmanian et al. (2013) and Butts et al. (2015) all suggest that email contains the information on work-nonwork conflict and work engagement escalation. Byron and Baldridge's (2007) experiment revealed that expression methods (using capitalization and emoticons) of emails have significant influence on the email recipients' impressions of senders' likability. Similarly, Lim and Teo (2009) and Francis et al. (2015) identified incivility behavior from emails and analyzed its relationship with work-load and work attitude. Although these studies focus on specific behaviors and hence are fragmented, they imply that email contains abundant behavioral cues, which may have profound implications for individual work performance (Lim and Teo, 2009; Fragale et al., 2012). Extending these findings, we aim to explore how the abundant information in email can be integrated to provide insights for work performance, and particularly, identifying top performers.

2.2 Email communication indicators

Compared to traditional questionnaire surveys, email archives provide more authentic recordings of real-world communication behaviors free from retrospective bias (Ahuja et al. 2003; Aral and Van Alstyne, 2007). Several studies (e.g. Gloor et al., 2011 and Gloor et al., 2011) developed a system of email indicators, known as "honest signals", to quantify email communication behaviors. However, many more studies explored similar email communication indicators without explicitly referring to the notion.

To develop a set of email indicators based on a comprehensive review of the extant literature, we conducted extensive literature searches in major databases, including INFORMs, Web of Science, Elsevier ScienceDirect and ProQuest. Using "email/e-mail network", "email/e-mail communication" and "electronic communication" as key words, we found 61 studies related to email communication. Based on the 61 studies, we further identified 8 studies relevant to this topic that are cited by them or citing them. This procedure was repeated until no more relevant reference could be found (Webster and Watson, 2003). As an additional validation, we compared the retrieved studies with the reference lists of two relevant meta-analyses (Mesmer-Magnus et al., 2011; Marlow et al., 2017) and found no other relevant study to be supplemented. In this way, totally 73 studies were retrieved, among which 18 studies developed indicators for email data in real-world work settings. Consistent with the theoretical implications in 2.1, most of the indicators in email communication were found to be significantly associated with information sharing, collaboration and various other outcome variables. We construct a Sankey diagram to illustrate which *study* analyzed which *dimension* using which *indicator* to explain which *outcome variable* (see the Appendix Table 1 for a detailed list).

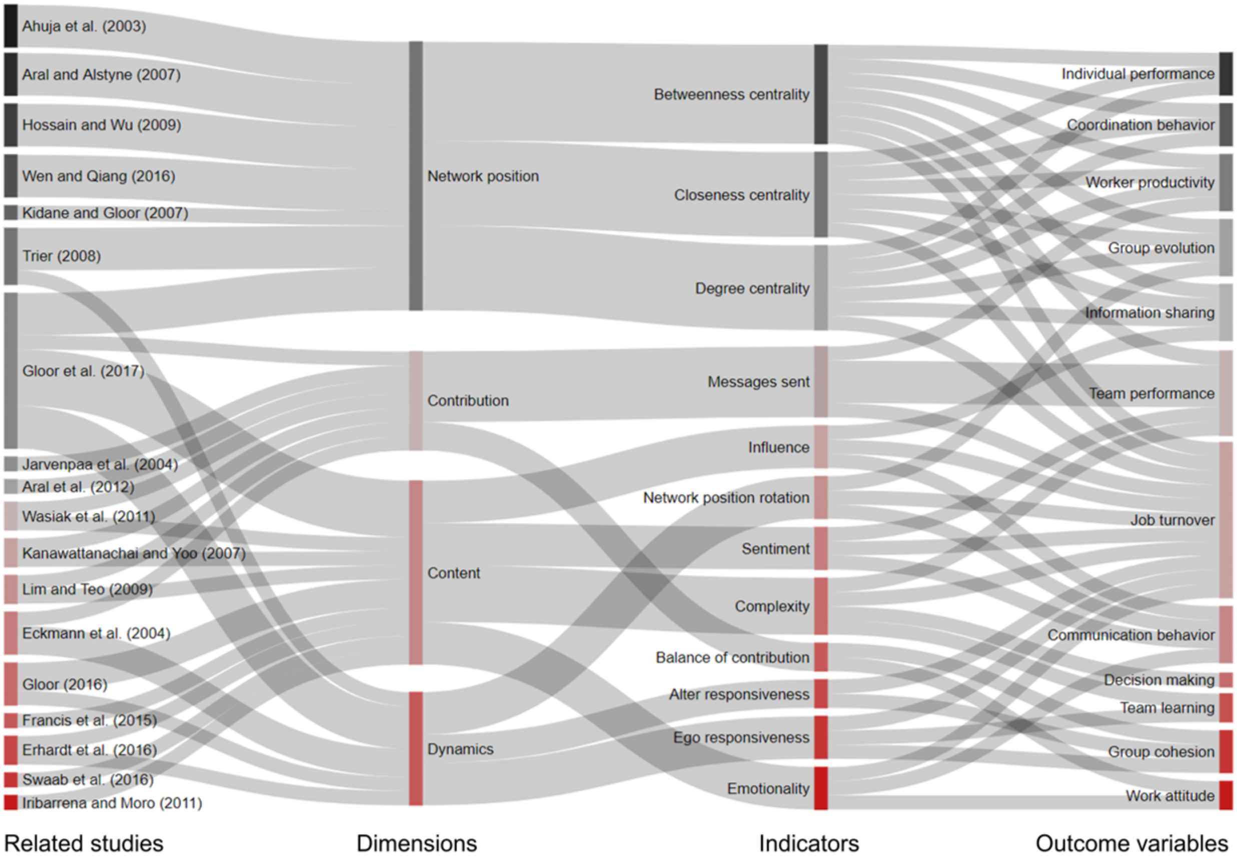


Figure 1 The Sankey diagram of existing studies using email communication indicators

A one-unit width of the “flow” in the Sankey diagram corresponds to one email indicator adopted in one previous study. In this way, the size of the rectangle corresponding to each study is proportional to the number of email indicators used in that study, and the width of the flow from each study to each dimension is proportional to the number of indicators that study used in that dimension. The size of each indicator reflects how many times the indicator was used in previous studies, and the width of the flow from each indicator to each outcome variable reflects the number of times that indicator was used to predict that outcome variable.

As shown in Figure 1, email network position is the most frequently studied dimension, with the three centrality indicators receiving similar research attention. The contribution dimension attracted much fewer research efforts, the majority of which were devoted to analyzing the number of emails sent. A wide range of studies investigated email content indicators with similar emphasis on the four indicators. The dynamics of email communication is less frequently studied, especially the responsiveness of co-workers (alter responsiveness).

The links between indicators and outcome variables are rather dispersed. This reflects the fact that there is a lack of wide consensus on which indicator has particularly good predictive power over an outcome variable, and many studies are still exploring using a variety of indicators (e.g. Gloor et al. 2017). Only a few studies explicitly examined individuals' performance using email communication indicators, they primarily focused on the effects of network position (Ahuja et al. 2003; Aral and Van Alstyne, 2007). This coincides with the argument that most previous studies on information advantages in communication networks are "content agnostic" (Aral and Van Alstyne 2011). Therefore, existing studies indicate the potential of email indicators in predicting individuals' work performance, whereas how the indicators can be used to identify top performers still remains to be explored.

3. Methods

In order to empirically explore how the email indicators can be used to predict individual performance, we designed an empirical analysis procedure as illustrated in Figure 2. In step one, the data is collected from the company's e-mail server at the end of each month for the duration of six months. For privacy reasons the collection process is executed on the company's server. In step two, using the dynamic semantic social network analysis software Condor (Gloor, 2017), the different social network metrics described in the remainder of this section are calculated twice, at the beginning and the end of the observation period, and exported as a table. In step three, regression analysis is applied to obtain interpretable association between each email indicator and individual work performance. In step four, machine learning is employed to further explore the predictive power of the email indicators.

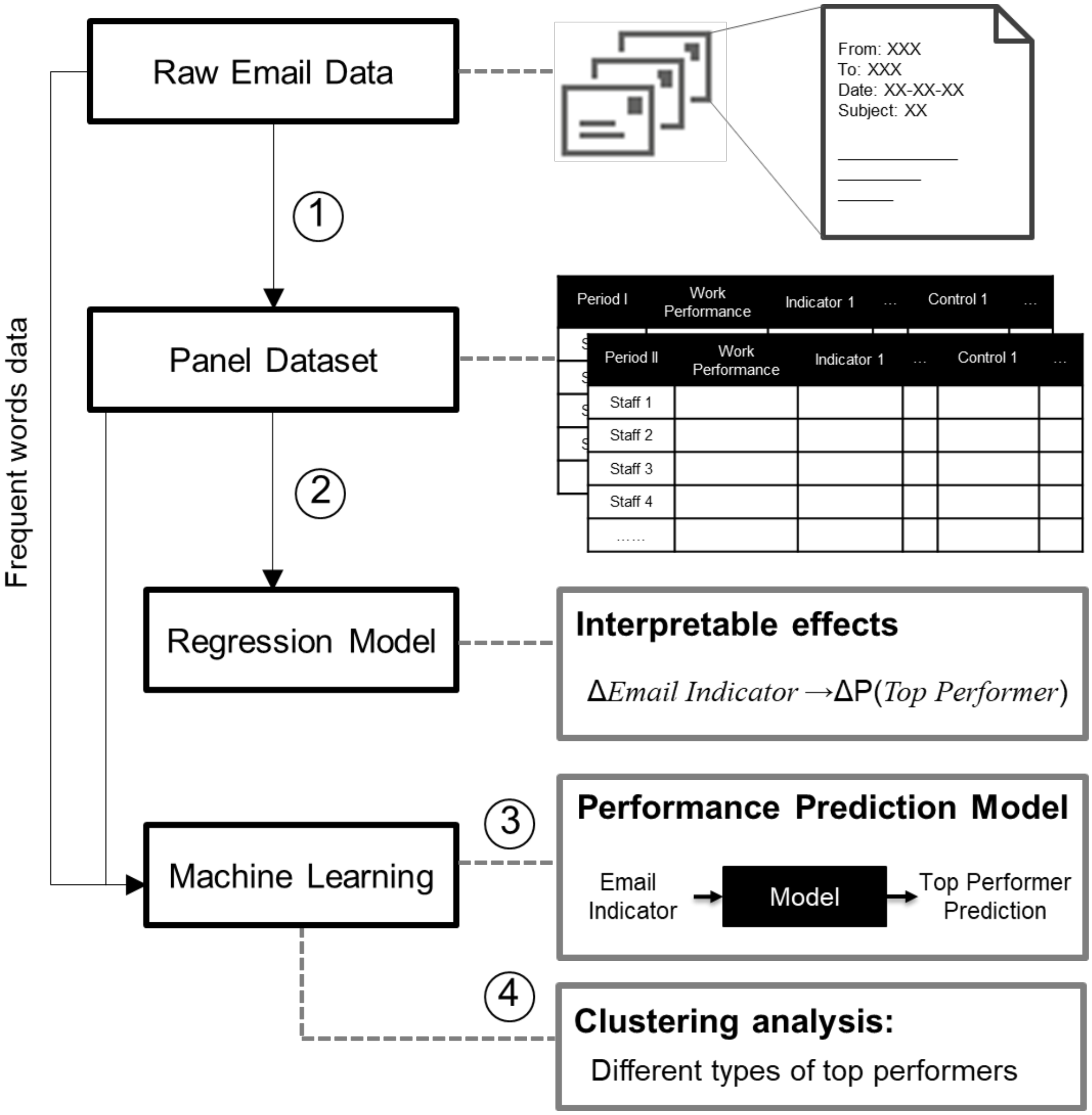


Figure 2 The explorative empirical analysis procedure

3.1 Data collection and indicator calculation

We collected email data from the top 578 executives of a global services company with over 70,000 employees, out of which 578 were included in our analysis. The email dataset is ideal for the purpose of this study since the 578 executives coordinate the company's global operation and use email as the primary communication channel. With the help of the software tool Condor, we were able to calculate email communication indicators from raw email data on the company's server without directly reading email contents (Gloor, 2017). This approach protects staffs' privacy and, at the same time, collects the data ready for further analysis (Kramer et al., 2014). Corresponding to the two sets of performance rating data, we included two waves of email indicator data calculated from the more than 2 million emails between the 578 executives and other staff members into our analysis, at the beginning and the end of the observation period To take full account of intra-organization email communication relationships, we set the boundary of the email network as the whole company and derived the network indicators based on the whole network (Eckmann et al., 2004). The details on indicator calculation methods are described in the next paragraph.

3.1.1 Performance rating as the dependent variable

The performance of the 578 executives was assessed and rated by their leaders and the human resource managers together, with the ratings impacting their bonus and promotion. The ratings are binary, indicating whether each executive was a top performer in that period or not. We obtained two sets of performance ratings in Jan-Feb 2017 and Apr-May 2017 from the company.

3.1.2 Network position indicators

A network tie from A to B is constructed if A sends at least one email to B. Betweenness, closeness and degree centrality indicators are among the most widely used indicators of network position in email communication analysis (Borgatti, 2013). For example, centrality in email communication networks has been found to influence productivity (Aral and Van Alstyne 2007), predict job turnover (Gloor et al. 2017) and mediate the relationship between formal position and performance (Ahuja et al. 2003). Thus, we expect that higher centrality in email communication network is associated with superior work performance. The three centrality indicators were calculated with Condor using the following formula (Freeman, 1978; Wasserman and Faust, 1994) where $\#d_{ij}$ is the number of shortest email communication paths from i to j and $\#d_{ij}(t)$ is the number of those paths that pass through *t*; d_{ij} is the distance from i to j.

$$Betweenness\ Centrality = \sum_{\{i,j\neq t\}} \frac{\#d_{ij}(t)}{\#d_{ij}}$$

Betweenness centrality corresponds to the probability of being on the shortest path in the network and is commonly taken as a proxy for power and influence of a person in the network.

$$Closeness\ Centrality = \frac{N-1}{\sum_j d_{ij}}$$

Closeness centrality describes the average number of steps one has to take to reach any other person in the network and is a proxy for the embeddedness of a person in the network.

$$Degree\ Centrality = Number\ of\ email\ communication\ relationships$$

Degree centrality describes the number of nearest neighbors of a person in the network and can be taken as a proxy for information diversity that a person is exposed to.

3.1.3 Network contribution indicators

The amount of information contributed by an individual to the whole network can be approximated by the number of emails sent by the individual. Some researchers suggest that email content should also be considered to eliminate non-work-related emails (Ahuja et al. 2003; Kanawattanachai and Yoo, 2007). However, human coding would require a huge effort for our dataset. So we adopted the total number of emails sent as a proxy herein and address email contents with content indicators and automatic content analysis in the machine learning part.

Besides the absolute contribution, how much an individual contributes information compared to how much s/he receives information from others is also an important relative indicator (Erhardt et al., 2016; Aral and Van Alstyne 2011; Lim and Teo 2009). An intuitive indicator of relative information contribution is (Gloor, 2017):

$$Contribution\ Index = \frac{Number\ of\ message\ sent - Number\ of\ messages\ received}{Number\ of\ message\ sent + Number\ of\ messages\ received}$$

3.1.4 Network dynamic indicators

Email communication is an inherently dynamic process, in which individual network positions may be continuously changing over time. At the team-level, existing studies suggest that oscillation in network positions enables leadership rotation and is beneficial for mobilizing diverse participants' advantages over time (Davis and Eisenhardt, 2011). However, it remains to be examined whether these findings can be generalized to the individual level. Following Gloor et al. (2017), we specifically focus on the oscillation in betweenness centrality and calculate it as the number of times that an individual changed her/his role from local maxima or minima back and forth on a weekly basis (Kidane and Gloor, 2007).

Another important email communication dynamics is responsiveness. Reciprocity is the premise of effective communication (Fragale et al., 2012; Barry and Crant, 2000), especially for asynchronous email communication (Eckmann et al., 2004). Some previous studies found that higher responsiveness represents a positive signal of respect and is associated with superior team learning (Erhardt et al. 2016) and lower job turnover tendency (Gloor et al., 2017). However, other researchers argued that promptly responding to emails can cause distractions (Jackson et al., 2003), increase stress level (Barley et al., 2011; Kushlev and Dunn, 2015) and hence negatively influences individuals' performance (Bellotti et al., 2005). Therefore, there is a need to empirically examine how an individual's work performance is influenced by his/her own and his/her co-workers' responsiveness. Specifically, we consider the average response time (ART) and the "nudges", defined as the average number of pings (emails sent) needed to get a response (Gloor et al., 2017). For each individual, *ego ART* is the average time needed for the individual to respond, *ego nudges* is the average number of emails needed to get a response from the individual, *alter ART* is the average time the co-workers take to respond to the individual, and *alter nudges* is the average number of emails the individual needs to send in order to get a response from co-workers.

3.1.5. Network content indicators

Due to privacy restrictions we were not able to directly read the full emails. For the purpose of this analysis, we relied on the content indicator calculation program in Condor to obtain the content indicator data. The sentiment of email content reflects the sender's mood state and also potentially influences the receiver's mood state (Gloor and Giacomelli, 2014). Some empirical evidences suggest that email sentiment is predictive of individual job turnover (Gloor, 2016; Gloor et al., 2017) and team performance (Wasiak et al. 2011). There are many email sentiment calculation approaches, including manual rating, lexicon-based methods or machine learning models. The Condor software provides a built-in sentiment analysis function based on a Bayesian classifier, which has been trained on billions of tweets and achieves over 80% accuracy on many English email corpora (Brönnimann, 2014). We directly utilized this function to get the sentiment score of each email varying from 0-negative to 1-positive and then averaged over each individual's emails to derive a sentiment score that reflect her/his average sentiment level. We also calculated the deviation from neutral sentiment of each individual's emails as the emotionality indicator (Gloor et al., 2017). The idea behind this indicator is that a language that contains less neutral, more strongly positive or negative, expressions is more emotional.

The informativeness of the email content is another frequently studied aspect (Aral and Van Alstyne, 2011). The complexity indicator is calculated based on the likelihood distribution of words within an email, i.e. the probability of each word to appear in the text based on the well-known term frequency/inverse document frequency (TF-IDF) information retrieval metric (Brönnimann, 2014).

$$Complexity\ index = \frac{1}{n}\sum_{w \in V} q(w) log \frac{1}{p(w)}$$

where n is the total number of words within an email, V is the vocabulary of words that appear in the email corpus, $q(w)$ is the frequency of word w in the email, $p(w)$ is the probability of word w to appear in an email and $log\frac{1}{p(w)}$ is the inverse document frequency of word w in the corpus.

In this way, the complexity indicator is the opposite of the log-likelihood of the email text. So it measures the extent to which an email uses rare (complex) words and introduces non-redundant information. It can also be regarded as a word-level analogy to the previous measure of information diversity in emails (Aral and Van Alstyne's 2011).

Beside the complexity of an email itself, the extent to which co-workers will adopt new ideas and reuse the words that identify them is calculated as the influence indicator. Each time a receiver receives an email, his/her subsequent emails - sent within four days[1] - are retrieved and combined to derive a word distribution (i.e. how many times each word appears in the emails content). This word distribution is transformed into a vector of the dimensionality of vocabulary (V) using TF-IDF and is compared with the TF-IDF vector of the original email based on the cosine similarity measure, which is widely used in text mining (Tata and Patel 2007):

$$Influence\ index = cos(\frac{\sum_{w \in V} TI_s(w) TI_r(w)}{\sqrt{\sum_{w \in V} TI_s(w)^2}\sqrt{\sum_{w \in V} TI_r(w)^2}})$$

where $TI_s(w)$ is the TF-IDF value of word w in the sender's original email, $TI_r(w)$ is the TF-IDF value of word w in the receiver's subsequent emails.

[1] The number of days was determined by trial and error by the algorithm developer to find the influence indicator with the strongest explanatory power (Brönnimann, 2014).

If an individual's unique words (with high TF-IDF values) were adopted by co-workers in subsequent emails, we can expect that the idea expressed by these words was influential and being spread across the network (Iribarren and More, 2011). However, the timing of emails is the only mechanism controlling the direction of influence. There may be some confounding factors (e.g. common experience or face-to-face communication) that randomly cause co-workers to use the same words. As we averaged the influence score of all the emails for each individual, we assume that such error will cancel out in the large email dataset. On the other hand, we acknowledge this as a limitation, concerning which the findings on the influence index should be interpreted with caution.

3.1.6 Control variables

We collected basic personal information from the company's human resource system to construct control variables. The collected information was matched with other variables extracted from the company's HR database. Specifically, we controlled the effects of age, formal position (*band*, a binary variable indicating whether an individual holds a higher level formal position or not), tenure in the company and length of time since last promotion (measured in months). These were all the variables that the company was willing to share based on their privacy policy. These variables potentially influence both individuals' email communication behaviors and work performance. So we included them in the regression model to eliminate alternative interpretation of the effects of email indicators.

3.2 Partial least square regression model

As widely reported in many previous studies, there exist strong correlations among email network indicators, especially the network centrality indicators (Krackhardt, 1990), raising concerns of a multi-collinearity problem.

Some previous studies directly conducted OLS regression and interpret the result based on the indicators that appear to be significant (e.g. de-Marcos et al., 2016). The regression coefficients obtained by simple regression were lacking stability and robustness. Many studies avoid the collinearity among network indicators by retaining only one network indicator that is most theoretically interpretable (Krackhardt, 1990; Owen-Smith and Powell, 2004). However, in view of the explorative nature of this study, a well-established theoretical foundation is not in place to help decide which indicator should be chosen.

To empirically explore the effects of different network indicators, one stream of studies select the indicator that generates the best statistical result. For example, Powell et al. (1996) employed stepwise regression to determine which variable to enter in the final model. Kao et al. (2017) utilized the feature clustering method that groups strongly correlated variables into clusters and chose one representative variable from each cluster. Gloor et al. (2017) explored the effects of three centrality indicators in three separate models to choose the best model based on the information criterion. However, the presence of strong correlation effects may be attributable to the underlying links among indicators, which is not captured when selecting any single representative indicator. From a network theoretical lens, Kilduff and Tsai (2003) also recommended to control the effects of the other centrality indicators when analyzing a particular centrality indicator. Another stream of studies propose to introduce latent variables that model the underlying links. For example, de Andrade and Rêgo (2018) and Chang et al. (2017) extracted principal components from network centrality indicators for subsequent analysis. Ahuja et al. (2003) further suggest to use a component-based estimation strategy to incorporate the strong correlations among network centrality indicators, and adopted partial least square (PLS) regression to examine the effects of network centrality on work performance.

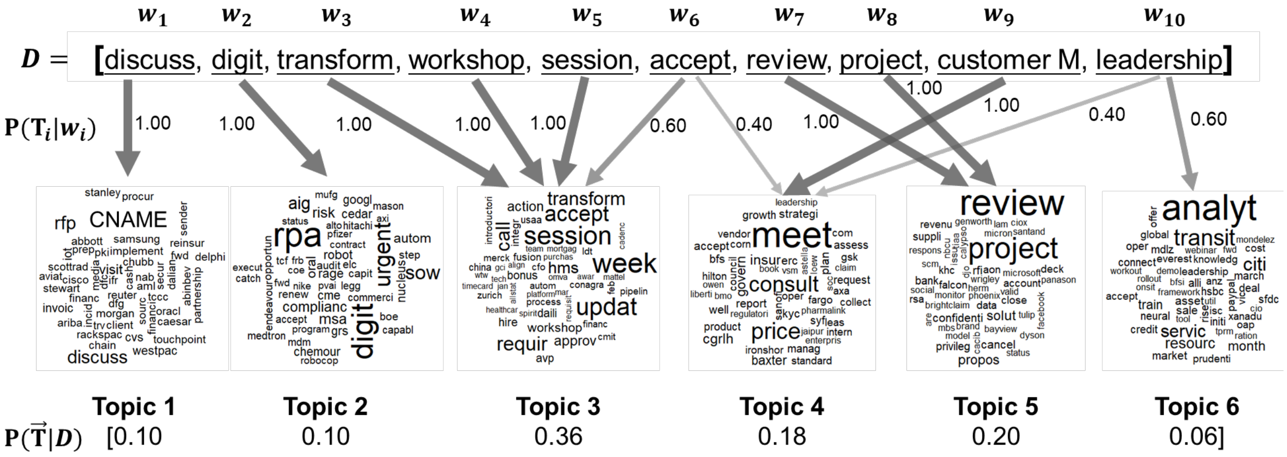
Combining the insights from these two streams of research, we explored the effects of email communication indicators with generalized PLS regression, which is suitable for our research purpose for two reasons. First, PLS regression addresses the multi-collinearity problem by grouping independent variables into latent components that capture the underlying links among independent variables. Second, it discovers component structure by maximizing model explanatory power rather than putting restrictions on the way independent variables are combined before running the regression. So it serves the explorative purpose of this study. To implement generalized PLS regression with binary dependent variable, we utilized the *plsRglm* package in R. To further address the concerns about multi-collinearity, we also followed Gilsing et al. (2008) to estimate multiple sub-models and assess the robustness of model coefficients. Besides, previous studies recommend to use multilevel data to alleviate the multi-collinearity problem of network indicators (Powell et al., 1996; Flynn and Wiltermut, 2010), so we performed regression analyses using two-wave panel data.

3.3 Machine learning models

In order to fully explore the predictive power of email communication indicators beyond linear regression and incorporate unstructured data into the prediction, we also employed machine learning models that allow more complex interactions between independent variables.

Email contents are not available to the researchers to protect privacy. As a compromise, we extracted the top 10 words that appeared in each individual's emails from the company's server using Condor's influence algorithm described in section 3.1. These keywords enable a high-level understanding of the topics that each individual talks about most frequently (Aral and Van Alstyne, 2011). We utilized latent Dirichlet allocation (LDA) to extract content features from this unstructured text data. LDA is a probabilistic model that estimates the probability distribution ($\mathbf{P}(\vec{\mathbf{T}}|\mathbf{D})$) of documents ($\mathbf{D}$) over topics ($\vec{\mathbf{T}}$), which are defined as probability distributions over words ($\mathbf{P}(\vec{\boldsymbol{w}}|\mathbf{T}_{\boldsymbol{i}})$ for each topic *i*). It has been found to be effective in identifying major topics from various kinds of text data, including emails (Sharaff and Nagwani, 2016). For our email keyword dataset, we treated each individual in one period as a unit of analysis (document $\mathbf{D}$) and estimated its topic distribution ($\mathbf{P}(\vec{\mathbf{T}}|\mathbf{D})$) to reflect how much the individual talked about each topic in that period. The meanings of the identified topics can be inferred based on their word distributions as shown in Figure 3. In this way, we transformed the raw keywords into 6 content features.

Combining the email indicators, the control variables and the content features, machine learning models can be trained to identify top performers. Extending the above logistic regression for the binary performance variable, we trained a logistic regression-based Adaboost model that considers more complex interaction among independent variables and uses a resampling strategy to reduce prediction errors (Friedman et al., 2000). The model parameters were tuned with cross-validation method to select the model with the strongest predictive power. We also performed cluster analysis to explore different types of top performers.



w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10}
$D =$ [discuss, digit, transform, workshop, session, accept, review, project, customer M, leadership]
$P(T_i|w_i)$
1.00
1.00
1.00
1.00
1.00
0.60
0.40
1.00
1.00
1.00
0.40
0.60
rfp CNAME discuss
rpa digit urgent
session week updat transform accept
meet consult price
review project
analyt transit citi servic
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
$P(\vec{T}|D)$
[0.10
0.10
0.36
0.18
0.20
0.06]

Figure 3 An illustrative example of the text feature derivation process based on the top 10 words

Note: The words and values in the figure are just indicative and do not reflect the word usage behavior of any specific individual. CNAME represents the company's name.

4. Results

The summary statistics and correlations of the variables are presented in Table 1. Many email indicators have strong correlations with the dependent variable (19. *top performance*), indicating potential predictive power that remains to be further explored.

Table 1 Summary statistics and correlations

	min	max	mean	SD	1	2	3	4	5	6	7	8	9	10
1. Betweenness	300.80	826181.04	35022.59	51562.03	-									
2. Closeness	0.27	0.50	0.37	0.01	0.25***	-								
3. Degree	28	924	232.10	120.05	0.75***	0.28***	-							
4. Messages sent	15	8064	1159	1142.93	0.17***	0.59***	0.22***	-						
5. Contribution	-0.91	0.89	-0.23	0.27	-0.14***	0.38***	-0.19***	0.26***	-					
6. Influence	0.21	0.73	0.18	0.33	0.19***	0.12***	0.31***	0.21***	0.03	-				
7. Sentiment	0.28	0.81	0.63	0.04	-0.15***	0.03	-0.18***	0.03	0.09***	-0.19***	-			
8. Complexity	7.17	10.11	8.26	0.32	0.01	0.04	0.02	0.07*	-0.16***	0.38***	-0.13***	-		
9. Emotionality	0.18	0.36	0.27	0.02	0.20***	-0.01	0.24***	0.00	-0.15***	0.10***	-0.11***	0.15***	-	
10. Bet OSC	14	37	25.83	3.65	-0.07*	0.02	-0.07*	-0.04	0.05	-0.02	0.08*	-0.04	-0.05	-
11. Alter nudges	1	6	1.53	0.31	0.06	0.13***	0.04	0.08*	-0.52***	-0.19***	0.00	0.09***	0.10***	-0.03
12. Alter ART	1.92	63.45	21.42	6.66	0.14***	0.26***	0.13***	0.12***	-0.36***	-0.28***	0.04	0.07*	0.14***	0.01
13. Ego nudges	1.04	3.52	1.47	0.02	0.01	0.09***	-0.02	0.14***	0.59***	0.27***	-0.24***	0.03	-0.03	0.04
14. Ego ART	6.66	40.10	20.76	5.42	-0.02	0.00	0.02	0.01	0.01	-0.05	0.03	-0.04	0.00	-0.01
15. Age	33	67	44.82	6.13	0.41***	0.14***	0.57***	0.09***	0.08	0.17***	-0.18***	-0.07**	0.14***	-0.02
16. Band	0	1	0.26	0.44	-0.02	-0.03	-0.04	-0.01	-0.04	0.01	-0.03	0.03	-0.05	-0.01
17. Tenure	4	323	99.75	71.86	0.34***	0.10***	0.39***	0.11***	-0.26***	0.02	-0.05	0.05	0.19***	-0.08**
18. TSLP	0	132	30.84	25.08	-0.29***	-0.21***	-0.23***	-0.17***	0.11***	-0.23***	0.11***	-0.03	-0.13***	0.07
19. Top performer	0	1	0.35	0.48	0.21***	0.49***	0.23***	0.26***	-0.03	0.32***	0.26***	0.17***	-0.13***	0.01

Table 2 Summary statistics and correlations (continued)

	11	12	13	14	15	16	17	18
12. Alter ART	0.66***	-						
13. Ego nudges	-0.43***	-0.36***	-					
14. Ego ART	-0.02	-0.01	0.02	-				
15. Age	-0.05	0.03	0.08**	0.03	-			
16. Band	0.04	0.00	0.02	-0.01	-0.02	-		
17. Tenure	0.16***	0.28***	-0.09***	0.03	0.19***	0.01	-	

4.1 Regression analysis results

First, we estimated a model with only control variables as the baseline. As shown in Table 1, all the four control variables appear to be significantly related to work performance. Second, we ran a regression for each group of email indicators separately with control variables included. Third, a full model including all variables was estimated to compare with the sub-models in the second step and assess the robustness of the model coefficients. For each model, we performed Wald tests on whether the period fixed effects are significant and whether each model coefficient varies significantly across the two periods. If the model coefficients are invariant across periods and the period fixed effect is insignificant, the model is essentially equivalent to a pooled regression. If the model coefficients are invariant but the period effect is significant, a fixed effect model is estimated. If both the model coefficients and the period effect vary significantly across periods, we estimate a variable coefficient model. The types of models used are listed in Table 1.

The comparison between the full and the sub-models suggests that the majority of model coefficients are consistent and robust, and the model coefficients are visualized in Figure 4. The communication network centrality indicators are all positively related to top performance with *closeness centrality* having the strongest relationship. The coefficients of contribution indicators vary significantly between the full model and the sub-model and across the two periods, indicating that the effects of contribution indicators are not robust. As for content indicators, *influence*, *sentiment* and *complexity* indicators all have significant positive relations with top performance, while *emotionality* is negatively related to top performance. Among the dynamic indicators, only *ego nudges* has a consistently significant relation with top performance, and lower ego responsiveness (higher *ego nudges*) is associated with a lower probability of being a top performer.

Table 3 Regression analysis results

	Baseline	**Position**	**Contribution**	**Content**	**Dynamics**	**Full**
Position						
Betweenness		0.028(0.005)***				0.045(0.008)***
Closeness		0.124(0.009)***				0.234(0.038)***
Degree		0.032(0.004)***				0.030(0.010)**
Contribution						
Message sent$_1$			0.185(0.046)***			-0.013(0.021)
Message sent$_2$			-0.189(0.039)***			-0.028(0.021)
BOC$_1$			-0.101(0.014)***			-0.022(0.010)*
BOC$_2$			-0.036(0.019)			0.011(0.021)
Content						
Influence				0.135(0.010)***		0.180(0.014)***
Sentiment				0.141(0.016)***		0.109(0.008)***
Complexity				0.057(0.006)***		0.051(0.008)***
Emotionality				-0.068(0.007)***		-0.087(0.011)***
Dynamics						
BetOsc$_1$					0.001(0.002)	0.003(0.004)
BetOsc$_2$					0.008(0.002)***	0.008(0.004)*
Alter nudges					-0.034(0.015)*	-0.023(0.014)
Alter ART					-0.037(0.017)*	-0.005(0.009)
Ego nudges					-0.169(0.020)***	-0.174(0.029)***
Ego ART$_1$					-0.045(0.009)***	-0.021(0.005)***
Ego ART$_2$					0.053(0.017)***	0.028(0.006)***
Controls						
Age	0.019(0.004)***	-0.009(0.006)	0.033(0.006)***	0.026(0.002)***	0.036(0.003)***	0.026(0.008)**
Band	-0.010(0.003)**	-0.005(0.001)***	-0.011(0.001)***	-0.011(0.003)***	-0.002(0.006)	0.004(0.004)
Tenure	0.037(0.013)**	0.008(0.002)***	0.005(0.013)	0.046(0.008)***	0.038(0.010)***	-0.014(0.008)
TSLP	-0.073(0.008)***	-0.018(0.004)***	-0.053(0.010)***	-0.070(0.004)***	-0.065(0.013)***	0.008(0.008)
AIC	1218.757	1189.940	1139.242	925.529	1063.884	796.536
Pseudo R^2	0.182	0.200	0.233	0.377	0.280	0.452
Components	2	3	4	3	6	11
Model type	Fixed	Fixed	Variable	Fixed	Variable	Variable
N	1156	1156	1156	1156	1156	1156

Note: BOC, BetOsc, ART and TSLP stand for balance of contribution, betweenness centrality oscillation, average response time and time since last promotion respectively. Variables that have coefficients significantly different in the two periods are listed with subscripts indicating the period.

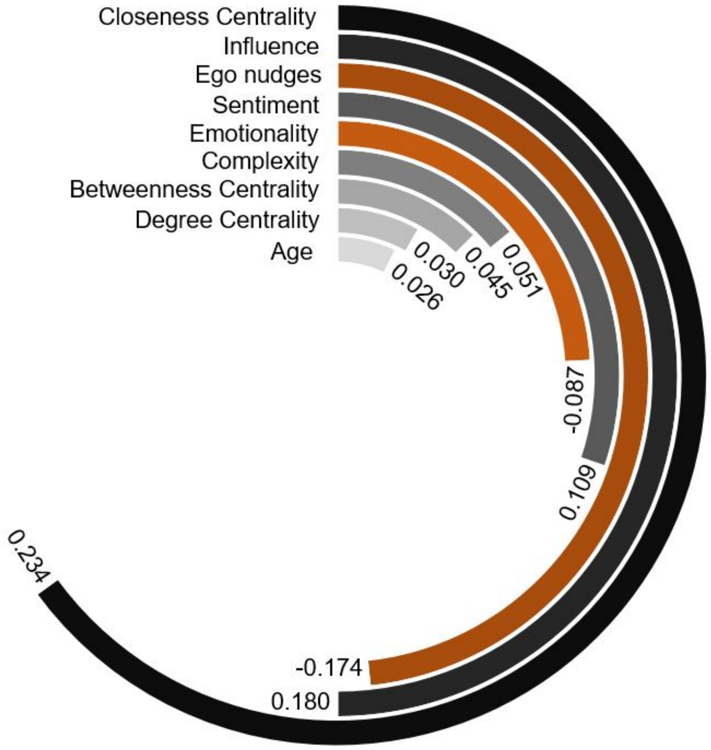


Figure 4 Significant coefficients in the full model

4.2 Performance prediction models

To better exploit the predictive power of email indicators and allow nonlinear interactions among them, we trained performance prediction models using logistic regression-based Adaboost algorithms. The extracted content features were also included in model training to explore their predictive power.

Logistic regression-based Adaboost models were trained using cross-validation to select the best model. The model accuracy was evaluated using leave-one-out cross-validation (LOOC) to avoid the randomness in splitting training and testing datasets (Wong, 2015). With text features included, the best model achieves 83.56% accuracy (Kappa coefficient is 0.620), while the best accuracy is 79.41% without text features (Kappa coefficient is 0.496).

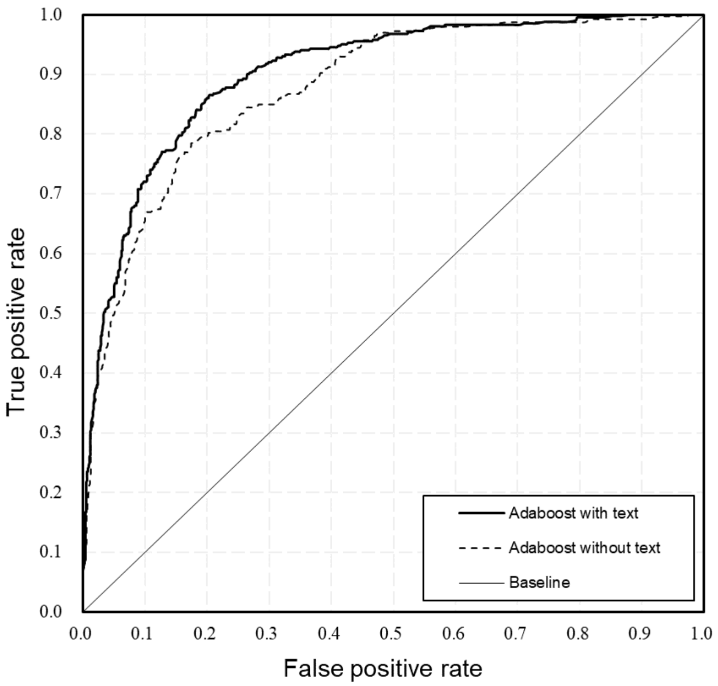


Figure 5 The ROC curve of the different machine learning models

The sample is approximately balanced with 400 top performers (positive samples) and 756 non-top performers (negative samples), so the performance of the models should also be evaluated with respect to the sensitivity and specificity of prediction besides accuracy. In the context of this study, the *sensitivity* of the performance prediction model refers to the proportion of top performers that are correctly predicted. It is also known as true positive rate (TPR), i.e. truly positive as predicted. *Specificity* measures the proportion of non-top performers that are correctly predicted, and 1-*specificity* is also known as false positive rate (FPR), i.e. falsely predicted as positive. The direct outputs of a machine learning model are decision values for each test sample, and the threshold level should be set to give a prediction for each test sample. A lower threshold level makes it easier for truly positive samples to be predicted as positive and hence increases TPR. But this also makes more negative samples to be falsely predicted as positive, and hence increases FPR. A higher threshold level causes the contrary. Thus, there is a trade-off between increasing TPR and controlling FPR, and the predictive power of a model can be evaluated by the extent to which the increase in TPR can be achieved without too much increase in FPR. This can be directly evaluated by plotting FPR on the x-axis and TPR on the y-axis to construct a Receiver Operating Characteristic (ROC) curve. The more an ROC curve fills the Area Under the Curve (AUC), the higher the predictive power of a model. The ROC curves of the six models are plotted in Figure 5.

The Adaboost model with text features has the strongest predictive power with the ability to correctly identify more than 80% of the top performers and, at the same time, keeps the mistakes of incorrectly identifying non-top performers at less than 20%. Although interpreting the effects of predictors based on machine learning models is less straightforward, it is clear that the email indicators are highly predictive signals of individuals' work performance. The text features can also provide additional predictive power in identifying top performers.

5. Discussions

5.1 Predictive email communication indicators

According to Table 3, three network position indicators (betweenness, closeness and degree), one dynamic indicator (ego nudges), and four content indicators (influence, sentiment, complexity and emotionality) have significant associations with top performance. The machine learning model results in 4.2 further suggest that these indicators can act as the predictive signal of top performers with considerable predictive power.

Organizational email communication networks have been considered as highly constrained by predefined formal organization structures (Brass et al., 2004). Thus, it remains to be examined whether email network position indicators can act as predictive signals of individuals' performance or are merely "shadows" of formal organization positions (Ahuja et al., 2003). The empirical findings of this study support the relevance of the three centrality indicators in predicting work performance. The predictive power can be interpreted from three perspectives. First, central email network positions enable individuals to obtain a wide range of information, exercise control over information flows and have timely access to information (Freeman, 1978; Trier, 2008). These information advantages can translate to superior work performance. Second, voluntary informal interaction with coworkers beyond predefined formal organization structure is an important dimension of organization citizenship behavior, which contributes to high work performance (Konovsky and Pugh, 1994). Email provides an easy access to coworkers across the organization regardless of temporal or geographic restrictions and hence has a large potential to facilitate additional informal interactions (Barry and Crant, 2000; Mazmanian et al., 2013). Such informal interactions are critical for accumulating social capital and completing collaborative tasks (Bolino et al., 2002). Third, central communication network positions imply more collaborative experience with coworkers, which according to the transactive memory theory, improves future collaboration efficiency (Lewis, 2004; Kanawattanachai and Yoo, 2007). Individuals that are familiar with each other were found to outperform unacquainted individuals in collaborative tasks (Parise and Rollag, 2010). Therefore, individuals with central network positions are at an advantage in utilizing abundant transactive memory to achieve better performance.

Compared with questionnaire surveys based on respondents' memory, email archives provide a unique opportunity to obtain additional information on the contents of communication in real-world organizations. The four content indicators calculated from raw email texts are significant predictors of top performance. The more unique the information in an individual's email contents is, i.e. the higher his/her language complexity, the higher the individual's probability of being a top performer. This result is consistent with previous findings that email information diversity is positively associated with work performance (Aral and Van Alstyne, 2012). The present study extends these findings in that the more coworkers adopt such diverse information in subsequent emails (higher influence indicator), the more likely the individual will be a top performer. Besides the informativeness of email contents, positive sentiment and low variability in sentiment (low emotionality) are associated with superior performance. This supports the notion that business communication should have less fluctuation in emotion, and expressing ideas with positive sentiment is beneficial (Byron, 2008).

Ego nudges, the average number of emails that need to be sent to an individual in order to get the individual's response, is the only dynamic indicator significantly associated with top performance. On the one hand, timely responses to emails are conducive to keeping coworkers in synchronization and coordinating collaborative tasks (Mazmanian et al., 2013). This is especially important for tasks with strong interdependencies (Kanawattanachai and Yoo, 2007). On the other hand, as email overload becomes increasingly prevalent in modern organizations, individuals may find it hard to absorb information from huge amount of emails and give timely responses (Barley et al., 2011). Beyond the time consumed by emails, emails may cause interruptions to individuals' work and lower work efficiency (Kushlev and Dunn, 2015). Therefore, an adequate email processing strategy is necessary to maintain the balance between giving timely responses and staying concentrated on tasks (Gupta et al., 2011). Our empirical analysis results suggest that being responsive to more emails (*ego Nudges*) seems to be a more effective strategy than responding in shorter time (*ego ART*).

Besides, email text features also facilitate better prediction accuracy, suggesting that it is also important to select appropriate topics to discuss in emails. Taken together, these email indicators have considerable predictive power on work performance. Many existing studies imply that individuals' communication competence is a valuable skill in the digital communication environment compared to other communication channels (Robert et al., 2008; Makarius and Larson, 2017). However, the majority of existing studies on communication competence rely on self-report, which may miss many behavioral cues underpinning communication competence (Hwang, 2011). The email communication patterns of top performers identified in this study also act as the foundation for operationalizing the construct of email communication competence.

Compared to previous studies at the team-level (Marlow et al., 2017), some enablers of superior team performance can be generalized to individual-level, such as high network centrality, high responsiveness and positive sentiment. However, not all team-level phenomena are supported at the individual-level. For example, relational leadership rotation (*betweenness oscillation*) and coworkers' high responsiveness (*alter ART* and *alter nudges*) are expected to promote team creativity and efficiency (Kidane and Gloor, 2007), but do not show significant association with individual performance. These results indicate that a top performer is not completely equivalent to an effective team collaborator, and imply different causal mechanisms to be further examined.

5.2 Versatile top performers

The predictive power of machine learning models suggests that there may exist complex interaction patterns among email indicators not captured by regression model. Therefore, we performed clustering analysis to further reveal top performers' email communication patterns. K-means clustering was employed to explore the underlying types of top performers based on their email communication indicators. The optimal number of clusters was determined by the “elbow” criterion (Ketchen and Shook, 1996), which balances the number of clusters (interpretability of clusters) and between-group variance maximization (explanatory power of clusters). The results suggest that three clusters of top performers emerged in the sample. In order to test the robustness of the clustering result, we also performed clustering analysis using the Gaussian mixture model based on the Expectation Maximization algorithm (Fraley and Raftery, 1999). The Kappa inter-rater agreement coefficient between the two sets of clustering results is 0.954 (*p*<0.001) indicating strong consistency. Therefore, we consider the clustering results consistent and eligible for further interpretation. To visualize the clustering results, we conducted dimension reduction using principal component analysis (PCA) and plotted top performers on the coordinates of the first two principal components (Figure 5). The values of the three cluster centers in the dimension of each email indicator are presented using radar charts as the representative email communication profiles of the three kinds of top performers.

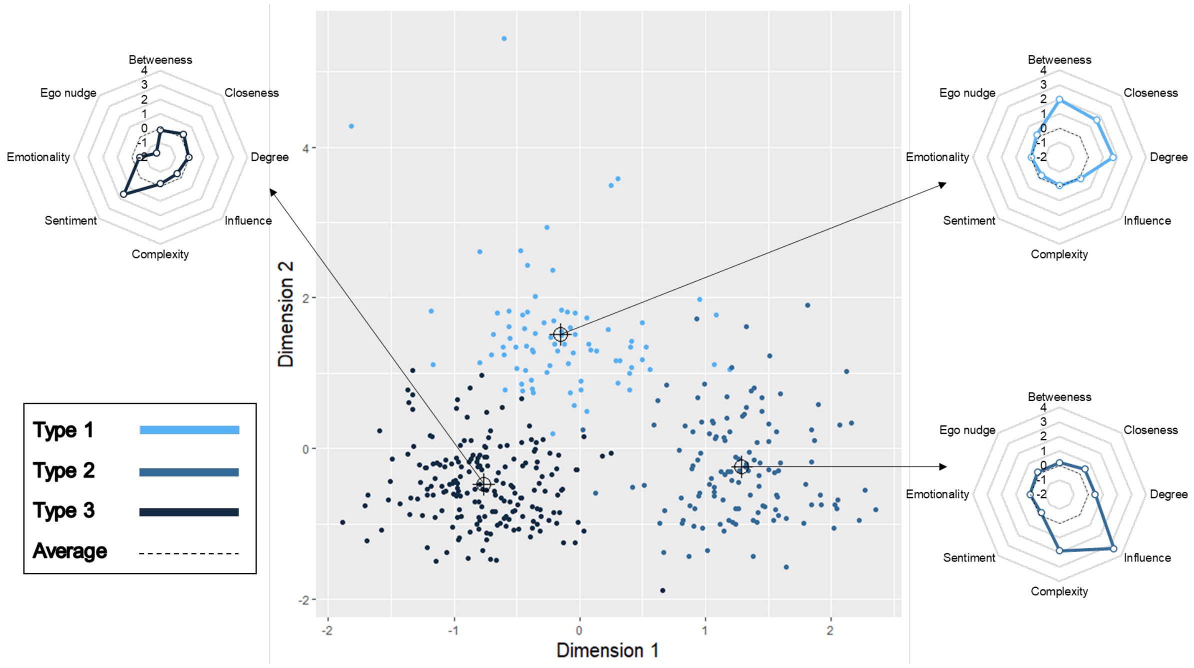


Figure 6 The email communication profile of the three types of top performers

Note: The values in the radar charts are measured in terms of standard deviations from the sample mean.

As shown in Figure 6, three kinds of top performers have rather distinct email communication profiles, which can be interpreted as: “networkers”, ‘influencers”, and “positivists”. Type 1 top performers locate at the center of the whole network with high network centrality values. Their superior social capitals in the email communication network enable them to act as the relational leaders that facilitate information diffusion in the organization. Type 2 top performers are only slightly higher in network centrality than average but have much higher influence and complexity indicator values. This suggests that their emails introduce more novel information (complexity) and tend to be followed by coworkers (influence). Thus, they can be regarded as the opinion leaders of the organization. Type 3 top performers stand out for their strongly positive sentiment, low emotion fluctuation and high responsiveness (low *ego nudges*) to coworkers' emails. They appear to be the emotional leaders that emit positive power and strengthen cohesiveness in the organization. The versatility of top performers' email communication profiles provides two fundamental implications for email communication competence.

First, not all top performers have the same pattern of email communication. There appears no unified set of optimal email communication pattern, and various patterns can be associated with top performance. This can be understood with respect to the fact that different roles, such as relational leaders ("networkers"), opinion leaders ("influencers") and emotional leaders ("positivists"), are needed to fulfill different organizational functions (O'Reilly et al., 1991; Soltis, 2015). This finding also implies the need to rethink the definition of email communication competence, which may be better operationalized as the combination of the strength in several communication dimensions instead of a unified construct.

Second, not all favorable communication behaviors (e.g. high responsiveness, positive sentiment) reside in one kind of top performers. This, on the one hand, reflects the inherent trade-off between the focus on different communication behavior given an individual's limited time and energy. For example, it may be hard for an individual to be highly responsive (type 3) and at the same time keep introducing novel and influential ideas (type 2). On the other hand, as shown in Figure 6, the three kinds of top performers are rather specialized in a few dimensions of email indicators and are average (or even below average) in other dimensions. This can also be viewed as top performers' adaptation to organizational needs instead of being dominant in every respect.

6. Conclusion

This study reveals the email communication indicators that are predictive of individuals' work performance. The panel regression models provide interpretable results that top performance is associated with central network position, positive sentiment, low emotionality, high complexity, more adoption of influential words by coworkers and high responsiveness in email communication. The machine learning models that allow more complex interactions among independent variables indicate the high predictive power of email indicators with the best model achieving over 80% accuracy. The cluster analysis results further reveal that the top performers can be generally classified into three types that have advantages in different dimensions.

For theory development, the findings provide implications for operationalizing the construct of email communication competence with subjective measures from real-world email data. The variation in communication style of top performers further implies that email communication competence might be better defined as a combination of several supportive factors instead of a unified construct. For management practices, the identified email indicators can be used to suggest improvements in email communication skill development training. Furthermore, as fine-grained email communication data becomes increasingly available, the analyses performed in this study can also be replicated in different organizations to understand individuals' contributions to organization communication.

The implications of this study should be viewed with respect to the following limitations, which leave room for improvements in future studies. First, the data are from only one company. Although the fact that the company operates internationally alleviates this limitation to some extent, caution should still be taken when generalizing the findings to other organizational settings. Future studies to test the robustness of the findings in a broader range of cultural and organization environments are needed. Second, because of privacy restrictions we cannot directly observe the email contents, which may provide additional information. Third, the findings of this study cannot support causal interpretation. Future research may use causal inference techniques, such as natural experiments, to precisely estimate the effects of influential email indicators.

References

Ahuja, M.K., D.F. Galletta, K.M. Carley. 2003. Individual centrality and performance in virtual R&D groups: An empirical study. Management science. 49(1) 21-38.

Aral, S., E. Brynjolfsson, M. Van Alstyne. 2012. Information, technology, and information worker productivity. Information Systems Research. 23(3-part-2) 849-867.

Aral, S., M. Van Alstyne. 2007. Network structure & information advantage. Proceedings of the Academy of Management Conference, Philadelphia, PA.

Aral, S., M. Van Alstyne. 2011. The diversity-bandwidth trade-off. American Journal of Sociology. 117(1) 90-171.

Barley, S.R., D.E. Meyerson, S. Grodal. 2011. E-mail as a source and symbol of stress. Organization Science. 22(4) 887-906.

Barry, B., J.M. Crant. 2000. Dyadic communication relationships in organizations: An attribution/expectancy approach. Organization Science. 11(6) 648-664.

Barry, B., I.S. Fulmer. 2004. The medium and the message: The adaptive use of communication media in dyadic influence. Academy of Management Review. 29(2) 272-292.

Bellotti, V., N. Ducheneaut, M. Howard, I. Smith, R.E. Grinter. 2005. Quality versus quantity: E-mail-centric task management and its relation with overload. Human-computer interaction. 20(1) 89-138.

Brandts, J., D.J. Cooper, R.A. Weber. 2014. Legitimacy, communication, and leadership in the turnaround game. Management Science. 61(11) 2627-2645.

Brass, D.J., J. Galaskiewicz, H.R. Greve, W. Tsai. 2004. Taking stock of networks and organizations: A multilevel perspective. Academy of management journal. 47(6) 795-817.

Brönnimann, L. (2014). Multilanguage sentiment analysis of Twitter data on the example of Swiss politicians. University of Applied Sciences Northwestern Switzerland. M.Sc. Thesis.

Butts, M.M., W.J. Becker, W.R. Boswell. 2015. Hot buttons and time sinks: The effects of electronic communication during nonwork time on emotions and work-nonwork conflict. Academy of Management Journal. 58(3) 763-788.

Byron, K. 2008. Carrying too heavy a load? The communication and miscommunication of emotion by email. Academy of Management Review. 33(2).

Byron, K., D.C. Baldridge. 2007. E-mail recipients' impressions of senders' likability: The interactive effect of nonverbal cues and recipients' personality. The Journal of Business Communication (1973). 44(2) 137-160.

Chan, S.H.J., H.Y.I. Lai. 2017. Understanding the link between communication satisfaction, perceived justice and organizational citizenship behavior. Journal of Business Research. 70 214-223.

Cornelissen, J.P., R. Durand, P. Fiss, J. Lammers, E. Vaara. 2015. Putting communication front and center in institutional theory and analysis. Academy of Management Review. 40(1) 10-27.

Cross, R., J.N. Cummings. 2004. Tie and network correlates of individual performance in knowledge-intensive work. Academy of management journal. 47(6) 928-937.

Davis, J.P., K.M. Eisenhardt. 2011. Rotating leadership and collaborative innovation: Recombination processes in symbiotic relationships. Administrative Science Quarterly. 56(2) 159-201.

de-Marcos, L., E. García-López, A. García-Cabot, J.-A. Medina-Merodio, A. Domínguez, J.-J. Martínez-Herráiz, T. Diez-Folledo. 2016. Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. Computers in Human Behavior. 60 312-321.

Eagle, N., A.S. Pentland, D. Lazer. 2009. Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences. 106(36) 15274-15278.

Erhardt, N., J. Gibbs, C. Martin-Rios, J. Sherblom. 2016. Exploring affordances of email for team learning over time. Small Group Research. 47(3) 243-278.

Fragale, A.R., J.J. Sumanth, L.Z. Tiedens, G.B. Northcraft. 2012. Appeasing equals: Lateral deference in organizational communication. Administrative Science Quarterly. 57(3) 373-406.

Fraley, C., A.E. Raftery. 1999. MCLUST: Software for model-based cluster analysis. Journal of classification. 16(2) 297-306.

Freeman, L.C. 1978. Centrality in social networks conceptual clarification. Social networks. 1(3) 215-239.

Friedman, J., T. Hastie., R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2) 337-407.

Gajendran, R.S., A. Joshi. 2012. Innovation in globally distributed teams: The role of LMX, communication frequency, and member influence on team decisions. Journal of Applied Psychology. 97(6) 1252.

George, G., M.R. Haas, A. Pentland. 2014. Big data and management. Academy of Management Journal. 57(2) 321-326.

Gloor, P.A. 2016. What email reveals about your organization. MIT Sloan Management Review. 57(2) 8.

Gloor, P.A. 2017. Sociometrics and Human Relationships: Analyzing Social Networks to Manage Brands, Predict Trends, and Improve Organizational Performance. Emerald Publishing Limited.

Gloor, P.A., A.F. Colladon, F. Grippa, G. Giacomelli. 2017. Forecasting managerial turnover through e-mail based social network analysis. Computers in Human Behavior. 71 343-352.

Gloor, P.A., K. Fischbach, H. Fuehres, C. Lassenius, T. Niinimäki, D.O. Olguin, S. Pentland, A. Piri, J. Putzke. 2011. Towards "honest signals" of creativity–identifying personality characteristics through microscopic social network analysis. Procedia-Social and Behavioral Sciences. 26 166-179.

Gloor, P.A., G. Giacomelli. 2014. Reading Global Clients Signals. MIT Sloan management review. 55(3) 23.

Golden, T.D., J.F. Veiga, R.N. Dino. 2008. The impact of professional isolation on teleworker job performance and turnover intentions: Does time spent teleworking, interacting face-to-face, or having access to communication-enhancing technology matter? Journal of Applied Psychology. 93(6) 1412.

He, S., T. Offerman, J. van de Ven. 2016. The sources of the communication gap. Management Science.

Hwang, Y. 2011. Is communication competence still good for interpersonal media?: Mobile phone and instant messenger. Computers in Human Behavior. 27(2) 924-934.

Iribarren, J.L., E. Moro. 2011. Affinity paths and information diffusion in social networks. Social networks. 33(2) 134-142.

Jackson, T., R. Dawson, D. Wilson. 2003. Reducing the effect of email interruptions on employees. International Journal of Information Management. 23(1) 55-65.

Jarvenpaa, S.L., T.R. Shaw, D.S. Staples. 2004. Toward contextualized theories of trust: The role of trust in global virtual teams. Information systems research. 15(3) 250-267.

Johns, T., R. Laubacher, T. Malone. 2011. The age of hyperspecialization. Harvard Business Review. 89(7-8) 56.

Kanawattanachai, P., Y. Yoo. 2007. The impact of knowledge coordination on virtual team performance over time. MIS quarterly 783-808.

Kao, T.-W.D., N. Simpson, B.B. Shao, W.T. Lin. 2017. Relating supply network structure to productive efficiency: A multi-stage empirical investigation. European Journal of Operational Research. 259(2) 469-485.

Ketchen Jr, D.J., C.L. Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strategic management journal 441-458.

Kidane, Y.H., P.A. Gloor. 2007. Correlating temporal communication patterns of the Eclipse open source community with performance and creativity. Computational and mathematical organization theory. 13(1) 17-27.

Kilduff, M., W. Tsai. 2003. Social networks and organizations. Sage.

Konovsky, M.A., S.D. Pugh. 1994. Citizenship behavior and social exchange. Academy of management journal. 37(3) 656-669.

Krackhardt, D. 1990. Assessing the political landscape: Structure, cognition, and power in organizations. Administrative science quarterly 342-369.

Kruger, J., N. Epley, J. Parker, Z.-W. Ng. 2005. Egocentrism over e-mail: Can we communicate as well as we think? Journal of personality and social psychology. 89(6) 925.

Kushlev, K., E.W. Dunn. 2015. Checking email less frequently reduces stress. Computers in Human Behavior. 43 220-228.

Lewis, K. 2004. Knowledge and performance in knowledge-worker teams: A longitudinal study of transactive memory systems. Management science. 50(11) 1519-1533.

Lim, V.K., T.S. Teo. 2009. Mind your E-manners: Impact of cyber incivility on employees' work attitude and behavior. Information & Management. 46(8) 419-425.

Lomi, A., D. Lusher, P.E. Pattison, G. Robins. 2013. The focused organization of advice relations: A study in boundary crossing. Organization Science. 25(2) 438-457.

Marlow, S.L., C.N. Lacerenza, J. Paoletti, C.S. Burke, E. Salas. 2017. Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. Organizational Behavior and Human Decision Processes.

Mazmanian, M., W.J. Orlikowski, J. Yates. 2013. The autonomy paradox: The implications of mobile email devices for knowledge professionals. Organization science. 24(5) 1337-1357.

Mazmanian, M., J. Yates, W. Orlikowski. 2006. Ubiquitous Email: Individual Experiences and Organizational Consequences of Blackberry Use. Academy of Management Conference.

Mesmer-Magnus, J.R., L.A. DeChurch. 2009. Information sharing and team performance: A meta-analysis. Journal of Applied Psychology. 94(2) 535.

Mesmer-Magnus, J.R., L.A. DeChurch, M. Jimenez-Rodriguez, J. Wildman, M. Shuffler. 2011. A meta-analytic investigation of virtuality and information sharing in teams. Organizational Behavior and Human Decision Processes. 115(2) 214-225.

Ocasio, W., J. Loewenstein, A. Nigam. 2015. How streams of communication reproduce and change institutional logics: The role of categories. Academy of Management Review. 40(1) 28-48.

O'Reilly, C.A., J. Chatman, D.F. Caldwell. 1991. People and organizational culture: A profile comparison approach to assessing person-organization fit. Academy of management journal. 34(3) 487-516.

Pan, W., G. Ghoshal, C. Krumme, M. Cebrian, A. Pentland. 2013. Urban characteristics attributable to density-driven tie formation. Nature communications. 4.

Parise, S., K. Rollag. 2010. Emergent network structure and initial group performance: The moderating role of pre - existing relationships. Journal of Organizational Behavior. 31(6) 877-897.

Pentland, A., T. Heibeck. 2010. Honest signals: how they shape our world. MIT press.

Pentland, A., D. Lazer, D. Brewer, T. Heibeck. 2009. Using reality mining to improve public health and medicine. Stud Health Technol Inform. 149 93-102.

Powell, W.W., K.W. Koput, L. Smith-Doerr. 1996. Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. Administrative science quarterly 116-145.

Sarker, S., M. Ahuja, S. Sarker, S. Kirkeby. 2011. The role of communication and trust in global virtual teams: A social network perspective. Journal of Management Information Systems. 28(1) 273-310.

Sharaff, A., N.K. Nagwani. 2016. Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. Journal of Information Science. 42(2) 200-212.

Shore, J., E. Bernstein, D. Lazer. 2015. Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. Organization Science. 26(5) 1432-1446.

Soltis, S.M. 2015. Person-Organization Fit and Employee Performance: A Social Network Perspective. Academy of Management.

Sosa, M.E., M. Gargiulo, C. Rowles. 2015. Can informal communication networks disrupt coordination in new product development projects? Organization Science. 26(4) 1059-1078.

Tata, S., J.M. Patel. 2007. Estimating the selectivity of tf-idf based cosine similarity predicates. ACM SIGMOD Record, 36(2), 7–12. https://doi.org/10.1145/1328854.1328855

Wasiak, J., B. Hicks, L. Newnes, C. Loftus, A. Dong, L. Burrow. 2011. Managing by e-mail: what e-mail can do for engineering project management. IEEE Transactions on engineering management. 58(3) 445-456.

Webster, J., R.T. Watson. 2002. Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly xiii-xxiii.

Wong, T.-T. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition. 48(9) 2839-2846.

Woolley, A.W., C.F. Chabris, A. Pentland, N. Hashmi, T.W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. science. 330(6004) 686-688.