

# Towards Finding Valuable Topics

Zhen Wen \*

Ching-Yung Lin †

## Abstract

Enterprises depend on their information workers finding valuable information to be productive. However, existing enterprise search and recommendation systems can exploit few studies on the correlation between information content and information workers' productivity. In this paper, we combine content, social network and revenue analysis to identify computational metrics for finding valuable information content in people's electronic communications within a large-scale enterprise. Specifically, we focus on two questions: (1) how are the topics extracted from such content correlate with information workers' performance? and (2) how to find valuable topics with potentially high impact on employee performance? For the first question, we associate the topics with the corresponding workers' productivity measured by the revenue they generate. This allows us to evaluate the topics' influence on productivity. We further verify that the derived topic values are consistent with human assessor subjective evaluation. For the second question, we identify and evaluate a set of significant factors including both content and social network factors. In particular, the social network factors are better in filtering out low-value topics, while content factors are more effective in selecting a few top high-value topics. In addition, we demonstrate that a Support Vector regression model that combines the factors can already effectively find valuable topics. We believe that our results provide significant insights towards scientific advances to find valuable information.

## 1 Introduction

It is increasingly important for enterprises to improve information workers' productivity [13], which has sparked an interest in tools assisting information workers to search relevant information such as expertise [23, 25]. However, it is not immediately clear how information found in these tools can impact workers' productivity. To help employees to find valuable information that improves their productivity, it requires a comprehensive understanding of the intrinsic characteristics of information content within enterprises and how this is

related to worker productivity. To this end, recent empirical work has started to capture people's electronic communications (e.g., email, text messaging, and document repositories) [6, 25], as well as productivity metrics (e.g., revenues and performance ratings) [36]. In particular, such electronic communications data have the advantage of wide coverage and minimal need of human involvement.

Based on people's relationships revealed by the captured data, research has shown the benefit of social networks on information worker productivity within an organization [6, 36]. However, most existing studies focus on social network topologies and node properties. Little research investigates the correlation between productivity and the ample yet diverse information content created by people's communications. On the other hand, prior works on information content in social networks have been focused on topics extraction [16, 31], content relevancy to given user queries [25], or content quality based on user interactions [2]. Nevertheless, such analyses do not directly explain how content can impact information workers' productivity, an issue especially important to enterprises. For example, a "hot" topic may appeal to employees' personal life, but may not improve their productivity. As a result, how to discover valuable topics in large-scale enterprise environments that impact worker productivity still remains a challenge.

This paper presents a study toward addressing this challenge. In particular, we focus on content of people's communications in a large enterprise. Our study is based on a privacy-preserving organizational social network analysis system [25] that gathers, crawls and mines various types of data sources within an organization, including email, instant message communications, calendars, and hierarchical structure of the organization. The system is deployed within a large-scale corporation in more than 70 countries for over 3 years. After anonymizing the identity and the content of these social communications, we are able to quantitatively infer the social networks of 400,000 employees.

Overall, our study aims to answer two questions:

1. How are the topics extracted in such enterprise environments correlate with information workers' performance?

\*IBM T.J. Watson Research Center

†IBM T.J. Watson Research Center

2. How to find valuable topics with potentially high impact on employee performance when the performance data are not available yet?

To address the first question, we collect performance data of a subset of the employees (e.g., the revenue they generate) from the corporate project and personal revenue database. Then, we study the correlation between topics extracted from employees' communication content and their performance. To answer the second question, we assess the value of the topics by their influence on worker performance. We further examine a set of content and social network factors to identify significant factors that can be exploited in enterprise content retrieval and recommendation systems for improving employee productivity [25, 31]. For example, enterprises need to continuously educate their information workers to keep productive. In an environment with dynamically changing technologies, new learning materials have to be constantly produced. However, measuring value of learning materials is one of the top challenges [33]. Utilizing the identified factors, a system may help an employee to find material ranked by value to improve productivity, or help corporate learning executives [11] to monitor the trend of high-value topics and keep training material up-to-date. Because it may be difficult to collect sufficient worker performance data in many such scenarios, the identified significant factors in our study can provide a vehicle for the systems to optimally find valuable topics.

To the best of our knowledge, this is the first study that quantitatively investigates the impact of information content in social networks on worker productivity within large enterprises. This paper offers two unique contributions. First, we present a quantitative study in large-scale on the correlation between topics and worker performance. Second, to find valuable topics without having to access performance data, we identify a set of social network factors including *graph entropy*, *network out-degree*, *number of managers* and *betweenness centrality*. In addition, we find these social network factors complement content features. Specifically, the social network factors outperform content features in filtering out low-value topics, while content features are more effective in finding a few top high-value topics.

The rest of the paper is organized as follows: we first provide a brief discussion of related work. We then describe how we extract topics and measure their values in Section 3. In Section 4, we investigate factors that may be used to find valuable topics. Next, in Section 5 we present a regression model to predict topic values and use the prediction to classify high-value topics and rank topics by value. We discuss practical usage scenarios for these factors in Section 6. Finally, conclusions and

future work are addressed in Section 7.

## 2 Related Work

Our study on finding valuable topics is related to several areas of research including social network analysis, prior work in content analysis and previous approaches that combine content analysis and social network. We review a brief sampling of related work in each of these areas.

**2.1 Social Network Analysis** Various studies of organizational social networks have been conducted to understand the relationship between social network and productivity. Burt [9] has shown the important influence of social network topologies on productivity. In addition, knowledge and expertise accumulated within one's social contacts can have a significant consequence on worker performance [29]. Recently, researchers [36] analyzed a large scale organizational social network and found both network topology and node attributes (e.g., strong ties to appropriate human capital) can be beneficial to worker performance.

Lately, mining various types of social networks is becoming an important research area. Lin *et al.* [25] mined social relationships from people's electronic communication data for expertise search. In addition to mining network structure, researchers have examined node profiles [21] and network dynamics such as node interactions [21] and network evolution [24, 35, 37].

In contrast to previous studies, which mostly concentrate on network node and link attributes, our study focuses on the ample and diverse unstructured content of people's electronic communications (e.g., email and instant messages) in a large scale enterprise. In particular, the goal of our study is to identify computational social network metrics for assessing the value of content.

**2.2 Content Analysis** Topic extraction and tracking [4, 10] is an important area of content analysis. To better model topic distribution and dynamics, a number of projects [7, 15, 31] have used various statistical models based on Probabilistic Latent Semantic Analysis (PLSA) [18] or Latent Dirichlet Allocation (LDA) [8] to extract topics from textual content. Topic extraction in general suffers a same problem: it is difficult to decide which topics are good and useful, and which topics are not. Lack of good sets of unsupervised clustered topics usually is a major factor preventing topic extraction being effectively used in practical applications. Recently, AlSumait *et al.* [5] proposed to measure topic significance by comparing a topic distribution to a "junk distribution". However, it is not immediately clear whether such statistical significance of topics can be related to the value in a particular practical appli-

cation, such as improving information workers' performance.

Many prior topic modeling works focus on a particular type of content such as research paper abstracts [16]. In contrast, email content is usually more diverse and less structured. Previous research on email content performed considerable cleaning and filtering [31]. Compared to prior works, our study focuses on much larger scale, more diverse content in a large enterprise with more than 400,000 employees. In addition, a unique contribution of our work is to associate topics with worker performance so that we can assess the value of topics and identify effective factors for finding valuable topics.

**2.3 Hybrid Approaches** Recently, researchers started to combine social networks and content analysis to improve community discovery. The author-recipient-topic (ART) model [26] extracts topics based on communication between people. Built upon ART, Pathak *et al.* [28] propose a social topic model that incorporates both the link and content information in the social network and use the model to extract community. Recently, Sun *et al.* [34] proposed a tensor-based model to perform content-based social network clustering. On the other hand, social feedback has also been shown effective to measure quality of social media [2]. Another related work is that of Glance *et al.* [14], which provides a system for users to interactively analyze online discussion based on content and social network metrics. These works are concerned with either improving community discovery by using content similarity, or evaluating content by user interactions (e.g., find "hot" topics by user visits). In contrast, our study focuses on revealing the relationship between content and worker productivity, and evaluating metrics for finding content with high-impact on productivity.

### 3 Topics and Their Values

Our study aims to understand how to identify valuable topics in enterprises. In particular, we measure the value of topics by the revenue generated by related workers. We examine factors that may be associated with topic values, such as social network topology properties (e.g., network *betweenness*) and information retrieval measures (e.g., TF-IDF).

**3.1 Dataset** We analyze the information content in an electronic communication social network of nearly 9000 employees over 3 years. The data contain email and instant messaging activities inside a global information technology firm with more than 30 product di-

visions. We collected detailed electronic communication records of 8952 volunteer employees in more than 70 countries. To preserve privacy, the original textual content of an email or text message is not saved. Instead, the content is represented as a vector containing the terms appeared in the text as well as their counts after stop-words removal and stemming. From the 8952 volunteers, we derive a social network of more than 400,000 people within the firm. While efficient algorithms for computations on large networks are beyond the scope of this paper, extracting characteristics of a 400,000 node network can be very time consuming. Therefore, we compared the network characteristics (e.g., social network features described in Section 4.1.2) of the volunteers with the whole population of employees. We found minimal differences [36] so that in this analysis we may speed up the computation of social network factors by constraining it to focus on the sub-network for the 8952 volunteers with complete electronic communication data.

To understand how the content of the communications is related to work performance, we also collected detailed financial performance data of more than 10,000 consultants. These consultants generate revenue by logging "billable hours". Previous study has found that a consultant ability to generate revenue is the most appropriate productivity measure [36]. Therefore, in our study the performance measure we use is the total US dollars a consultant generated from June, 2007 to July, 2008. Combining the financial data with social network data yields a total of 1029 consultants whom we have both network and financial performance data. To protect the privacy of the participants, their identities are replaced with hash identifiers. To match the timing of the content and the performance data, we apply the same time window of the performance data for the content. In addition, to construct a view of the network that reflects the real communications, we eliminate spam and mass email announcements and are left with 2.1 million emails and text messaging chats.

In addition, we looked up the positions of the consultants in the organizational hierarchy of the enterprise. We classified the consultants into two groups: non-managers (69.4% of them) and managers.

**3.2 Topics Extraction** We apply Latent Dirichlet Allocation (LDA) [8], a generative probabilistic model to extract topics. Given a document corpus, LDA models each document  $d$  as a finite mixture over an underlying set of topics, where each topic  $t$  is characterized as a distribution over words. A posterior Dirichlet parameter  $\gamma(d, t)$  can be associated with the document  $d$  and the topic  $t$  to indicate the strength of  $t$  in  $d$ . As

a result, the document  $d$  can be reduced to a vector  $\vec{\gamma}_d = \langle \gamma(d, t_1), \gamma(d, t_2), \dots, \gamma(d, t_T) \rangle$ , where  $T$  is the total number of topics.

To remove extremely rare terms (e.g., misspelled words) and common terms (e.g., the name of the enterprise), we use TF-IDF to get a 40000-term vocabulary. Next, we choose the number of topics  $T$  to be 100, in order to balance the need to cover the diverse content and the need for sufficient revenue observations per topic.

From our observations, the extracted topics are very diverse. There are topics ranging from daily greetings, travel arrangements to business processes and project discussions. Such diversity is expected, since the social network is much larger and the content is not cleaned compared to previous corpus like the Enron email corpus [1].

**3.3 The Value of Topics** Because email and instant text messaging are two main medium for consultants to communicate for their daily work, we hypothesize that the content of emails and chats must be correlated to the consultants' productivity. To measure the value of topics, we relate them to the productivity measure of the corresponding consultants, which is the revenue they generate. We consider a topic to be of high value, if it has high influence on the consultants' productivity.

First, we define a matrix  $\mathbf{S}$  to describe the relationship between consultants and topics, where an element  $s_{ij}$  denotes the degree the  $i$ -th consultant is involved in the  $j$ -th extracted topic. We compute  $s_{ij}$  by aggregating the strengths of the topic in all of the  $i$ -th consultant's emails and chats. Specifically, we have

$$(3.1) \quad s_{ij} = \sum_{d \in D_i} \gamma(d, t_j)$$

where  $D_i$  is the set of content by the  $i$ -th consultant and  $\gamma(d, t_j)$  is a posterior Dirichlet parameter describing the  $j$ -th topic strength in a document  $d$  (see Section 3.2). After that, we normalize  $\mathbf{S}$  by  $s_{ij} = \frac{s_{ij}}{\sum_j s_{ij}}$ , such that  $s_{ij}$  represents the percentage of the  $i$ -th consultant's communication efforts spent on the  $j$ -th topic.

Next, we employ a linear regression model to examine the effect of topics on revenue.

$$(3.2) \quad r_i = q_0 + \sum_{j=1}^T q_j \cdot s_{i,j}, \quad (i = 1, \dots, M)$$

where  $r_i$  is the revenue generated by the  $i$ -th consultant,  $q_0$  is a constant revenue that is independent of topics,  $q_j$  are the coefficients that indicate the effects of topics on revenues,  $M$  is the number of consultants ( $M = 1029$ ). Intuitively,  $q_j$  is the amount of US dollars that a consultant can make in a year in addition to  $q_0$  for

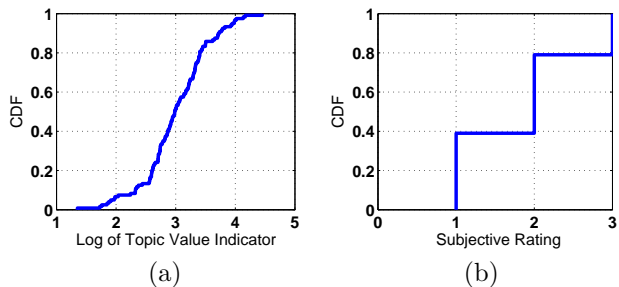


Figure 1: The Cumulative Distribution Function (CDF) of (a) log of topic value indicator, and (b) assessor subjective rating.

every percent of his communication efforts on the  $j$ -th topic.

For the linear regression results, the R-square statistics equals to 0.47 and  $p < 0.01$ , suggesting that topics indeed have statistically significant effects on revenues. There are positive values as well as negative values among the coefficients  $q_j$ . The positive coefficients imply that the corresponding topics have positive effects on revenues, and negative coefficients imply negative impact. Because it is valuable for enterprise to find topics with both positive and negative impact on revenues, we use the magnitude of coefficients  $|q_j|$  as the indicator of topic value.

**3.3.1 Preliminary Subjective Evaluation** To test whether revenue-based topic value analysis is consistent with human assessor perception, we also conduct a preliminary study with human assessors.

We recruit five employees of the enterprise as assessors to evaluate the extracted 100 topics. They are asked to rate the value of the topics to the enterprise on a scale of 1 to 3, where 1 means low value and 3 means high value. To rate a topic, they can utilize their general knowledge of the various aspects of the enterprise. In addition, they are asked to search Internet and the enterprise intranet to find relevant information if needed.

We average the ratings from the five assessors, and then compute the linear correlation coefficient between the average ratings and the log of the topic value indicator. The coefficient is 0.48, suggesting a strong correlation between the revenue-based indicator and the subjective evaluation.

We further compare the distributions of topic value indicator and human subjective rating. Figure 1 shows their Cumulative Distribution Functions (CDF). Based on the CDFs, there are 40% low-value topics by subjective evaluation, and the corresponding topic value

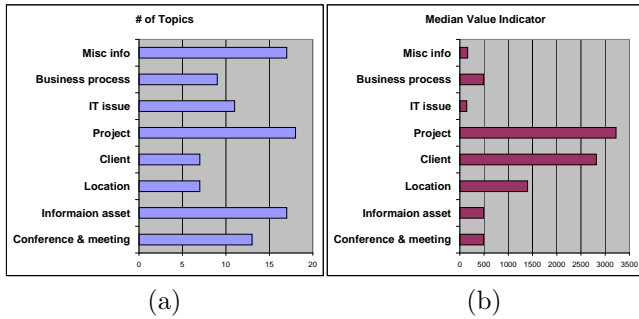


Figure 2: The topic categories: (a) the number of topics in each category, and (b) the median topic value indicator of each category.

indicator threshold is approximately  $|q| < 500$ . That means each percent of a consultant’s communication effort spent on these low value topics can only make less than \$500 a year. On the other hand, about 20% of topics have high value ( $|q| > 2000$ ) with each percent of communication leads to more than \$2000 a year.

In addition, to obtain a more intuitive overview of the topics, we ask the assessors to group the extracted topics into eight categories. The number of the topics and the median value indicator in each of the category is shown in Figure 2. Among the eight categories, the topics in “Location” are concerned with geographical locations world wide that are of interests to this global enterprise. The topics in “Information asset” are on various external and internal information portals frequently used by the consultants, such as Google or the internal consultant knowledge portal. Figure 2(b) illustrates the value distribution over the topic categories, which shows the chosen topic value indicator agrees with our intuition. For example, the topics with high value indicator mostly focus on information of concrete projects and clients, while the topics with low value indicator tend to be about common routines such as travel arrangements. In Section 4, we will more rigorously examine a set of factors that can be related to  $|q|$  and thus can be used to find potentially valuable topics when the worker performance data are not available yet.

#### 4 What Do Topic Values Depend On

In large enterprises, it is highly desirable to discover high-impact topics to facilitate effective content search. To this end, our analysis tries to identify factors that can be used to rank topics by value. In particular, we focus on two types of factors. First, we examine how well common content factors may determine topic values. Second, we study the effects of a set of social network factors, including *graph entropy*, network *in-degree*, *out-degree*, *betweenness centrality*, *network constraint* and

*number of managers*.

Our hypothesis is that the network properties of consultants involved in a topic may have effects on the topic value. For example, we expect social networks of high-impact topics to have strong clusters because strong relationships among topic participants can be conducive to productivity in the topic-related projects. In contrast, networks of low-impact topics may be scattered since the topics may be general and can involve a diverse population in the enterprise. Figure 3 illustrates such an example. The two displayed topic networks<sup>1</sup> have similar TF-IDF but different topic value indicators as computed in Section 3.3. The two topics have only 0.1% difference in their TF-IDF values. In contrast, the topic value indicator (see Section 3.3) of the topic in Figure 3(a) is 205 times higher than the topic in Figure 3(b). The high-impact topic (Figure 3a) is about concrete projects and its network is much more clustered. On the other hand, the low-impact topic (Figure 3b) is about general project management process and its network appears scattered.

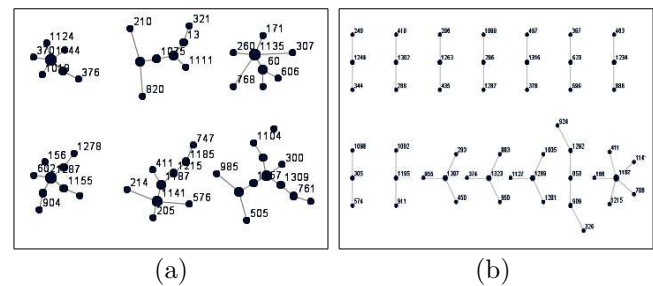


Figure 3: Two topic networks that have similar TF-IDF but different network properties. Only the top-50 connected nodes are shown in each of the network.

### 4.1 Factor Definitions

**4.1.1 Content Factor: TF-IDF** In this paper, we choose to first focus on TF-IDF for two practical reasons. First, TF-IDF and its variants are closely related to measures of significance in probabilistic models and information theory [3]. Moreover, they are the most common content factors used in search and recommendation systems. Thus, insights derived on TF-IDF can be generalized and widely applicable. Second, in global enterprises there could exist vastly different (e.g., non-probabilistic or probabilistic) topic representations on diverse information. TF-IDF is desirable because it may be used to characterize the topics represented by different topic models.

<sup>1</sup>The extraction of topic networks is described in Section 4.1.2

Since LDA models a topic  $t$  as a distribution over words, we compute the TF-IDF metric of a topic by aggregating each word's TF-IDF measure weighted by the word probability. Formally, the metric is computed by  $TFIDF(t) = \sum_{w \in W_t} P_t(w) \cdot tfidf(w)$ , where  $W_t$  is the set of words for topic  $t$ ,  $P_t(w)$  is the probability of the word  $w$  under topic  $t$ , and  $tfidf(w)$  is the TF-IDF measure of the word  $w$ .

**4.1.2 Social Network Metrics** To compute social network metrics of a topic, we first extract the topic sub-network for consultants involved in the topic, using the consultant-topic relationship matrix  $\mathbf{S}$  defined in Section 3.3. Specifically, for the sub-network of the  $j$ -th topic, we include consultants whose corresponding topic strength is larger than the median value of the  $j$ -th topic strength. Here the *median* measure is used instead of *mean* because it is a statistically robust measure, especially when the distribution of the values is skewed [19]. Formally, the node set of the  $j$ -th topic network is  $V_j = \{i \mid s_{ij} > median_j\}$ , where  $median_j$  is the median value of the  $j$ -th column in  $\mathbf{S}$ . An edge between two nodes in the network indicates that there is communication on this topic between the two people. The edge is also assigned a weight based on the amount of communication [36].

**Graph entropy.** We use a graph entropy metric [22, 30] to measure how well the consultants involved in a topic are clustered. First, we partition the topic network  $G = \langle V, E \rangle$  into  $K$  sub-graphs, using a spectral factorization method [17]. For a node in the topic network  $v \in V$ , let  $p_k$  denote the probability that  $v$  belongs to the  $k$ -th partition. Next, we use graph entropy to measure the node distribution  $\mathbb{P}$  by:

$$(4.3) \quad H(G, \mathbb{P}) = \sum_{k=1}^K p_k \cdot \log \frac{1}{p_k}$$

Intuitively, large values of  $H$  mean that all the partitions have similar sizes. In contrast, small values of  $H$  indicate that nodes are concentrated in a few partitions, i.e., well clustered. In our study, we choose the number of graph partition  $K = 100$ . The impact of the value of  $K$  on graph entropy will be investigated in Section 4.3.2.

**Network in-degree and out-degree.** The two metrics measure the size of communication between a topic network and the outside network. Topic network in-degree is measured as the number of email sent into the topic network while out-degree is measured as the number of emails sent out from the topic network. Because a consultant is involved in multiple topic networks, we weight the in-degree and out-degree of the

$j$ -th topic network by the degree he is involved in:

$$(4.4) \quad inDegree_j = \sum_{i \in V_j} s_{ij} \cdot I(i)$$

$$(4.5) \quad outDegree_j = \sum_{i \in V_j} s_{ij} \cdot O(i)$$

where  $V_j$  is the node set for the  $j$ -th topic network,  $I(i)$  is the number of emails coming from the outside network into the  $i$ -th consultant,  $O(i)$  is the number of emails coming out to the outside network, and  $s_{ij}$  represents how strongly the  $i$ -th consultant is involved in the  $j$ -th topic (see Section 3.3).

**Betweenness centrality.** For an individual node  $i$  in a social network, the betweenness centrality  $b(i)$  measures the relative importance of the node in the information flow within the network [12]. Specifically,  $b(i)$  is defined as the probability that node  $i$  will fall on the shortest path between any two other individuals in a network:

$$(4.6) \quad b(i) = \sum_{l < m} [\bar{\delta}_{lm}(i) / \bar{\delta}_{lm}]$$

where  $\bar{\delta}_{lm}(i)$  is the number of shortest geodesic paths from  $l$  to  $m$  that pass through a node  $i$ , and  $\bar{\delta}_{lm}$  is the total number of shortest geodesic paths from  $l$  to  $m$ . For the topic network of the  $j$ -th topic, we compute the overall betweenness centrality by aggregating the values of the individual nodes within the network as:

$$(4.7) \quad B(j) = \sum_{i \in V_j} s_{ij} \cdot b(i)$$

where  $s_{ij}$  represents how strongly the node  $i$  (i.e., the  $i$ -th consultant) is involved in the  $j$ -th topic, and  $V_j$  is the node set for the  $j$ -th topic.

**Network constraint.** We also use a network constraint metric to measure the degree to which nodes have diverse contacts (i.e., neighbors) in a network. For an individual node  $i$ , the contacts of  $i$  are considered diverse if they are not connected to each other [9]. By having diverse contacts, an individual may tap into diverse and novel information sources and thus improve his productivity. Formally, the network constraint of  $i$  is defined as:

$$(4.8) \quad c(i) = \sum_l (\rho_{il} + \sum_{m \neq i, l} \rho_{im} \cdot \rho_{ml})^2$$

where  $\rho_{il}$  is the proportion of node  $i$ 's network efforts invested in communicating with node  $l$ . Large values of  $c(i)$  indicate that the neighbors of node  $i$  are highly connected to each other. Thus node  $i$ 's contacts are considered concentrated instead of diverse. For the

Factors	TF-IDF	Graph entropy	In-degree	Out-degree	Betweenness	Constraint	# of Manager
$Corr_t$	0.17	-0.21	-0.10	-0.21	-0.19	0.10	0.22
$Corr_h$	0.11	-0.21	-0.10	-0.17	-0.08	-0.03	0.16

Table 1: Linear correlation coefficients (1) between the factors and the topic value indicator ( $Corr_t$ ), and (2) between the factors and the human subjective rating ( $Corr_h$ ).

topic network of the  $j$ -th topic, we also compute the overall network constraint by aggregating the values of the individual nodes as  $C(j) = \sum_{i \in V_j} s_{ij} \cdot c(i)$ .

**Number of managers.** Finally, we compute the number of managers in a topic network, to examine the effect of organizational leadership on topic values.

**4.2 Analysis of Correlation and Variance** Our analysis aims to investigate how well the factors can be applied to ranking topics by value, as it is an important step in content search and recommendation applications.

To this end, we first need to examine the correlation between the factors and the topic value indicator  $|q|$  (see Section 3.3). Table 1 shows the computed linear correlation coefficients  $Corr_t$ . The correlation coefficients show that high-impact topics have high *TF-IDF* and low *graph entropy*, which agree with the common practice in the fields of information retrieval and information theory. In addition, the positive coefficient of *number of managers* implies that high-impact topics tend to involve more managers, suggesting that management support is important for greater value. The coefficient of *network constraint* is also positive. That means topic networks with high constraint values (i.e., more concentrated networks) are more valuable. In contrast, large values of *in-degree*, *out-degree* and *betweenness centrality* indicate low-impact topics. One explanation is that a topic is related to many other topics if its network has a large amount of communication with the outside network. Therefore, such a topic is very likely to be a general topic that may have little predictable effect on revenue generation. Finally, Table 1 also shows that all social network factors are better correlated with  $|q|$  than TF-IDF, except for network in-degree and constraint.

We then conduct one-way analysis of variance (ANOVA) to examine the significance of these factors more rigorously. The analysis is performed for discrete values of factors obtained by log binning and rounding of the original values to limit the number of values for factors. For all factors except in-degree and network constraint, ANOVA analysis results in  $p$ -values smaller than 0.05. This result indicates that all factors are statistically significant except network in-degree and constraint. For network constraint, this result corroborates

the findings in early work [36], which indicated that network constraint is effective in characterizing the social capital of an individual person but not a whole sub-network such as a project-level network. Based on such results, we will leave out the network in-degree and constraint in the rest of the analysis.

#### 4.2.1 Analysis of Correlation and Variance Based on Subjective Rating

Additionally, we examine the factors' correlation and variance using human assessors' subjective topic rating obtained in Section 3.3.1. The linear correlation coefficients  $Corr_h$  between the factors and the human rating are shown in Table 1, where a similar trend as in  $Corr_t$  can be observed. Again, we can see many social network factors (e.g., graph entropy, network out-degree, and number of managers) are better than TF-IDF.

We also examine the statistical significance of these factors by ANOVA analysis. The  $p$ -value of graph entropy is smaller than 0.01. For all other factors except network in-degree and constraint, the  $p$ -values are smaller than 0.1. The subjective-rating-based results further confirm the finding that both TF-IDF and the social network factors (except network in-degree and constraint) are statistically significant for ranking topics by value.

**4.3 Rank Topics** Based on the correlation coefficients, a value-based ranking of topics may be approximated using these factors, among which graph entropy, out-degree and betweenness need to be negated.

To examine how well a factor  $x$  can rank topics, we define a ranking quality measure as the percentage of total topic value that is associated with the top- $N$  topics in the ranked list. Formally, it is computed by

$$(4.9) \quad m_x(N) = \frac{\sum_{j=1}^N |q_j|}{\sum_{j=1}^T |q_j|}$$

where  $|q_j|$  is the topic value indicator of the  $j$ -th topic and  $T$  is the total number of topics. A large value of  $m_x(N)$  indicates that a large portion of valuable topics are ranked in top- $N$ . Thus, the factor  $x$  is effective in ranking topics.

Figure 4 shows the ranking quality curves of the factors. We use TF-IDF as the baseline for comparison

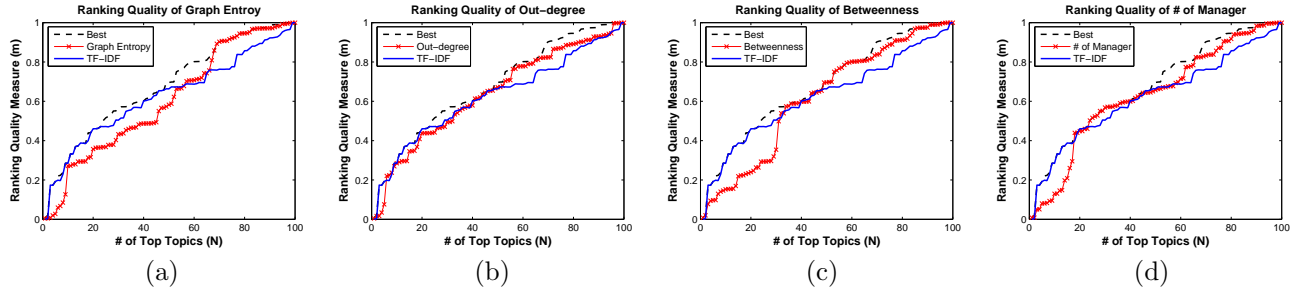


Figure 4: The revenue-based ranking quality curves of the factors.

Range of $N$	1 ~ 20	21 ~ 40	41 ~ 60	61 ~ 80	81 ~ 100
Best factor	TF-IDF	# of Managers	Betweenness	Graph entropy	Graph entropy
Avg improvement over TF-IDF	0	6.2%	6.4%	11.7%	6.0%

Table 2: The factors with the best ranking quality in different ranges of  $N$ .

in Figure 4(a-d). In addition, we display a reference “best” ranking quality curve, computed by taking the highest value of all the factors’ ranking quality measure value at a given  $N$ :

$$(4.10) \quad best(N) = \max_{x \in F} (m_x(N))$$

where  $F$  is the set of five factors:  $\{TF-IDF, graph\ entropy, out-degree, betweenness, number\ of\ managers\}$ .

We observe that TF-IDF slightly outperforms the social network factors when the value of  $N$  is small (e.g.,  $N < 20$ ), which implies that TF-IDF is better at finding a small number of high-value topics. In contrast, the social network factors such as graph entropy significantly work better than TF-IDF when  $N$  is larger (e.g.,  $N > 60$ ). That means the social network factors are more effective in retaining as many high-value topics as possible and filtering out low-value topics. As a result, the social network factors are more useful in application scenarios such as user modeling [32], where it is more important to obtain a broad set of valuable topics by filtering out the “noise”. Furthermore, Figure 4 shows that these factors can complement each other. Specifically, TF-IDF performs the best when  $N$  is small (e.g.,  $N < 20$ ). Next, the most effective factors in the middle range are number of managers (e.g.,  $20 < N < 40$ ) and betweenness (e.g.,  $40 < N < 60$ ). Finally, graph entropy works the best in filtering the “noise” (e.g.,  $N > 60$ ). Table 2 summarizes the best ranking factors in different ranges of  $N$ . We believe such findings can help guide the design and development of effective composite metrics for finding valuable topics.

**4.3.1 Ranking Quality Based on Subjective Rating** To verify the factors’ ranking quality against human evaluation, we also define a subjective quality

measure using assessors’ subjective ratings obtained in our preliminary subjective topic evaluation:

$$(4.11) \quad u(N) = \frac{\sum_{j=1}^N rating_j}{\sum_{j=1}^T rating_j}$$

where  $rating_j$  is the human rating of the  $j$ -th topic.

Figure 5 shows the subjective ranking quality curves for the factors. We again observe a similar trend that TF-IDF slightly outperforms the social network factors for small  $N$ , and the social network factors work better than TF-IDF when  $N$  is larger. In particular, graph entropy appears to be the best factor for filtering out low-value topics, which is consistent with the findings using revenue-based ranking quality measure (Section 4.3). Specifically, for  $N > 40$ , the average improvement of graph entropy over TF-IDF is 10.7%. The consistency between the revenue-based results and subjective-rating-based results demonstrates that the findings obtained from our revenue-based study can be generalized.

### 4.3.2 Parameter Sensitivity of Graph Entropy

Among the social network factors, graph entropy uses an important parameter  $K$ , which is the number of graph partition. Here we evaluate the impact of the parameter values on the quality of topic ranking results. The default value of  $K$  is 100. We vary  $K$  from 20 to 200. Then, we compute the corresponding topic ranking quality measure  $m(N)$  defined in Section 4.3. To examine the impact of  $K$  in different ranking scenarios such as filtering out low-value topics (i.e.,  $N$  is large) or selecting a few top-value topics (i.e.,  $N$  is small), we divide the values of  $N$  into five ranges: 1 ~ 20, 21 ~ 40, 41 ~ 60, 61 ~ 80, 81 ~ 100. For each range, we compute the average ranking quality measure. Figure 6 shows

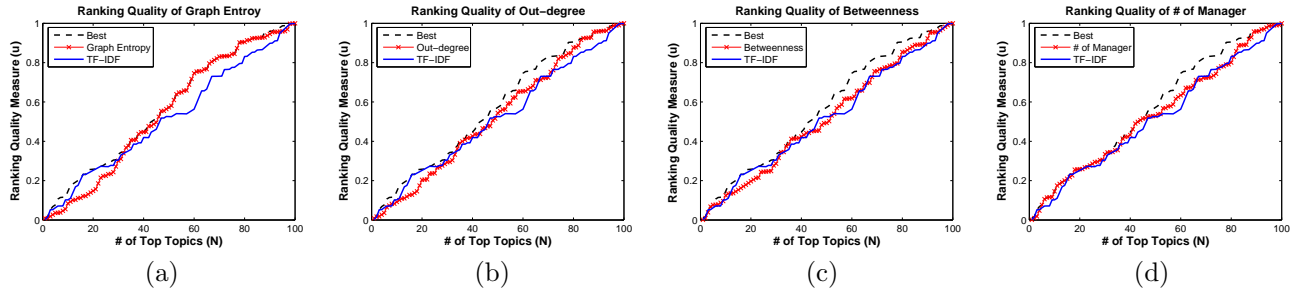


Figure 5: The assessor-subjective-rating-based ranking quality curves of the factors.

the impact of  $K$  on the five average ranking quality measure.

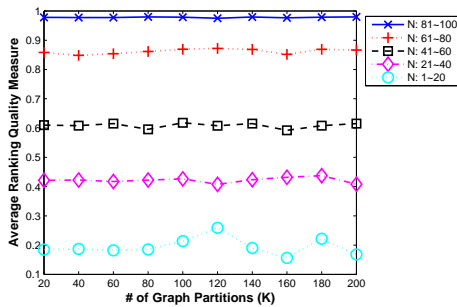


Figure 6: The relationship between average topic ranking quality measures and graph partition number  $K$ .

We observe that the average ranking quality has larger variation across different  $K$  values when  $N$  is small, especially when  $1 < N < 20$ . That finding indicates that the choice of  $K$  value may have a larger impact on ranking quality when the scenario is to select a few top- $N$  topics. As the ranking goal becomes to filter out a few low-value topics and retain more valuable topics (i.e.,  $N$  becomes larger), the impact of  $K$  gets much smaller to a point that is negligible. Therefore, the parameter  $K$  has little impact on the performance of graph entropy, especially when it is the best performing factor (e.g., when filtering out low-value topics).

**4.3.3 Concavity of Number of Managers** Our analysis shows that the factor *number of managers* works the best in the middle range (see Figure 4d and Table 2). This fact corroborates early results [36], which observed more managers in a project are associated with greater revenue to a point, after which there are negative returns to increased number of managers. To verify such concave relationship in topic networks, we plot number of managers and topic value indicator in Figure 7. The confirmed concavity as shown may be explained that a topic network involving too many managers is likely

to be a high-level topic with few concrete results. It also suggests the need to search for an optimal number of managers in the design of composite topic ranking metrics.

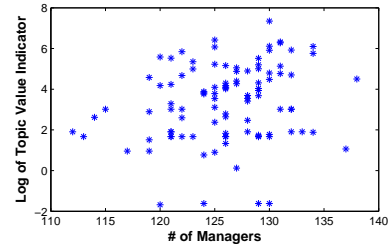


Figure 7: The concave relationship between topic value indicator and number of managers.

## 5 Predict Topic Value

In practice, many applications may not be able to have access to sufficient worker performance data. Therefore, it is highly desirable to train topic value prediction models based on limited performance data. The findings in our study can help guide the design and development of effective topic value prediction schemes using content and social network features. In this section, we present our exploration of combining the identified five significant factors to predict topic values. We then evaluate the prediction effectiveness by using the predicted values to classify and rank topics.

**5.1 Predict Value By Regression** Regression is a classic statistical problem which tries to determine the relationship between two random variables  $x = (x_1, x_2, \dots, x_F)$  and  $y$ . In our scenario, independent variable  $x$  can be the vector of the five factors identified in Section 4:  $x = (TF-IDF, graphentropy, out-degree, betweenness, numberofmanagers)$ , and dependent variable  $y$  can be the log of topic value indicator  $\log(|q|)$ .

Given the five factors, we first use a linear regression

model to predict topic value as:  $y = y_0 + \sum_{j=1}^F a_j \cdot x_j$ . However, this linear regression model gives R-square statistics equal to 0.08, indicating that this problem is highly non-linear.

Due to the strong non-linearity, we choose to use support vector regression (SVR). In SVR, the input  $x$  is first mapped onto a high dimensional feature space using a nonlinear mapping, and then a linear model is constructed in this feature space. SVR uses a so called  $\varepsilon$ -insensitive loss function:

$$(5.12) \quad L_\varepsilon = \begin{cases} 0 & \text{if } |y - f_\omega(x)| < \varepsilon \\ |y - f_\omega(x)| & \text{otherwise} \end{cases}$$

where  $\varepsilon$  is a predefined deviation threshold, and  $f_\omega(x)$  is the regression function to predict  $y$  which has a parameter  $\omega$ . Then the regression is formalized as the following minimization problem:

$$(5.13) \quad \begin{aligned} & \text{minimize} && \frac{1}{2} \|\omega\|^2 + A \sum_{i=1}^l (\xi_i + \xi_i^*) \\ & \text{subject to} && \begin{cases} y_i - f_\omega(x) \leq \varepsilon + \xi_i \\ f_\omega(x) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned}$$

where  $A$  is a constant, and  $\xi_i, \xi_i^*$  ( $i = 1, \dots, l$ ) are slack variables introduced for the optimization to measure the deviation of training samples outside  $\varepsilon$  insensitive zone.

In our study, we collect ground-truth training data for SVR by estimating the value indicators of topics extracted from the 2.1 million consultants' electronic communication data (Section 3.3). To obtain sufficient observations for regression, we extract multiple sets of topics with different sizes ( $size = 50, 60, 70, \dots, 150$ ), which describe the content of the emails at different granularities. As a result, we obtain 1100 samples of topic value indicator, as well as corresponding content and social network features.

We use the support vector regression implementation in SVM-Light [20]. In our support vector regression experiments, we use sigmoid kernel function  $\tanh(\mathbf{s} \cdot x_i^T \cdot x_j + \mathbf{c})$  with parameter  $\mathbf{s} = 0.2$ . Other parameters such as  $\varepsilon$  are set to default. Next, we evaluate the effectiveness of the regression by using the prediction to classify and rank topics by value.

**5.2 Classify Topics Using Predicted Value** The predicted topic value can be used to classify high value topics. In practice, the threshold for high value can vary according to application scenarios. For example, an application may just need to filter out low-value topics, or find topic "above average", or a few top value topics. In our experiments, we use three criteria for "high-value" topics illustrated by the CDFs in Section

3.3.1: (1) C500: more than \$500 a year for every 1% of a consultant's communication; (2) C1000: more than \$1000 a year for every 1%; and (3) C2000: more than \$2000 a year for every 1%.

To classify "high-value" topics, we test whether the prediction is larger than a threshold  $TH$ . The precision and recall for a particular  $TH$  can be defined as:

$$(5.14) \quad \text{precision}_{TH} = \frac{|E_H \cap E_{TH}|}{|E_{TH}|}$$

$$(5.15) \quad \text{recall}_{TH} = \frac{|E_H \cap E_{TH}|}{|E_H|}$$

where  $E_H$  is the ground truth set of "high-value" topics whose value indicators satisfy our preselected criteria, and  $E_{TH}$  is the set whose predictions are larger than  $TH$ . A precision-recall curve can be derived by varying the threshold  $TH$ .

We randomly partition the ground-truth topic value data into 3 parts and use three-fold cross validation to evaluate the average classification performance based on support vector regressions. The precision-recall curves of the classifications are shown in Figure 8. As a baseline for comparison, we also show the precision-recall curve of the classification results using *TF-IDF* feature only for regression and criteria C2000. We

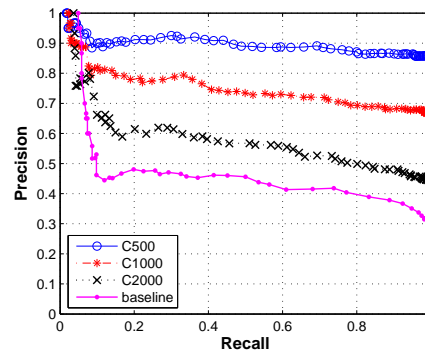


Figure 8: The precision-recall curves for high-value topic classification.

observed significant improvement on the classification performance by leveraging social network features. For example, for the same criteria C2000 at  $recall = 0.5$ , the precision by combining social network features is 58%, outperforming the precision (44%) by using *TF-IDF* alone.

**5.3 Rank Topics Using Predicted Value** The topic value prediction based on combined features can also be applied to topic ranking by value. To examine how much the value prediction can improve topic

ranking, we use the same revenue-based ranking quality measure defined in Equation 4.9. Figure 9 shows the ranking quality measure of the scheme that uses predicted value based on combined features (denoted by SVR-Rank), compared with the baseline TF-IDF based method (denoted by TF-IDF).

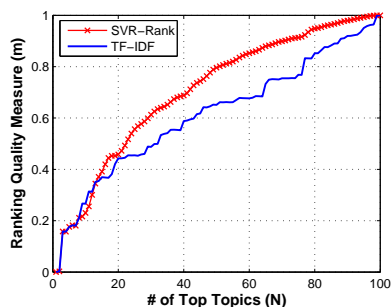


Figure 9: The ranking quality of a ranking scheme that uses predicted value to rank topics .

We observe that the topic ranking quality can be significantly improved by SVR-Rank. Because social network features can better filter out low-value topics, more high-value topics can be retained in the top- $N$  topics by SVR-Rank. Figure 9 shows that SVR-Rank starts to have more high-value topics in top- $N$  than TF-IDF since  $N = 20$ . Specifically, compared to TF-IDF, the average improvement of ranking quality measure for SVR-Rank is 15.6%. In addition to the revenue-based ranking quality measure, we examine SVR-Rank method using the assessor subjective topic rating obtained in Section 3.3.1. The results are consistent, where the average improvement of ranking quality measure for SVR-Rank method is 6.2%.

## 6 Discussion

Our study aims at identifying a set of significant factors that can be used to find valuable topics. In this section, we briefly discuss practical scenarios where our findings can be utilized.

Materials that improve information workers' productivity nowadays are mostly from informal sources (e.g., Web, colleagues) rather than traditional classrooms [11]. Therefore, enterprises are facing increasing need for systems that collect these materials and measure their value. For such systems, it is desirable that they can not only rank information by relevance to employee queries but also by value to enterprises. However, these systems often only have access to content or content plus employee networks, but not sufficient employee performance data. In such scenarios, these systems can utilize our findings to find relevant and valuable content

in three steps: (1) extract latent topics from the content; (2) rank the latent topics using the identified factors in our study; and (3) incorporate the ranking scores of the content's latent topics into the overall content ranking scores. In general, both content and social network factors should be used to optimally find valuable content. Nevertheless, certain applications may restrict the size of returned top content (e.g., mobile applications with limited screen real estate). In such scenarios, applications may just use content factors to select content corresponding to a few top topics.

## 7 Conclusion and Future Work

We present an analysis of information content in one of the largest organizational electronic communication networks collected and combined with detailed performance data. We focus on evaluating topics by worker performance and understanding which factors influence the topic values. Our study uncovers two key results. First, we observe that topics are significantly correlated with worker performance. We derive a performance-based topic value indicator and verify that the indicator concurs with human assessors' judgments. Second, to rank topics by value, we also identify a set of significant factors, including content features and several social network factors. We demonstrate that content features and the social network factors are complementary to each other.

In practice, we understand many applications may not be able to access all three types of major data in our study: content, network, and performance data. Many scenarios may only have access to content or content plus network. Therefore, the insights provided by this paper offer a vehicle for researchers or developers to optimally find valuable topics without having to access performance data. In particular, we present a regression model that combine the factors to predict topic value, and demonstrate that it can already effectively classify high-value topics and rank topics by value.

We are planning to adopt scalable techniques for network factors extraction (e.g., [27]), and incorporate the identified significant factors to improve enterprise expert finding and content recommendation systems. In addition, we intend to study topic temporal evolution and dissemination inside the network. Such study can increase our understanding of how information content in social networks is related to performance, and lead to improved employee learning systems.

**Acknowledgement** This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053.

## References

- [1] Enron email corpus. <http://www.cs.cmu.edu/~enron>.
- [2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proc. of WSDM*, pages 183–194, 2008.
- [3] A. Aizawa. An information-theoretic perspective of tfidf measures. *Information Processing and Management*, 39(1):45–65, 2003.
- [4] J. Allan. *Topic detection and tracking: event-based information organization*. Kluwer, 2002.
- [5] L. AlSumait, D. Barbara, J. Gentle, and C. Domeniconi. Topic significance ranking of lda generative models. In *Proc. of ECML/PKDD*, pages 67–82, 2009.
- [6] S. Aral, E. Brynjolfsson, and M. V. Alstynne. Information, technology and information worker productivity: Task level evidence. In *Proc. of the 27th Annual International Conference on Information Systems*, 2006.
- [7] D. Blei and J. Lafferty. Dynamic topic models. In *Proc. of ICML*, pages 113–120, 2006.
- [8] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [9] R. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
- [10] C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. *IEEE Trans. Knowl. Data Eng.*, 16(8):949–964, 2004.
- [11] J. Cross. *Informal learning: rediscovering the natural pathways that inspire innovation and performance*. John Wiley and Sons, 2006.
- [12] L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.
- [13] Gartner. The knowledge worker investment paradox, July 2002.
- [14] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proc. of KDD*, pages 419–428, 2005.
- [15] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou. Topic evolution in a stream of documents. In *SDM*, pages 859–872, 2009.
- [16] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of National Academy of Science*, pages 5228–5235, 2004.
- [17] J. P. Hespanha. An efficient matlab algorithm for graph partitioning. Technical report, University of California, Oct. 2004.
- [18] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
- [19] P. Huber. *Robust Statistics*. Wiley, 1981.
- [20] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. B. Schlkopf and C. Burges and A. Smola. (ed.). The MIT Press, 1999.
- [21] T. Karagiannis and M. Vojnovic. Behavioral profiles for advanced email features. In *Proc. of WWW*, pages 711–720, 2009.
- [22] J. Korner. Bounds and information theory. *SIAM Journal on Algorithms and Discrete Mathematics*, 7:560–570, 1986.
- [23] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *Proc. of KDD*, pages 467–476, 2009.
- [24] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. of KDD*, pages 462–470, 2008.
- [25] C. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges. Smallblue: People mining for expertise search. *IEEE Multimedia Magazine*, 15(1):78–84, 2008.
- [26] A. McCallum, C. Andres, and X. Wang. Topic and role discovery in social networks. In *Proc. of IJCAI*, pages 786–791, 2005.
- [27] S. Papadimitriou and J. Sun. Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In *ICDM*, pages 512–521, 2008.
- [28] N. Pathak, C. DeLong, K. Erickson, and A. Banerjee. Social topic models for community extraction. In *Proc. of ACM Workshop on Social Network Mining and Analysis*, 2008.
- [29] S. Rodan and D. Galunic. More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, 25:541–556, 2004.
- [30] J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proc. of Intl. Workshop on Link Discovery (LinkKDD)*, pages 74–81, 2005.
- [31] X. Song, C. Lin, B. Tseng, and M. Sun. Modeling and predicting personal information dissemination behavior. In *Proc. of KDD*, pages 479–488, 2005.
- [32] X. Song, B. Tseng, C. Lin, and M.-T. Sun. Expertisenet: Relational and evolutionary expert modeling. In *Proc. of User Modeling*, pages 99–108, 2005.
- [33] B. Sugrue and D. Lynch. Profiling a new breed of learning executive. *Training and Development Magazine*, pages 51–56, Feb 2006.
- [34] J. Sun, S. Papadimitriou, C.-Y. Lin, N. Cao, S. Liu, and W. Qian. Multivis: Content-based social network exploration through multi-way visual analysis. In *SDM*, pages 1064–1075, 2009.
- [35] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proc. of KDD*, pages 677–685, 2008.
- [36] L. Wu, C. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.
- [37] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. A bayesian approach toward finding communities and their evolutions in dynamic social networks. In *SDM*, pages 990–1001, 2009.