Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Socio-temporal analysis of conversations in intra-organizational blogs

Nachiketa Sahoo
Tepper School of Business and iLab, Carnegie Mellon University

Ramayya Krishnan
Heinz College and iLab, Carnegie Mellon University

Christos Faloutsos
Computer Science Department and iLab, Carnegie Mellon University

Blogs have been popular on the Internet for a number of years and are becoming increasingly popular within organizations as well. The analysis of blog posts is a useful way to understand the nature of expertise within the firm. In this paper we are interested in understanding the *topics of conversations* that evolve through blog posts and replies. While keywords within blog posts can be used to characterize the topic being discussed, their timestamps permit one to monitor how the intensity of the topic has changed over time, and the author information permit the social nature of the topics to be monitored. Based on this observation we define *topics of conversation* using *keywords*, *author & recipient*, and *timestamps* of the blog posts & replies. We use tensors to capture these multiple modes of the blog data. With this rich representation of the multi-modal data we identify significant topics and key entities in those topics. This is done by generalizing the idea of *significance by association*, that has been extensively used in social network analysis, to multi-modal network data. We show that such significance in blogs can be calculated by tensor factorization. This method is illustrated by applying it to a dataset extracted from the blog network within a large globally distributed IT services firm. We discuss implications of this work for monitoring opinion developments and detecting opinion leaders within the organization. We find that the central bloggers identified by tensor factorization are more "on topic" with respect to the topic of discussion in their responses than the central bloggers identified by HITS algorithm. Finally a tensor factorization based clustering method is designed to discover communities from the online social conversations. The effectiveness of this method is measured with the help of author provided community labels on the conversations.

Key words: Blogs, Tensors, Online Social Networks, Dynamic Network Analysis

1. Introduction

Increasingly organizations are creating private blogs to promote peer-to-peer communication and collaboration among employees. With the advent of Enterprise 2.0, employees at the grass root level who used to be end users of information, are playing a larger part in generating and disseminating valuable information using blogs and user forums. The activities in the blog network permit monitoring of employee opinion, identification of leaders or experts in different topics and enable an organization to develop a map of the expertise that is available within the organization. The automated analysis of large scale blog data to gather organizational intelligence is a topic of considerable interest to managers and decision makers.

However, blogosphere is an environment of complex interaction between entities of different types, e.g., the blog authors, responders, keywords occurring in the blog posts etc. Each of these entities can potentially provide information on a different aspect of any blog post. Datasets that are described by different types of entities are known as *multi-modal datasets*. By analyzing the relations between such entities in the collection one can potentially get insights into the broader conversations occurring in the blogosphere. The current work develops methods to identify such broader conversation patterns in online social networks in general and blogs in particular.

Objective
The objective of this paper is to detect significant topics in conversations occurring in an online social network taking into account the relation between the different types of entities such as authors, responders, keywords, and time periods.

The contributions of this work are:

1. Representation of the multi-modal data that constitute social conversation as a tensor. This overcomes the shortcomings of the matrix based framework for analyzing multi-modal interactions with more than two types of entities
2. An interpretation of tensor factorization that defines significance of entities in such a tensor. This helps us identify the most significant authors and topics in the social network
3. Evaluation of tensor factorization approaches for community discovery and tracking in blogosphere

2. Literature Review

The Knowledge Management community has developed quite a few Expert Finder systems to identify experts within the information system of an organization (Yimam 2000, Becerra-Fernandez 2006). The dominant theme of the Expert Finder systems is indexing expertise in a database and providing an interface to the managers to query it. However, it has grown to include user referrals (Kautz and Selman 1998) and personalized agents to identify experts (Vivacqua 1999). The input to such systems are user responses to expertise surveys, user posts in discussion groups(Krulwich et al. 1996), and the technical documents users produce (Streeter and Lochbaum 1988a,b). In recent years the Text REtrieval Conference community has taken a renewed interest in the task of finding experts (Craswell et al. 2005, Soboroff et al. 2006). The task which is part of the TREC Enterprise Track uses a dataset consisting of corporate Intranet pages, emails, and document repositories. Several teams have proposed information retrieval approaches for identifying experts in predefined topics. This is also known as query specific expert search. In addition to the innovations in language modeling techniques, the key piece of information used in several of the proposed approaches is the email messages, replies, and discussion threads with which individuals are associated. This suggests that the extent to which an individual is involved in the social process of knowledge sharing is a useful indicator of her expertise.

On the other hand sociologists have been interested in identifying people with higher status, prestige, and influence in intra-organizational social networks (Brass 1992, Bonacich 1987, Bonacich and Lloyd 2001). They have proposed a number of centrality measures such as eigenvector centrality, betweenness, closeness etc. Most such measures are computed on networks with ties of one type that varies in intensity for different pairs of nodes, e.g., number of times a person seeks advice from another, strength of friendship between two actors, etc. Only a few have looked at differing content of the ties in the network (Burt and Schøtt 1985).

Due to the adoption of online social media at the enterprise level, employees are participating in creating and disseminating a wide variety of content. The topics include topics of professional interest such as organizational practice, technical developments and social topics such as sports and politics. Therefore, to identify expertise present in the organization and experts in those topics one needs to look into the content of the social exchange taking place in the online social network within the organization. This is the domain of text data mining literature.

The text data mining community has made considerable progress over last decade in analyzing and tracking topics in text posts. The Topic Detection and Tracking initiative (Allan et al. 1998, Yang et al. 1998, Franz et al. 2001, Doddington et al. 2000), extension of the Latent Dirichlet Allocation (LDA) for temporal topic modelling (Wang and McCallum 2006), construction of patterns of statistically significant word occurrences (Swan and Jensen 2000) in news streams are important examples of work in this area. A second set of works have ventured beyond using just the word occurrences in text documents to include the author and the recipient of the documents. The author-recipient-topic model (ART) extends LDA to incorporate sender and recipient information for modeling email topics (McCallum et al. 2004). The content-community-time model is a two step probabilistic clustering approach for identifying time bound news topics in blogosphere (Qamra et al. 2006). The modeling of co-voting records by various senators is yet another example of socio-textual analysis for detecting groups of actors associated with certain topics (Wang et al. 2005).

A third set of research has taken a matrix/tensor factorization approach to uncovering topics and trends in online social media. Eigen-trend is a method of tracking the importance of a keyword in the blogosphere taking into account weights of different blogs in which they are mentioned (Chi et al. 2006). They also propose a higher order singular value decomposition approach to compute hubs and authority scores of a set of blogs specific to a particular keyword. Zhu et al. (2007) has proposed a technique to simultaneously factorize a web linkage matrix and web-page content matrix to generate a parsimonious representation of the original matrices. They have shown on WebKB Craven et al. (1998) and Cora dataset McCallum (2000) that using this parsimonious representation of one can perform page classification with accuracy as good or better than other state-of-the-art classification algorithms. TOPHITS is an approach to compute topical hub and authority in a set of web pages by representing the hyperlinks labeled by anchor text in a $from \times to \times keyword$ tensor and factorizing the tensor(Kolda and Bader 2006).

Dynamic modeling of relations in a social network has seen recent interest as well. Chi et al. (2007) present a non-negative matrix factorization approach to model interaction between blogs over time as sum of interaction within communities of blogs whose intensities vary over time. A three way nonnegative tensor factorization approach has been applied in Bader et al. (2007) for tracking relationship between employee from the Enron email dataset using the number of emails exchanged between them. They have also applied tensor factorization on international trade data for tracking trade relation between countries. A latent space model to track relation between authors in NIPS publication network has been presented in Sarkar and Moore (2005).

Despite recent progress in expert search, expert *discovery* without any pre-specified query, has been a relatively unexplored area. Such discovery would be very helpful for an organization with a vast collection of user generated content and little idea of the expertise present therein. Despite progress in tracking of social interaction and tracking topic developments in news streams, *topical analysis of social interaction over time* has not been well explored. Topical analysis of dynamic social interaction is of considerable value for detecting and tracking significant topics of *conversation* between actors in a social network[1]. Since, the significance of a topic of conversation depends not only on the content of the conversation, but, on the significance of the people participating in the conversation, we need an analysis framework that can handle actors as the source and the target of a message, text content, and time stamp on the conversation. Author-recipient-topic model is based on one such framework, but, at the current stage it does not track topics or actors over time.

The current paper aims to fill these gaps. The proposed approach discovers expertise and experts in an unsupervised and query independent manner. It takes into account the social interaction between the actors as well as the content they exchange in determining the experts. It also tracks how expertise has developed over time. We use a tensor based framework to achieve this as opposed to the probabilistic graphical modeling framework that the Author-Recipient-Topic model is based on.

In Section 3 the dataset that is collected for this study is described. In Section 4 an interpretation of tensor factorization is developed that defines significance of entities such as actors and keywords. In Section 5 a set of simulation studies are performed to examine the performance of the tensor

[1] We define conversation as a set of messages exchanged between two or more people that discusses one common topic of interest.

Blog post and reply data	
Bloggers	4.8K
Commentors	16K
Blogs	4.7K
Blog posts	71.5K average length 300 words
Comments	286K average length 33 words
Date range	Jan '07–Oct '08

Table 1 Blog data description

Total # of posts 71.5K			
In community			Outside
44K			27.7K
	Work	Non-Work	
# posts	15K	29K	
# communities	11	14	

Table 2 Posts in different communities

Non work			Work		
Community Name	Posts	Replies	Community Name	Posts	Replies
Arts	93	1415	BusinessDevelopment	500	858
Books	220	1283	CorporateFunctions	190	1796
Corporate Social Responsibility	625	17324	Domains	955	501
Education Motivation	1677	3557	Feedback	166	2737
Fun	4861	38671	Free Linux Open Source Software	477	3159
Geographies	47	128	KM	845	5652
History-Culture	153	262	Practices Programs Accounts	189	1300
Miscelleneous	7759	23286	Project Management	173	262
Movie-TV-Music	573	4378	Senior Management	34	243
Photography	2235	14137	Technology	9882	4447
Poetry-Stories	6239	36879	Testing	1662	1030
Puzzles	1927	5705			
Religion-Spiritual-Culture	729	3689			
Sports	1910	7129			

Table 3 Different communities

factorization as an expert discovery tool under different data generation conditions. In Section 6 the tensor factorization is applied to the blog dataset and the discovered topics, experts, and trends are discussed.

3. Description of an enterprise blog

The data for this study has been extracted from an employee-only blog network in a large IT services firm. It contains the blog posts and replies indexed by employee id of the author along with timestamps and demographic information about the bloggers. The dataset was collected over a period of 1 year and 10 months from Jan 2007 to Oct 2008. A subset of the blog articles are posted in one of the 25 communities by their authors. These 25 communities have been further classified as work related and non-work related by experts at the firm(Table 3). It is expected that there should be more replies which are shorter than blog articles (usually much longer). Often a

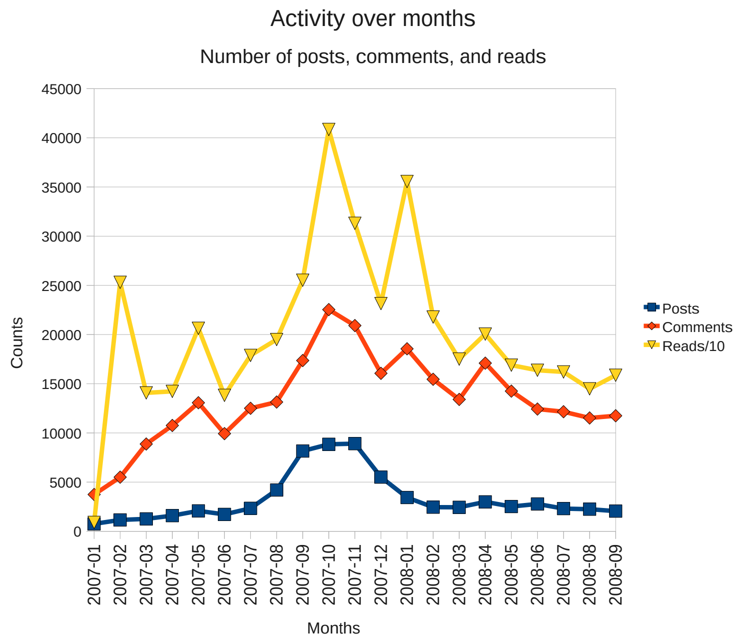


Figure 1 **Blog posting, replying, and reading activity over the duration of data collection**

blog article attracts multiple replies. However, in some very technical topics such as "Technology" and "Testing" (automated software testing) there are fewer responses than blog articles suggesting reduced involvement of the community of blog readers in these topics.

We have also collected web access logs from the blog hosting server. The log contains the reading activity of the employees in the firm during the same 1 year and 10 month time period over which the blog posts and responses were collected. The activity levels in the blogosphere is shown in Figure 1.

The blogs can be thought of as an evolving network of interaction between employees. Each interaction could be an activity between a pair of bloggers. For example it could be a reply to a post written by another blogger, or a citation to another post written by a different bloggers. Such interactions are characterized by the text associated with them. We give a few examples in Table 4 and 5[2]. As we can see from these discussion threads employees use blogs to carry out group conversations. In Table 4 a person from senior management has written about the company's new office in China. Other employee in the company in their response to the article have congratulated (post #1), have been critical (post #32), and also have added to the discussion from their own experiences (post #53). In Table 5 a post in "Corporate Social Responsibility" is shown along with readers' responses to it. The replies can be targeted to the original author of the blog article, or as in the case of Table 5 posts #26 and #27, they could be towards any particular person participating in the conversation. With the help of the timestamps of the messages, the Blogosphere consisting of bloggers and exchanges of replies among them, can be characterized as an evolving online social network.

From the examination of the blog posts we found that, unlike the posts in Table 4 and 5, a majority of the posts are chatters: collection of jokes, quotes, and interesting news items etc. that form no consistent pattern and generate interest for only a short duration. Often they are of limited value to the organization. In a random sample of 50 blog posts we found 31 to be collected joke,

[2] All the names of person and organizations occurring in the data shown have been replaced to protect the privacy of the individual and the organization

EmployeeID	Timestamp	Text
		Article
xxx358	2007-05-24 01:23:57	One more successful trip for me. Sanghai, the ninth largest city in the world and a critical center of communication for China. Tall buildings, great infrastructure, fascinating architecture, all speaking progress and world class standards. The place that hosts our China Development center, and where I celebrated the 2nd Annual Day on May 18 2007, with a bunch of very inspiring people. ...
		Replies
(#1) xxx944	2007-05-24 03:05:18	It is nice and heartening to know that our operations in China has scaled new heights. Congratulations to the entire team who made it happen...
	...	
(#32) xxx548	2007-05-26 02:12:57	Congratz to team in China & people who are pioneering this effort. However there is always a question in my mind, what makes us pursue China in a big way (even asking business strategist to market out China operations). Is it strength of numbers or people skills? Is it inrastructure of China? Is it clients in China itself? These are few of questions which I think our strategist would have thought over, but still it will be helpfull if they can put in one of their subsequent blogs
	...	
(#53) xxx667	2008-01-18 15:09:23	China team is definitely rocking. I was lucky enough to work with one of the teams for around 3 - 4 months and I was very pleased with the productivity of the team. Other things I liked about them are: Attention to detail & Process adherence. Above all, the collaborative business model (India and China teams working hand-in-hand) is here to stay & deliver commendable results.

Table 4 **Post from a senior manager about a new office in China and subsequent replies to it**

EmployeeID	Timestamp	Text
		Article
xxx865	2007-04-24 17:47:37	Hi All, First of all I thank each and every one of you for your overwhelming response towards the contribution for the Orphanage construction work. We went to that Orphanage two weeks before for gathering their requirements. As they were in a hurry to shift the orphanage at the earliest, they have already started the construction work with the fund donated by some private company people. So, the orphanage In-charge requested us to buy a water tank and an electric motor for their use. We promised him to give both at the earliest and we provided them the same last week. The water tank, electric motor and its accessories costs Rs.13000. (Rs.8000 from the recently collected fund and Rs.5000 from our existing fund.) Please find the attached document regarding the expenditure details - Expenditure.doc Now, it is the time for us to visit that Orphanage and spend our time with the kids. This home takes care of 29 children. So, I invite 15 volunteers to join us this Saturday (Apr-28) and make their time and ours more joyful. Note: We dont need any contribution from you this time. We are looking forward for your presence. ...
		Replies
(#2) xxx745	2007-04-24 18:04:57	Hi friends, I am very much interested in serving poor people. So i wont fail to attend this visit. Also i feel happy that there is a team involved in this divine activity. I feel proud to be a member of this excellent team. Meet u guys on Apr 28
	...	
(#12) xxx883	2007-04-25 08:12:34	hi, i will be more than glad to be a part of this wonderful event. Our visit to "old age home" last week was simply great and i am sure that this visit will also bring smiles on the faces of all 29 little kids present there.. kindly inform me about details like where we have to meet and at what time.. thanks.
	...	
(#26) xxx883	2007-04-27 13:31:52	hi Karth... Apart from the important work (water tank, electric motor and its accessories) which we have already done at this orphanage, it will be great if we can carry some sweets or candies for them... i am sure they will love it.... what are your views ?
(#27) xxx865	2007-04-27 14:11:16	Hi Nim, We have planned to provide fruits to the kids ont Saturday. We avoid giving sweets to the kids in orphanage, since many people visiting the orphanage use to provide the same. So, we prefer fruits which is good for their health too.
	...	
		(32 replies to the article)

Table 5 **Post from a person in middle management in "Corporate Social Responsibility" topic and subsequent replies from other employees**

Total # of posts made	Average # of unique readers per post
639	3
534	4
532	3
512	1
480	3

Table 6 **Average number of people who read posts of the most frequent bloggers. The average for each of the five most frequent bloggers shown. Average over the entire blog dataset is 22**

Id: xxx081 Date: 2007-09-05	**Id: xxx991 Date: 2007-11-09**	**Id: xxx368 Date: 2007-10-10**
`Diodes can be classified by the functions of the circuit in which it is being used, or more commonly, by the shape that is demanded by the size of the products in which it will be mounted. The complicated point is that there is no direct relation between the two and you must keep them both in your mind at all times. ...` (125 more posts by xxx081 in next ten days on "voltage", "diodes" and "semiconductors")	`Benefits of Automated Testing. If you have ever tested applications or Web sites manually, you are aware of the drawbacks. Manual testing is time-consuming and tedious, requiring a heavy investment in human resources. Worst of all, time constraints often make it impossible to manually test every feature thoroughly ...` (150 more posts by xxx991 in next eight weeks on "software", "test", "automation")	`20 Minute Home Work Out. If you are busy, not able to get up early morning or have no time for gym just follow this 20 minute home work out to stay healthy and fit. 1. Jog in one place for 3 minutes. Simple light jogging on the spot. 2. Jumping jacks: 25 repeats. When landing, bend your knees slightly to reduce the impact on knee ...` (190 more posts by xxx368 in next hundred days on "exercise", "muscle", "weight")

Table 7 **Some of the topics in a blog network along with posting pattern of people behind them.**

quote, or news item and 16 to be original and work related (2 of the remaining were non-English, 1 was empty post). In fact the bloggers who post most frequently exhibit this behavior. However, their posts are not read by many (Table 6). Therefore, the bloggers who are the most active in the blogosphere are not necessarily the experts or the most popular bloggers[3]. In order to determine the expertise of a person one needs to consider the content of her posts along with other signals from the blogosphere such as the people from whom the post derives a response.

The methods described in this paper use such signals to identify significant authors and topics. The result of such analysis reveals that buried in the chatters there are long running topics driven by one or a small group of bloggers who could be considered authorities in the subject (See Table 7). In the remainder of the paper we develope this method (Section 4). Then we discuss the results obtained by applying this method to the blog social network data and evaluate the proposed method by comparing it to other similar methods (Section 6).

4. Importance of entities in intra-organizational blogs

Importance by association is behind the calculation of many importance measures.

Eigen centrality has been used to defined status in a social network (Wasserman and Faust 1994). The centrality of an actor is determined by the centrality of other actors in the network that it is connected to. This follows the intuition that if a person is affiliated with other people who has high status in the network the person also has high status because of it. The adjacency matrix representation of such social networks are symmetric and its elements are positive. On such a matrix the centralities of the actors are given by the leading eigenvector. The leading eigenvector can be computed by singular value decomposition.

Singular value decomposition of an adjacency matrix ($from \times to$) of a network of directed hyperlinks between a set of web pages produces the hub and authority scores of the pages(Pagerank Brin

[3] Spam was not an issue in this dataset, since, no one outside the firm had access to the blog network

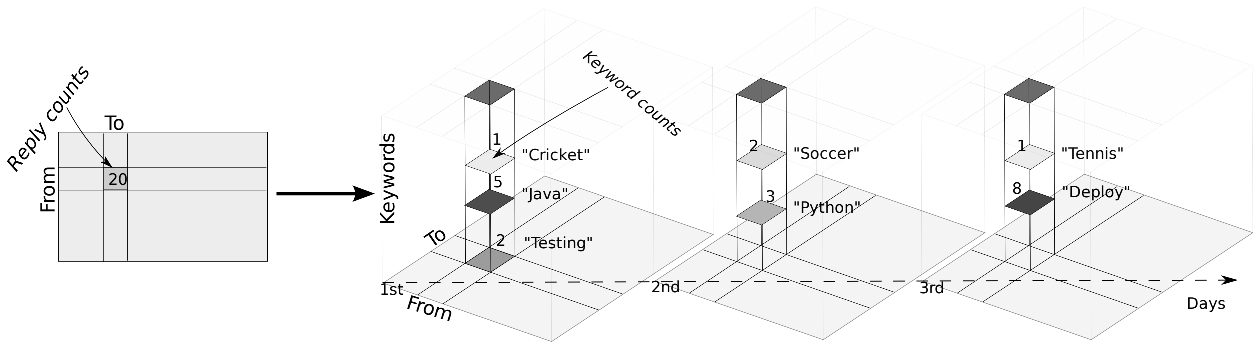


Figure 2 Tensor representation of the conversation

and Page (1998), HITS (Kleinberg 1999)). The leading left singular vector gives the hub scores whereas the leading right singular vector gives the authority scores. The node with high hub scores are the ones that link to nodes with high authority scores and the nodes with high authority scores are the ones linked to by nodes with high hub scores. Usually the leading singular vector pair is used since they explain most of the variations in the data, however subsequent singular vector pairs can also be used if their singular values indicate that they explain substantial portion of the data as well. Subsequent pairs have the same relation between the hubs and the authorities. Each pair corresponds to a different community sub-structure over the nodes. The first k pairs of singular vectors provide a decomposition of the two dimensional data matrix into k rank-1 matrices. This method is unsupervised: topics are determined solely from the co-occurrence patterns in the data. Although, HITS and Pagerank use SVD to identify authoritative pages in a set of hyperlinked webpages, they differ in their interpretation and in the adjacency matrices they operate on. HITS takes a more traditional network factorization approach to identify hubs *and* authorities. It focuses on a smaller sub-network identified by local crawling around a root web-page. However, Pagerank is computed on the entire WWW network. The authority of a webpage is computed as the probability of a random surfer being at the webpage after a long time. The random surfers transition from one page to another can be modeled by a Markov process. The authorities of the web page is the stationary probabilities at the webpages. This is obtained by SVD on a transition probability matrix over the entire network.

Outside the network analysis literature SVD has been applied to the co-incidence matrix of documents and terms. This is known as Latent Semantic Indexing (LSI)(Deerwester et al. 1990). LSI produces a parsimonious representation of the original document-term matrix where each document has certain weight on a small number of semantic topics and each word has some weight on those semantic topic. In each topic, the weight of the document is determined by the weight of the keywords in the document and the weight of each keyword is determined by the weight of the documents in which it occurs.

Not all datasets can be satisfactorily represented by a two dimensional matrix. In a blog network where relations are indicated by citations and replies, encoding the relation by a single number would lose the content and the context of the relation. Or, in the case of an evolving network, where a timestamp is associated with each edge, a two dimensional representation of the relational data would have to be at the expense of temporal information. Such data is better represented and analyzed in a tensor. For example text messages exchanged between people in the social networks can be represented in a $from \times to \times keyword$ tensor. To capture the dynamic nature of the exchange, one can form such tensor for each time period. This would lead to a $from \times to \times keyword \times time$ tensor (Figure 2).

Each cell of the tensor contains co-occurrence count, or a value derived thereof, of the three corresponding *author*, *recipient*, and *keyword*. The cell value indicates the strength of association between the three. Other similar examples can be found in TOPHITS (Kolda and Bader 2006) and three way DEDICOM (Bader et al. 2007). The current work builds on this literature using tensors for encoding semantic graphs and focuses on identification of significant themes along with important actors and important dates in each as part of the larger investigation into mapping expertise within an enterprise blog network.

4.1. Summary of notations

Here is a list of notations used in this paper.

$a, b, c \dots$ are used to represent scalars.

$\mathbf{a}, \mathbf{b}, \mathbf{c} \dots$ are used to represent a vector.

$\mathbf{A}, \mathbf{B}, \mathbf{C} \dots$ are used to represent two dimensional matrices.

$\mathbb{A}, \mathbb{B}, \mathbb{C} \dots$ are used to represent tensors that have more than two dimensions or *modes*.

$\mathbf{a} \circ \mathbf{b}$ is the outer product between vector $\mathbf{a}$ and vector $\mathbf{b}$. The result is a matrix whose ith row and j th column contains $a_i b_j$. This can be extended to outer product between more than two vectors. Outer product of three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ would result in a tensor whose i, j, kth element contains $a_i b_j c_k$.

$\| \cdot \|_F$ represents the Frobenious norm of a matrix or a tensor. This is equal to square root of the sum of square of the elements in the matrix or the tensor.

$\times_k$ is the k mode multiplication of a tensor with a vector (Kolda and Bader 2008). It is defined as

$$\mathbb{Y} = \mathbb{X} \times_k \mathbf{v}$$
$$\Leftrightarrow \mathbb{Y}_{i_1, i_2, \dots, i_{k-1}, i_{k+1}, \dots, i_M} = \sum_{i_k} x_{i_1, i_2, \dots, i_{k-1}, i_k, i_{k+1}, \dots, i_M} \times v_{i_k}$$

Notice that (1) the length of $\mathbf{v}$ must equal to the size of the kth mode of $\mathbb{X}$ and (2) $\mathbb{Y}$ has $M-1$ modes. This is similar to multiplication of a matrix with a vector: multiplying dimensions of the matrix and the length of the vector must match; and the result is a vector of length equal to the non-multiplying side of the matrix.

The last one is the Kruskal operator $[\![\cdot]\!]$ defined by Kolda (2006) as

$$[\![\boldsymbol{\lambda}; \mathbf{A^{(1)}}, \dots, \mathbf{A^{(M)}}]\!] = \sum_{r=1}^{R} \lambda_r \times \mathbf{a_r^{(1)}} \circ \dots \circ \mathbf{a_r^{(M)}}$$

Each $\mathbf{A^{(k)}}$ matrix has R columns. The matrices are of equal width, but, they need not have equal height. The Kruskal operator adds together R outer products of M vectors to produce a M mode tensor. The length of the kth side of the tensor is equal to the height of $\mathbf{A^{(k)}}$.

4.2. Importance definition for multi-modal data

Extending the idea of significance by association we propose that significance can be considered to be assigned from actors to other *objects* they are affiliated with. For example, consider an email network. Lets assume we have only the dataset of who emailed who and how many times. From this dataset we can identify central hubs and central authorities in the network depending on their email patterns: centrality of the hubs are determined by the centrality of the authorities they are sending emails to and the centralities of the authorities are determined by the centralities of the hubs they are receiving emails from. Lets again consider that we now have the content of those emails and our task is to identify central topics in this email dataset. To guide our search for important topics one heuristic might be to look for what is being discussed between the central hubs and authorities in the email network. Thus we can consider status being assigned from high status *hubs* and *authorities* to the *topics* they discuss. This can be seen in many other context as well. We often judge importance of a new topic of research by the prominence of researchers working on it. Product endorsement by star players is based on endorsement of status by affiliation. Significance by association has a broader domain of application than only two dimensional network data. Guided by this intuition we propose the following definition.

DEFINITION 1. Generalizing *importance by association* to tensors the *importance* of an entity in a multi-modal dataset represented by a co-incidence tensor depends on the importance of the entities in the other modes it is co-incident with.

If we assume that each mode makes a multiplicative contribution of importance for each co-incidence, as is done in many matrix based importance calculations, we can formalize the definition in the following way.

For a co-incidence tensor $\mathbb{X}$ of M modes let the importances of entities along the kth mode be $\mathbf{a^{(k)}}$. Each $\mathbf{a^{(k)}}$ satisfies the following condition

$$a_j^{(k)} = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_{k-1}} \sum_{i_{k+1}} \cdots \sum_{i_M} x_{i_1,i_2,\ldots,i_{k-1},j,i_{k+1},\ldots,i_M} a_{i_1}^{(1)} a_{i_2}^{(2)} \ldots a_{i_{k-1}}^{(k-1)} a_{i_{k+1}}^{(k+1)} \ldots a_{i_M}^{(M)}; \forall k \tag{1}$$

Using tensor multiplication notation the Equation 1 can be compressed to

$$\mathbf{a^{(k)}} = \mathbb{X} \prod_{i \neq k} \times_i \mathbf{a^{(i)}} \tag{2}$$

Notice that after the sequence of multiplications $\mathbb{X}$ reduces to a vector of length equal to the kth side of the tensor $\mathbb{X}$. This gives the weight of the entities along kth mode of the tensor $\mathbb{X}$.

Applied iteratively for all k, $\mathbf{a^{(k)}}$s converge to minimize $\|\mathbb{X} - \mathbf{a^{(1)}} \circ \cdots \circ \mathbf{a^{(M)}}\|_F$ (De Lathauwer et al. 2000). In other words

$$\mathbb{X} \approx \mathbf{a^{(1)}} \circ \cdots \circ \mathbf{a^{(M)}} \tag{3}$$

or, $\mathbf{a^{(1)}} \circ \cdots \circ \mathbf{a^{(M)}}$ is the rank-1 approximation of $\mathbb{X}$. $\{\mathbf{a^{(1)}}, \ldots, \mathbf{a^{(M)}}\}$ is the most dominant factor in $\mathbb{X}$.

One can compute the best rank-R approximation of $\mathbb{X}$ by using parallel factorization of the tensor (Harshman 1970)—often abbreviated to PARAFAC. Denoting the kth mode vector of rth factor by $\mathbf{a_r^{(k)}}$, the rank-R approximation can be expressed as:

$$\mathbb{X} \approx \sum_{r=1}^{R} \lambda_r \times \mathbf{a_r^{(1)}} \circ \cdots \circ \mathbf{a_r^{(M)}} = [\![\boldsymbol{\lambda}; \mathbf{A^{(1)}}, \ldots, \mathbf{A^{(M)}}]\!] \tag{4}$$

λ_r is a normalizer to keep norm of each weight vector $\mathbf{a_r^{(k)}}$ equal to 1. Each of R sets of $\{\mathbf{a^{(1)}}, \ldots, \mathbf{a^{(M)}}\}$ importance weights satisfy 2. Each $\mathbf{A^{(m)}}$ is a modal matrix with R columns. The rth column is $\mathbf{a_r^{(m)}}$. Therefore, $\mathbf{A^{(m)}}$ contains all the factorization weights for entities along the mode m.

The popular approach to compute PARAFAC is based on Alternating-Least-Square error minimization (ALS). The error $\|\mathbb{X} - [\![\boldsymbol{\lambda}; \mathbf{A^{(1)}}, \ldots, \mathbf{A^{(M)}}]\!]\|_F$ is minimized by successively optimizing one of the M matrices while keeping the remaining $M-1$ matrices constant. The detailed ALS algorithm can be found in Kolda and Bader (2008). An implementation is available in their TensorToolbox matlab package (Bader and Kolda 2007).

We illustrate two applications of this method for blog data analysis next.

4.3. Blog post developments

One view of the blogs is that they are self-publication media where bloggers write on the topics of their interest. If the goal is to identify important developments of a topic in the posts, we posit that we need to look beyond the word occurrences in the blog posts. We also need to consider the importance of the author of the post. The post made by an authority in a subject is a stronger signal of a development in the topic, than a post that is made by someone who is not an authority in the subject. Therefore, to identify different topic developments in blog posts, the relevant variables are the authors, timestamps and keywords of the blog posts. This data can be represented as a $author \times keyword \times timestamp$ co-incidence tensor $\mathbb{X}$, where, each cell of the tensor contains a weight derived from the counts of the word occurrences (Section 6.1). This value indicates the strength of association of the three variables. Spelling out Definition 1 for $author \times keyword \times timestamp$ tensor we obtain the following reinforcing definition of *authority* of bloggers, *importance* of keywords and *intensity* of a topic at a given time period for a particular topic:

1. The *authority* of a blogger in a topic can be judged from her participation during the period when the intensity of the topic is high and from her use of important keywords.
2. The *importance* of a keyword in a topic can be judged from its use by the authorities in the topic and from its use during the period when the intensity of the topics is high.
3. The *intensity* of a topic during a time period can be measured from the number of posts made by authorities and the presence of important keywords in the topic.

This is a higher order extension of hub and authority computation. When we want to identify only the most dominant topic, the importance of pth blogger in this topic can be calculated according to the definition as:

$$a_p = \sum_q \sum_r x_{pqr} k_q t_r \Longleftrightarrow \mathbf{a} = \mathbb{X} \times_2 \mathbf{k} \times_3 \mathbf{t} \tag{5}$$

Similarly

$$\mathbf{k} = \mathbb{X} \times_1 \mathbf{a} \times_3 \mathbf{t} \tag{6}$$

$$\mathbf{t} = \mathbb{X} \times_1 \mathbf{a} \times_2 \mathbf{k} \tag{7}$$

where, $\mathbb{X} \in \Re^{|\mathbf{a}|\times|\mathbf{k}|\times|\mathbf{t}|}$; $\mathbf{a}$, $\mathbf{k}$, and $\mathbf{t}$ are the vectors of importance of the authors, keywords and time periods; $\times_j$ is the j-mode product of a vector with a tensor. Applied iteratively the vectors $\mathbf{a}$, $\mathbf{k}$, and $\mathbf{t}$ converge to minimize the error $\|\mathbb{X} - \mathbf{a} \circ \mathbf{k} \circ \mathbf{t}\|_F$. Thus $\mathbf{a} \circ \mathbf{k} \circ \mathbf{t}$ is the rank-1 approximation of the tensor $\mathbb{X}$. Extending from one dominant topic to R topics and using a set of normalizers $\boldsymbol{\lambda}$ to make each vector of unit length, the approximation can be expressed as sum of R rank-1 tensors:

$$\mathbb{X} \approx \sum_r^R \lambda_r \times \mathbf{a_r} \circ \mathbf{k_r} \circ \mathbf{t_r} = [\boldsymbol{\lambda}; \mathbf{A}, \mathbf{K}, \mathbf{T}] \tag{8}$$

where, $\mathbf{A}$,$\mathbf{K}$,and $\mathbf{T}$ are the three modal matrices each with R $\mathbf{a_r}$, $\mathbf{k_r}$, and $\mathbf{t_r}$ as column vectors respectively.

4.4. Blog conversation development

In this extension we take into account the conversational nature of the blog posts. A comment to a blog post or a post with a citation has an author and a recipient. Subject of the post not only depends on who is making the post but also who it is targeted to. To capture this we represent the blog data in a fourth order tensor ($author \times recipient \times kewords \times timestamp$). The idea behind evaluating the importance of a variable is similar to that in blog topic development analysis. The extension is that the importance of the recipient of the conversation influences the importance of the variables in other modes.

4.5. Comparison with the existing methods

The HITS algorithm (Kleinberg 1999) separates a blog network into multiple layers of network. But, it does so based on the pattern of links—not taking into account the content. One could envision an approach where first the blog posts are clustered into topics based on their text content and then HITS is performed in each to find important people in the group. Although, this approach separates conversations into topics based on the content of the documents, it does not take into account the importance of the words said in computing the importance of the bloggers. This Blog conversation development work has more similarities with the TOPHITS (Kolda and Bader 2006) where a $from \times to \times term$ tensor was constructed for hyperlinked web pages. TOPHITS uses the anchor text as the description of the link. We use text in the blog posts and replies that are much longer; and require more cleanup and normalization of the term vectors. Our work is also different in its extension with a time dimension to track topic intensities over time.

5. Simulation Examining Expert Identification by Tensor Factorization

In this section we simulate the blogging behavior of the experts in a blogosphere. The generated multi-modal data is represented in a tensor. Then we use tensor factorization to detect the experts and their topics of expertise. The model of the user behavior is simplified considerably so that the properties of the factorization can be well understood. Two scenarios are simulated. In the first, the bloggers are observed to make posts in different topics at different points of time. In the second, the bloggers are observed to exchange messages with others depending their interest in different topics. Tensor factorization is applied in each scenario to detect experts and their topics of expertise. The performance is measured at different parameter values to find out when the algorithm works well and when it does not.

5.1. Detecting Experts from their Individual Posting Behaviors

In the first scenario we observe only the publication of the experts and not their interaction with others. In such a scenario a person is considered to be an expert in a topic if she posts repeatedly in a topic with a certain set of keywords over a long period of time. This is contrasted to a non-expert's who has posted as many times but in no specific topic because he does not specialize in any topic. The simulation is set up as follows.

- There are K experts in K different topics
- There are N_{ne} individuals who are not expert in any topic
- There are V *ordinary words*
- Each topic has V_k different *keywords* that occur H times more frequently in the topic than they do outside the topic
- The data is available over T days
- Each individual writes N_d documents on N_d different days
- Each document has L words
- The documents written by the experts have the topic specific word distribution, i.e., when writing a document the experts choose a topic specific keyword with a higher probability than they choose an ordinary word. The document written by the non-experts do not have topic specific word distribution.

Random three dimensional tensors are created using the above setup. The three modes of the tensors are *author*, *word*, and *day*. The tensor is factorized into K factors. The authors with the highest score in each factor are extracted and compared with the list of experts. Similarly, the top V_k words in each factor are collectively compared to the set of keywords in all K topics to judge the accuracy of the topic discovery.

We begin by observing the performance of the tensor factorization at a set of reasonable parameter settings. They are $K = 5, N_{ne} = 10, V_k = 5, V = 100, H = 10, T = 100, N_d = 10,$ and $L = 100$.

Probability of a keyword occurring in a document on a topic (P_t) and in general (P_e)

Arts				*Business Development*				*Free Linux Open Source*			
Keyword	P_t	P_e	H	Keyword	P_t	P_e	H	Keyword	P_t	P_e	H
art	0.0645	0.0197	3.3	organize	0.0367	0.0041	9.0	linux	0.3944	0.0101	39.0
color	0.1613	0.0312	5.2	project	0.1568	0.0696	2.3	free	0.2608	0.0576	4.5
capture	0.0645	0.0184	3.5	process	0.1894	0.0892	2.1	bash	0.0216	0.0016	13.5
sketch	0.086	0.0029	29.7	economy	0.0448	0.0071	6.3	emacs	0.0151	0.0003	50.3
paint	0.2366	0.0133	17.8	recession	0.0163	0.0016	10.2	boot	0.1013	0.0064	15.8
Religion Spiritual Culture				*Technology*				*Software Testing*			
Keyword	P_t	P_e	H	Keyword	P_t	P_e	H	Keyword	P_t	P_e	H
spiritual	0.1874	0.0084	22.3	data	0.2724	0.0751	3.6	automate	0.1484	0.0143	10.4
god	0.3441	0.069	5.0	server	0.1382	0.0362	3.8	bug	0.1435	0.0161	8.9
soul	0.1441	0.0221	6.5	dsn	0.0126	0.0022	5.7	coverage	0.0715	0.0071	10.1
conscious	0.1021	0.0072	14.2	command	0.1237	0.033	3.7	regress	0.0812	0.0038	21.4
ritual	0.0503	0.0034	14.8	system	0.285	0.1082	2.6	testcase	0.0279	0.001	27.9

Average length of a document (in words) after removing stop-words and rare words = 117
Average number of articles written by a blogger = 14.8

Table 8 Parameters from the blog data

These are comparable to the real world observation of these parameters as verified from the blog dataset (Table 8).

The topic specific keywords of the 5 topics are indexed as 1,..., 5; 6,..., 10; 11,..., 15; 16,..., 20; and 21,..., 25. The words that are not keywords are indexed 26,..., 125. The 5 experts are indexed as 1,..., 5 and the non-experts in the system are indexed as 6,..., 15. The scores of the top authors and top keywords in each factor are displayed in Table 9. The 5 experts are correctly ranked as the top authors in 5 different factors. The 5 keywords in each of the 5 different topics are ranked as the top words in the corresponding factors. In addition, there is a sharp difference between scores of the expert and that of the other authors, as well as between the keywords and non-keywords. A perfect expert discovery would assign score 1 to all the experts and 0 to the other authors. In addition, since each column is normalized to have euclidean length of 1, it would assign score $\frac{1}{\sqrt{5}} = 0.447$ to each of the keywords and 0 to the other words in the vocabulary.

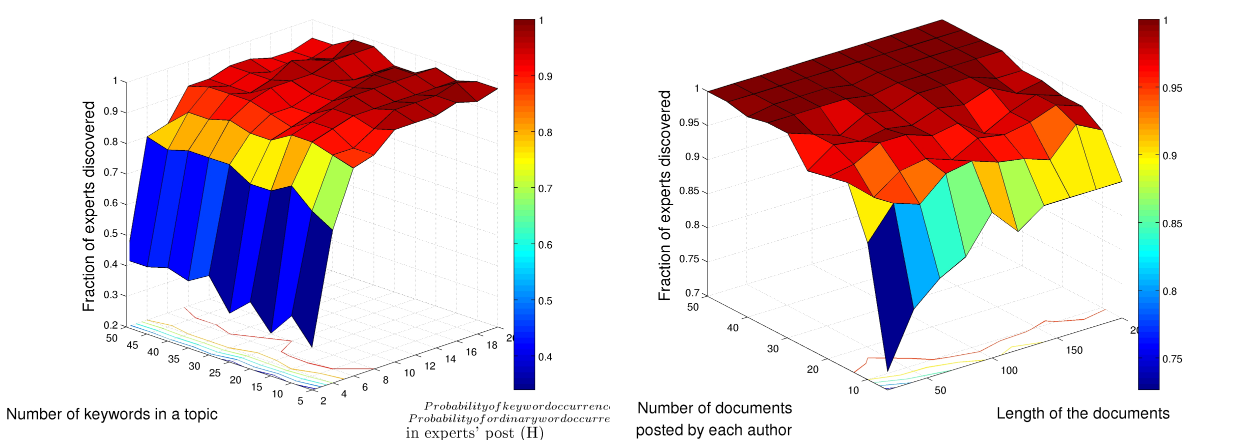
The accuracy of the expert discovery is measured by computing the fraction of experts that appear at the top of the list of authors in the K factors. Similarly the accuracy of topic discovery is measured by computing the fraction of the $K \times V_k$ keywords that are among the top V_k words of the K factors produced by the tensor factorization. From the Table 9 we can see that both these fractions are 1, i.e., all the experts and all the keywords have been successfully identified. However, it is not always so. Depending on the parameter values the tensor factorization can run into some difficulty in identifying experts and topic specific keywords.

One of the factors that helps us identify authors specializing in different topics is the unique distribution of words in their posts. However, the relative frequency of topic specific keywords with respect to the background word frequency can vary from one author to another and one topic to another. One would expect it to be easier to find experts when topics have markedly different frequency of keyword occurrence. In the described setup this relative frequency is the tunable parameter H. When H is greater than 5 the tensor factorization is able to identify experts with more than 90% accuracy even with a small number of keywords (Figure 3(a)) .

The length of the documents posted by the bloggers and the number of documents posted by them affect the accuracy of the expert discovery algorithm. If the documents posted are very short and if the bloggers have posted only a few documents each, then the tensor factorization has difficulty in identifying the experts among the authors. However, with the other parameters set

Words					Experts				
Word indices					Expert indices				
13	22	19	1	9	3	5	4	1	2
14	24	18	3	8	6	10	14	8	15
15	21	16	4	6	14	6	6	9	8
12	25	20	5	10	7	15	15	7	6
11	23	17	2	7	9	8	7	11	11
84	47	94	83	57	10	13	12	12	9
68	107	51	28	58	12	9	11	13	13
6	80	105	32	114	8	7	2	10	7
53	51	106	45	22	11	1	8	15	12
48	48	78	117	72	1	2	9	14	10
...					...				
Word scores					Expert scores				
0.4379	0.4344	0.4585	0.4417	0.4442	0.9942	0.9860	0.9932	0.9715	0.9886
0.4329	0.4240	0.4224	0.4123	0.4181	0.0575	0.1157	0.0684	0.1282	0.0862
0.4294	0.4001	0.3878	0.3834	0.3996	0.0471	0.0777	0.0657	0.1112	0.0542
0.3883	0.3716	0.3771	0.3483	0.3619	0.0395	0.0475	0.0354	0.1058	0.0430
0.3156	0.3457	0.3268	0.3093	0.3076	0.0358	0.0435	0.0322	0.0688	0.0427
0.0719	0.0931	0.1006	0.1013	0.0815	0.0347	0.0408	0.0306	0.0531	0.0417
0.0685	0.0743	0.0882	0.0825	0.0798	0.0276	0.0374	0.0295	0.0503	0.0409
0.0682	0.0742	0.0870	0.0802	0.0770	0.0253	0.0328	0.0156	0.0473	0.0409
0.0681	0.0690	0.0768	0.0794	0.0760	0.0242	0.0056	0.0063	0.0439	0.0406
0.0633	0.0670	0.0685	0.0763	0.0737	0.0052	0.0040	0.0052	0.0405	0.0397
...					...				

Table 9 **Typical performance of a tensor factorization. The experts (indexed 1–5) and the keywords (indexed 1–25) have been 'floated' to the top of their corresponding factors by the tensor factorization.**



(a)Sensitivity to the number of non-experts and the rate of keyword occurrence

(b)Sensitivity to the document length and the number of documents posted by the authors

Figure 3 **(a) When there are more non-experts posting, the task of expert discovery becomes harder. However, if the frequency of keyword occurrences in the posts made by experts is sufficiently different, then we can still identify the experts. (b) The tensor factorization does not identify experts very well when the document lengths are very short or when authors have written few documents. However, with increase in either of these the accuracy quickly improves significantly**

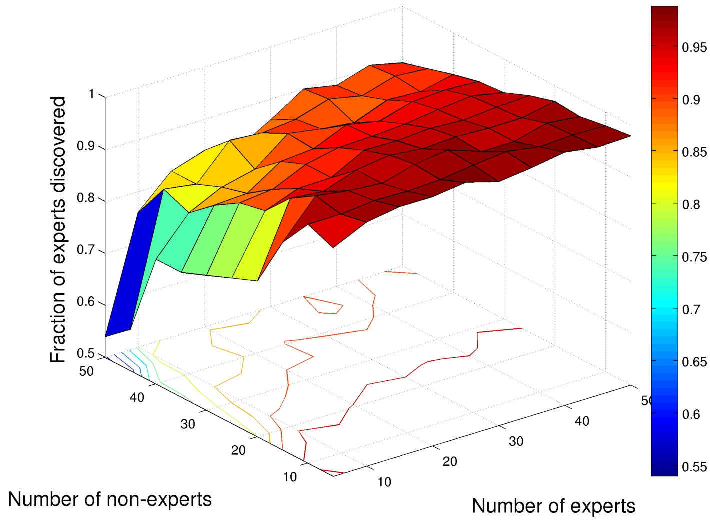


Figure 4 Accuracy of expert identification with different numbers of experts and non-experts.

as outlined in the baseline case when the authors write more than 10 posts or when they writes posts longer than 20 words each, the tensor factorization correctly identifies the experts among the writers (Figure 3(b)). The accuracy of *expertise* discovery as measured by the fractions of keyword discovered follows a very similar pattern. Therefore, it is omitted.

One might expect the number of non-experts present in the blog system to affect the performance of the method. As the non-experts modeled to be writing on no specific topic in particular the posts made by the non-experts acts as noise for an algorithm that is looking to discover experts and their expertise. Therefore, the task of discovering them becomes harder as there are more non-experts writing about various topics. In this simulation the number of non-experts N_{ne} is another tunable parameter. As the posts by the non-experts act as noise in the system, the presence of many experts each specializing in different topics makes it easier to identify the experts in the system. We contrast these two factors in Figure 4. By varying the number of experts and the number of non-experts between 5 to 50 at steps of 5 we see that as the number of experts per each non-expert increases the tensor factorization is able to identify the experts in the system more accurately. Even when the non-experts vastly outnumber the experts in the system, e.g., when there are 5 experts and 50 non-experts, the tensor factorization can detect the experts with approximately 55% accuracy. A random selection of users would pick experts with less than 10% accuracy.

5.2. Detecting Experts from their Social Exchanges

In the second scenario we observe the text of the exchanges between bloggers. A person can be considered an expert if she exchanges messages in a topic with other experts in the same topic and does so over a period of time. This agrees with the intuition that expertise often develops from a group of people discussing and working together on a common topic of interest. This is different from a group of friends who frequently talk to each other not because they are interested in a common topic, but, because they happen to know each other. Messages exchanged in such a group is likely to be not focused on any particular topic. An expert discovery method should be able to tell a group of experts apart from a group of otherwise friends.

This is tested by a simulation study that is set up as follows:

- There are K topics
- Each topic has N_e experts

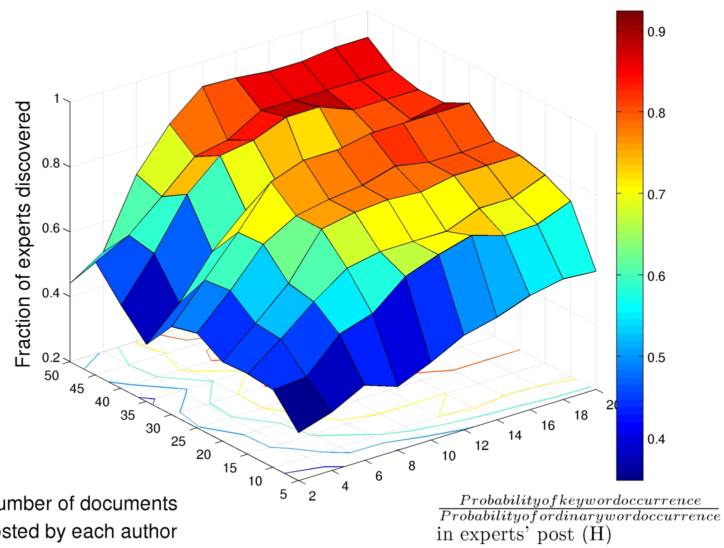


Figure 5 **Sensitivity analysis with respect to number of documents posted by each author and the rate of the keyword occurrence in topical documents**

- There are G groups of bloggers who are not experts but who communicate within the group
- Each of the G groups has N_g bloggers
- There are N_r bloggers who are not part of any group
- Each topic has V_k topic specific *keywords* that occur H times more frequently in the documents specific to the topic than they do in the documents outside the topic
- The data is available over T days
- Each blogger writes N_d documents on N_d different days targeted to bloggers chosen in the following manner
  — The experts in each topic write documents in their topic and send to a randomly selected expert in their group
  — The non-expert group members send each message to a randomly selected member of their group
  — The individuals who are not member of any group write messages targeted to randomly selected individuals from the entire set of users
- Each document has L words

Random four dimensional tensors are created with the data generated from this model. The modes of the tensor are *author*, *recipient*, *word*, and *day*. The tensor is factorized into K factors. Top authors and keywords in those factors are examined as outlined in Section 5.1. Again the typical performances look similar to Table 9, except that in this case with have 5 experts per topic that the factorization identifies.

We explore the expert and expertise discovery performance of tensor factorization at different parameter values. Although, it is able to discover experts accurately for most parameter settings the tensor factorization fails when the occurrences of topical keywords are not very different from ordinary words, or when bloggers have not posted many documents (Figure 5).

From Figure 5 we see that when the bloggers have posted very few documents or if the rate of keyword occurrence is not sufficiently different than the rate of the other word occurrences then tensor factorization is not able to identify the experts accurately. However, this improves with more documents per author and higher rate of keyword occurrence in the topical documents.

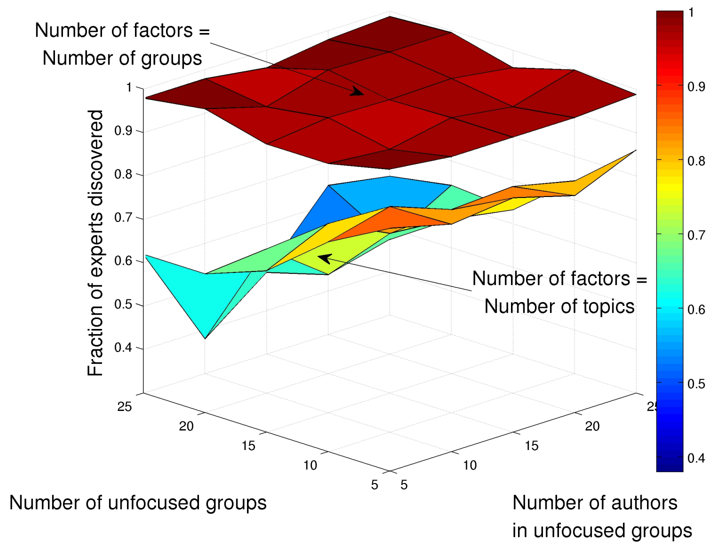


Figure 6 Sensitivity analysis with respect to the number of groups and sizes of the groups

We further explore the performance of the tensor factorization around a set of baseline parameter values. These are $K = 5, N_e = 5, G = 5, N_g = 5, N_r = 20, V_k = 5, V = 100, H = 10, T = 100, N_d = 30,$ and $L = 100$.

In Figure 6 we explore the two new parameters introduced in this scenario: the number of groups of non-expert bloggers and sizes of these groups. When the number of groups of bloggers who write on no specific topic is small, the tensor factorization is able to ignore even large groups of such bloggers to discover the few experts in the system. However, when there are many such *groups* then the tensor factorization fails to detect experts among them as accurately. We suspect that this behavior is due to the use of incorrect number of factors.

In Section 5.1 there was a natural way to select the number of factors. We set it to the number of topics. However, when there are groups of bloggers writing to each other, some focused on a topic and some not focused on any, the number of factors is less clear. In addition, since the PARAFAC factors are not orthogonal to each other, the addition of more factors do not produce the original factors. The factorization quality is thus dependent on the use of correct number of factors. This behavior has been observed in the literature and development of methods that are tolerant to incorrect specification of number of factors is an open research problem (Acar et al. 2009). We find that, in this set of simulations if we created as many factors as the number of groups and considered top authors in only the K most dominant factors we can identify experts significantly more accurately. This is shown in the top plot of Figure 6.

The response of the algorithm to the changes in document length and number of topical keywords is similar to what we observe in Section 5.1, i.e., The performance of the algorithm suffers if all the documents are less than 40 word long. The performance was largely independent of the number of topical keywords present, but, sensitive to how much more frequently they occur in topic compared to outside of the topic.

6. Tensor factorization of blog dataset

6.1. Data preparation

We used two subsets of the data for the two methods described in Section 4.3 and Section 4.4 for blog data analysis. For blog post development analysis we used the text of the blog posts, author-ids, and the timestamp on the post. We followed the standard text processing steps to convert the text of each post to a vector of term weights.

Number of replies	260K
to blog posts	176K
to other replies	84K
to multiple post/reply	12K

Table 10 Descriptive statistics of the reply network

1. Very common words (*stop words*) were removed. For this study we used a custom created list of words that are very frequent in this blog dataset
2. Words that occur in at least a minimum number of documents were kept (minimum number of documents was 10 for the reported results). This reduces the vocabulary size considerably while removing the words that are very rare
3. Frequency of word j in document i was counted and normalized using the formula proposed in Singhal et al. (1996) :

$$x_{ij} = \underbrace{\frac{1+log(TF_{ij})}{1+log(avg_i(TF))}}_{\text{Doc Term Weight}} \times \underbrace{\left(log\left(\frac{N}{DF_j}\right)\right)}_{\text{IDF}} \div \underbrace{\left(0.8+0.2\left(\frac{\#\text{of unique terms in document } i}{\text{average number of unique terms}}\right)\right)}_{\text{Doc Length Normalization}} \quad (9)$$

The first factor reduces the weight of terms that occur multiple times in a document. The second factor computes the information content of the term. The third factor normalizes for the length of the document.

The timestamps were coalesced to a granularity of one day. This data is stored in a $author \times keyword \times timestamp$ tensor. Each cell of the tensor contains total weight of a keyword used by an author on a particular day. This resulted in a $4.6K \times 650 \times 22.5K$ tensor with 4.4M nonzero entries ($sparsity = 6.6 \times 10^{-5}$).

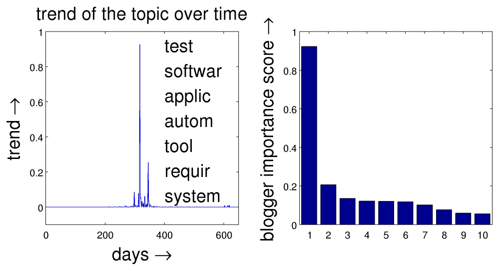
For the blog conversation development analysis we used only the words in the reply text, the author-id, the id of the target blogger and the timestamp. The reason for excluding the blog posts is that it is not clear who the initial blog post is targeted to. To determine the target of a reply we searched for mentions of names of authors who have written the blog post or replied to the blog post prior to the current reply in the reply thread. If one or more names are mentioned in the current reply, then the reply is taken to be targeted to latest replies by the named authors in the enclosing reply chain. Note that a reply could be targeted to multiple authors. If a reply does not contain any name then it is taken to be directed at no one other than the person who wrote the original blog article, since it is a reply to the original article. The resulting reply statistics is shown in Table 10.

We carried out the same transformation of text and the timestamp as done in the case of blog-post-development, but, this time we stored the data in a $author \times recipient \times keyword \times timestamp$ tensor. Each cell of this tensor contains the weight of a word said by the author to the recipient on a particular day. This resulted in a $16.8K \times 3.8K \times 900 \times 11.8K$ tensor with 1.6M nonzero entries ($sparsity = 3.4 \times 10^{-9}$).

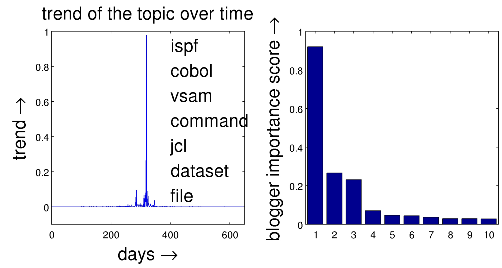
6.2. Illustrative results

Each tensor was decomposed into 25 rank-1 tensors using PARAFAC, because evidence from community labels suggest that there are 25 communities in the data. Some of the resulting factors are displayed in Figures 7 and 8. For each factor the top five keywords are shown along with the importance of the top-authors in descending order and the daily intensity of the topic.

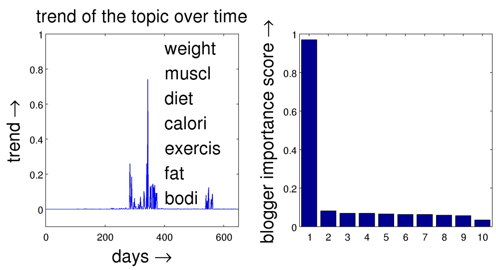
As we can see from Figure 7 the decomposition is separating activities in different topics. The intensities of the topics over time show us the trend of the topic. Posts about "software testing" (7(a)) have generated interest for much shorter period compared to the conversations about "new



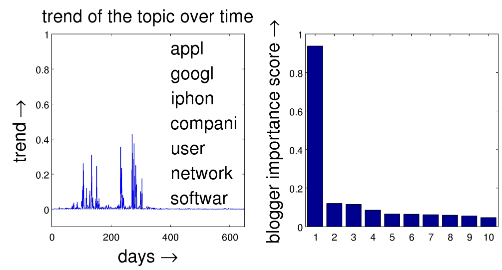
(a)Software testing



(b)Programming in COBOL



(c)Physical fitness



(d)New technology companies

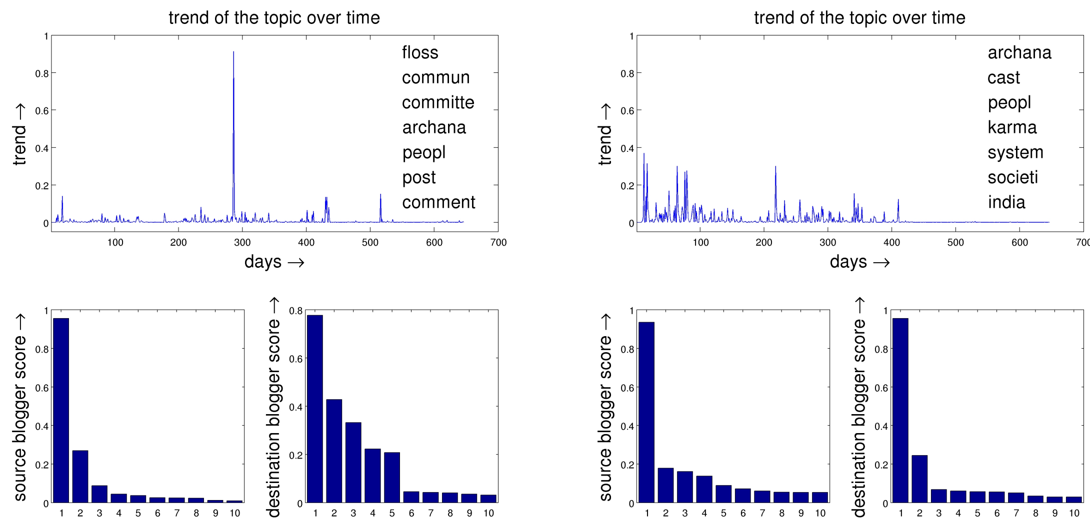
Figure 7 Trends of topics and importance of top bloggers in each topic. Four identified topics are illustrated. The histograms of the actor importances show different degree of participation of actors in the topic. For example, the topic of "physical exercise" can be seen to be driven primarily by one person.

technology companies" (7(d)). The importance scores of the top authors also give insight into the nature of the conversation. Posts about "physical exercise" have seen activity over about 100 days, but, they are primarily made by one person. On the other hand more people have contributed to "software testing" and "COBOL" topics, though they were active for a shorter period.

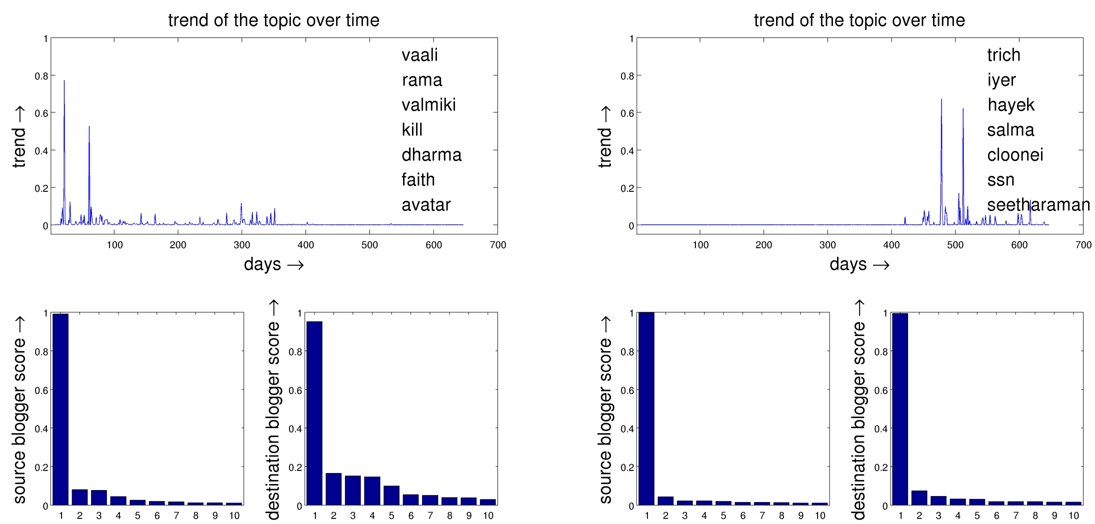
Analysis of comments on the blog posts reveals a different set of factors. These are usually the factors that generate more reactions from the bloggers, e.g., "Free Linux Open Source Software" (FLOSS, Figure 8(a)), "Cast system in India" (Figure 8(b)), "Mythology" (Figure 8(c)), and "Movies" (Figure 8(d)). In these sets of plots we have a set of most significant reply *recipients* in addition to the most significant *keywords*, *authors*, and *days*. The intensity of the topic over time shows how the discussion on the topic varied over time. E.g., software testing is discussed over the entire observation period, albeit at lower intensity except a peak around 290th day. Cast systems and FLOSS have been discussed during the first half of the observation period, where as movies and actor/actresses have been discussed during the second half of the observation period. The weights of the message senders and recipients in each topic tells us how wide-spread the discussion is. E.g., FLOSS is discussed by more people than movies are.

We further analyzed the posts made into individual communities separately. For illustration we present some of the sample topics detected in the community "Sports" (Figure 9), "FLOSS" (Free Linux Open Source Software, Figure 10), and "Corporate Social Responsibility" (Figure 11). The topics in the Sports community are largely event driven as can be observed from the intensity of the cricket and soccer topics. The spike in discussion in the Cricket topic during Feb 2008[4] coincides with Commonwealth Bank Series of matches between India, Sri Lanka, and Australia (3rd Feb '08–4th Mar '08). The top keywords detected in this topic are primarily the names of the cricketers from India and Sri Lanka. The spike in the Soccer topic occurs towards the end of May

[4] 1st Jan 2007 is the first day in the plot



(a)Free Linux Open Source Community (b)Cast system in India



(c)Mythology (d)Movies

Figure 8 **Trends of topics and importance of topic specific hubs and authorities in each. The histogram of source bloggers show the extent to which the top bloggers have posted in the topic. The histogram of target bloggers show the extent to which they are the recipient of messages in the topic—presumably because of their prior participation in the topic.**

2008 followed by a few small activities. The large spike is right before the 2008 UEFA European Football Championship (7th Jun '08–29th Jun '08).

We show two topics of conversation detected in the FLOSS community. Microsoft is discussed using the keywords competition, fine, payment etc. in the first week of Mar 2008. This discussion follows the record amount of fine imposed by the European antitrust regulators against Microsoft on 27th of Feb 2008. The other highlighted discussion in the FLOSS community is focused on open source web technologies.

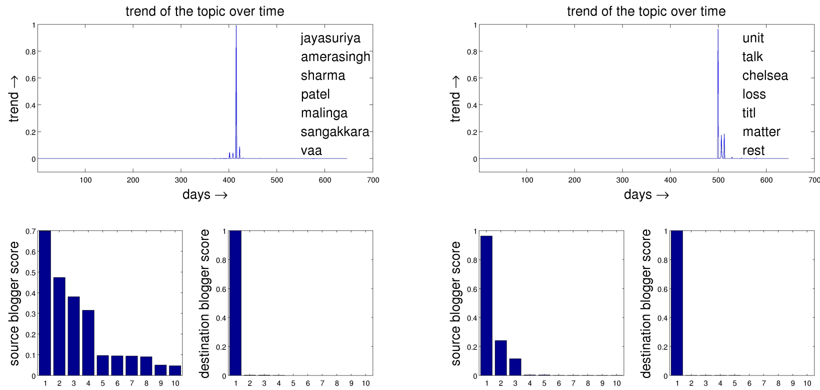


Figure 9 **Analysis of messages exchanged in the Sports community reveals Cricket and Soccer topics. The left topic is recognizable as Cricket from the names of the prominent Cricket players that are in the topic**

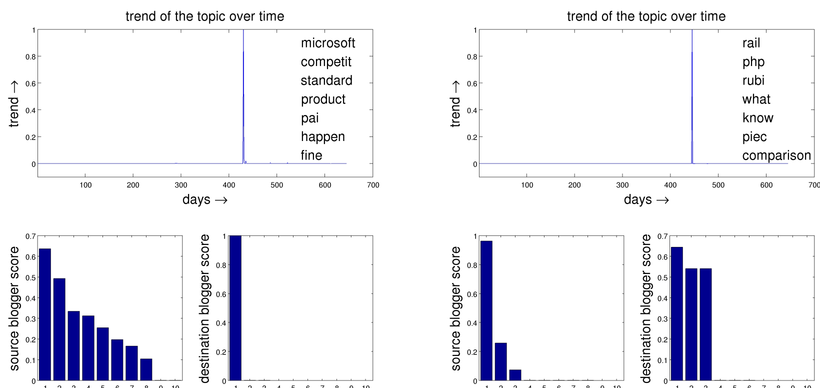


Figure 10 **Discussion of Microsoft and of open source web technologies in the "Free Linux Open Source Software" community**

Looking inside the "Corporate Social Responsibility" community with the help of tensor factorization we find two significant topics of discussion. The first one is discussing how a company can engage more with the young population in local community. The second topic is a campaign to build awareness about the effect of smoking on ones health. Unlike the discussions in the Sports community, these two discussions are likely to be triggered by factors within the company and not readily observable to someone outside. So, it is not possible to correlate these activity spikes with the help of external events.

6.3. Comparison with content independent hub and authority

We compare the hubs and authorities identified by tensor factorization with those identified by the HITS algorithm. Since, hubs are measure of the quality of the sender of the replies, we also compare these to the most frequent reply makers. For illustration we show the size of the overlap of the top-10 author list identified by these three methods. These are given in Table 11(a). As we can see HITS algorithm, that does not consider the content of the reply or the timestamp of the replies, tends to sample more from the set of most *vocal* bloggers than the PARAFAC tensor factorization algorithm does. Similar behavior is also observed when we compare the top-10 authorities identified by the two factorization algorithms with the top-10 bloggers who received most responses (Table 11(b)). This shows that the scores of the HITS algorithm that does not take into account the

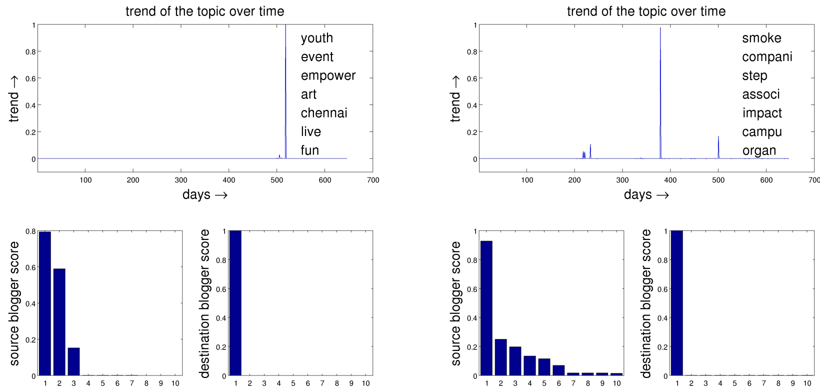


Figure 11 **Discussion of higher engagement in local community and campaign against smoking in "Corporate Social Responsibility"**

(a)Comparison of hubs.
H_H : top 10 hubs identified by HITS
H_P : top 10 hubs identified by PARAFAC
H_{voc} : top 10 bloggers who *wrote* most replies

Factor	$\lvert H_H \cap H_{voc}\rvert$	$\lvert H_P \cap H_{voc}\rvert$	$\lvert H_H \cap H_P\rvert$
1	8	4	4
2	5	4	4
3	5	2	3
4	6	2	2
5	4	3	4
6	4	2	2
7	7	2	4
8	4	4	0
9	3	2	1
10	0	1	0
11	6	6	3
12	0	3	0
13	4	3	1
14	2	0	0
15	3	3	0
16	2	3	0
17	2	2	2
18	2	7	0
19	2	6	0
20	3	1	2
21	4	6	2
22	0	1	0
23	2	4	1
24	0	3	0
25	0	4	0

(b)Comparison of authorities.
A_H : top 10 authorities identified by HITS
A_P : top 10 authorities identified by PARAFAC
A_{pop} : top 10 bloggers who *received* most replies

Factor	$\lvert A_H \cap A_{pop}\rvert$	$\lvert A_P \cap A_{pop}\rvert$	$\lvert A_H \cap A_P\rvert$
1	6	1	2
2	4	3	2
3	4	1	1
4	4	3	0
5	4	3	2
6	8	2	1
7	6	1	0
8	4	3	0
9	2	2	0
10	8	1	1
11	6	2	1
12	7	2	1
13	0	3	0
14	1	2	0
15	2	2	0
16	1	1	0
17	1	3	0
18	0	6	0
19	3	3	2
20	0	4	1
21	1	2	0
22	1	1	0
23	0	3	1
24	1	4	0
25	0	3	0

Table 11 **Comparison of hubs and authorities. The table cells show the number of common bloggers in a pair of top-10 bloggers.**

content of the interaction are influenced more by the volume of the interaction than the scores of the tensor factorization.

6.4. "On topic" quality of the top hubs' response

Tensor factorization simultaneously determines the significant words, authors and recipients in a blog reply network. We illustrate the benefits of such an approach over the existing methods that do not take into account the words exchanged between the authors. To illustrate this we look at the characteristic of the content produced by the top actors identified by each method. We decided to measure the quality of the hub by the closeness of the reply to the target post. The idea behind this measure is that a good responder would contribute positively to the topic of discussion by keeping his responses within the topic of discussion as opposed to someone who makes offtopic comments. To compute the distance of an author's responses from their targets we use the KL-divergence of the word distribution in the author's replies from the word distribution in their target posts.

We create two probability distribution for each author:
Q : Distribution over the words in all of the replies written by the author
P : Distribution over the words in all the posts and replies that are targets of the replies
where, the target of a reply is determined as described in Section 6.1.

The KL-divergence of probability distributions Q from probability distribution P is (Kullback and Leibler 1951) :

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{10}$$

A reply could be targeted to the blog article that started reply thread. It could also be targeted to one of the replies within the thread that occurred at a time before the current reply was written. If the name of one of the authors who wrote the blog article or any subsequent reply occurs in the current reply then the current reply is deemed to be targeted to that particular author.

Then the KL-divergence of Q from P are computed for each author. We compare this score for the top hubs identified by the Tensor factorization with the top hubs identified by the HITS algorithm over the entire reply network.

In addition we compare these to a simpler topic specific HITS algorithm. In this approach we identify keywords in the reply set by Latent Semantic Indexing (LSI). For each factor identified by LSI the top 10 words were chosen as the keywords. Hub and Authority scores were computed over the replies that contain these keywords. This approach produces a overlapping but different sets of Hubs and Authorities than the previous two approaches.

It is worth noting that these three approaches form a continuum of methodologies where keyword information in the replies are used to different extents. In the HITS over the entire reply network only the reply count between two authors is used. No keyword information is used in this approach. The simpler topic specific HITS is a two step approach. First a set of keywords are selected and then Hub and Authority is computed on the reply network on which those keywords occur. In the tensor factorization approach the keywords, hubs, and authorities are determined simultaneously.

For each of the three methods the average KL-divergence of the top-k hubs identified by all three methods is plotted in Figure 12. The confidence interval is drawn at two standard deviation from the mean using dashed lines. From this figure we can see that by using keywords in the responses while computing hubs and authorities we can identify hubs that are closer to the target of their response. For the hubs identified by HITS the KL divergences of their replies from the targets is the largest. Thus they are the most off topic in their replies. The KL divergences of the responses made by the hubs identified by the keyword specific HITS algorithm is lower. So, they can be thought of being closer to the target posts in their response. The hubs detected by the Tensor factorization have the lowest KL-divergence of response from the target posts. Therefore, they can be considered to be most on topic in their response to their target post or reply.

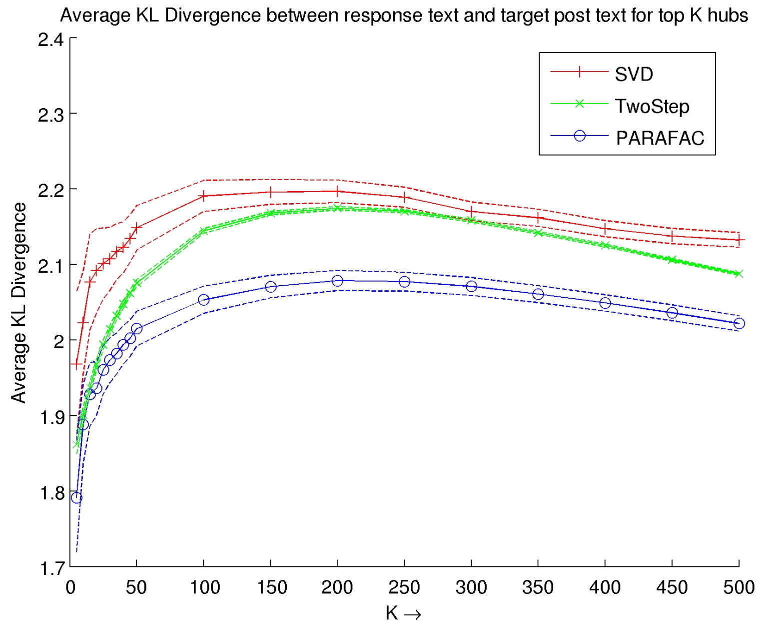


Figure 12 Average KL divergences of the top hubs identified by HITS, keyword specific HITS, and Tensor factorization

6.5. Community discovery

We evaluate the effectiveness of the tensor factorization by applying it to discover topical communities in the blog conversation data. Communities in an online social network are based on user interests that manifest in the text exchanged between the users over the network. We hypothesize that because of its use of multi-modal data a tensor based approach is better suited for discovering communities than methods that do not use the multi-modal data that constitute conversation within online communities.

Definition 2. We define a *conversation* in the blog network to be a set of replies to a blog post from one or more bloggers in the network. Conversation data consists of the text, author, target blogger, and timestamp of the replies.

Task : We define the *community discovery* task as given a set of conversations to discover the clusters of similar conversations in them.

To perform this task using tensor factorization we follow two steps :

1. Identify the predominant factors in the entire blog conversation dataset by tensor factorization (Section 4.4). Each factor consists of a set of weights over the authors, recipients, keywords, and days.

2. Map each conversation to the factor it most closely aligns with by an operation of *Tensor Query*

Tensor Query We pose each conversation as a query to retrieve the factor most similar to it. This is done by first approximating each conversation, represented as a tensor $\mathbb{C}$, by a linear combination of the R factors in the original tensor.

If $\mathbf{A^{(1)}}, \ldots, \mathbf{A^{(M)}}$ are the M modal matrices obtained from the factorization (Equation 4) then the approximation is

$$\mathbb{C} \approx \sum_r \beta_r \times \mathbf{a_r^{(1)}} \circ \cdots \circ \mathbf{a_r^{(M)}} \tag{11}$$

	Macro average		Micro average	
	Tensor factorization	Document clustering	Tensor factorization	Document clustering
P	0.4	0.46	0.25	0.37
R	0.73	0.35	0.71	0.27
F	**0.51**	**0.40**	**0.37**	**0.31**

Table 12 Precision recall of community discovery

Let $V(\cdot)$ be the vectorization operator that reshapes a tensor to a column vector in a given order. Let $\mathbf{v_r} = V(\mathbf{a_r^{(1)}} \circ \cdots \circ \mathbf{a_r^{(M)}})$. Then, Equation 11 can be expressed as:

$$V(\mathbb{C}) \approx \sum_r \beta_r \times \mathbf{v_r} \tag{12}$$

$$V(\mathbb{C}) \approx [\mathbf{v_1}, \ldots, \mathbf{v_r}] * \boldsymbol{\beta} \tag{13}$$

where, $*$ is the left product of a matrix with a vector. Equation 13 is a least square problem of which $\boldsymbol{\beta}$ is the solution. Magnitude of the elements of $\boldsymbol{\beta}$ indicate how closely $\mathbb{C}$ is aligned to the corresponding factors. The factor with the largest β_r is the factor that $\mathbb{C}$ aligns most closely with. Therefore, we assign the conversation to that factor.

Repeating this for each conversation in the dataset gives us a set of clusters of conversations. These clusters of conversations can be evaluated using one of many cluster evaluation techniques if we have some kind of true label on each conversation. We use the author provided "community labels" that is part of the blog dataset (Table 3) for this purpose.

6.5.1. Experiment setup and results For community discovery we used only the part of the dataset that has community labels on it, so that we can evaluate the resulting communities. Only the reply data is used to form conversations among bloggers. This dataset has 180K responses forming 21.5K conversations. We cluster these conversations into 25 communities using tensor factorization. To obtain a baseline to compare with we clustered the response text using repeated bisection document clustering algorithm (Zhao and Karypis 2004). Using the author provided community labels we computed the precision (P), recall(R), and the F-score for the clusters formed as described in Larsen and Aone (1999). The scores are shown in Table 12.

As we can see the tensor factorization has a lower P score meaning the communities it is uncovering have more of the other community conversations than the communities uncovered by document clustering algorithm. However, tensor factorization is keeping conversations in each community together in one cluster more so than document clustering algorithm (reflected in larger R value). This leads to a higher F score for the tensor factorization approach for community discovery. The results show that the conversation clusters generated by tensor factorization agrees with the author provided labels on the blog articles more than the clusters generated by document clustering do.

7. Conclusion

In this paper an importance definition is proposed for the multi-modal data that constitutes online social conversation. The definition is a higher order extension of importance computation approaches already seen in the literature, e.g. eigenvector centrality computation, latent semantic indexing, etc. This higher order computation requires a multi-modal representation of the data. We use tensors to represent the multi-modal data and compute importance scores on it. We show that the proposed importance of the entities in the multi-modal data can be computed by tensor factorization.

Using a simulation study we show that the tensor factorization can discover experts accurately at realistic parameter values. The instances where it does not discover experts accurately are when there are few documents posted by the bloggers, or when the documents are very short, or when the rate of occurrence of the topical keywords is not very different from ordinary words. The accuracy of the method also suffers when the number of factors created is very different from the number of groups present in the dataset.

The proposed approach is illustrated by applying it to a blog dataset collected from the corporate Intranet of a large IT services firm. We analyze the blog data in two different tensors representing two views of the blogosphere: 1. Blogs are self publication media, 2. Blogosphere is a channel for discussing topics and exchanging ideas among people with common interests. We apply tensor factorization to identify significant topics of conversation, important bloggers in each, and the development of the topic over time.

We also assess the quality of the hubs identified by tensor factorization vis-a-vis HITS on the entire network and HITS on a subnetwork containing content keywords. For each author we measure the closeness of her replies from the target posts using KL-divergence. The lower the KL-divergence the more on topic the authors' replies are. The hubs identified by Tensor factorization had the smaller KL-divergence from their targets than the hubs identified by keyword specific HITS. HITS on the entire reply network, without taking into account the content of the replies, produces hubs with highest KL-divergence from the target.

To evaluate the tensor factorization approach we use it to discover topical communities in the conversations occurring in the blog network. Conversations are defined as a set of messages exchanged in response to a blog post. If we can identify important actors in different communities and important keywords used by the people in those community we should be able to better identify different topical communities using these information. To identify topical communities using tensors we design a *tensor query* operation to retrieve the factor most similar to a conversation posed as a query. This operation takes into account the weight of entities in each mode of the data to compute the similarity of the conversation to a factor. Using author provided community labels on the conversations we show that tensor factorization performs better than document clustering on conversation text in identifying communities in the conversation.

Such a tool has significant utility to managers and information officers in organizations. As enterprises adopt Web 2.0 technology and culture, large collections of user generated content are being created in the new online social environments. This work provides a lens to discover expertise within such a collection. The trends of the topic provides insight into the dynamics of the discovered topics, and identification of key individuals provides actionable information to the managers.

One of the limitations of this study is that the claimed importance definitions have not been *directly* validated. Obvious importance measures, such as number of times a blogger's posts are read, importance of a blogger based on importance of people who read the blogger's post, are content independent. In our experiment we found them to be better predicted by simpler eigenvectors of the adjacency matrix containing number of replies between the bloggers. The importance of a blogger in a specific topic, although makes intuitive sense, is difficult to obtain as part of a dataset. One approach to obtain such information is by asking experts about the important topics in a given blog network and who are the important bloggers in them. We leave this measurement for a future study.

Tensors are a natural extensions of matrices for capturing and analyzing multi-modal datasets. We have shown their application for identifying important bloggers, topics, and trends in online conversations. There are several future research directions. One of the assumptions in this work is the multiplicative endowment of importance due to association. Although, it allows us to use existing tensor factorization techniques for importance computation the assumption may not be valid in all situations. For example Bonacich (1987) has shown that being surrounded by powerful actors in a network might *reduce* the leverage and power of an actor. Thus the tie with powerful neighbors might not increase one's power. If this is the case then one cannot use the existing tensor factorization directly. Another direction of research is the application of non-negative tensor factorization to obtain entity weights that are non-negative (Shashua and Hazan 2005). Non-negative weights are are easier to interpret in many real word scenarios.

References

Acar, Evrim, Tamara G. Kolda, Daniel M. Dunlavy. 2009. An optimization approach for fitting canonical tensor decompositions. Tech. Rep. SAND2009-0857, Sandia National Laboratories, Albuquerque, NM and Livermore, CA.

Allan, James, Ron Papka, Victor Lavrenko. 1998. On-line new event detection and tracking. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, 37–45. doi:http://doi.acm.org/10.1145/290941.290954.

Bader, Brett W., Tamara G. Kolda. 2007. Matlab tensor toolbox version 2.2. URL `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`.

Bader, B.W., R.A. Harshman, T.G. Kolda. 2007. Temporal analysis of semantic graphs using ASALSAN. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining-Volume 00*. IEEE Computer Society Washington, DC, USA, 33–42.

Becerra-Fernandez, I. 2006. Searching for experts on the Web: A review of contemporary expertise locator systems. *ACM Transactions on Internet Technology (TOIT)* **6**(4) 333–355.

Bonacich, P. 1987. Power and centrality: A family of measures. *American Journal of Sociology* 1170–1182.

Bonacich, P., P. Lloyd. 2001. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **23**(3) 191–201.

Brass, D.J. 1992. Power in organizations: A social network perspective. *The Political Consequences of Social Networks: The Political Consequences of Social Networks: 1992* **4** 295–323.

Brin, S., L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Seventh International World-Wide Web Conference (WWW 1998)*. URL `http://ilpubs.stanford.edu:8090/361/`.

Burt, R.S., T. Schøtt. 1985. Relation contents in multiple networks. *Social Science Research* **14**(4) 1985.

Chi, Yun, Belle L. Tseng, Junichi Tatemura. 2006. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 68–77. doi:http://doi.acm.org/10.1145/1183614.1183628.

Chi, Yun, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, Belle L. Tseng. 2007. Structural and temporal analysis of the blogosphere through community factorization. *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 163–172. doi:http://doi.acm.org/10.1145/1281192.1281213.

Craswell, N., A. de Vries, I. Soboroff. 2005. Overview of the trec-2005 enterprise track. *TREC 2005 Conference Notebook*. Citeseer, 199–205.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery. 1998. Learning to extract symbolic knowledge from the World Wide Web. *Proceedings of the National Conference on Artificial Intelligence*. Citeseer, 509–516.

De Lathauwer, L., B. De Moor, J. Vandewalle. 2000. On the Best Rank-1 and Rank-(R, R,..., R) Approximation of Higher-Order Tensors. *SIAM Journal on Matrix Analysis and Applications* **21** 1324.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* **41**(6) 391–407.

Doddington, George, Jaime Carbonell, James Allan, Jonathan Yamron, Umass Amherst, Yiming Yang. 2000. Topic detection and tracking pilot study final report.

Franz, Martin, Todd Ward, J. Scott McCarley, Wei-Jing Zhu. 2001. Unsupervised and supervised clustering for topic tracking. *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on*

Research and development in information retrieval. ACM Press, 310–317. doi:http://doi.acm.org/10.1145/383952.384013.

Harshman, R.A. 1970. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics* **16**(1) 84.

Kautz, H., B. Selman. 1998. Creating models of real-world communities with ReferralWeb. *Working notes of the Workshop on Recommender Systems, held in conjunction with AAAI-98, Madison, WI*.

Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* **46**(5) 604–632. doi:http://doi.acm.org/10.1145/324133.324140.

Kolda, Tamara G. 2006. Multilinear operators for higher-order decompositions. Tech. Rep. SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore, CA. URL `http://www.prod.sandia.gov/cgi-bin/techlib/access-control.pl/2006/062081.pdf`.

Kolda, Tamara G., Brett W. Bader. 2008. Tensor decompositions and applications. *SIAM Review* To appear (accepted June 2008).

Kolda, TG, BW Bader. 2006. The TOPHITS model for higher-order web link analysis. *Workshop on Link Analysis, Counterterrorism and Security*.

Krulwich, B., C. Burkey, A. Consulting. 1996. The ContactFinder agent: Answering bulletin board questions with referrals. *Proceedings of the National Conference on Artificial Intelligence*. 10–15.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statistics* **22** 79–86.

Larsen, Bjornar, Chinatsu Aone. 1999. Fast and effective text mining using linear-time document clustering. *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 16–22. doi:http://doi.acm.org/10.1145/312129.312186.

McCallum, Andrew. 2000. Cora. URL `http://www.cs.umass.edu/~mccallum/code-data.html`.

McCallum, Andrew, Andres Corrada-Emmanuel, Xuerui Wang. 2004. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *NIPS'04 Workshop on'Structured Data and Representations in Probabilistic Models for Categorization*.

Qamra, Arun, Belle Tseng, Edward Y. Chang. 2006. Mining blog stories using community-based and temporal clustering. *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 58–67. doi:http://doi.acm.org/10.1145/1183614.1183627.

Sarkar, P., A.W. Moore. 2005. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter* **7**(2) 31–40.

Shashua, A., T. Hazan. 2005. Non-negative tensor factorization with applications to statistics and computer vision. *Proceedings of the 22nd international conference on Machine learning*. ACM, 799.

Singhal, Amit, Chris Buckley, Mandar Mitra. 1996. Pivoted document length normalization. *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 21–29. doi:http://doi.acm.org/10.1145/243199.243206.

Soboroff, I., A. de Vries, N. Craswell. 2006. Overview of the trec 2006 enterprise track. *TREC 2006 Working Notes* .

Streeter, L.A., K.E. Lochbaum. 1988a. An expert/expert locating system based on automatic representation of semantic structure. *Proceedings of the fourth conference on artificial intelligence applications*. 345–350.

Streeter, L.A., K.E. Lochbaum. 1988b. Who knows: A system based on automatic representation of semantic structure. *RIAO*, vol. 88. 380–388.

Swan, R., D. Jensen. 2000. Timemines: Constructing timelines with statistical models of word usage. *KDD-2000 Workshop on Text Mining*. Citeseer, 73–80.

Vivacqua, A. 1999. Agents for expertise location. *Proc. 1999 AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace*. 9–13.

Wang, Xuerui, Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 424–433. doi:http://doi.acm.org/10.1145/1150402.1150450.

Wang, Xuerui, Natasha Mohanty, Andrew McCallum. 2005. Group and topic discovery from relations and text. *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. ACM, New York, NY, USA, 28–35. doi:http://doi.acm.org/10.1145/1134271.1134276.

Wasserman, S., K. Faust. 1994. *Social network analysis: Methods and applications*. Cambridge Univ Pr.

Yang, Yiming, Tom Pierce, Jaime Carbonell. 1998. A study of retrospective and on-line event detection. *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, 28–36. doi:http://doi.acm.org/10.1145/290941.290953.

Yimam, D. 2000. Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR approach. *ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop*. MIT Press, 276–283.

Zhao, Ying, George Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Mach. Learn.* **55**(3) 311–331. doi:http://dx.doi.org/10.1023/B:MACH.0000027785.44527.d6.

Zhu, Shenghuo, Kai Yu, Yun Chi, Yihong Gong. 2007. Combining content and link for classification using matrix factorization. *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 487–494. doi:http://doi.acm.org/10.1145/1277741.1277825.