Topic Evolution and Social Interactions: How Authors Effect Research

Ding Zhou[1], Xiang Ji[2], Hongyuan Zha[1,3], C. Lee Giles[3,1]

Department of Computer
Science and Engineering[1]
The Pennsylvania State
University, University Park, PA

Yahoo! Inc.[2]
701 First Avenue,
Sunnyvale, CA

Information Sciences and
Technology[3]
The Pennsylvania State
University, University Park, PA

ABSTRACT

We propose a method for discovering the dependency relationships between the topics of documents shared in social networks using the latent social interactions, attempting to answer the question: *given a seemingly new topic, from where does this topic evolve?*. In particular, we seek to discover the pair-wise probabilistic dependency in topics of documents which associate social actors from a latent social network, where these documents are being shared. By viewing the evolution of topics as a Markov chain, we estimate a Markov transition matrix of topics by leveraging social interactions and topic semantics. Metastable states in a Markov chain are applied to the clustering of topics. Applied to the CiteSeer dataset, a collection of documents in academia, we show the trends of research topics, how research topics are related and which are stable. We also show how certain social actors, authors, impact these topics and propose new ways for evaluating author impact.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithm, Experimentation, Human Factors

Keywords

Clustering, Social Network Analysis, Text Data Mining, Markov chains

1. INTRODUCTION

Mining text documents has many basic tasks including topic classification [17], topic novelty detection [16, 22], and topic summarization [2, 14]. In particular, document summarization is often based on the discovery of topics (or themes). Methods applied to topic discovery in static documents encompass probabilistic modeling [2], matrix factorization [9], and entropy optimization [1]. More recent work has been concerned with temporal documents using a collection of incremental and efficient algorithms [14, 16, 22].

While there are a rich set of choices regarding topic discovery in set of temporally related documents, our concern is when and where these topics evolve and how the topics relate, if any, dependencies with each other. In Fig. 1, for example, we illustrate the probability of appearance in documents in CiteSeer of four research topics discovered using Latent Dirichlet Allocation [2], which is similar to previous topic trend discovery [20].

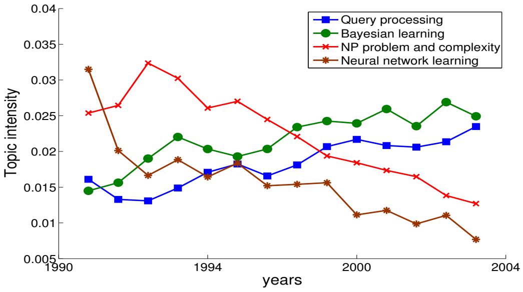


Figure 1: Probability of four research topics over 14 years in CiteSeer.

Some topics in Fig. 1 have been growing dramatically in popularity while other topics seem to be less popular, at least according to the CiteSeer database. A more interesting question is whether a newly emergent topic is truly new or rather a variation of an old topic? We address this by revealing the dependencies among the discovered topics. With the temporal dependencies among topics available, one can survey a research area from a genealogical graph of related topics instead of perusing the citation graphs.

One may assume that the discovery of pair-wise dependencies between topics can be achieved via document content analysis. A naive approach would be to cluster documents recursively yielding a hierarchical structure of topics. Alternatively, one might define a similarity metric between probability distributions (e.g. the Kullback-Leibler distance) if

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011 ...$5.00.

the topics discovered are represented in terms of probability distributions over words. In this paper, we introduce consideration of social factors into traditional content anlysis.

In order to interpret and understand the changes of topic dynamics in documents, we resort to discovering the *social reasons* of why a topic evolves and relates dependencies with others. We hypothesize that one topic evolves into another topic when the corresponding social actors interact with other actors with different topics in the latent social network. Consider an actor a_u associating a topic t_i at time k. For some reason, this actor meets and establishes a social tie with actor a_v who is mostly associated with a new topic t_j and they start to work on the new topic with a higher probability. At a later time $2k$, we observe that a_u is more likely to be concerned with t_j rather than the previous t_i. In return, t_j has received a higher popularity than t_i. When such a switch of topics in actors is statistically significant, we will observe an aggregate transition tendency from t_i to t_j in the topic dynamics, yielding the marginal probabilities of t_i and t_j moving towards different directions over time, as illustrated in Fig. 1. Such a trend is defined as transition between topics. The dependency of t_i on t_j is measured by the probability of transition from t_j to t_i. Abstracted from the above example, our goal seeks to estimate the dependencies between topics using social interactions.

We consider this work as an attempt to bridge the dynamics of topics in documents with the latent social network. In particular, we hypothesize the changes of topics in documents in a social network as a Markov chain process which is parametrized by estimating the social interactions among individuals conditioned on topics. The primary assumption is that topics in *social documents* evolve due to the development of the latent social network. This assumption has been supported in recent work [7] which experimentally shows that the information diffusion in a social network creates new topics in the Web blog space.

Our contributions are: (1) a model of the topic dynamics in *social documents* which connect the temporal topic dependency with the latent social interactions; (2) a novel method to estimate the Markov transition matrix of topics based on social interactions of different order; (3) the use of the properties of finite state Markov process as the basis for discovering hierarchical clustering of topics, where each cluster is a Markov metastable state; (4) a new topic-dependent metric for ranking social actors based on their social impact. We test this metric by applying it to CiteSeer authors.

The rest of the paper is organized as follows: in the next section, we introduce related research. The definitions and our problem framework are given in § 3. § 4 describes our probabilistic modeling of topic dynamics and the author/topic dependency network. In § 5 we introduce the maximum likelihood estimation algorithm for parameterizing the Markov chain model. § 6 interprets the metastable state discovery in the Markov chain as an approach for topic clustering. Experimental results are presented in § 7 ending the conclusions and future work in § 8.

2. RELATED WORK

2.1 Document Content Analysis

A variety of statistical approaches have been proposed to model a text document. The unigram model models each document with a multinomial distribution and the words in the document are independently drawn from the multinomial distribution [17]. It argued that each document in the document corpus has a distinct topic and developed mixture of unigrams based on the unigram. The mixture of unigrams models each document by considering the words in a document as generated from the conditional probability $p(w|t)$, where t is the topic of this document.

The probability latent semantic analysis (pLSA) [9] has each document generated by the activation of multiple topics, and each topic is modeled as multinomial distributions over words, which relaxes the mixture of unigrams model which considers each document as generated from only one topic. However, pLSA model uses a distribution indexed by training documents, which means the the number of parameters being estimated in a pLSA model must grow linearly with the number of training documents. This suggests that pLSA could be prone to overfitting in many practical applications.

Latent Dirichlet Allocation (LDA) [2] addresses the overfitting of pLSA by using the Dirichlet distribution to model the distribution of topics for each document. Each word is considered sampled from a multinomial distribution over words specific to this topic. As an alternative, the LDA model is a well-defined generative model and generalizes easily to new documents without overfitting. Recent work has been concerned with discovering dynamics in topics [14] and incremental discovery [16].

2.2 Social Network Analysis

The social network analysis is an established field which proposes to analyze the relationships between social actors in a social network [21]. The evolution of the Web and the wide usage of socially related e-formated communication media (e.g. emails, blogs) has promoted the interest in computational social network analysis, such as web community discovery [5]. In the Referral Web project, social networks were mined from online information and used to help users target experts who can answer their questions with geographical proximity [10].

The above research mostly focused on static properties of social networks. However, social networks are dynamic in essence and evolve over time. The dynamic property of a social network greatly impacts evolution of the communications content among its social actors, usually in terms of *social documents* [23].

The dynamics of social ties in a social networks can be shown by tracking the changes in large-scale data by periodically clustering data and examining the extracted temporal clusters [11]. Link structures can be used to predict future interactions among social actors [12]. Based on the assumption that the observed social networks are outcomes of a Markov process evolving in continuous time, models of the changes in the social ties [19] can be derived.

Despite of these traditional emphasis on structural approaches, content-based analysis of social networks is only a recent trend [23, 7]. In fact, the content of documents shared in social networks, or *social documents*, embraces the valuable information of both the changes of topics and the developments of social interactions. Mining *social documents* to interpret and understand the changes of dynamics in documents as well as the dynamics of social ties is becoming an important direction for computational SNA.

This paper attempts to bridge the changes in social ties with the changes in their communication patterns. We address the discovery of dynamics in social communications and how these dynamic changes link to the latent social interactions.

3. PROBLEM DEFINITION

3.1 Social Networks

The primary goal of this paper is the discovery of pair-wise dependencies in the topic dynamics from a *social document* corpus. The definition of a *social document* is based on a *social network*:

DEFINITION 1. *A Social Network (SN) is defined as a homogeneous network (graph), in which each vertex denotes an individual, also known as a Social Actor (SA), and each edge represents a certain relationship between two SA's.*

A typical SN instances encountered everyday include the SN of authors, Web blog owners or Email users. The social ties between two SA's can be recognized in a variety of ways depending on the application settings. For example, the collaboration between authors can be seen as one social tie between authors.

A SN in real world is not isolated. It is always preferable to perform the analysis of a SN by studying the corresponding information carrier (a Social Document), e.g. the Email text in Email SNs, which motivates and maintains the social ties in a SN. *Social documents* are defined as:

DEFINITION 2. *A social document is a document composed of a set of SAs in a SN for the purpose of exchanging information or soliciting future social ties.*

A *social document* collection embodies valuable information regarding the social ties in the hidden SN and defines the social ties for modeling the topic dynamics in a *social document* corpus.

With the many document summarization tools previously developed [17, 9, 2], each *social document* can be seen as a mixture of topics. By a topic, we mean a certain probability distribution over the document vocabulary. The topic dynamics refer to the series of topic with various strength of probabilities over time.

3.2 Problem Formalization

Given the definitions in § 3.1, we outline our problem setting and the solution framework. Consider the evolution of topics in a social document corpus as a Markov chain. Per definition, the description of every finite dimensional Markov chain includes the specification of a finite state set and a Markov transition matrix. In the *social document* setting, we first recognize the topics from the corpus and consider each topic as a state in the Markov chain whose marginal probability can be estimated. As such, we need to address the estimation of the Markov topic transition matrix in the topic dynamic system spanned by a *social document* stream.

For a formal definition of the problem, denote the *social document* stream as a matrix $\mathcal{DW} \in \mathbb{R}^{D\times W}$, where D is the number of documents and W the number of words. Define the matrix $\mathcal{DA} \in \mathbb{R}^{D\times A} = \{\lambda_{i,j}\}^{D\times A}$ denoting the creators of these documents, where A is the number of social actors and $\lambda_{i,j} = \mathbf{1}(d_i, a_j)$, an indicator function of whether document d_i is composed by actor a_j. Note that one document may have several actors. (For our experiments actors will be denoted as authors.)

Using the summarization tools (LDA [2]) we transform $\mathcal{DW}$ into $\mathcal{DT} \in \mathbb{R}^{D\times T}$, where T is the number of prespecified topics. We assume that $\mathcal{DT}$ is normalized by row such that each document is a distribution over topics.

Using the matrix $\mathcal{DA}$, a collaboration matrix $\mathcal{A}$ is obtained by setting $\{\alpha_{i,j}\}^{A\times A} = \mathcal{A} = (\mathcal{DA})^t DA$, where $\alpha_{i,j}$ denotes the number of collaborations between social actors a_i and a_j if $i \neq j$ and the number of composed documents by a_i if $i = j$. Let the author set be Λ. Using matrix $\mathcal{DT}$ and $\mathcal{DA}$, we obtain a set $\mathcal{Q} = \{\langle \mathbf{a}, \mathbf{t}\rangle | \mathbf{a} \subseteq \mathbf{\Lambda}, \mathbf{t} \in \mathbb{R}^{\mathbf{1}\times\mathbf{T}}\}$, where $\mathbf{a}$ is the set of authors on a document and $\mathbf{t}$ is the distribution over topic specifying this document. Here each element q_i in $\mathcal{Q}$ denotes an observation of a document.

Now the problem becomes, given a set $\mathcal{Q} = \{\langle \mathbf{a}, \mathbf{t}\rangle | \mathbf{a} \subseteq \mathbf{\Lambda}, \mathbf{t} \in \mathbb{R}^{\mathbf{1}\times\mathbf{T}}\}$ and $\mathcal{A} \in \mathbb{R}^{A\times A}$ that can be calculated from $\mathcal{Q}$, find a Markov transition matrix $\Gamma \in \mathbb{R}^{T\times T}$ that captures the dependencies among the discovered topics, i.e. determine a function Ψ such that $\Gamma = \Psi(\mathcal{Q}, \mathcal{A})$.

4. SOCIAL INTERACTIONS & MARKOV TOPIC TRANSITION

In our setting, where topics are those discovered from *social documents*, we propose a measurement method that accentuates the social interactions in the latent SN in order to estimate the topic transitions. The function Ψ determines the measurement of pair-wise dependencies between topics.

Namely, we limit our search for Ψ to consider only the social interactions mediating the evolution of topics. The assumption is supported by the intuition that topics created by close *social actors* in a SN sense[21] represent greater dependencies than those created randomly. For example, a topic t_a is more likely to be dependent on t_b if the social actors found in t_a are tightly connected to those social actors found in t_b. The idea here is similar to but different from that of *collaborative filtering* [15] in that now heterogeneous social ties are taken into consideration.

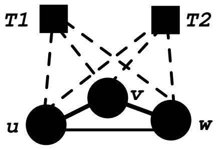


Figure 2: Different dependency networks among two sets of variables: *topics (squares)* and *social actors (circles)*.

The estimation of Markov topic transition matrix Γ breaks down to a set of estimation tasks each in the form of $P(t_i|t_j)$, which denotes the probability that topic t_j transits to t_i in the Markov chain process.

In order to estimate $P(t_i|t_j)$ using social ties in a SN properly, we first set up the probability independence among two sets of variables: topics and social actors. Let Fig. 2 illustrate our assumptions of the dependencies of variables. The topics are assumed to be of no direct dependency between each other. The social actors are assumed to be pair-wise dependent. For two social actors with no relationships, we can consider their dependency as zero. In Fig. 2, we show that three actors u,v and w are socially connected (solid lines). The two topics associated with them, T_1 and T_2 respectively, are linked (dashed lines) to all actors.

Following the above, consider the joint distribution $P(t_i, t_j)$ resulting from the interactions in the latent SN. In particular, we consider the social interaction bounded by order two, i.e. $P(t_i, t_j)$ is constrained by single self and pairs of social interactions only, respectively denoted by $P(t_i, t_j)^{(1)}$ and $P(t_i, t_j)^{(2)}$. This can be denoted by:

$$P(t_i, t_j) = \gamma P(t_i, t_j)^{(1)} + (1-\gamma)P(t_i, t_j)^{(2)} \quad (1)$$

$$= \gamma \sum_{1\leq u\leq A} P(t_i, a_u, t_j) + (1-\gamma) \sum_{1\leq u,v\leq A} P(t_i, a_u, a_v, t_j) \quad (2)$$

where a_u and a_v are social actors in the underlying SN. γ is a smoothing parameter that weighs the importance of 1st-order social interactions. Eq. 2 assumes independence when estimating $P(t_i, a_u, t_j)$ and $P(t_i, a_u, a_v, t_j)$.

Note the assumption above regarding the order of social interaction can be relaxed to deal with higher order. We leave it to the readers to generalize Eq. 2 in high order case.

4.1 Multiple orders of social interactions

Multiple types of social ties can be considered as a basis for determining the estimation of the topic transition probability. In this subsection, we provide a solution to the estimation problem based on social interactions, one typical social tie in a SN, with different orders. Denote the measurements based on 1st-order and 2nd-order social interactions respectively by $P(t_i, t_j)^{(1)}$ and $P(t_i, t_j)^{(2)}$. We focus on deriving $P(t_i, a_u, t_j)$ and $P(t_i, a_u, a_v, t_j)$ estimation formulas from our social interaction considerations.

First, we consider the estimation of $P(t_i, a_u, t_j)$ as a 1st-order social interaction. We illustrate the 1st-order probability dependence between topics and social actors in Fig. 3. The social actor u is present in both topics T_1 and T_2.

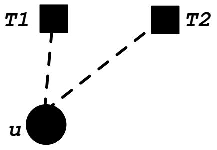


Figure 3: 1st-order probability dependence between topics and a social actor.

We can estimate $P(t_i, a_u, t_j)$ by Eq. 3:

$$P(t_i, a_u, t_j) = P(t_i|a_u, t_j)P(a_u|t_j)P(t_j) \quad (3)$$

We derive Eq. 3 using the chain rule for a joint probability. Based on the assumption of the dependency network among a_u, t_i and t_j as illustrated in Fig. 3, we obtain the joint probability $P(t_i, a_u, t_j)$ as a chain of probabilities:

$$P(t_i, a_u, t_j) = P(t_i|a_u)P(a_u|t_j)P(t_j) \quad (4)$$

The intuition behind the 1st-order social interaction is that a new topic may be initiated by the same actor who is present in an older topic but without collaboration with any other social actors.

Second, we discuss the estimation of $P(t_i, a_u, a_v, t_j)$ considering the 2nd-order social interaction (a dyad in SN notation [21]). The 2nd-order probability dependency between topics and social actors is presented in Fig. 4. Here we introduce the pair-wise interaction in the latent SN as the motivation for the evolution of topics.

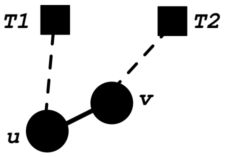


Figure 4: 2nd-order probability dependence between topics and social actors.

Again, consider the joint distribution $P(t_i, t_j)$ as being constrained by the relationship between two social actors a_u and a_v to the 2nd-order, as measured by $P(t_i, a_u, a_v, t_j)$. The constraint is captured in Eq. 1 by $P(t_i, t_j)^{(2)}$. These 2nd-order SN interaction constraints can be seen as the sum of the joint probabilities $P(t_i, a_u, a_v, t_j)$, which is represented as:

$$P(t_i, t_j)^{(2)} = \sum_{1\leq u,v\leq A, u\neq v} P(t_i, a_u, a_v, t_j). \quad (5)$$

We factorize the joint probability $P(t_i, a_u, a_v, t_j)$ in Eq. 6 to Eq. 8 using chain rule:

$$P(t_i, a_u, a_v, t_j) = P(t_i|a_u, a_v, t_j)P(a_u, a_v, t_j) \quad (6)$$

$$= P(t_i|a_u, a_v, t_j)P(a_u, a_v|t_j)P(t_j) \quad (7)$$

$$= P(t_i|a_u, a_v, t_j)P(a_u|a_v, t_j)P(a_v|t_j)P(t_j) \quad (8)$$

Based on the independence assumption in Fig. 4, we arrive at a new form of $P(t_i, a_u, a_v, t_j)$ as:

$$P(t_i, a_u, a_v, t_j) = P(t_i|a_u)P(a_u|a_v)P(a_v|t_j)P(t_j) \quad (9)$$

where $P(a_u|a_v)$ can be seen as the conditional probability that social actor a_u interacts with another social actors a_v.

Note that the idea of relating the evolution of topics in SNs with the various orders of social interactions naturally coincides with the assumption that "collaborations bring about new topics".

The assumptions we made about the independence networks in 1st- and 2nd-order social interactions help the derivations of Eq. 4 and Eq. 9. In traditional topic discovery methods where social factors are not considered (e.g. LDA), topics are assumed to be unconditionally independent from each other. Thus we see the assumptions of our approach as weaker and relaxed in conditions.

4.2 Markov transition

With the derivation for the joint probability of two topics with 1st- and 2nd-order social interactions, the estimation for Markov transition matrix, Γ, becomes straightforward. In particular, we define $\Gamma \in \mathbb{R}^{T\times T}$, where each element $\Gamma_{i,j}$ quantifies the transition probability from t_j to t_i. Then, $\Gamma_{i,j}$ is the conditional probability $P(t_i|t_j)$:

$$\Gamma_{i,j} = P(t_i|t_j), \quad \text{where } \Gamma \in \mathbb{R}^{T\times T}. \quad (10)$$

where the transition probably is a directional estimate such that $\Gamma_{i,j}$ does not necessarily equal to $\Gamma_{j,i}$. We assume Γ will be normalized by row such that the row elements sum to one.

Next we revisit the estimation of the joint probability $P(t_i, t_j)$ in Eq. 1 for the estimation of $P(t_i|t_j)$. Using Bayes rule, we have $P(t_i|t_j) = \frac{P(t_i,t_j)}{P(t_j)}$. Substituting this into Eq. 1, we obtain $P(t_i|t_j) = \frac{\gamma P(t_i,t_j)^{(1)} + (1-\gamma) P(t_i,t_j)^{(2)}}{P(t_j)}$. According to Eq. 3 and Eq. 8, we rewrite Eq. 2 as:

$$P(t_i|t_j) = \frac{\gamma \sum_u P(t_i, a_u, t_j) + (1-\gamma) \sum_{u,v} P(t_i, a_u, a_v, t_j)}{P(t_j)} \quad (11)$$

$$= \gamma \sum_{1 \le u \le A} P(t_i|a_u) P(a_u|t_j) + (1-\gamma) \sum_{1 \le u,v \le A, u \ne v} P(t_i|a_u) P(a_u|a_v) P(a_v|t_j) \quad (12)$$

So far we have given analytical formulas for $P(t_i|t_j)$ which are required for deriving the Markov transition matrix Γ. However, we still cannot write out the closed form solution from the observations for $P(t_i|t_j)$ because of the unknown quantities in Eq. 12, such as $P(t_i|a_u)$, $P(a_u|t_i)$ and $P(a_u|a_v)$. In the next section, we derive the algorithmic estimation of these quantities.

5. MODEL ESTIMATION

Here we estimate $P(t_i|a_u)$, $P(a_u|t_i)$ and $P(a_u|a_v)$ required to obtain the Markov transition matrix Γ. Remember that our goal is to estimate Γ given the document set $\mathcal{Q}$, which we assume is already sorted by time.

An algorithm that estimates a solution to the function Ψ uses the Maximum Likelihood Estimation (MLE) for the collaboration matrix $\mathcal{A}$ as well as the set of conditional probabilities required for Eq. 12. Let Λ be the author set. Fig. 5 illustrates the algorithm which has three phases: (1) initialization; (2) training, and (3) estimation.

In particular, $\mathcal{A}$ and $\mathcal{T}$ are matrices recording the co-occurrence of authors and author-topic pairs. We update $\mathcal{T}$ by adding the probabilities of each topic to the corresponding row of $\mathcal{T}$ (in step 5). The same idea is applied when incrementing the "count" for each topic in step 7. Step 6 increments the count of author a_i. Author collaborations are recorded by step 8 - step 9. Note that the estimation phase (step 10 - step 12) is not needed in training and can be carried out whenever estimations are required. We design the training algorithm in an incremental counting manner so that online estimations becomes easy to compute.

Let us now consider the analytical complexity of training and estimation. For training, step 3 - step 9 consumes a computation complexity of $O(NLT + NL^2)$ on average, where L is the average length of author list on a document, usually a small integer. N and T are number of documents and number of topics. The overall computational complexity to obtain the estimation of $P(t_i|t_j)$ via Eq. 12 is further enhanced by a ratio of A for 1st-order and A^2 for 2nd-order dependence estimation. The $P(t_i|t_j)$ then costs $O((NLT+NL^2)(A+A^2))$, which is bounded by $O(A^2NLT)$.

```
Input: observation set $\mathcal{Q} = \{\langle \mathbf{a}, \mathbf{t} \rangle | \mathbf{a} \subseteq \mathbf{\Lambda}, \mathbf{t} \in \mathbb{R}^{\mathbf{1} \times \mathbf{T}}\}$
Output: estimations of $P(t_i|a_u)$, $P(a_u|t_i)$ and $P(a_u|a_v)$

     //initialization
(1)  $\mathcal{A} \leftarrow 0^{A \times A}$, $\mathcal{T} \leftarrow 0^{A \times T}$
(2)  $\mathbf{c_A} \leftarrow 0^{1 \times A}$, $\mathbf{c_T} \leftarrow 0^{1 \times T}$

     //training
(3)  for $\langle \mathbf{a}, \mathbf{t} \rangle$ in $\mathcal{Q}$ //topic-author counting
(4)      for $a_i$ in $\mathbf{a}$
(5)          $\mathcal{T}_{i,:} \leftarrow \mathcal{T}_{i,:} + \mathbf{t}$
(6)          $\mathbf{c_A}^{\{i\}} \leftarrow \mathbf{c_A}^{\{i\}} + 1$
(7)          $\mathbf{c_T} \leftarrow \mathbf{c_T} + \mathbf{t}$
(8)          for $a_j$ in $\mathbf{a}$ //author-author counting
(9)              $\mathcal{A}_{i,j} = \mathcal{A}_{i,j} + 1$

     //estimation
(10) $P(t_i|a_u) \leftarrow \mathcal{T}_{u,i} / \mathbf{c_A}^{\{u\}}$
(11) $P(a_u|t_i) \leftarrow \mathcal{T}_{u,i} / \mathbf{c_T}^{\{i\}}$
(12) $P(a_u|a_v) \leftarrow \mathcal{A}_{u,v} / \mathbf{c_A}^{\{v\}}$
```

Figure 5: MLE for $P(t_i|a_u)$, $P(a_u|t_i)$ and $P(a_u|a_v)$

6. MARKOV METASTABLE STATE DISCOVERY

Now we have topic and topic-topic dependencies respectively estimated as the system states and the stochastic transition probability of a Markov chain. We will explore other topic discovery using well established methods in Markov analysis [18]. This section describes the discovery of metastable states [3] in a Markov chain as an approach to identifying hierarchical clustering of topics.

Consider a Markov chain with its transition matrix P, state set S with the marginal distribution of S as π. Let $A \subseteq S$, $B \subseteq S$ be two subsets of S. Then the transition probability from B to A with respect to π is defined as the conditional probability from B to A:

$$\omega_\pi(A|B) = \frac{\sum_{a \in A, b \in B} \pi_a p_{a|b}}{\sum_{b \in B} \pi_b} \quad (13)$$

where a,b are dummy variables denoting the states in S.

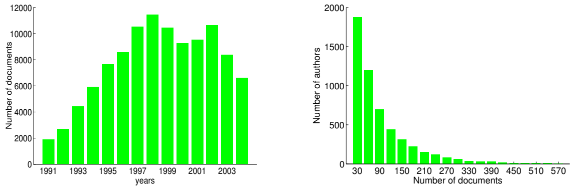
Let A_1,..., A_K be disjoint K subsets of S. We define a new $K \times K$ transition matrix $W = \{\omega_\pi(A_i|A_j)\}_{ij}$ as described above. Thus we arrive at another Markov chain with dimensionality reduced to K in which each state now is an aggregate of the unit states from the previous state space.

Markov chains are called *nearly uncoupled* if its state space can be decomposed into several disjoint subsets $\mathbf{A}$ such that $\omega_\pi(A_i|A_j) \approx 1$ for $i = j$ and $\omega_\pi(A_i|A_j) \approx 0$ for $i \ne j$. Each aggregate in a *nearly uncoupled* Markov chain M is called a *metastable state* of M. In our setting, a metastable state in Γ is a cluster of topics. Recursively discovering the metastable states[3], we may obtain a hierarchical clustering of topics that capture their taxonomy. Identification of the metastable states in a Markov chain has been studied extensively [4, 3]. In numerical analysis, the identification can be viewed as a process which seeks the matrix permutation such that the transition matrix is as block diagonal as possible; a method [4] we also use.

7. EXPERIMENTS

7.1 Data preparation

For experiments, we use data from CiteSeer [6], a popular online search engine and digital library which currently has a collection of over 739,135 academic documents in Computer Sciences, most of which were obtained by web crawling and the others by author submissions. The documents have 418,809 distinct authors after name disambiguation. Each document is tagged with a time-stamp giving the parsed time of the first crawled date.



(a) Number of documents versus year acquired

(b) Authors associated with number of documents.

Figure 6: Statistics of the sample CiteSeer.

We associate each document with the list of disambiguated authors [8]. Then we construct a co-authorship graph where two nodes share an edge if they ever coauthored a document. Next we perform breadth-first-search search on the co-authorship graph from several predefined well known author seeds until the graph is completely connected or there are no new nodes. For seeds selection, we choose two researchers with a large number of publications in CiteSeer, Michael Jordan and Jiawei Han, from statistical learning and data mining and database respectively. The constructed subgraph of authors is further pruned by eliminating the authors with less than 50 publications in CiteSeer over the last fourteen years. We end up with a sampling of CiteSeer containing 3,974 authors and 108,676 documents spanning from 1991 to 2004. The number of documents acquired w.r.t years is illustrated in Fig. 6(a). We observe that the number of documents written by individual authors follows a *power law* distribution (Lotka's law) [13].

7.2 Discovered topics

We train a Latent Dirichlet Allocation (LDA) model over our entire sample collection of CiteSeer by setting the topic number as $T = 50$, resulting in 50 discovered topics illustrated in Table 1. The setting of desired topic number is small because we only work on a small subset of authors in CiteSeer (3,974 authors out of 418,809). Due to the limited space, we cannot present all the automatically extracted top words for all topics. Instead, we manually tag all the topics with labels using ranked keywords in the words.

For a more detailed description of some topics, in Table 4, we give a sample of six topics from Table 1 and their top words. Here the last row is manually labeled to summarize the topics. We are able to observe that LDA easily discovers the topics from a variety of areas [1].

After the models are trained, we re-estimate each document with the LDA model to obtain the mixture of topics for each document. We further normalize the weights of the mixture components. It should be noted that this permits us to track the topic over time using some recently proposed online methods(e.g. [16]).

7.3 Topic trends

We visualize the four topic dynamics w.r.t. time in Fig. 7. Given a year, the strength of a topic is calculated as the normalized sum of all the probabilities of this topic inferred for all documents in this year. The topics trend is an indicator of the trend of interests in *social documents* and in our setting, the research interest trends.

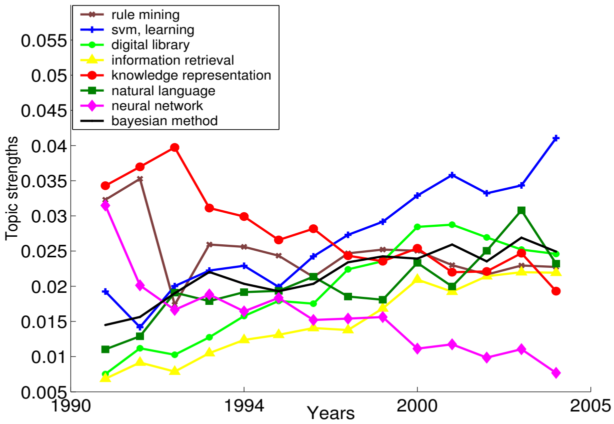


Figure 7: Topic probability of eight topics over 14 years in CiteSeer.

The eight topics we choose to plot are (1) query processing (Topic 02); (2) svm learning (Topic 08); (3) digital library (Topic 019); (4) information retrieval (Topic 026); (5) knowledge representation (Topic 032); (6) natural language processing (Topic 038); (7) neural network learning (Topic 046), and (8) Bayesian learning (Topic 050) (similar results are in[20]).. This raises the question of *where do the researchers in a declining trend go (ex. neural networks)? Do they switch to new topics, and which topics?* Our next goal is to automatically extract the dependencies among these discovered topics.

7.4 Markov topic transition

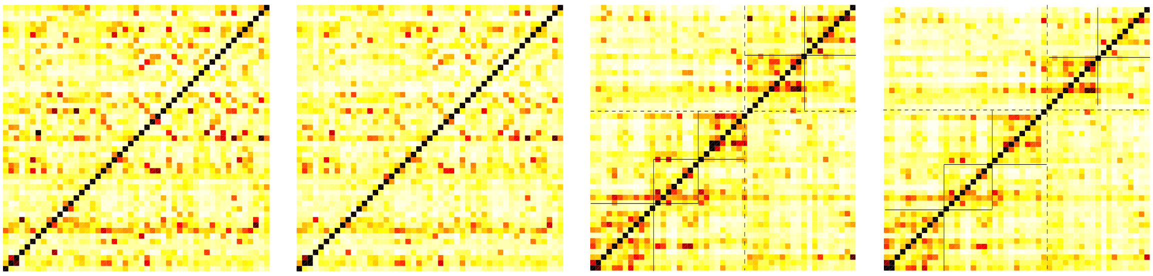
In order to explore the temporal dependencies among a group of discovered topics, we identify the Markov topic transition matrix via maximum likelihood estimation of the 1st- and 2nd-order constraints brought about by the hidden social interactions of authors (interactions of single social actors, or collaboration between social actor pairs).

The Markov transition matrices Γ are shown in Fig. 8(a) and Fig. 8(b) to highlight the extraction of metastable topic states. The values of matrix entities are scaled with the color intensity with the darker color denoting large value. Fig. 8(a) and Fig. 8(b) visualize the Γ with 1st-order and 2nd-order social relationship, before block diagonalization.

[1] note that Topic 17 denotes the affiliation and venues in which the keywords are *university, department, email, conference, proceedings, etc* which are also considered as topics since there was no deliberate removal of such information from the title/abstracts in CiteSeer.

Table 1: Topics discovered with manual labels.

Topic #	manual namings	Topic #	manual namings
0	real-time system, performance	25	network traffic congestion control, protocols
1	rule mining, database	26	document retrieval, search engine
2	database query processing	27	language, automation machine
3	communication, channel capacity	28	mathematical derivation, proof
4	information theory	29	image segmentation, computer
5	programming language, compiler	30	multimedia, video streaming
6	scheduling, queueing	31	statistical learning theory
7	software engineering, system development	32	knowledge representation, learning
8	svm, learning, classification	33	protein sequence, dna structure
9	signal processing	34	robotics
10	ai, planning	35	system kernel, unix
11	matrix analysis, factorization	36	security, cryptography
12	dynamic flow control	37	mobile network, wireless protocols
13	dimension reduction, manifold	38	natural language, linguistic
14	decision tree, learning	39	np problem, complexity
15	numerical optimization	40	network package routing
16	mobile network, energy	41	user agents, interface
17	affiliation and venues	42	geometry, spatial objects
18	object oriented design	43	parallel processing
19	digital library services, web	44	distributed computing, network infrastructure
20	os cache strategy design	45	system architecture
21	circuit design	46	neural network, learning
22	concurrent control, distributed system	47	graph algorithms, coloring
23	game and marketing	48	linear programming
24	algorithm complexity	49	bayesian method, learning



(a) Transition under 1st-order interaction

(b) Transition under 2nd-order interaction

(c) Transition under 1st-order interaction after block diagonalization

(d) Transition under 2nd-order interaction after block diagonalization

Figure 8: Markov transition matrices before and after block diagonalization

From Fig. 8(a) and Fig. 8(b), we observe that Γ is a sparse matrix, with large values in diagonal elements. The sparseness shows that these topics are separate though some transitions among them exist. The large diagonal values indicates that the discovered topics in our case are relatively stable with mostly transitions to themselves. Authors in our CiteSeer sample prefer to remain in their own topics rather than switching between topics.

While the separateness among topics is for future investigation, we now take a closer look at the diagonal elements. Diagonal elements in Γ indicate the probability that an author (and author pair collaboration as well) will continue to work on the same topics over time. This *self-transition probability* shown in Fig. 9 allows us to rank the topics according to the authors reluctance to change topics.

Note that Topic 17 (affiliation and venue info.) is that with largest self-transition probability. The rational is obvious since most authors tend to continue including their af-

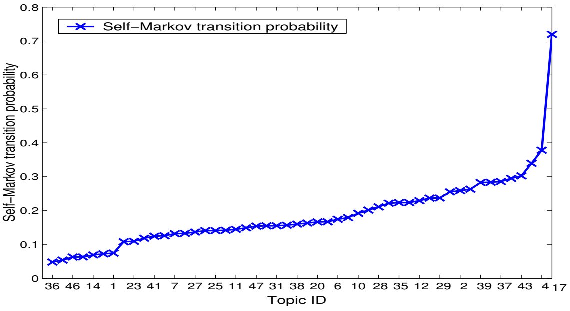


Figure 9: The self-transition probability ranking of topics. Topics with high probability are more stable.

Table 2: Discovery of mTopics via block diagonal Markov transition matrix.

#	Topic IDs										
mT_1	1	2	8	10	18	19	23	26	32	38	41
mT_2	0	5	7	20	21	22	27	28	35	43	45
mT_3	6	25	30	36	37	40	44				
mT_4	13	14	15	17	24	31	33	39	42	47	48
mT_5	3	4	9	11	12	16	29	34	46	49	
mT_1	data management, data mining										
mT_2	system, programming language, and architecture										
mT_3	network and communication										
mT_4	numerical analysis, machine learning										
mT_5	statistical methods and applications										

filiation/venue information which was part of the meta data used. In addition, we can see that generally the topics with heavy methodology requirements (e.g. np problem, linear system) and/or popular topics (e.g. mobile computing, network) are more likely to remain stable. By contrast, topics closely related to applications are more likely to have higher transition probabilities than other topics (e.g. data mining in database, security) all things being equal.

Second, in order to investigate the sparseness in matrix Γ, we perform metastable state recognition (introduced in $ 6), viewing Γ as the adjacency matrix of the Markov transition graph. In particular, we permute Γ in such a way that Γ is approximated by a block diagonal matrix. The resultant $\hat{\Gamma}$ is illustrated in Fig 8(c) and Fig. 8(d), on 1st-order and 2nd-order consideration of social relationship respectively.

The metastable states have in effect reduced the original Markov transition process to a new Markov process with fewer states and each diagonal block can be seen as a metastable state [3] which is a cluster of topics with tight intra-transition edges.

From Fig 8(c) and Fig. 8(d), we are able to initially break the two $\hat{\Gamma}$ into two major blocks, as noted by the dashed lines. Recursively, we can arrive at five smaller blocks, illustrated by solid lines, with each block as a metastable topic (or mTopic). Even though there exists a transition between topics, the transitions are more likely to occur within a metastable topic rather than between them. Table 2 gives the list of mTopics and the corresponding topics.

Comparing Table 2 with Table 1, we observe that the topic descendants discovered readily capture natural intuitions of the relationships among topics. Specifically, mTopics mT_1 includes topics on data management and data mining; mT_2 includes programming language, system and architecture; mT_3 covers network and communication; mT_4 covers machine learning and numerical analysis; mT_5 are mainly statistical methods.

7.5 Transition within metastable topics

With metastable topics (mTopic) discovered according to the approach introduced in § 6, we are able to compute the accumulated transition probability among mTopics. Fig. 10 illustrates the uncoupled Markov transition graph among five mTopics we have discovered from the original stochastic matrix. Transitions with probability lower than 0.16 are hidden from the graph to clarify the major transition among the five mTopics. Such transition probabilities among metastable topics are useful information for understanding the trends of topics and their dependencies in social document corpora.

Comparing Fig. 10 with the descriptions of each mTopic in Table 2, we can outline the major dependencies between mTopics. Out data indicates that mT_4 (numerical analysis) has been essential in these mTopics. And there is a transition to mT_5 (statistical methods) and which is tightly coupled with research in mT_1 (data management and data mining). Results also imply that researchers in mT_3 (networks) will be concerned with mT_2 (systems) and that data management research is coupled with systems issues due the high mutual transition probability between mT_1 and mT_2.

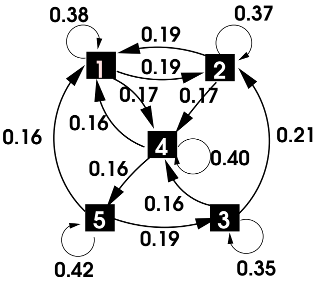
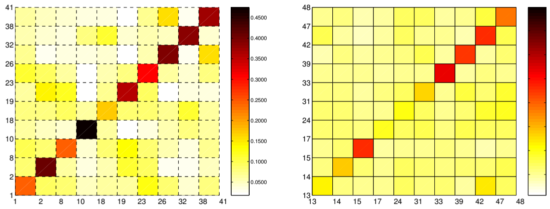


Figure 10: The Markov transition graph among mTopics. Transitions with probability lower than 0.16 are not shown.

Next we look at the transitions within these metastable topics. Now that we know the topics within a metastable topic (mTopic) are very less likely to jump across mTopics, questions may be asked about how tightly the topics in the same metastable state are aggregated. We present the stochastic matrices of mT_1 and mT_4 in Fig. 11(a) and Fig. 11(b).



(a) mTopic: mT_1

(b) mTopic: mT_4

Figure 11: The Markov transition structure in metastable topics

We observe that diagonal elements show the existence of high self-transition probabilities and that both matrices are almost symmetric, meaning the pairwise transition between topics in the same mTopic are largely balanced.

7.6 Who powers the topic transition

If one accepts the above interpretation of Markov transitions among the topics discovered in social document collections, a natural question to ask is what author or authors cause such a transition between topics, evaluating their roles as prominent social actors.

In particular, in the CiteSeer data setting, we seek to provide rank of authors based on their impact on the transition from one topic to another. We give a new metric $\delta(a_u)$ for the author impact ratio of a_u as measuring the difference between the obtained $P(t_i|t_j)$'s, with and without a_u.

Table 3: Top ranked authors according to their impact on three topic transitions.

T2→T1	T49→T26	T1→T33
Jiawei Han	W. Bruce Croft	Mark Gerstein
Jennifer Widom	David Madigan	Heikki Mannila
Timos Sellis	Norbert Fuhr	Mohammed Zaki
Dimitris Papadias	Andrew Mccallum	Limsoon Wong
Hans-peter Kriegel	James Allan	George Karypis
H. V. Jagadish	Thomas Hofmann	Jiawei Han
Jeffrey Naughton	John Lafferty	Susan Davidson
Divesh Srivastava	Hermann Ney	Dennis Shasha
Amr El Abbadi	Michael I. Jordan	Serge Abiteboul
Philip S. Yu	Ronald Rosenfeld	Jignesh M. Patel

Formally, consider how the transition probabilities change if an author a_u does not exist. Denote the estimation of $P(t_i|t_j)$ without a_u as $P(t_i|t_j)_{-a_u}$. One can then measure the importance of a_u w.r.t. topic $t_j \rightarrow t_i$ as $\delta(a_u, t_j \rightarrow t_i)$:

$$\delta(a_u, t_j \rightarrow t_i) = P(t_i|t_j) - P(t_i|t_j)_{-a_u}. \tag{14}$$

The new author ranking differs from previous ranking by citation counting, currently done in CiteSeer, Google Scholar, and ISI, by now incorporating social interactions while ranking social actors. In addition, the new ranking is dependent on the specified topic pairs thus quantifying the impact of social actors w.r.t. certain field(s).

Next, we choose all pairs of topics from the 50 discovered topics in our data and test our hypothesis. This ranking of social actors captures common knowledge of the importance of these social actors w.r.t. to different fields. Due space limitations, we select three topic transition instances ($T2 \rightarrow T1$, $T49 \rightarrow T26$, and $T1 \rightarrow T33$) and present the corresponding top ten ranked CiteSeer authors in Table 3.

We observe that many commonly believed influential researchers in data management to data mining ($T2 \rightarrow T1$), Bayesian learning to search engine ($T49 \rightarrow T26$), and rule mining to bioinformatics ($T1 \rightarrow T33$) are well ranked [2].

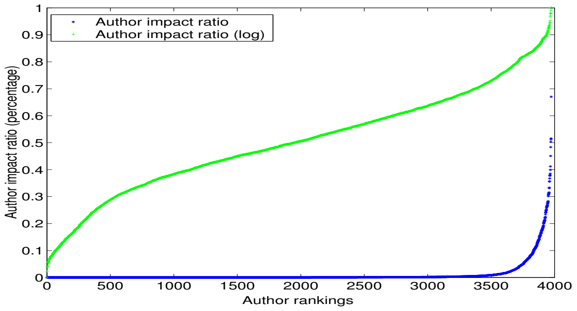


Figure 12: Ranking of authors w.r.t. their impact on the transition from Topic 02 to Topic 01.

Finally we give the distribution of impact over all authors in Fig. 12 for the transition of topic 02 to 01. The impact distribution is a power law, indicating that only a few social actors have large effects over a certain topic transitions.

7.7 Order of social interactions

It is interesting to consider what can be obtained by investigating social relationships with higher order. For that purpose, we compare the entropy of matrices [20] based on 1st- and 2nd-order social interactions. We can see that as the orders increase, the entropy of the matrix increases as well. This observation shows that the separation in social actors generally decreases as the collaboration social network increases. In addition, we can see that high-order effects of the social ties does not really help identify topic transitions. Here we limited our consideration of social orders to two.

	1st order	2nd order
Entropy	259.35	399.82

8. CONCLUSION

We develop new methods for relating social actors to their associated social topics and use them to derive topic trends. We show that certain topics are stable while others have a tendency to change over time. Certain social actors can be shown to play more important roles than others in topic transitions.

In particular, we model the topic dynamics in a social document corpus as a Markov chain and discover the probabilistic dependency between topics from the latent social interactions. We propose a novel method to estimate the Markov transition matrix of topics using social interactions of different orders. With the properties of Markov process with finite states, we apply the application of Markov metastable states as an approach for discovering the hierarchical clustering of topics and new topics. In addition, we give an experimental illustration of our methods using Markov transitions of topics to rank social actors by their impact on the CiteSeer database. An initial evaluation of our methodology on authors as social actors presents other methods for author impact besides citation counting. Future work could refine our estimation of topic dependency, use larger data sets, derive social ranking of actors independent of topics, explore better estimation methods and generate new communities of actors. We believe our approach of introducing social factors to traditional document content analysis could be useful in other discovering the impact of social actors in other areas.

9. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval.* Addison Wesley, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schutte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315(1–3):39–59, 2000.

[4] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. Spectral min-max cut for graph partitioning and data clustering. In *Proceedings of intl. conf. on data mining*, pages 107–114, Nov. 2001.

[2] some researchers are not on the list because of no (indirect) collaboration with our seed authors and/or having the number of papers in CiteSeer necessary for our cut.

Table 4: Six topics discovered by LDA on CiteSeer subset. Each topic is described using 20 words with highest probabilities.

Topic 00	Topic 01	Topic 02	Topic 03	Topic 05	Topic 07
real	data	queries	channel	program	software
time	database	data	coding	code	systems
system	mining	join	rate	analysis	development
simulation	spatial	patten	performance	java	tools
fault	relational	matching	bit	compiler	process
tolerance	query	clusters	capacity	data	engineering
embedded	temporal	analysis	transmission	language	components
events	large	algorithms	fading	programming	application
timing	rules	hierarchical	receiver	source	design
synchronization	association	large	interference	execution	component
execution	information	incremental	decoding	fortran	framework
scheduling	management	space	frequency	run	modeling
dynamic	discovery	aggregation	low	machine	specification
performance	support	evaluation	cdma	automatic	case
response	sql	views	distortion	compilation	study
distributed	frequent	cost	signal	optimization	reuse
task	patterns	efficient	systems	runtime	management
events	dbms	compression	block	dynamic	evaluation
clock	integration	approximate	modulation	static	object
period	schema	text	time	loops	oriented
real-time system performance	**association rule mining**	**query processing**	**communication capacity**	**program lang. compiler**	**software engr. system**

[5] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.

[6] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.

[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[8] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE joint conference on Digital libraries*, 2004.

[9] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[10] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Comm. ACM*, March 1997.

[11] J. Kubica, A. Moore, J. Schneider, and Y. Yang. Stochastic link and group detection. In *Proceedings of the 2002 AAAI Conference*, pages 798–804, 2002.

[12] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM Press.

[13] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 1926.

[14] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceeding of the eleventh intl. conf. on Knowledge discovery in data mining*, 2005.

[15] B. Miller and J. Riedl. A hands-on introduction to collaborative filtering. In *Proc. of the ACM conf. on computer supported cooperative work*, 1996.

[16] S. Morinaga and K. Yamanishi. Tracking topic dynamic trends using a finite mixture model. In *KDD '04: Proceedings of the tenth intl. conf. on Knowledge discovery and data mining*, pages 811–816, 2004.

[17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134, 2000.

[18] E. Seneta. *Non-negative Matrices and Markov Chains.* Springer-Verlag, 1981.

[19] T. Snijders. Models for longitudinal network data. *Book Chapter in Models and methods in social network analysis*, 2004.

[20] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM Press.

[21] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, New York, USA, 1994.

[22] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *KDD '02: Proceedings of the eighth intl. conf. on Knowledge discovery and data mining*, pages 688–693, 2002.

[23] D. Zhou, E. Manavoglu, J. Li, L. Giles, and H. Zha. Probabilistic models for discovering e-communities. In *WWW'06: Proceedings of the 15th ACM International World Wide Web Conference*, 2006.