Mining Social Dependencies in Dynamic Interaction Networks

Freddy Chong Tat Chua* Hady W. Lauw* Ee-Peng Lim*

Abstract

User-to-user interactions have become ubiquitous in Web 2.0. Users exchange emails, post on newsgroups, tag web pages, co-author papers, etc. Through these interactions, users co-produce or co-adopt content items (e.g., words in emails, tags in social bookmarking sites). We model such dynamic interactions as a user interaction network, which relates users, interactions, and content items over time. After some interactions, a user may produce content that is more similar to those produced by other users previously. We term this effect *social dependency*, and we seek to mine from such networks the degree to which a user may be socially dependent on another user over time. We propose a *Decay Topic Model* to model the evolution of a user's preferences for content items at the topic level, as well as a *Social Dependency Metric* that quantifies the extent of social dependency based on interactions and content changes. Our experiments on two user interaction networks induced from real-life datasets show the effectiveness of our approach.

1 Introduction

1.1 Motivation User interactions in a dynamic social network provide insights for the evolution of relationships among a set of users. The user interactions in this dynamic social network lead to the *production* or *adoption* of content items covering a set of evolving latent factors. Using these evolving latent factors, we aim to derive the social dependency relationships among the users. We define **social dependency** as a temporal correlation between (a) the latent factors in the current time step of the target user, and (b) the latent factors in the previous time step of other users she interacts with. The degree of correlation capture the extent to which the target user depends on the other users, which explains the change in her latent factors.

Social dependency can be useful in many different applications including diffusion of innovations, recommendation of new products, measurement of influential users, and prediction of item adoptions, etc. [2, 3, 18, 23, 25, 32]. The strength of social dependency links also allows us to determine the cohesiveness of users, which can be used to divide users into smaller communities [14, 22, 28].

1.2 Social Dependency Modeling in User Interaction Networks User Interaction Network. A user interaction network consists of interactions that produce new content items over time. We consider a general approach of defining an interaction d (e.g., an email exchange, a published paper) as a tuple $\langle A_d, W_d, \tau_d \rangle$ where A_d, W_d and τ_d denote the set of users, content items (e.g., words in an email or paper), and time point of the interaction respectively. We represent a set of interactions over a time period as a graph called *user interaction network*, as shown in Figure 1. Users, interactions, and content items are the vertices in the user interaction network example. An edge connects a user a to an interaction d taking place at time τ, which a participates in. Similarly, we draw an edge from d to each content item w produced through d. This network has three interactions: $d_1 = \langle \{a_1, a_2\}, \{w_1, w_2\}, \tau_1 \rangle$, $d_2 = \langle \{a_1, a_3\}, \{w_3, w_4\}, \tau_2 \rangle$, and $d_3 = \langle \{a_2, a_3, a_4\}, \{w_1, w_2\}, \tau_3 \rangle$.

The user interaction network or interaction network can be found in many situations involving user communication of one form or another. In an email-based user interaction network, users produce email content as they interact with other email users by replying to email threads. In a newsgroup-based user interaction network, users submit news posts as they respond ("interact") to other users' news posts. As Web 2.0 and social media sites become very popular, we can find even more interaction networks.

Social Dependency. From the interaction linkages among users and their evolving latent factors, one can observe the dependencies among users. An email user may change her email content after exchanging emails with another email user. Similarly, a newsgroup user may change news content in her posts after reading news posts from another user. In both cases, we say the first user is *socially dependent* on the second user if the former produces content that is more similar to the latter after some interaction between them.

We use the scenario in Figure 1 to illustrate the notion of social dependency. Suppose that the three

*Singapore Management University

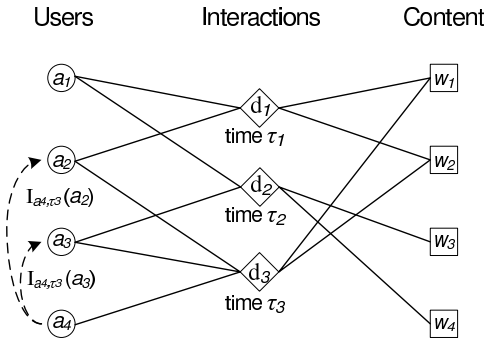


Figure 1: User Interaction Network

interactions occur at different time points $\tau_1 < \tau_2 < \tau_3$. At τ_3, the interaction between a_2, a_3, and a_4 result in the co-production of content items w_1 and w_2. We are interested in whether a_4 is socially more dependent on a_2 or on a_3 for producing the items w_1 and w_2. The dotted lines represent the social dependency links, the direction implies who is dependent on whom, and the weight signifies the extent of dependency. To answer this question, it is instructive to look at the previous time points τ_1 and τ_2. It is evident that since a_2, but not a_3, has been previously associated with w_1 and w_2 before τ_3, so it is likely that a_4 is socially dependent on a_2 for the production of w_1 and w_2, rather than on a_3.

Social dependency is therefore defined based on two key criteria: (a) interactions between two users; and (b) content changes of the user who depends on the other user. As interactions can be ordered by time, we study precedence between interactions by considering a snapshot representation of interactions by sampling the network at different time points. From the snapshots, we derive the *set of interactions occurring at time step* t by $D_t = \{d|\tau_d \in t\}$. For a sequence of multiple time steps T, we have interactions $D_T = \bigcup_{t \in T} D_t$.

The second criteria, content change, can be modeled in different ways. A straightforward approach is to model content as a bag of words and content change is then measured by difference in word usage. This approach however does not work well as word usage can be noisy. Instead, we adopt the topic modeling approach which determines the latent factors as topics behind the observed words. Content change can therefore be measured by a change in topics.

Problem Statement. The research problem of modeling social dependency is thus defined by: *Given a set of users with interactions D_T over a sequence of time steps T, determine the social dependency between a_i on another user a_j at time step t, $I_{a_i,t}(a_j)$, for every $a_i, a_j \in A$ and every $t \in T$.* A is the set of all users in D_T. $I_{a_i,t}(a_j) \in [0,1]$ such that 0 and 1 represent no dependence and complete dependence respectively. Social dependency is time step specific so as to capture its evolution. The social dependencies may exist among users at a time step only when these users have interactions within the same time step. Otherwise, they are deem to be socially independent of one another.

Modeling social dependency comes with the following research challenges.

- *Dynamic changes in topics of interaction content.* The existing topical models are designed primarily for static content. To cope with emerging new interactions and users, we need to develop new and efficient topic models that can model dynamically changing interaction content.
- *Missing user interaction data.* User interactions do not occur with the same intensity in all time steps. They may be dense in some time steps, but sparse or even missing in others. Even in the case of missing data for a given user in a time step, we still need to model how the user's topic preferences are related to those of other users.
- *Smooth transition of user topic preferences.* Users normally do not change their topical preferences abruptly. Hence, the challenge is how to model the smooth transition in user's topical preferences.
- *Dependency weight assignment.* It is expected that a user may be dependent on more than one other user, each potentially with a different weight. Thus, we need to develop the principles in which these weights can be derived from the interactions.

Contributions. In the following, we summarize our research contributions to the social dependency modeling problems as follows:

1. We model how a user's latent factors may change over time. Our proposed model, called *Decay Topic Model*, measures the personal topic preferences of a user at every time step. This model is novel in that unlike previous topic models (see Section 2) where documents have fixed topic distributions and only the topics may change, in our model users may have different affiliations to topics over time. Furthermore, a decay factor is included in the topic model to moderate the rate of change in topic preferences of users so as to create smooth transition of topic preferences as well as to address missing user interaction data.
2. Given the interaction links among users and topic preferences determined by decay topic model, we propose a *Social Dependency Metric* that measures

how user a depends on other users in producing or adopting content. The social dependency metric is topic-based and it considers the topic preferences of a in the current time step and other users' in the previous time step. This notion of changing social dependencies of a user that also takes into account the changing topic preferences of others that the user depends on, is a novel concept.

3. We apply social dependency to the prediction of future user topic preferences on two real datasets extracted from DBLP [15] and ACM Digital Library [1]. Compared with a baseline method, our proposed prediction method using social dependency derives more accurate prediction of future topic preferences.

Organization. The rest of the paper is organized as follows. Section 2 will discuss the past research done on modeling the temporal dynamics of content and user interactions. In Section 3, we describe the decay topic model and our measure of dependency. We then proceed to evaluate our method in Section 4. Finally we conclude our paper in Section 5.

2 Related Work

Social correlation has been studied with regards to different kinds of activities or interactions. Fond and Neville [13, 21] explained that social correlation was a result of alternating transition between homophily and influence among users. Crandall et al. [9] earlier showed that there are feedback effects alternating between social influence and social selection. McPherson et al. [19] surveyed articles establishing that homophily involved similarity factors such as socio-demographic attributes. Singla and Richardson [26] also established the correlation of search queries among instant messaging friends.

There are various related works that study the notion of "influence", although this term is not always used in the same way. For instance, the works by Yang and Leskovec [33], as well as Nallapati and Cohen [20] associate influence to a node, rather than to an edge as we do. The notion of k-exposure [8, 9] assumes that the probability of a user adopting an item is proportional to the number of neighbors who have previously adopted the item. This does not take into account that a user may depend on other users with different weights and on different topics. Goyal et al. [11] estimates influence probability between a pair of nodes in the context of information diffusion, in terms of explicit adoption of items rather than at the level of topics as discussed in our work. Tang et al. [27], Liu et al. [17] and Dietz et al. [10] attempted to measure influence at the topic level where the directionality is given (from cited publication to citing publication), but they did not take into account the temporal evolution of user's topic distributions or social dependency over time.

There are existing works extending topic models to include the notions of users or time, but none really captures all the aspects to be addressed in our work. Rosen-Zvi et al. proposed the Author Topic Model [24] to discover the topic distribution of authors of a document. However, it assumes each word in a document comes only from one author, who independently generates topics without any dependency on another author. This is different from our case, where authors co-produce or co-adopt these content items (words) in interactions, and become socially dependent on one another through these interactions. For analyzing evolving text documents, Blei proposed Dynamic Topic Model (DTM) [4]. DTM was concerned with the evolution of words within topics. Canini et al. [7] addressed yet a different aspect, that of learning topic models for a document collection that grows over time. Neither case addressed our concern about the evolution of users' topic distributions.

3 Dynamic Social Dependency

In this paper, we are interested in modeling the evolution of user interaction networks so as to derive social dependency. In particular, we observe that there are two main components in the evolution of user interaction networks, namely: 1) the change in user preferences for different content items over time, as well as 2) the change in social dependency between users over time. Each of these two components can be represented formally as networks induced from the original user interaction network as follows.

Content Network. This network relates users to content items that they produce or adopt through interactions. For a given set of interactions D_t occurring at time t, an edge (a, w) exists if $\exists d \in D_t,\ a \in A_d \wedge w \in W_d$. Figure 2 illustrates three content networks over three time steps $t_1 = \{\tau_1\}, t_2 = \{\tau_2\}, t_3 = \{\tau_3\}$, induced from the interactions in Figure 1.

Social Dependency Network. This network relates users to other users whom they may socially depend on. A directed edge from a_i to a_j exists if a_i has social dependency on a_j. The edge weight $I_{a_i,t}(a_j)$ reflects the degree to which a_i is socially dependent on a_j at time step t. A loop indicates a user a_i's self-dependency with weight $I_{a_i,t}(a_i)$. In this work, we assume social dependency can be inferred from interactions. Therefore, we only draw an edge from a_i to a_j at time t, if both participate in at least one interaction at time t, i.e., $\exists d \in D_t,\ a_i, a_j \in A_d$. Figure 3 illustrates how the social dependency network evolves

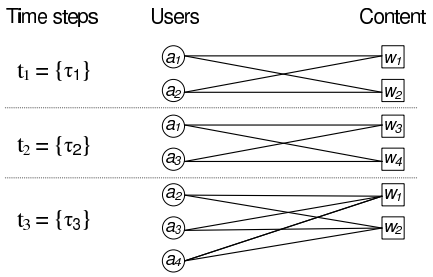


Figure 2: Evolving Content Network

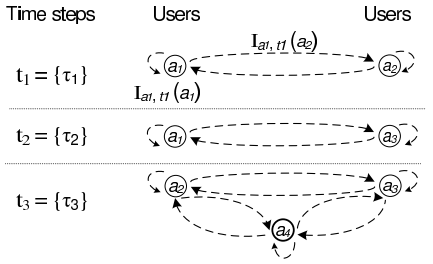


Figure 3: Evolving Social Dependency Network

over three time steps, induced from the interactions in Figure 1.

Given a user interaction network spanning the time period T, the problem we address here is determining the social dependency metric $I_{a_i,t}(a_j)$ for every $a_i, a_j \in A$, and $t \in T$. In the following sections, we will describe how we can model users' content changes at the topic level from the evolving content network. We will then show how the temporal correlation of content changes between users reveals the edge weights in the social dependency network over time.

3.1 Topic Models for Evolving Content Network While a content network reveals the various content items produced by a user, it may not show the user's underlying topic preferences that give rise to the production of those content items. The reason is that content items may be noisy. For instance, in different interactions, a user may produce different words (e.g., "Porsche", "Ferrari") that actually refer to the same topic (e.g., luxury cars). This motivates us to model a user a's content as a topic distribution $\theta_{a,t}$ derived from the content network at time t. As the content network evolves, so does a's topic distribution, i.e., $\theta_{a,t}$ varies with t. In the following, we will first model a user's topic distribution in a static manner, before moving on to our proposed temporal-based *Decay Temporal Model*.

3.1.1 Static Topic Model We observe that the bipartite structure of the content network resembles the relationship between documents and words. Just as a document contains a bag of words, a user is associated with a bag of content items from various interactions. As a naive baseline, we consider topic modeling techniques for text documents in order to model the static topic distribution of users. One such technique is Latent Dirichlet Allocation or LDA [5].

LDA can be adapted to our context as follows. To facilitate the presentation of our model, we introduce a set theoretic notation to explain the variables. Let Z denote the set of topics. For each $z \in Z$, ϕ_z denotes the topic z's item distribution. Each ϕ_z is modeled as a Dirichlet Distribution of V dimensions where V is the total number of unique content items (non-stop words) in the interaction network (corpus).

Let A denote the set of users. For each $a \in A$, θ_a denotes a's topic distribution. Each θ_a is modeled as a Dirichlet Distribution of K dimensions, where K is the number of topics in the set Z. To put it more formally, we have:

$$\phi_z \sim Dirichlet(\beta), \quad \beta \text{ is a constant}$$
$$\theta_a \sim Dirichlet(\alpha), \quad \alpha \text{ is a constant}$$

Each user $a \in A$ participates in a set of interactions denoted by $D_a \subseteq D$, where D is the set of all interactions. Each interaction $d \in D_a$ contains a set of items W_d. Then each w is generated by a topic $z \in Z$, and z is in turn generated by the topic distribution θ_a of user a.

$$z \sim Multinomial(\theta_a)$$
$$w|z \sim Multinomial(\phi_z)$$

In this static formulation, the problem is to find the posterior distribution $P(\phi_z|D, \beta), \forall z \in Z$ and $P(\theta_a, |D, \alpha), \forall a \in A$ given the set of interactions D.

3.1.2 Decay Topic Model The above static model assumes that a user's topic distribution remains the same over time. However, in an evolving content network, a user may produce content items of different topics over time. We extend the above notations to model the notion of temporality. Let T denote an ordered set of discrete time steps with order relation $<$ such that $\forall t_1, t_2 \in T$, $t_1 < t_2$ implies that t_1 is earlier than t_2. $\forall a \in A$, each user a has a topic distribution $\theta_{a,t}, \forall t \in T$, where $\theta_{a,t}$ is modeled as a Dirichlet Distribution.

$$\theta_{a,t} \sim Dirichlet(\{\alpha_{a,t,z}\}_{z \in Z})$$

Unlike the static topic model, each time step t has a Dirichlet distribution for the topic of user a parameterized by a set of parameters specific to the respective

user and time. Since our focus here is on the evolution of users' topic distribution over time, to isolate its effects, we keep topic item distribution ϕ_z the same over time.

Each user $a \in A$ participates in a set of interactions in time step t as denoted by $D_{a,t} \subseteq D_t$, where D_t represents the set of interactions in time t. The interaction $d \in D_{a,t}$ contains a set of items W_d. Then each $w \in W_d$, w is generated by a topic $z \in Z$ and z is in turn generated by the topic distribution of user a at time t.

$$z \sim Multinomial(\theta_{a,t})$$

Hence, what is of interest to us now is the posterior distribution in each time step t, $P(\theta_{a,t}|D_t, \alpha), \forall a \in A, \forall t \in T$.

Generative Process. To arrive at this posterior distribution, we propose the *Decay Topic Model*, which we illustrate using the following generative process.

1. At time t, each user a samples their prior topic distribution $\theta_{a,t}$ from Dirichlet distribution with parameters $\{\alpha_{a,t,z}\}_{z \in Z}$.

2. User a samples the topic distribution $\phi_z, \forall z \in Z$ from Dirichlet distribution with symmetric parameters β.

3. For each interaction $d \in D_{a,t}$, there are a set of content items W_d. In turn, for each of the $|W_d|$ items:

   (a) User a generates a topic z_w from $\theta_{a,t}$ for the item w.

   (b) User a generates an item w from the topic item distribution ϕ_z.

4. Update the parameters of $\phi_z, \forall z \in Z$.

5. Update the parameters of $\theta_{a,t}$ to obtain the posterior topic distribution of a at time t. The posterior distribution also follows a Dirichlet distribution with parameters $\{\alpha_{a,t,z} + n_{a,t,z}\}, \forall z \in Z$, where $n_{a,t,z}$ denotes the number of items that user a produced in time t that belongs to topic z.

6. For every $a \in A$, let the prior topic distribution of $t+1$ be the posterior distribution of t with the parameters multiplied by a decay factor, δ, such that $0 \leq \delta \leq 1$. i.e., $\alpha_{a,t+1,z} = \delta \times (\alpha_{a,t,z} + n_{a,t,z}), \forall z \in Z$, then the prior distribution $\theta_{a,t+1} = Dirichlet(\{\alpha_{a,t+1,z}\}_{z \in Z})$.

7. Repeat steps 1 to 6 for all the time steps.

Decay Factor. The decay factor δ in step 6 helps to moderate the rate of change in topic preferences of users by balancing the contributions of the past time steps versus the current time step. $\delta = 1$ implies no decay. $\delta = 0$ implies that we expect the authors to change their topic distribution at every time step. In other words, by setting $0 \leq \delta \leq 1$, we want to adjust the importance of content produced earlier compared with the recent content for determining the topic distribution of a. For instance, $\delta = 0.5$ means the preferences of a accumulated over time drops by half at every time step, i.e., the half life is one time step. The right setting of δ may differ in different scenarios. In the experiments, we conduct parameter sensitivity test to help determine the best δ setting. In the case where a user has no interaction at time t, her topic distribution will still remain the same as at previous time step $t-1$.

In this work, δ applies to the whole network. While it may be argued that δ may vary from user to user, and from time step to time step, in practice that would generate too many variables, which we may not be able to learn effectively.

3.2 Social Dependency Metric

Having modeled a user's changing topic distribution over time, we now investigate how to model a user's evolving social dependency on other users. This evolving social dependency has been shown in the example as shown in Figure 3. In our formulation, the key idea is that, for user a to depend heavily on another user c at time t, the following criteria have to be met:

- **Interactions.** User a participates in one or more interactions with c at time t. We assume that when an interaction between two users is observed at time t, the actual interaction would have taken place before t. This is reasonable given that our model works on time steps that combine interactions from several time points.

- **Content change.** User a's topic distribution grows to resemble c's topic distribution in the previous time step, i.e., between time steps $t-1$ and t, a's topic is becoming more similar to c's.

Based on the above principles, we propose the *Social Dependency Metric* in the form of a vector $\mathbf{I}_{a,t}$, which is computed as follows.

Given :

1. The set of interactions $D_{a,t}$ that user a participates at time t.
2. Topics associated with the content items, i.e., $\{z_w \mid w \in \bigcup_{d \in D_{a,t}} W_d\}$.

3. Topic distribution $\theta_{c,t-1}$ of every user $c \in \bigcup_{d \in D_{a,t}} A_d$, who has participated in at least one interaction with a in the previous time step $t-1$.

Find : Dependency vector $\mathbf{I}_{a,t}$, where each element $I_{a,t}(c)$ is the dependency of a on user $c \in \bigcup_{d \in D_{a,t}} A_d$.

Algorithm :

1. Initialize the array $\mathbf{I}_{a,t}$ with zero elements.
2. For each interaction $d \in D_{a,t}$, content item $w \in W_d$, and user $c \in A_d$,
   (a) We determine the generation of topic z_w by a user c as follows:
   $$P(z_w|c, \theta_{c,t-1}) \propto \theta_{c,t-1,z_w}$$
   (b) Then update array $\mathbf{I}_{a,t}$ as follows,
   $$\begin{aligned} I_{a,t}(c) &= I_{a,t}(c) + P(z_w|c, \theta_{c,t-1}) \\ &= I_{a,t}(c) + \frac{\theta_{c,t-1,z_w}}{\sum_{c \in A_d} \theta_{c,t-1,z_w}} \end{aligned}$$
3. Normalize the array $\mathbf{I}_{a,t}$ to sum to one for easy interpretation.

Step 2(b) calculates the contribution of each item w and its corresponding topic z_w to user a's dependency on user c. The higher the probability of c generating this topic, the higher is the value of $I_{a,t}(c)$. We assume that the generation of topic z_w comes from a linear combination of a's friends and a herself. The dependency of a on c should be proportional to how much c is likely to generate the topic z_w. The dependency also accounts for the frequency of interaction, i.e. the more interactions a has with c, the higher is the value of $I_{a,t}(c)$.

We run this computation chronologically for every time step $t = 1$ to T to obtain the social dependency values $I_{a,t}(c)$ for each a and c across different time steps $t \in T$.

We explore how the social dependency metric is affected by the topic modeling of content network. At each time step t, we want to compare the changes of a's topic distribution $\theta_{a,t}$ and the changes of c's topic distribution $\theta_{c,t-1}$ for every c in $\bigcup_{d \in D_{a,t}} A_d$. Note that this set of users that a interacts with also contains a herself. Without any decay factor in the topic modeling, the accumulative effect over time will favor larger self-dependency values for a. The decay factor acts to reduce the importance of topics in previous time steps, allowing new interactions to change the topic distribution of a in t significantly enough, so as to better detect a's social dependency on others.

4 Experiments

While user interaction networks model many kinds of interactions, there are only a limited number of datasets available for research, which track those interactions over a significant period of time. We work with two such datasets derived from DBLP and ACM Digital Library (ACM DL). We model co-authorship as a user interaction network, where a publication d is an interaction between one or more authors a_i(s) in the year t. The content items w associated with d are words in the titles/abstracts. In this setting, we say author a has social dependency on author c, if a and c co-author a paper (interact) on topics that a is unlikely, but c is likely, to publish. We assign social dependency to co-authors of a based on the likelihood of the co-authors generating the topics in the papers that a publishes.

After describing the datasets, we will first evaluate the *Decay Topic Model* by comparing two settings (decay vs. non-decay) on the task of predicting an author's observed topic distribution in the next time step. We then evaluate: *Social Dependency Metric*, by conducting two prediction tasks. The first task is similar to the above but with a different approach. Instead of using an author's own topic distribution, we use her co-authors', weighted by the author's social dependency on each co-author. The second task predicts an author's ranking of her co-authors by topic similarity at the next time step using social dependency at the current time step.

4.1 Datasets For experiments, we use a subset of publications from DBLP and ACM DL. To ensure a wide coverage of fields in Computer Science, we use papers published in the reputable Journal of ACM (JACM) as a seed set. We grow this seed set by including other non-JACM publications by authors who has at least one JACM publication. We extend this further to also include the co-authors of JACM authors, and their publications as well.

Table 1: Dataset Sizes

	#authors	*#papers*	*#unique non-stop words*	*period*
DBLP	268,299	546,500	83,440	1936–2011
ACM	157,693	188,086	217,667	1952–2011

The sizes of our datasets are given in Table 1. DBLP has almost three times as many publications as ACM DL. One reason is the longer history of publications maintained by DBLP (since 1936). Another is the larger scope, since ACM DL focuses mainly on ACM-related publications. However, ACM DL has many more unique words than DBLP, because ACM DL has both titles and abstracts, whereas DBLP only has titles. In

both cases, the datasets are significantly large, with hundreds of thousands of nodes, with more than 10 million author-word links for DBLP and 46 million author-word links for ACM DL.

Table 2: Top Words for Sample Topics

Web Systems and Algorithms	*Computational Biology*	*Database Systems and Theory*
DBLP		
web	protein	data
information	gene	database
semantic	analysis	query
based	data	xml
retrieval	database	processing
ACM		
web	data	data
information	gene	query
search	protein	database
content	biological	xml
user	expression	processing

To show that topic modeling on these datasets would discover the latent topics effectively, we produce three sample topics, and the top words for each topic of DBLP and ACM in Table 2. Notably, the top words (e.g., web, information, retrieval) capture well the essence of the topics (e.g., Web systems and algorithms). Moreover, both DBLP and ACM DL discover similar topics with similar top words, even when DBLP has only titles and ACM DL has both titles and abstracts. During experiments, we observe that both datasets result in similar observations. From here onwards, we will use the larger dataset DBLP as the main dataset to discuss our results.

4.2 Evaluating Decay Topic Model The topic modeling step seeks to arrive at $\theta_{a,t}$, the topic distribution of each author a at time t. We hypothesize that the decay factor allows it to better adapt to the author's changing preferences over time. Without decay, the accumulative effect tends to overweigh the older topics more heavily. To test this hypothesis, we compare the topic distributions $\delta < 1$ ($\theta^{decay}_{a,t}$) and $\delta = 1$ ($\theta^{non-decay}_{a,t}$) to the observed topic distribution at the next time step ($\theta^{obs}_{a,t+1}$). For $\theta^{decay}_{a,t}$, we vary δ from 0.2 to 0.8 to determine the optimal setting of δ.

Both $\theta^{decay}_{a,t}$ and $\theta^{non-decay}_{a,t}$ incorporate information from the first time step to the current time step t. We compare them to $\theta^{obs}_{a,t+1}$, which is derived independently using only the set of documents published by a at time $t+1$. If $\theta^{decay}_{a,t}$ is more similar to $\theta^{obs}_{a,t+1}$ than $\theta^{non-decay}_{a,t}$, it shows that the decay approach is better adapted to the preferences in $t+1$.

To measure similarity between two probability distributions p and q, we use the following function:

$$Sim(p, q) = 1 - D_{JS}(p, q),$$

where D_{JS} is the Jensen-Shannon Divergence [16]. Sim ranges from 0 (different) to 1 (identical).

We use this Sim function to measure the similarity between $\theta^{decay}_{a,t}$ and $\theta^{non-decay}_{a,t}$ respectively to $\theta^{obs}_{a,t+1}$. To compare the decay vs. non-decay setting directly, we then take the ratio of the two similarity values as follows:

$$\text{Sim Ratio } \Phi(a,t) = \frac{Sim(\theta^{delta}_{a,t}, \theta^{obs}_{a,t+1})}{Sim(\theta^{non-decay}_{a,t}, \theta^{obs}_{a,t+1})}$$

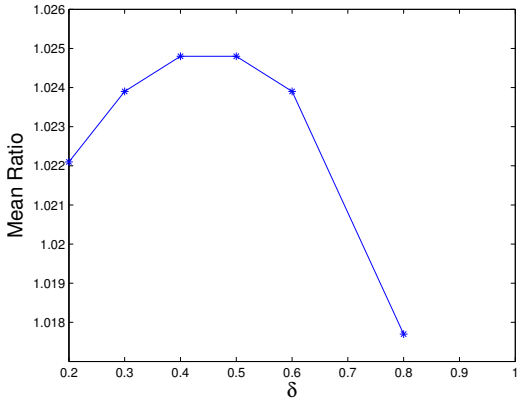


Figure 4: DBLP: Mean of Sim Ratio $\Phi(a,t)$

$\Phi(a,t)$ ratio > 1 indicates that having the decay is better than no decay. Figure 4 shows the mean of values given by the Sim Ratio $\Phi(a,t)$ with respect to the different δ values. Given that the values lie above 1, it indicates that that having some decay is better than no decay at all. From the various choices of δ, we can see that the optimal value of δ lies between 0.4 to 0.5. For the rest of our experiments we therefore use $\delta = 0.5$.

4.3 Prediction of Author's Topic Distribution We now show that our dependency values at t can also be used for the prediction of author a's topic distribution in $t+1$. In this case, the predicted topic distribution for a at time $t+1$ will be a linear combination of the topic distributions at time t of her co-authors $c \in \bigcup_{d \in D_{a,t}} A_d$, weighted by the social dependency values $I_{a,t}(c)$.

Hence, if one set of dependency values arrive at a better estimation of the author's topic distribution than another set of dependency values, it implies that the former more accurately estimate the social dependency weight of each co-author.

Due to the way in which we extract the subset of data from DBLP and ACM DL, we can only evaluate

for the authors who have at least one JACM paper. For these authors, we have the complete co-authors information, while for the rest of the other authors, we have only partial information.

Decay vs. Non-decay. To evaluate our topic prediction, we use the Sim Ratio $\Phi(a,t)$ as defined earlier to compare against the observed topic distribution at $t+1$ (based on only the documents published at time $t+1$) as ground truth. The first comparison is again for decay vs. non-decay, but this time the prediction is based not on the author's own topic distribution, but rather on her co-authors'. We derive two predicted topic distributions at $t+1$, $\theta^{dep-d}_{a,t+1}$ and $\theta^{dep-nd}_{a,t+1}$. $\theta^{dep-d}_{a,t+1}$ is computed using the dependency value $I^{decay}_{a,t}(c)$, for each $c \in \bigcup_{d\in D_{a,t}} A_d$ by the decay topic distribution. $\theta^{dep-nd}_{a,t+1}$ is computed using the dependency values $I^{non-decay}_{a,t}(c), c \in \bigcup_{d\in D_{a,t}} A_d$ by the non-decay topic distribution.

User a's predicted preference for topic z at time $t+1$ is computed as follows.

$$\theta^{dep-d}_{a,t+1,z} = \sum_{c\in\bigcup_{d\in D_{a,t}} A_d} I^{decay}_{a,t}(c) * \theta^{decay}_{c,t,z}$$

$$\theta^{dep-nd}_{a,t+1,z} = \sum_{c\in\bigcup_{d\in D_{a,t}} A_d} I^{non-decay}_{a,t}(c) * \theta^{non-decay}_{c,t,z}$$

We then compute the Sim Ratio $\Phi_1(a,t)$ as follows.

$$\text{Sim Ratio } \Phi_1(a,t) = \frac{Sim(\theta^{dep-d}_{a,t+1}, \theta^{obs}_{a,t+1})}{Sim(\theta^{dep-nd}_{a,t+1}, \theta^{obs}_{a,t+1})}$$

$\Phi_1(a,t)$ ratio > 1 would indicate that the dependency values computed by decay topic distribution give a better prediction than the dependency values computed by the non-decay topic distribution. Figure 5(a) shows a histogram of $\Phi_1(a,t)$ values. The x-axis of the histogram are bins with boundaries given by the value of $\Phi_1(a,t)$. The y-axis of the histogram indicate the frequency of author a and time point t pairs falling into the respective bins. For Figure 5(a), 68% of the (a,t) pairs have $\Phi_1(a,t) > 1$, 1% have $\Phi_1(a,t) = 1$ and 31% have $\Phi_1(a,t) < 1$. This suggests that incorporating the decay factor results in an improvement for the large majority of (a,t) pairs.

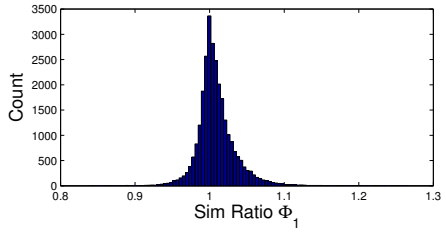
Dependency vs. Co-authorship Count. As another baseline, we use a naive way of computing social dependency weight $I^{base}_{a,t}(c), c \in \bigcup_{d\in D_{a,t}} A_d$, which considers a's dependency on c at t as the count of papers co-authored by a and c, normalized by the total count of a's papers, at time t. Using such social dependency weights, we compute the predicted topic distribution $\theta^{base}_{a,t+1}$, and compare this to the dependency-based prediction $\theta^{dep-d}_{a,t+1}$.

$$\theta^{base}_{a,t+1,z} = \sum_{c\in\bigcup_{d\in D_{a,t}} A_d} I^{base}_{a,t}(c) * \theta^{decay}_{c,t,z}$$

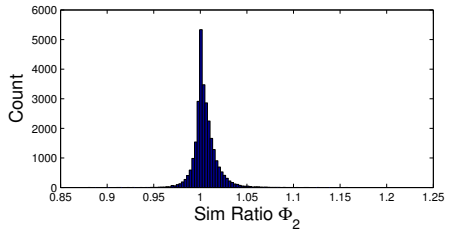
For this comparison, we compute the Sim Ratio Φ_2 as follows.

$$\text{Sim Ratio } \Phi_2(a,t) = \frac{Sim(\theta^{dep-d}_{a,t+1}, \theta^{obs}_{a,t+1})}{Sim(\theta^{base}_{a,t+1}, \theta^{obs}_{a,t+1})}$$

$\Phi_2(a,t) > 1$ indicates that the dependency method outperforms the baseline. Figure 5(b) shows a histogram of $\Phi_2(a,t)$ values. In the figure, 62% have $\Phi_2(a,t) > 1$, 1% have $\Phi_2(a,t) = 1$ and 37% have $\Phi_2(a,t) < 1$. This implies that in most cases, the dependency method tends to arrive at a better prediction than the co-authorship baseline.



(a) DBLP: Histogram for Sim Ratio $\Phi_1(a,t)$



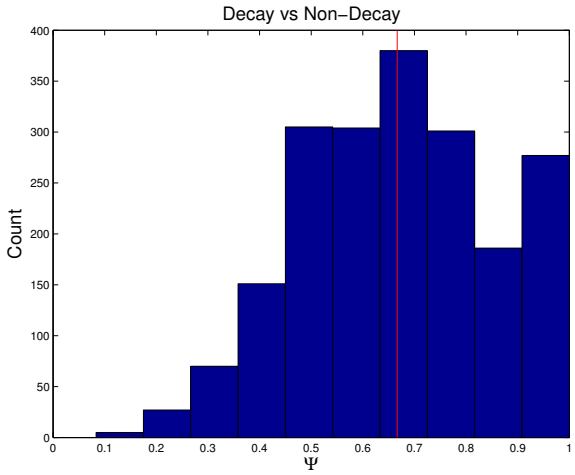
(b) DBLP: Histogram for Sim Ratio $\Phi_2(a,t)$

Figure 5: DBLP

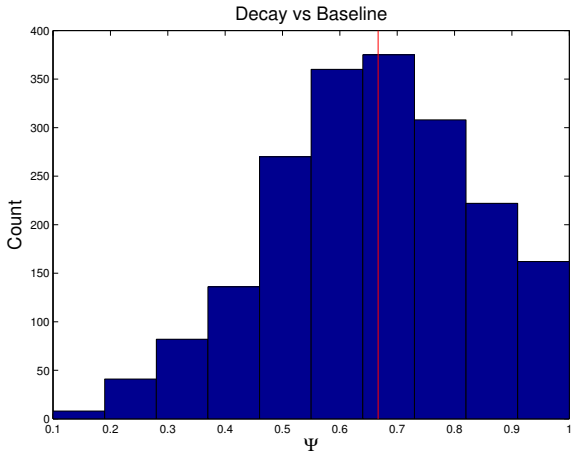
Result Analysis. We now seek to understand better the profiles of users for which our method works especially well. As mentioned previously, Sim Ratio $\Phi(a,t) > 1$ indicates that our method performs better than the baseline at predicting an author's topic distribution. Most authors are active for more than one year, and each year gives a different Sim Ratio. Hence, the proportion of years in which $\Phi(a,t) > 1$ for a given author indicates the degree to which the user has benefited consistently from our proposed method.

To measure this, we introduce the following metric:

$$\Psi(a) = \frac{\text{number of years where } \Phi > 1 \text{ for } a}{\text{number of years } a \text{ publishes}}$$



(a) DBLP: Histogram for $\Psi(a)$



(b) DBLP: Histogram for $\Psi(a)$

Figure 6: DBLP

Figure 6(a) shows the histogram of $\Psi(a)$ values for various users, for a comparison against the non-decay baseline (i.e., $\Phi_1(a,t) > 1$). Figure 6(b) shows the corresponding histogram of $\Psi(a)$, for a comparison against the co-authorship baseline (i.e., $\Phi_2(a,t) > 1$). The red line in both figures indicate the median value of $\Psi(a)$ among the authors. In both cases, the median lies close to 0.7, which implies that a majority of users benefit from our proposed method at least two thirds of the time. In order for us to understand why we are able to predict the topic distribution of some authors and not the others, we examine the $\Psi(a)$ of each author with respect to some factors. Figures 7, 8 and 9 show the boxplots of $\Psi(a)$ with respect to their number of active years, the total number of papers published and the number of co-authors they worked with over the entire duration of their careers. The bins in the boxplots

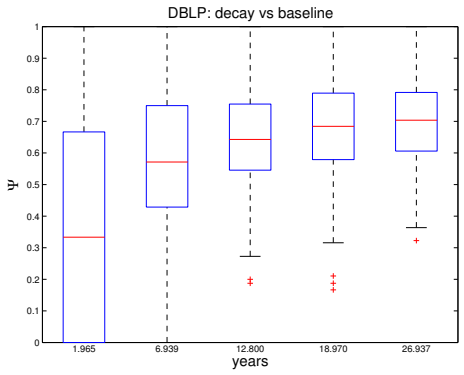


Figure 7: DBLP: $\Psi(a)$ vs Number of Active Years

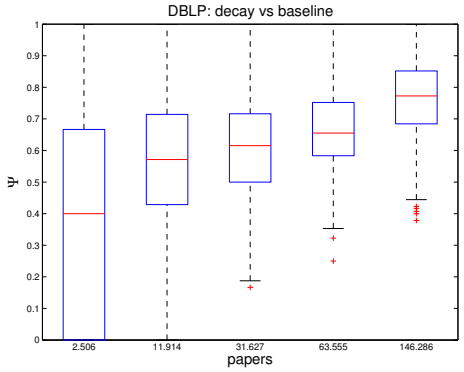


Figure 8: DBLP: $\Psi(a)$ vs Number of Published Papers

are determined by having equal number of data points and the labels on the x-axis represent the mean value of the data points in each bin. The figures collectively tell the story of better performance for authors with higher number of active years, papers, and co-authors. This suggests that we tend to do better when there is more information for a given author. The consistency and the degree to which an author interacts with others allow better inference of not just their topic distributions, but also their social dependency values.

4.4 Prediction of Co-Author's Topic Similarity Ranking In this section, we perform co-author's topic similarity ranking prediction at time $t+1$ using social dependency at time t. At time t, an author a has social dependency value of $I_{a,t}(c)$ on a co-author c. Assuming that a usually does not change the social dependency on her co-authors drastically over two time steps, we expect the ranking of her co-authors by social dependency at time t would be a good predictor for the ranking at time $t+1$. Since a does not necessarily have identical sets of co-authors in t and $t+1$, the ranking prediction will only involve the co-authors appearing in both t and $t+1$.

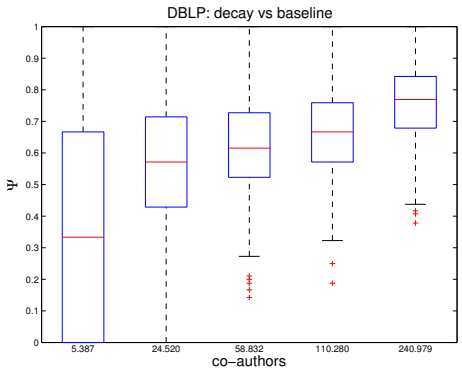


Figure 9: DBLP: $\Psi(a)$ vs Number of Co-Authors

Similar to the previous experiment, we only evaluate for authors who have at least one JACM paper.

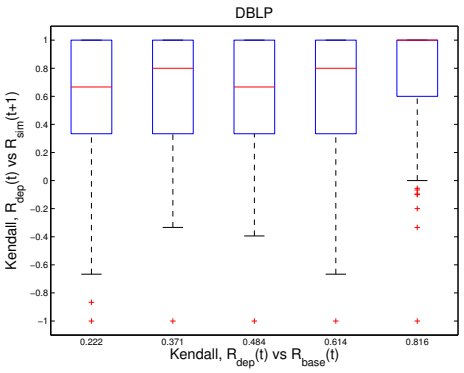
In this task, we denote the *ground truth* ranking of co-authors by topic similarity as R_{sim}. We derive R_{sim} for an author a at time $t+1$ as follows. For each co-author c of a, we obtain the "observed" topic distribution $\theta^{obs}_{c,t+1}$ using only publications by c at time step $t+1$. We then compute the similarity between c's topic distribution $\theta^{obs}_{c,t+1}$ with author a's observed topic distribution $\theta^{obs}_{a,t+1}$ using the Sim function as defined earlier in Section 4.2. Finally, we obtain the ranked list by sorting a's co-authors in descending order of the similarity values.

We compare the R_{sim} of a at time $t+1$ (ground truth) with the following two ranked lists:

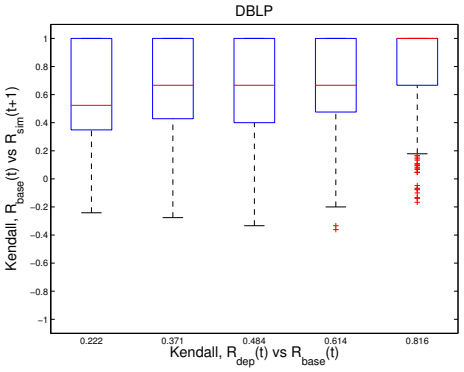
1. *Social Dependency.* R_{dep} ranks co-authors in terms of $I_{a,t}(c)$.

2. *Co-authorship Baseline.* R_{base} ranks co-authors in terms of the number of co-authored papers at time t.

We derive the pair-wise rank correlations between R_{dep} (or R_{base}) and R_{sim} using Kendall Tau Rank Correlation Coefficient (tau coefficient) [12], which is a measure of correlation between two ranked lists where 1 represents full positive correlation, -1 represents full negative correlation and 0 represents no correlation. Hence, if $tau(R_{dep}, R_{sim})$ is higher than $tau(R_{base}, R_{sim})$, it implies that the proposed social dependency metric has higher predictive value than the baseline co-authorship method. To perform this comparison, we first compute the R_{sim}, R_{dep}, and R_{base} for all authors. We then bin the authors into five equisized bins according to their $tau(R_{dep}, R_{base})$. The bin with the smallest values group authors for which R_{dep} and R_{base} are most different. The bin with the highest values group authors for which R_{dep} and R_{base} are most similar.



(a) Evaluating Dependency-based Ranking

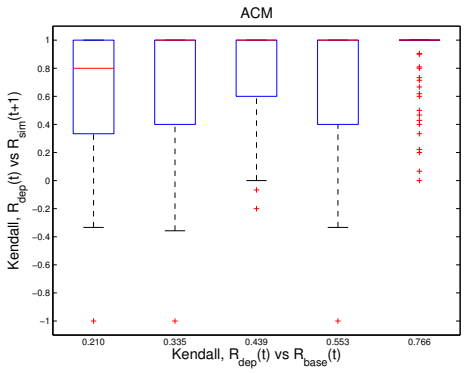


(b) Evaluating Baseline Co-authorship-based Ranking

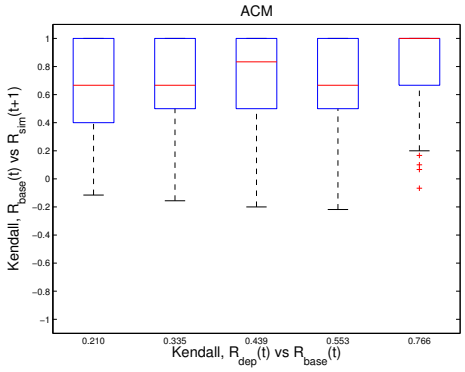
Figure 10: DBLP

We then look at the distribution of $tau(R_{dep}, R_{sim})$ values in each bin. Figure 10(a) shows a boxplot representation of $tau(R_{dep}, R_{sim})$ distributions (y-axis) for each of the five $tau(R_{dep}, R_{base})$ bins (x-axis). The number shown in the x-axis is the mean within each bin. The red line in each box represents the median value, edges of the blue box represents the 25^{th} and 75^{th} percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Figure 10(b) shows the corresponding boxplot representation for the baseline $tau(R_{base}, R_{sim})$.

Comparing Figure 10(a) (proposed) and Figure 10(b) (baseline), we observe that for each bin, the boxplots in Figure 10(a) consistently show higher medians (higher similarity to the ground truth) than the boxplots in Figure 10(b). As the previous figures capture only the DBLP dataset, we repeat a similar experiments for the ACM dataset as well. The results for ACM are given in Figures 11(a) and 11(b), where similar observations can be made to support the higher prediction performance of R_{dep}, as compared to the baseline R_{base}.



(a) Evaluating Dependency-based Ranking



(b) Evaluating Baseline Co-authorship-based Ranking

Figure 11: ACM

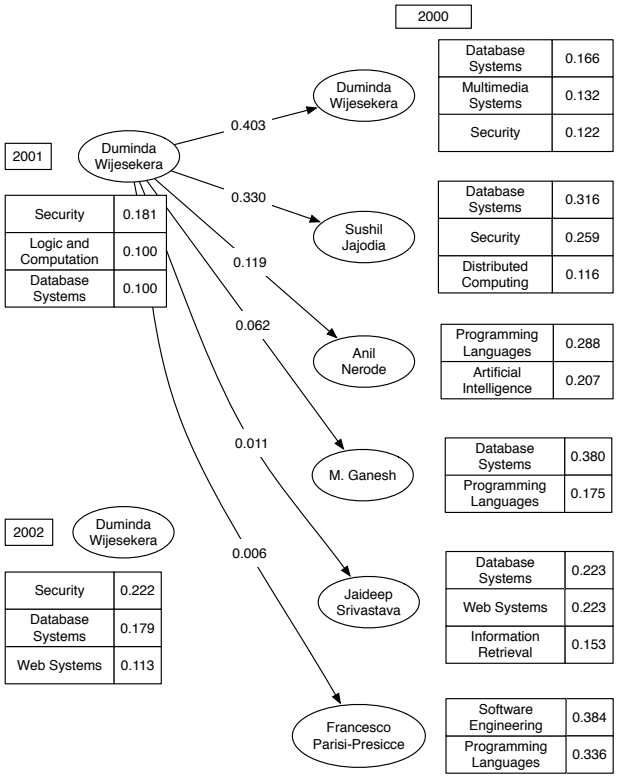


Figure 12: DBLP Case Study

4.5 Case Study Using DBLP, we provide a case study to help illustrate the workings of our proposed social dependency model. For this case study, we use the profile of Associate Professor Duminda Wijesekera. Figure 12 shows the social dependencies of Duminda Wijesekera for the year 2001. The directed edges show Duminda Wijesekera's dependencies on his co-authors who publish with him in the year 2001. Next to these co-authors are their respective topic distributions for year 2000. From the year 2000 to 2001, we observed that Duminda Wijesekera's topic in Security has increased from third position to first position [30]. Based on the dependencies, we observe that he depended on Sushil Jajodia most as compared to other co-authors (excluding himself). Based on the co-authors topic distribution, Sushil Jajodia's topic in Security is the highest which explains why Duminda Wijesekera's dependency on Sushil Jajodia is the highest [6]. In 2002, Duminda Wijesekera continues to increase his topic in Security [29, 31]. This illustrates how social dependency works based on the two components of interactions as well as content change.

5 Conclusion

In this paper, we address the problem of modeling the evolution of user interaction networks, in order to determine the social dependency weights among users at various time steps. We identify two primary factors to social dependency, namely: interactions between users, and temporal correlation between the users' topic distributions. We propose a *Decay Topic Model* to model a user's evolution of content at the topic level, as well as a *Social Dependency Metric* to determine the degree to which a user is dependent on another user. Comprehensive experiments on real-life co-authorship datasets DBLP and ACM show that our proposed models perform well against the baseline (co-authorship count) in two predictive tasks: predicting an author's ranking of co-authors by social dependency, as well as predicting the author's topic distribution in the next time step. This validates our hypothesis that we also need to take into account the changing topic preferences of users beyond just interactions (which the co-authorship baseline only models indirectly). For future work, we aim to incorporate additional factors to further improve the model. One is to learn the decay factor δ automatically. Another is to incorporate the

magnitude (e.g., number of interactions) in addition to topic distribution in determining social dependency between users.

6 Acknowledgement

We acknowledge ACM for providing the ACM digital library data for this research. This work is supported by Singapore's National Research Foundation's research grant, NRF2008IDM-IDM004-036.

References

[1] Association for Computing Machinery. 2011. ACM Digital Library.

[2] Frank M. Bass. A new product growth for model consumer durables. *Manage. Sci.*, 50:1825–1832, December 2004.

[3] Chumki Basu, Haym Hirsh, and William Cohen. Recommendation as classification: using social and content-based information in recommendation. AAAI '98/IAAI '98, pages 714–720, 1998.

[4] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, 2006.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] Alexander Brodsky, Csilla Farkas, and Sushil Jajodia. Secure databases: Constraints, inference channels, and monitoring disclosures. *IEEE Trans. Knowl. Data Eng.*, 12(6):900–919, 2000.

[7] Kevin R. Canini, Lei Shi, and Thomas L. Griffiths. Online Inference of Topics with Latent Dirichlet Allocation. In *Proceedings of AI Stats*, 2009.

[8] Dan Cosley, Daniel P. Huttenlocher, Jon M. Kleinberg, Xiangyang Lan, and Siddharth Suri. Sequential influence models in social networks. In *International Conference on Weblogs and Social Media*, 2010.

[9] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. KDD '08, pages 160–168, 2008.

[10] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. ICML '07, pages 233–240, 2007.

[11] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. WSDM '10, pages 241–250, 2010.

[12] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):pp. 81–93, 1938.

[13] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW*, pages 601–610, 2010.

[14] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. WWW '08, pages 695–704, 2008.

[15] Michael Ley. DBLP Computer Science Bibliography, 2005.

[16] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[17] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining topic-level influence in heterogeneous networks. In *CIKM*, 2010.

[18] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. CIKM '08, pages 931–940, 2008.

[19] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2008.

[20] R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of ICWSM*, 2008.

[21] Jennifer Neville and David Jensen. Relational dependency networks. *J. Mach. Learn. Res.*, 8:653–692, 2007.

[22] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[23] Huiming Qu, Jimeng Sun, and Hani T. Jamjoom. Scoop: Automated social recommendation in enterprise process management. In *ICSC*, 2008.

[24] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. UAI '04, pages 487–494, 2004.

[25] Upendra Shardanand and Pattie Maes. Social information filtering: algorithms for automating "word of mouth". CHI '95, pages 210–217, 1995.

[26] Parag Singla and Matthew Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW*, pages 655–664, 2008.

[27] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. KDD '09, pages 807–816, 2009.

[28] Ken Wakita and Toshiyuki Tsurumi. Finding community structure in mega-scale social networks. WWW '07, pages 1275–1276, 2007.

[29] Lingyu Wang, Duminda Wijesekera, and Sushil Jajodia. Towards secure xml federations. In *DBSec*, 2002.

[30] Duminda Wijesekera and Sushil Jajodia. Policy algebras for access control: the propositional case. CCS, 2001.

[31] Duminda Wijesekera and Sushil Jajodia. Policy algebras for access control the predicate case. CCS, 2002.

[32] Xin Xin, Irwin King, Hongbo Deng, and Michael R. Lyu. A social recommendation framework based on multi-scale continuous conditional random fields. CIKM '09, pages 1247–1256, 2009.

[33] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, 2010.