Which Topic will You Follow? *

Deqing Yang[1], Yanghua Xiao[1], Bo Xu[1]
Hanghang Tong[2], Wei Wang[1], and Sheng Huang[3]

[1] School of Computer Science, Fudan University, Shanghai 200433, P.R.China
{yangdeqing,shawyh,xubo,weiwang1}@fudan.edu.cn
[2] IBM T.J. Watson Research Center, USA. htong@us.ibm.com
[3] IBM China Research Lab, P.R.China. huangssh@cn.ibm.com

Abstract. Who are the most appropriate candidates to receive a call-for-paper or call-for-participation? What session topics should we propose for a conference of next year? To answer these questions, we need to precisely predict research topics of authors. In this paper, we build a MLR (Multiple Logistic Regression) model to predict the topic-following behavior of an author. By empirical studies, we find that social influence and homophily are two fundamental driving forces of topic diffusion in SCN (Scientific Collaboration Network). Hence, we build the model upon the explanatory variables representing above two driving forces. Extensive experimental results show that our model can consistently achieves good predicting performance. Such results are independent of the tested topics and significantly better than that of state-of-the-art competitor.

Keywords: topic-following, social influence, homophily, SCN

1 Introduction

User behavior understanding and prediction are important tasks in social computing. One of the typical tasks is to underhand the author behavior from the public publication records and one of the most interesting author behaviors is topic-following. In general, among all possible topics, an author may select one or several as his future research topics due to his limited time and efforts. Then, a problem will naturally arise: *Can we predict the topic of the next paper for an author?* More specifically, given the historical publications of an author, can we predict the most possible topic of his next papers? In this paper, we answer this question with a positive answer by successfully modeling the topic-following behavior of authors.

* This work was supported by NSFC under grant Nos 61003001, 61033010, 60673133 and 60703093; Specialized Research Fund for the Doctoral Program of Higher Education No. 20100071120032; and partly supported by Zhejiang Provincial NSFC (LY12F02012). The fourth author is support in part by DAPRA under SMISC Program Agreement No. W911NF-12-C-0028 and by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the offcial policies, either expressed or implied, of DARPA, ARL, or the U.S. Government. Correspondence author: Yanghua Xiao.

One may directly use the historical topics of an author to predict the topic of his/her next papers. However, such information in general is insufficient for acceptable accuracy of prediction. Because an author's topic-following behavior is subject to many other factors, such as the influence from his/her collaborators, the current popular topics, historical topics etc. These factors are usually mixed together to affect an author's topic-following behavior.

In this paper, by empirical studies, we found that the topic-diffusion on the co-author networks has significant influence on authors' topic-following behavior. Hence, our basic idea is first constructing a scientific collaboration network, and then model the users' topic-following behaviors by explanatory variables observed from the topic-diffusion among authors in the network.

1.1 Applications

Our research is driven by the following real applications:

1. *Call for participation or paper submission*. When a workshop for a certain topic is announced, delivering the call-for-paper or call-for-participation to the most appropriate candidates who are interested in the topic is critical for the success of the workshop.
2. *Proposal of session topic*. Suppose we need to organize a conference of the next year. What topics should be proposed as sessions of the conference to attract as many attendees as possible? If an accurate topic-following model is available, we can easily summarize the amount of potential audience for sessions of different topics.

The model can also find more applications, such as advertisement, friend recommendation etc. For example, in online social networks, by identifying the topics of posts or comments produced by users, we can deliver advertisements of the topic to potential users who are recognized by the topic-following model [1]. In addition, we can recommend the users who will follow the same topic as the friends of the objective users [2].

1.2 Topic Diffusion

Topic diffusion in *Scientific Collaboration Network* (SCN) is one of important processes that may influence the topic-following behavior of an author. SCN is a co-author network, in which each vertex is an author and each edge represents a co-author relationship. Intuitively, if a topic is diffused to an author from many of his coauthors in the SCN, it is of high probability that he will adopt the topic in his future publications. In this paper, *we build our topic-following model based on the topic-diffusion principles in SCN*.

Generally speaking, there are two typical ingredients which impact information diffusion among individuals in a social network: *social influence* and *homophily* [3, 4]:

1. Social influence means that an individual tends to adopt behaviors of his neighbors or friends.

2. Homophily is the tendency of individuals to choose friends with similar characteristics [3, 5].

Social influence depends on the structure of the social network. In contrast, homophily focuses on the attribute similarity among individuals, in other words, it does not matter whether they are connected to each other. These two factors have been widely investigated as the underlying mechanisms accounting for linkage formation in large social networks [5].

1.3 Challenges and Contributions

Thus, we first need to understand the effect of *social influence* and *homophily* on topic diffusion in SCN before we can precisely model authors' topic-following behavior. Unfortunately, most previous research work on SCN mainly focus on macroscopic structure of the whole network [6, 7], collaboration pattern [8] or community evolving [9], leaving topic diffusion in SCN rarely investigated. Many findings about information propagation[4] on other social networks have been discovered, but the diffusion laws observed on these networks in general do not necessarily hold in SCN any more.

Hence, the purpose of this paper is two-fold. First, understanding the effects of *social influence* and *homophily* on topic diffusion in SCN. Second, developing an effective topic-following model based on the above findings. However, there still exist many challenges that remain unsolved.

- First, it is difficult to distinguish impacts of social influence and homophily from each other. Because they are often mixed together [10] to affect topic diffusion. Furthermore, quantifying their impacts on research topic-following is subjective.
- Second, it is hard to accurately define topics for papers due to the uncertainty and multiplicity of topic identification. Since topic-following behaviors of authors are topic-sensitive, precisely defining the topic of a paper is critical for the model's performance.
- Third, sample sparseness poses a great challenge. Many scientists have quite small number of papers and many topics have only a few papers, which generally bring great challenges to accurately predict the authors' topic-following behavior.

In this paper, we address above challenges and make the following contributions:

- First, we uncover the effects of social influence and homophily on topic diffusion in SCN by extensive empirical studies.
- Second, we propose a *Multiple Logistic Regression* (MLR) model based on the empirical results to predict topic-following of authors.
- Third, we conduct extensive experiments with comparison to the state-of-the-art baseline to show the advantage of our proposed model in prediction performance.

Although our model is proposed for SCN, it can also be used in other social settings, for example, predicting buyer behavior in e-commerce, topic prediction in microblogging etc.

[4] In the following texts, *information propagation* or *information spreading* may also be used interchangeably with *information diffusion*.

1.4 Organization

The rest of this paper is organized as follows. Sect. 2 is a brief review of related work. We introduce the basic concepts and try to identify the effects of social influence and homophily on topic diffusion in SCN in Sect. 3. In Sect. 4, we present empirical results about driving forces of topic propagation in SCN. Based on the findings in empirical analysis, in Sect. 5, we propose a MLR model to predict topic-following of authors with the comparisons to the baseline approach. At last, we conclude our work in Sect. 6.

2 Related Work

We review the related works from the following three aspects: *information diffusion*, *scientific collaboration network*, and *user behavior modeling*.

Information diffusion Topic diffusion can be regarded as a special case of information/idea propagation on social networks, which has already been studied in sociology, economics, psychology and epidemiology [11–13]. Many research work of information diffusion focused on concrete object propagation on online Web media, such as article diffusion on Wikipedia [4], picture diffusion on Flickr [3], post diffusion on Blogsphere [14, 15] and event diffusion on Twitter [16], but for topic addressed in this paper, it is rarely explored in terms of information diffusion on social networks. Although D.Gruhl *et al.* studied topic diffusion [15]. But their focus is the social network in Blogsphere other than SCN. Research topics of SCN have also been investigated in [17, 18]. But they focused on detecting topic evolution and transition over time. Social influence and homophily have been regarded as two major causal ingredients [3, 10, 4] of information diffusion on social networks. It is widely established that it is social influence and homophily as well as their interactions that determine the linkage formation of individuals [19, 13] or interplays between two individuals in social networks [20]. It is a traditional belief that social influence accounts for the information diffusion on typical social networks [11, 14]. However, recent study in sociology argues that homophily also plays an important role for individuals to follow others' idea or adopt others' behavior [21]. As a result, some literatures [22, 4] studied the cumulative effects of social influence and homophily on diffusion cascading. However, to the best of our knowledge, the effects of social influence and homophily on research topic diffusion in SCN have rarely been reported.

Scientific collaboration network As a typical social network, SCN was systematically investigated by Newman *et al.* [6, 8]. But they only focused on the structural properties of the network without exploring topic diffusion. The SCN constructed in [23] is identical to the one used in this paper, but it studied the evolution of collaboration between individuals instead of research topic diffusion. Tang *et al.* [24] also investigated topic-level social influence on SCN, but they did not take homophily's influence into account.

User behavior modeling Information spreading can be considered as one kind of user behavior. Many user behavior models have been proposed in previous studies. For example, some works [25, 26] modeled retweet patterns of users in Twitter. Others [27, 3] modeled the user interaction pattern in Flickr and MySpace. All these works did not model topic-following behavior in SCN.

3 Preliminaries

In this section, we first review the preliminary concepts about topic diffusion in SCN. Then we propose our solution to quantify the influence of social influence and homophily on topic diffusion in SCN.

3.1 Basic Concepts

We first formalize SCN and explain the rationality to use SCN, then formalize the concept of author's topic-following behavior in SCN.

Scientific Collaboration Network (SCN) is a co-author network. We give the formal definition of SCN in Definition 1. Notice that SCN is evolving over time, the snapshot of SCN at time t is denoted by G_t. Its vertex set and edge set are denoted by V_t and E_t, respectively. G_t encodes all coauthor relationships till time t. In other words, if $t_1 \leq t_2$, G_{t_1} is a subgraph of G_{t_2}, i.e., $V_{t_1} \subseteq V_{t_2}, E_{t_1} \subseteq E_{t_2}$. And for an edge $e_{u,v} \in E_{t_1}$, we have $w_{t_1}(e_{u,v}) \leq w_{t_2}(e_{u,v})$.

Definition 1 (SCN). *A Scientific Collaboration Network is an undirected, edge-weighted graph $G = (V, E, w)$, where node set V represents authors, edge set E represents coauthor relationships, and $w : E \to \mathbb{N}$ is the weight function of edges. For each edge $e_{u,v} \in E$, $w(e_{u,v})$ is defined as the number of papers that u and v have ever coauthored.*

Rationality to use SCN In this paper, we mainly focus on SCN constructed from DBLP data set. The reason is two-fold.

- First, it is a good approximation of social networks in real life since most coauthors are acquainted to each other. SCN shares many generic properties of a social network. Most principles guiding the users' behavior on social networks still hold true.
- Second, SCN extracted from DBLP contains enough clean information. DBLP contains plenty of computer science publication records, each of which includes title, author list, venue information and publishing year. These information allows us to explore the topic-following behavior of authors. DBLP dataset is cleaned before its publication. Some noise in the data, such as name ambiguity, has been preprocessed. Thus, the extracted SCN is free of such noise.

Topic diffusion Given a topic s, we say an author u followed s if u has published at least one paper of s. Moreover, the set of authors who published at least one paper of topic s in year t is denoted as U_t^s whose size is $|U_t^s|$. Similarly, authors who published papers of s up to year t are denoted by $U_{\leq t}^s$. Then, *the popularity of topic s in year t* can be measured by $|U_t^s|$. The *diffusion of topic s* is a dynamic process which can be observed from the evolution of $|U_t^s|$ along time t. DBLP only records the year when a paper was published, thereby the unit of one time step is defined as one year when we study temporal properties of topic diffusion.

Dataset description and topic extraction We select the papers published up to year 2011 from the seven major categories[5], e.g., *database*, *data mining*, *World Wide Web*, to construct SCN. The resulting SCN contains 193,194 authors and 557,916 co-authoring relationships. Identifying the topic of each paper is a preliminary step for the study of topic diffusion. We select 25 representative topics, such as *Query Processing, Privacy and Security, and Social Networks, etc*. Then, we build a SVM [28] classifier trained on a manually-labeled dataset to classify each paper into the 25 topics.

3.2 Social Influence and Homophily

In this subsection, we present our solution to evaluate the effects of social influence and homophily on topic propagation in SCN. We first show that it is intractable to precisely distinguish them from each other. Hence, we turn to a qualitative way to evaluate the two factors' effects. In general, it was well established that the more neighbors adopting an idea, the more possible himself will follow the idea. Thus, we evaluate the effect of social influence by the number of neighbors who have adopted a topic before. In DBLP, the *topic similarity* is a good indicator of the homophily between two authors. Thus, we use topic similarity to evaluate the effect of homophily.

Intractability In general, it is intractable to precisely distinguish the effects of social influence and homophily from each other [10, 3]. We illustrate this by Figure 1. In the graph, a dark node represents an author who has followed a certain topic (say s). In G_{t-1}, only one author a has ever published a paper of topic s. Then in G_t, author d, h, f, g also adopt the topic. Since d and h are the direct neighbors of a, we can assume that they are infected by a's influence through social ties between them. However, we can not exclude the possibility that the topic-following behavior of d and h is due to their own interests on the topic. f and g have no direct links to a. They are linked to a only by two-step paths. Hence, we may assume that their topic-following behaviors are mainly due to homophily since in general social influence through indirect links is weak. However, it is also possible that e learned about topic s from a and then recommended it to his neighbors f and g. Hence, it is hard to precisely quantify the effect of social influence and homophily.

Social influence in SCN Social influence refers to the process in which interactions with others cause individuals to conform, e.g., people change their attitudes to be more similar to their friends [22, 4]. In the context of topic diffusion in SCN, social influence can be characterized as the tendency of an author adopting the same topic as his neighbors. In general, the more neighbors infected by the topic, the more tendency he will adopt the same topic of his neighbors. Thus, the effect of social influence can be directly evaluated by the the number of neighbors who have published papers of a certain topic [10]. In other words, if an author followed a topic at t and a significant number of his neighbors (i.e., coauthors) had ever published papers of this topic before year t, it would be of high confidence that the author's topic-following behavior is affected by social influence.

[5] http://academic.reserach.microsoft.com

Homophily in SCN In our study, we use topic similarity among authors to approximate homophily in SCN. Homophily can be regarded as demographic, technological, behavioral, and biological similarities of individuals [10, 5]. It is intractable to precisely define homophily in SCN due to the limited information available from DBLP dataset. One good approximation in SCN is *topic similarity*. We use the *history topic vector* to represent an author's research interests, which can be formally defined as $\boldsymbol{u} = [n_1, n_2, ..., n_{25}] \in \mathbb{N}^{25}$, here each n_i is the number of the author u's papers belonging to i-th topic. Then, the topic similarity between author u and v can be given as follows:

Definition 2 (Topic Similarity). *Given two authors u and v, the topic similarity of author u and v is defined as,*

$$sim(u, v) = cosine(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\| \boldsymbol{u} \| \| \boldsymbol{v} \|} \tag{1}$$

Note that $\boldsymbol{u}$ and $sim(u, v)$ are time-dependent variables, which are calculated within a time window. Recall that $|U_t^s|$ varies as time elapses. By summarization, we find that most topics' $|U_t^s|$s keep above 80% of peak value only for three years, indicating most scholars retain their interests of one topic for about three years. Hence, we will count n_i according to the papers published in a three-year time window $[t-3, t-1]$ when computing $\boldsymbol{u}$ at time t.

Further we define the topic similarity between one author u and a group of authors U. Similarly, we first define a history topic vector for U as $\boldsymbol{\mathcal{U}} = [N_1, N_2, ..., N_{25}]$, where N_i is the total number of papers of i-th topic composed by any one in U, then we have $sim(u, U) = cosine(\boldsymbol{u}, \boldsymbol{\mathcal{U}})$. $\boldsymbol{\mathcal{U}}$ is also calculated within three-year window.

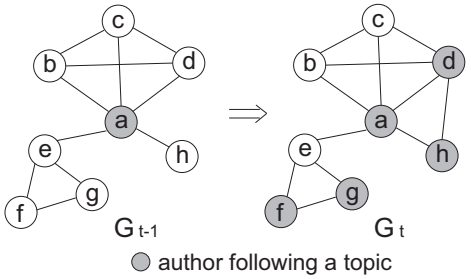

Fig. 1. Illustration of topic diffusion. Social influence and homophily are mixed together to affect topic diffusion.

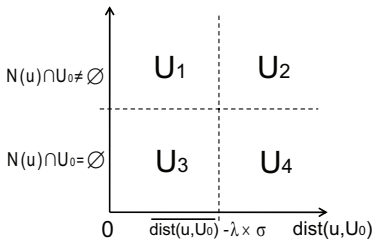

Fig. 2. The division of $\bar{U}_0$.

4 Empirical Study

In this section, we present the empirical study results. Our purpose of empirical study is two-fold. First, in Sec 4.1 we show that social influence and homophile are two fundamental driving forces of topic diffusion in SCN. Second, we reveal the way that social influence affects topic diffusion in Sec 4.2.

4.1 Driving Forces of Topic Diffusion

We first give the detail of our experiment design, then give the results.

Experiment design Let $U_0 = U^s_{\leq t_0}$, i.e., the set of authors who have published papers of topic s till t_0. We will focus on $\bar{U}_0 = V_{t_0} - U_0$, i.e., those who have not published any papers of topic s till t_0. $\bar{U}_0$ can be divided into four disjoint subsets U_1, U_2, U_3 and U_4 [6], according to two conditions. This division of $\bar{U}_0$ is illustrated in Figure 2, where $N(u)$ is the neighbor set of u.

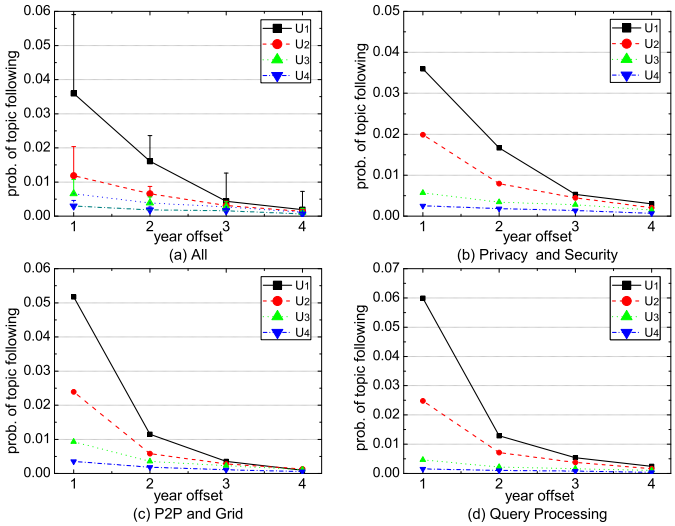

Fig. 3. Evolution of driving forces of individuals' topic-following. Both social influence and homophily are effective on topic diffusion in SCN. And both of them decay in an exponential way.

The first condition is whether there exist neighbors that belong to U_0. If exist, this author may publish a paper of topic s due to social influence [19, 3]. The second is the discrepancy of an author's topic vector to U_0's. Specifically, for each author $u \in \bar{U}_0$, we can calculate $dist(u, U_0) = 1 - sim(u, U_0)$, i.e., the distance between topic vectors of u and U_0. Then, we compute the standard deviation σ for $\{dist(u, U_0) | u \in \bar{U}_0\}$. We further set a threshold $\tau = \overline{dist(u, U_0)} - \lambda \times \sigma$, where $\overline{dist(u, U_0)}$ is the mean value and $0 < \lambda < 1$ is a tuning parameter. Any author $u \in \bar{U}_0$ with $dist(u, U_0) \leq \tau$ will be identified as the one whose topic vector is sufficiently similar to U_0. These authors may publish a paper of topic s after t_0 driven mainly by homophily [3].

Accordingly, if there are authors in U_i $(1 \leq i \leq 4)$ publishing a paper of topic s after t_0, those in U_1 may be affected by social influence as well as homophily; those in U_2 are affected merely by social influence; those in U_3 may be affected merely by homophily. While U_4 represents the remaining authors who are not influenced by the two forces with high probability.

[6] We may also use $U_i(s)$ to denote each U_i when topic s needs to be specified explicitly.

For each U_i ($1 \leq i \leq 4$), we count the number of authors who publish the paper of topic s after t_0 for each s. Then, we calculate the proportion of authors within each U_i that follow the topic. This proportion can be regarded as the probability that an author within each group will follow the topic. Each proportion is normalized over all topics. For example, the proportion of authors within U_1 over all topics is normalized as: $\frac{\sum_s |U_1'(s)|}{\sum_s |U_1(s)|}$, where $U_1(s) \subset V_{t_0} - U^s_{\leq t_0}$ and $U_1'(s)$ is the set of authors in $U_1(s)$ that followed topic s after t_0.

Results The experimental results are shown in Fig. 3, where $t_0 \in [2005, 2007]$ and $\lambda = 0.8$. Fig. 3(a) shows that authors exhibiting more topic similarity to U_0 or having more neighbors in U_0 are more probable to follow the topic than those without these characteristics. We also can see that the cumulative effect of social influence and homophily on topic diffusion is more significant than either one of these forces. Moreover, it can be observed that the effects of social influence, homophily and their mixture are decaying in an exponential way as time elapses. Generally, three or four years later after t_0, minor effects can be observed (Similar results can be observed when we vary the year window to compute $dist(u, U_0)$). These facts indicate that social influence and homophily are generally time-sensitive. When we compare social influence to homophily, we find that social influence is more sensitive to time. These findings provide additional evidence for the time-sensitivity of social influence, which was first discovered in the study of product-adopting behavior [10].

All above findings are generally consistent with those found in specific topics, e.g., *Privacy and Security, P2P and Grid* and *Query Processing*, as shown in Fig. 3(b)∼(d). Different topics only show minor difference on the decaying speed.

4.2 Social Influence

In this subsection, we show that the number of infected neighbors and relationship strength have positive influence on topic diffusion.

Dependency on the number of infected neighbors It has been shown that the probability that an individual joins a group depends on the number of his friends in this group [19]. Then, *does an individual's topic-following behavior also depend on the number of his neighbors who have followed the topic before?* We get a *positive* result from the following studies.

We first summarize the probability p with which an author follows his neighbor's research topic, as the function of the number or the proportion of his neighbors that have followed the topic. Let U_x be the set of authors that have x neighbors who have ever published papers of a given topic before. In U_x, some of them will follow the behavior of their neighbors to publish papers of the same topic. The set of such authors is denoted by U_x'. Thus, for each value of x, we can define $p(x)$ as:

$$p(x) = \frac{|U_x'|}{|U_x|} \tag{2}$$

$p(x)$ can be similarly defined when x is the proportion of neighbors who have ever published papers of a certain topic before.

The correlation between $p(x)$ and x is shown in Fig. 4. It is clear that for either case when x is the number (Fig. 4(a)) or proportion (Fig. 4(b)), $p(x)$ generally increases with

x, strongly suggesting the probability that an author will follow a topic heavily depends on the number/proportion of his neighbors who have followed the topic.

All above results about neighbor's influence are consistent with classical diffusion theory. It was shown in [11] that innovation decision is made through a cost-benefit analysis where the major obstacle is uncertainty. Similarly, in topic diffusion, when more neighbors have followed a topic, other authors will be more certain about the benefit of following a certain topic, and consequently it is quite probable that an individual is persuaded to accept it.

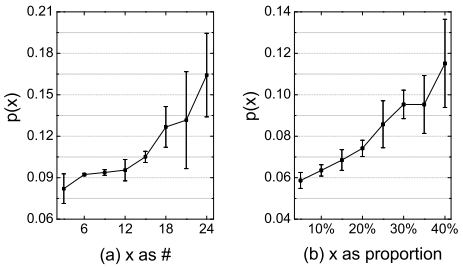

Fig. 4. p(x) vs x, shows that an author's topic-following behavior depends on the number/proportion of his neighbors who have followed the topic.

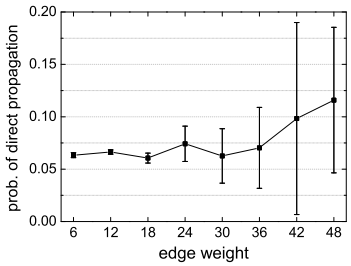

Fig. 5. Prob. of being infected vs edge weight. This figure shows that the strength of coauthoring is influential on the direct propagation from an author to his neighbors.

Dependency on strength of coauthoring Recall that in SCN (Definition 1), each edge is assigned a weight indicating the number of coauthored papers. Thus, *whether the strength of the edge is influential on the direct propagation from an author to his neighbors?* The answer is *Yes* based on the following studies.

To answer the question, we summarize the correlation between the strength of coauthoring and the probability that a topic is propagated from an author to his neighbors. In general, more coauthored papers imply more common research interests, or other similarities between authors. Hence, it is expected that the probability of direct propagation is positively correlated to edge weight. The plot shown in Fig. 5 verifies our conjecture. In the figure, the *probability of direct propagation* is measured by *the proportion of edges on which direct propagation happens*, and is plotted as a function of edge weight. It is evident from the figure that direct propagation probability increases with the growth of edge strength. In other words, an individual is more likely to follow the research topics of his neighbors who have tighter relationships with him. This observation is consistent with our intuition that one person is likely to share his friend's interests or follow his friend's ideas.

5 Modeling Topic Diffusion in SCN

Based on the previous empirical results, in this section, we will propose a MLR model to predict the topic-following behavior of authors in SCN. Next, we will present the detail to build the model and evaluate predicting performance of the model.

5.1 Model Selection

Recall that for the authors who have not yet published any papers of a given topic before t, we intend to accurately estimate the number of them that will/or not follow the topic in t and future. This problem can be casted as a typical binary classification problem. *Logistic Regression Model* is one of the most widely used binary classifiers, which has been widely used in applications of medicine, genetics, business and etc. In this paper, since the behavior of individuals' topic-following is driven by more than one force, we adopt *Multiple Logistic Regression* (MLR for short)[29] to predict topic diffusion in SCN.

In MLR, dependent variable Y is a binary response variable with only two possible values, 1 or 0, which respectively represents whether an author will or not follow a certain topic if he has not adopted it before. And the value of Y relies on the multiple explanatory variables x_i, each of which represents an influential factor that affects an author's topic-following behavior. Let $\pi(x) = P(Y = 1)$ be the probability that an author will follow a certain topic. In MLR model, a linear relationship is established between *logit* function of $\pi(x)$ (or log odds of $\pi(x)$) and p explanatory variables. The detailed model can be described as the following equation:

$$logit[\pi(x)] = \ln \frac{\pi(x)}{1-\pi(x)} = \alpha + \Sigma_{i=1}^{p} \beta_i x_i \tag{3}$$

By simple transformation, we can calculate $\pi(x)$ by the following equation:

$$\pi(x) = 1/(1 + e^{-(\alpha + \Sigma_{i=1}^{p} \beta_i x_i)}) \tag{4}$$

where we have $0 \leq \pi(x) \leq 1$. Both α and β_i are the parameters that can be estimated by training the model. Since MLR is used as a binary classifier, we still need a cutoff value (cv for short) to help us classify each author into two categories. The simplest rule to use cv for classification is: *if* $\pi(x) \geq cv$, $Y = 1$*; otherwise* $Y = 0$. Typically, $cv = 0.5$ is used.

5.2 Explanatory Variables

Previous empirical studies suggest two explanatory variables representing social influence and homophily, respectively, to model the probability of topic-following. As we can see in Fig. 3, topic-following behavior of an individual in SCN varies as time elapses. So the two explanatory variables are time-dependent and are always discussed w.r.t. year t.

For Social Influence We have shown that the probability that an author follows a topic is positively correlated to the number of his neighbors who have already followed the topic, as well as the strength of social ties between them. Similar to belief propagation model on factor graph [30], we quantify social influence as follows. For an author u, the probability that u follows the topic s at year t can be given as:

$$F_{SI}(u, s, t) = \sum_{v \in N'(u)} \frac{w(e_{u,v})}{\sum_{v \in N'(u)} w(e_{u,v})} \times f(v, s, t-1) \tag{5}$$

where $N'(u)$ is the neighbors of u who have followed topic s before u, $w(e_{u,v})$ is the weight of edge $e_{u,v}$ and $f(v, s, t-1)$ quantifies the influence from u's neighbor v in

$t-1$. The function $f(\cdot)$ can be precisely defined as,

$$f(v,s,t) = \delta F_{SI}(v,s,t) + \frac{n_t^s}{n_t} \tag{6}$$

where $0 < \delta < 1$ is a punishing parameter, n_t is the number of u's publications at t among which n_t^s papers belong to topic s.

In the definition of $f(v,s,t)$, $\delta F_{SI}(v,s,t)$ summarizes the influence inherited from v's direct neighbors and indirect neighbors. As we have discussed in the example of Fig. 1, indirect neighbors may also have potential influence on topic-following by some intermediate authors. However, generally such indirect social influence degrades in an exponential way as the propagation length increases [31]. Hence, we need δ (δ=0.5 in our experiments) to punish the influence from faraway neighbors. The ratio $\frac{n_t^s}{n_t}$ accounts for v's interest on topic s in year t.

The computation starts from $F_{SI}(u,s,t_0)$ for each author u with $t_0 = 2002$. The initial value is set as $\frac{n_{<t_0}^s}{n_{<t_0}}$. Then, the computation proceeds iteratively for each year raning from t_0+1 to t. As above, Equation 5 will produce large F_{SI} when an individual has many infected neighbors and retains strong relationships to these neighbors, which confirms the findings about driving effects of social influence.

For Homophily Homophily indicates that an author u tends to follow the topic of those whose research topics are similar to himself. This factor can be captured by $F_{TS}(u,s,t)$, which can be directly defined as the topic similarity between an author u and the group of authors who have ever published paper of the same topic before year t:

$$F_{TS}(u,s,t) = sim(u, U_{<t}^s) \tag{7}$$

Finally, Equation 3 can be rewritten as,

$$logit[\pi(x)] = \alpha + \beta_1 F_{SI} + \beta_2 F_{TS} \tag{8}$$

We use maximum likelihood method to estimate all parameters, i.e., α and each β_i.

5.3 Sample Preparation

In this subsection, we introduce our sample selection for model training and testing.

Preparing the samples To build the MLR model, we collect publications in year $[2004, 2008]$ as the training data and the publications in year 2009 as the testing data. Note that we build MLR model for each topic since the parameters are topic sensitive.

Suppose now we need to generate samples for a certain topic s. In general, the topic-following behavior of authors who seldom publish papers in one year is subject to randomness, and hence their behaviors tend to be outliers. Therefore, we will only consider those authors who published *significant* number of papers in one year as the *valid* training samples. In our experiments, the threshold is set to *3* papers. Then, all valid authors will be collected for each year t in $[2004, 2008]$. Thus, each pair $< u, t >$ ($2004 \leq t \leq 2008$ and u is a valid sample) will be regarded as one training pair sample for topic s.

Now, for each pair sample $< u, t >$, we need to assign a value of 0 or 1 to the binary response variable Y. We process Y as follows: $Y = 1$ *if author* u *publishes at least one*

paper of topic s or other topics closely related to s during the three-year time window $[t, t+2]$; otherwise $Y = 0$. The setup of three-year time window is due to the following two reasons. It generally takes one or two years (or even more) for an author to follow a certain topic. It also takes time for a topic to be diffused to more authors especially for a new topic.

Relaxing the topics In the computation of the response variable, topic s is relaxed to be itself or some other related topics, which is due to the fact that many topics are closely related to each other. For example, the topic *XML* is closely related to *Query Processing* since one of the core tasks in XML data management is XML query processing. As a result, many authors may publish papers containing more than one topic and usually change their research interests from one topic to another related one. Given an author u, we say a *topic transition* $s_1 \rightarrow s_2$ happens in year t_2 if u first published a paper of topic s_1 in year t_1 and then published a paper of topic s_2 in year t_2 such that $t_2 > t_1$. Based on it, we can define *topic transition probability* from s_1 to s_2 before year t as follows,

$$P(s_1 \rightarrow s_2, t) = \frac{\sum_{t_1<t_2<t} |U_{t_1}^{s_1} \cap U_{t_2}^{s_2}|}{|U_{<t}^{s_1}|} \tag{9}$$

Next we use the following equation to identify topic s' that is *closely related to* s:

$$\frac{P(s \rightarrow s', t)}{P(s \rightarrow s, t)} \geq \gamma \tag{10}$$

where γ is a threshold parameter defining the *topic closeness*. The rationale is that if a topic s transits to another topic s' with a high probability which is close to that of s transiting to itself, s and s' are supposed to be closely related to each other. In our experiment, we set γ as 0.65. We found that the $\gamma = 0.65$ can find intuitively appropriate related topics. For example, *P2P and Grid* is related to *Web Service and Semantics*, *Frequency Mining* is related to *Classification and Learning*.

Balanced sampling We found that the training samples are imbalanced distributed over two classes. For example, for topic *XML*, there are 9,127 negative samples ($Y = 0$) and 2,517 positive samples ($Y = 1$). Traditional classification model aims to minimize the number of errors made during training under the assumption of balanced data distribution over classes. They are therefore not suitable for class-imbalanced data. Hence, we undersample negative samples [32] to ensure the balanced distribution of positive and negative samples.

5.4 Model Evaluation

In this section, we will evaluate the predicting performance of our model. For comparisons, the regression model proposed in [3] is also tested as the baseline. The baseline model also tried to predict the probability of an individual's topic-following action. But the model uses only one variable a, i.e., the number of already-active friends. The baseline model is formulated as

$$logit[\pi(x)] = \alpha + \beta ln(a + 1) \tag{11}$$

Clearly, the baseline approach only considers the effect of social influence.

We first justify the rationality of the selected explanatory variables. For topic *XML*, we give the parameters of MLR and the baseline model estimated by maximum likelihood method in Table 1. From the table, we can see that in MLR, all the predictors can explain the response variable (all estimated β_is are significant enough ($Sig. < 0.05$)), hence should be imported into MLR model as explanatory variables. Furthermore, we can see that F_{TS} is more influential to response variable than F_{SI} since β_2 as well as its $Wald$ is larger than β_1. Similar results can be obtained on other topics.

Table 1. Parameter estimation. *S.E.* is standard error of coefficients, *Wald* and *Sig.* are Wald Chi-square and P-value that test the null hypothesis of coefficient, respectively.

Model	Para.Name	Value	S.E.	Wald	Sig.
MLR	α	-1.620	0.064	631.5	0.00
	β_1	4.440	0.509	76.08	0.00
	β_2	9.566	0.346	763.6	0.00
baseline	α	-0.472	0.040	142.3	0.00
	β	0.808	0.044	338.4	0.00

Table 2. Predicting performance of MLR and the baseline model on *XML*, $\beta = 1.1$ in F_β computation.

Metrics	MLR	baseline
recall/sens.	72.9%	70.3%
precision	57.9%	47.3%
F_β	65.3%	57.6%
specificity	65.4%	48.7%
accuracy	68.4%	57.2%

Predicting performance of the model In general, the performance of a binary classifier can be measured by $sensitivity$, $specificity$, $precision$ and $accuracy$ [32], where $sensitivity$ (or $recall$) is the proportion of positive samples that are correctly predicted by the model, $specificity$ is the proportion of negative samples that are correctly predicted, $precision$ is the proportion of instances classified as positive that are really positive, and $accuracy$ is the proportion of samples that are correctly predicted either positive or negative. We give these metric results in Table 2 which shows that for all the tested accuracy indicators, MLR is prior to the baseline model. In some applications, for example, finding potential participants of a conference, we hope that more person who are really interested in a certain topic can be found. In other words, in these cases, improving *recall* and *precision* are more preferred. Hence, we also use F_β measure to evaluate the combined score of recall and precision [32].

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall} \tag{12}$$

We set $\beta = 1.1$ to favor recall a little.

Table 2 summarizes the prediction performance of MLR and the baseline model against test samples. We find that MLR outperforms its competitor for each metric. Specially, MLR outperforms the baseline model by about 20% with regard to accuracy, and by 13% with regard to F_β. MLR achieves almost 70% *accuracy* and F_β, which suggests that MLR is practically valuable in real applications.

We further give *accuracy* and F_β on each topic. As shown in Fig. 6 and Fig. 7, the advantage of MLR model over the baseline model can be consistently observed independent on all the tested topics. Fig. 8 further shows the ROC (receiver operating characteristic) curves [33] of MLR and the baseline model, where the area under MLR's ROC curve is 0.743 (area > 0.7 generally implies good predicting performance) suggesting our model is more effective to predict topic-following than the baseline (whose area is 0.638).

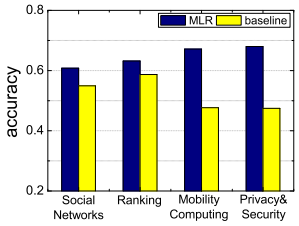

Fig. 6. Comparison of predicting accuracy.

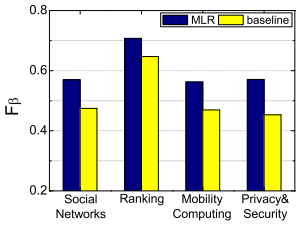

Fig. 7. Comparison of predicting F_β.

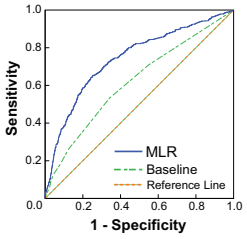

Fig. 8. ROC curves show that MLR is more effective than the baseline.

6 CONCLUSION

Motivated by many real applications, such as call for participation or paper submission, we build a Multiple Logistic Regression model (MLR) to predict the topic that an author will adopt. We build the model upon our understanding about the topic diffusion in Scientific Collaboration Network (SCN). We find that social influence and homophily are mixted together to affect topic-following behavior of authors in SCN through empirical studies. We also uncover the characteristics that social influence affects topic diffusion. By extensive experimental studies, we show that our model can consistently achieves close to 70% accuracy and good F_β. Such results significantly outperform the state-of-the-art competitor model and can be applied in real applications.

References

1. Provost, F.J., Dalessandro, B., Hook, R., Zhang, X., Murray, A.: Audience selection for on-line brand advertising: Privacy-friendly social network targeting. In: Proc. of SIGKDD. (2009)
2. Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., Merom, R.: Suggesting friends using the implicit social graph. In: Proc. of SIGKDD. (2010)
3. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In: Proc. of SIGKDD. (2008)
4. Crandall, D., Dan Cosley, J.K., Huttenlocher, D., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proc. of SIGKDD. (2008)
5. McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a feather: Homophily in social networks. Annual Review of Sociology **27** (2001) 415 - 445
6. Newman, M.E.J.: The structure of scientific collaboration networks. PNAS **98**(2) (2001) 404 - 409
7. Newman, M.E.J.: Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical Review E **64**(016132) (2001) 1 - 7
8. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. PNAS **101** (2004) 5200 - 5205
9. Wu, B., Zhao, F., Yang, S., Suo, L., Tian, H.: Characterizing the evolution of collaboration network. In: Proc. of SWSM. (2009)
10. Aral, S., Muchnika, L., Sundararajan, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS **106** (2009) 21544 - 21549
11. Rogers, E.: Diffusion of Innovations. Free Press (1995)
12. May, R.M., Lloyd, A.L.: Infection dynamics on scale-free networks. Physical Review E (2001)

13. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: Proc. of SIGKDD. (2010)
14. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: Proc. of ICDM. (2010)
15. D.Gruhl, R.Guha, Liben-Nowell, D., A.Tomkins: Information diffusion through blogspace. In: Proc. of SIGKDD. (2004)
16. Lin, C.X., Zhao, B., Mei, Q., Han, J.: Pet: A statistical model for popular events tracking in social communities. In: Proc. of SIGKDD. (2010)
17. Zhou, D., Ji, X., Zha, H., Giles, C.L.: Topic evolution and social interactions: How authors effect research. In: CIKM. (2006)
18. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, C.L.: Detecting topic evolution in scienti?c literature: How can citations help? In: Proc. of CIKM. (2009)
19. Backstrom, L., Huttenlocher, D., Kleinberg, J.M., Lan, X.: Group formation in large social networks: Membership, growth, and evolution. In: Proc. of SIGKDD. (2006)
20. Scholz, M.: Node similarity is the basic principle behind connectivity in complex networks. arXiv:1010.0803[physics.soc-ph] (2010)
21. Benjamin Golub, M.O.J.: How homophily affects diffusion and learning in networks. arXiv:0811.4013[physics.soc-ph] (2008)
22. Fond, T.L., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: Proc. of WWW. (2010)
23. Huang, J., Zhuang, Z., Li, J., Giles, C.L.: Collaboration over time: Characterizing and modeling network evolution. In: Proc. of WSDM. (2008)
24. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: Proc. of SIGKDD. (2009)
25. Peng, H.K., Zhu, J., Piao, D., Yan, R., Zhang, J.Y.: Retweet modeling using conditional random fields. In: Proc. of ICDM Workshop. (2011)
26. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: Proc. of ICWSM. (2011)
27. Choudhury, M.D., Sundaram, H., John, A., Seligmann, D.D.: Contextual prediction of communication flow in social networks. In: Proc. of WIC. (2007)
28. Harris, D., C., C.J., Linda, K., J., S.A., Vapnik, V.: Support vector regression machines. NIPS (1996) 155 - 161
29. Agresti, A.: Categorical data analysis. Wiley, Berlin, Ger. (2002)
30. Kschischang, F.R., Member, S., Frey, B.J., andrea Loeliger, H.: Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory **47** (2001) 498 - 519
31. Goetz, M., Leskovec, J., McGlohon, M., Faloutsos, C.: Information propagation and network evolution on the web. In: Proc. of CWSM. (2009)
32. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3^{rd} ed. Morgan Kaufmann (2006)
33. Mark, G.: Receiver operating characteristic (roc) plots: Fundamental evaluation tool in clinical medicine. Clin Chem **30(4)** (1993) 561 - 567