Towards Topic Following in Heterogeneous Information Networks

Deqing Yang*, Yanghua Xiao*, Hanghang Tong†, Wanyun Cui* and Wei Wang*
*School of Computer Science, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
*Email: {yangdeqing,shawyh,cuiwanyun,weiwang1}@fudan.edu.cn Tel: (86)21-51355555
†Arizona State University, AZ, USA. Email: hanghang.tong@gmail.com

***Abstract*—Who are the best targets to receive a call-for-paper or call-for-participation? What kind of topics should we propose for a workshop or a special issue of next year? Precisely predicting author's topic following behavior, i.e., publishing papers of a certain research topic in future, is essential to answer these questions. In this paper, we aim to model and predict author's topic following behavior in a *heterogeneous* information network. The heart of our methodology is to evaluate the author-author similarity through informative meta paths in the network. The models we propose in this paper can predict not only whether a given author will follow a certain topic but also the topic distribution over all publications in the next year. Extensive experimental evaluations justify that the prediction performance of our approach outperforms the existing approaches across various topics.**

***Keywords*—topic following, heterogenous information networks, meta path**

I. Introduction

In this paper[1], we aim to understand and predict the topic following behavior of individuals in a heterogeneous setting. Specifically, on the *micro-level*, we want to predict whether an author in a bibliographic network will follow a certain topic in the next year, i.e., publish papers of a certain research topic; on the *macro-level*, we try to forecast the topic distribution of overall publications in the next year.

Generally speaking, a user's topic following behavior might be dependent on many different factors, e.g., whether or not he/she has cited the papers on that topic, whether or not he/she has published/attended the venues which also attract papers on the same or similar topic, etc. However, the existing works focus on the *homogeneous* co-authorship networks and therefore largely over-simplify or even ignore these important factors. To address this issue, we aim to model and predict the author's topic following behavior in a more realistic, *heterogenous* information network. Topic following behavior of authors can be understood as a result of topic diffusion in social networks. Understanding the driving forces of topic diffusion in social networks is a fundamental step towards modeling topic following. Most previous works focused on the topic/idea diffusion in the context of *homogeneous* social interaction. For example, an author tends to adopt his/her coauthors' topics [1]. A user tends to buy the mobile phones that his/her friends are using [2], [3].

All these existing works study the diffusion in a *homogenous* network (e.g., coauthor network or friendship network, etc). But in the real world, topics/products/ideas are always diffused in a heterogeneous network. A heterogeneous network contains different types of objects and diverse relationships. For example, a bibliographic network is indeed a heterogenous network consisting of multiple typed objects, such as authors, papers, venues (conferences or journals) and etc., as well as multiple relations such as coauthoring, co-citation, co-occurrence in the same venue. An author's decision to follow a certain topic is probably a consequence of multiple interactions between different objects. For example, in Fig. 1, author a may follow topic *Social Network Mining* because author b whose paper2 is cited by a has adopted this topic before. Author c may also be influenced by author e because they have both attended ASONAM. Such important clues are, however, missed in a homogeneous coauthor network. By shifting our focus from a homogeneous network to a more realistic, heterogeneous setting, we conjecture that we might further improve the prediction accuracy by retrieving such informative factors. Specifically, the central question we aim to answer in this paper is, *how to model the topic following behavior in a real heterogeneous network to maximally boost the prediction performance?*

In this paper, we study how to model topic following behavior in a *Heterogenous Information Network* (HIN for short) that is constructed from publication dataset (e.g., DBLP records and citation data). Currently, modeling author's topic following behavior in HIN was rarely studied and the state-of-the-art user behavior models in HIN cannot be directly used to solve this problem. Most previous works about user behavior modeling in HIN focused on linkage modeling, e.g., coauthor link prediction [4] and citation link prediction [5], where the two ends of these links are both persons. In contrast, our objective is to predict whether an author will follow a topic. It is a kind of link between a person and a topic, e.g., the dashed arrow from author a to topic *Social Network Mining*, or to topic *Graph Mining* in Fig. 1.

In general, we need effective features to model the topic following behavior of authors in HIN. In [4], [5], *Meta paths* were proposed as features to characterize the person-to-person relationships. For example, in Fig. 1, the path $a - paper1 \xrightarrow{cite} paper2 - b$ represents the citation relationship between author a and b. Clearly, there are many such kinds of meta paths connecting two persons and different metal paths play different roles in predicting person-to-person relationships. Meta-paths have been successfully used for measuring the similarity between two persons [4] and entity linking [6]. It is also promising for topic following modeling.

But when we follow the meta-path idea, we still face the following obstacles:
1. it is not trivial to capture the informative meta paths for topic-following modeling. In general, different meta paths have different semantic implications and play different roles for a concrete task.
2. it needs some features that are extracted from meta paths, to accurately model topic following behavior. In the previous works [4], [5], the instance numbers of different meta paths are the important features for modeling author similarity in HIN. However, only counting the number of path instances is not enough to measure the tendency of an author to follow a certain topic.

To address above challenges, we extract a heterogenous bibliographic information network from DBLP publications and citation dataset. This HIN framework allows to encode

[1]This paper was supported by the National NSFC (No.61472085, 61171132, 61033010), by National Key Basic Research Program of China under No.2015CB358800, by Basic research project of Shanghai science and technology innovation action plan under No.15JC1400900, and by Shanghai Science and Technology Development Funds (13dz2260200, 13511504300).

much more information in publications. We further extract the meta paths that are influential on author's decision to follow a topic. Based on these meta paths, we propose a unified measurement to quantify the influences passing along these meta paths. Finally, we build *Multiple Logistic Regression* (MLR for short) model and *Support Vector Machine* (SVM for short) model to predict whether an author will follow a certain topic. Based on our models, we further propose a solution to predict the topic distribution of overall papers in next year.

In general, the more meta paths an author has to the more authors who have followed a topic before, the more likely the author is to follow this topic too. To solve the second challenge, we first identify all candidate meta paths and then cast them into our models. We use a series of statistical indicators to filter out ineffective features. We will elaborate the details in the evaluation section. We propose a unified measure to quantify the influence passing along each meta path. In our measure, we not only consider the number of path instances but also the tendency of the author to follow a certain topic. Evaluation results show that our method is significantly better than the measure only considering path instance number.

The contributions of this paper can be summarized as:
1. We reduce the problem of topic following prediction into an author-to-author similarity evaluation problem in HIN. Under this framework, we propose a unified measure to quantify the influence passing along each meta path.
2. We not only utilize classification model built upon the key features from HIN to realize micro-level prediction of an author's topic following behavior, but also propose a solution for macro-level prediction, i.e., the topic distribution prediction of overall papers in the next year. These predictions are essential to many real applications.
3. We conduct extensive evaluations of our models. The results show that our models can consistently achieve high prediction performance across different topics, highlighting their effectiveness and practicability.

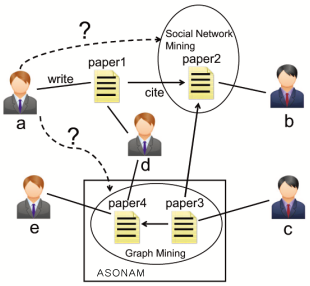

Fig. 1. Illustration of a bibliographic network.

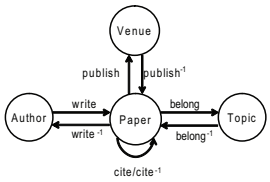

Fig. 2. Schema of bibliographic HIN.

II. Modeling

A. Preliminaries

Definition 1 (HIN): A *h*eterogenous information network is a directed graph $G(V, E)$ with a type mapping function $\phi : V \rightarrow \mathcal{A}$ and an edge mapping function $\psi : E \rightarrow \mathcal{R}$, where each object $v \in V$ belongs to one particular type $\phi(v) \in \mathcal{A}$, and each edge $e \in E$ belongs to a particular relation type $\psi(e) \in \mathcal{R}$.

In this paper, we use DBLP and Arnetminer[2] [7] citation dataset to construct a bibliographic HIN. The schema of the bibliographic HIN is shown in Fig. 2. There are four entity types, namely paper, author, topic and venue (denoted as P, A, T and V in short, respectively). There exists a relationship between each pair of entity type. For example, links between authors and papers represent *'write'* or *'written by'* (denoted as $write^{-1}$) relation. A paper *'belongs to'* a certain topic (the reverse relation is denoted as $belong^{-1}$). A paper may *'cite'* or be *'cited by'* (denoted by $cite^{-1}$) another paper. A papers is *'published in'* (the reverse is denoted by $publish^{-1}$) a certain venue. In general, there exist many different meta paths in the schema of a HIN. Each meta path explicitly or implicitly represents a certain semantic relationship between the two ending typed objects. The concept of meta path is formally defined as follows.

Definition 2 (Meta Path): A meta path is a path defined on the graph of network schema and denoted as $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_n} A_{n+1}$, which defines a composite relation between node A_1 and A_{n+1}.

For example, the coauthor relation in heterogenous bibliographic network can be described as the meta path $A \xrightarrow{write} P \xrightarrow{write^{-1}} A$. To simplify the notation, we just use $A-P-A$. Similarly, the meta path $A - P \rightarrow P - A$ represents the citation relationship between two authors, e.g., the relationship between author a/c and author b in Fig. 1.

For a pair of authors, there may exist different meta paths connecting them. For each of these meta paths, the number of its path instances in the bibliographic network is a good indicator of the similarity between these two authors [8]. In general, the author similarity is positively correlated to the number of meta paths and the number of path instances. This underlies the author-to-author link prediction in heterogeneous bibliographic networks. In this paper, we identify seven types of meta paths indicating author similarity, which are listed in Table I.

TABLE I. Meta paths indicating similarity between authors.

meta path	semantic meaning	feature
$A-P-A$	u and v are coauthors	f_{SI}
$A-P-A-P-A$	u and v have the same coauthor	
$A-P \rightarrow P-A$	u cites v's paper	f_{CI}
$A-P \rightarrow P \leftarrow P-A$	u and v co-cite the same paper	f_{CCI}
$A-P \leftarrow P \rightarrow P-A$	u and v are co-cited by the same paper	f_{CCD}
$A-P-V-P-A$	u and v have papers in the same venue	f_{VT}
$A-P-T-P-A$	u and v write papers of the same topic	

B. Feature Measurement

According to *Homopily* [3], in HIN, two authors connected through many important meta paths are believed to be similar to each other in terms of research interests [8]. In a more precise sense, *an author u has high tendency to follow a topic s, if u is linked to the authors who have adopted the topic s by many meta paths*. In general, such tendency is positively correlated to both meta paths' diversity and the number of path instances of each meta path. Thus, we reduce the seven types of meta paths into five key features to model topic following, as listed in Table I. Then, we propose how to quantify the five features.

Formally, given a research topic s, for each author u in HIN, we need to quantity his/her tendency to follow s in next year in terms of different features. Let $f_P(u)$ be the score function that characterize the motive of u to publish a paper of topic s in the future in terms of meta path P. The larger $f_P(u)$ is, the more likely that u will follow topic s. Then, $f_P(u)$ in general can be defined as[3]

$$f_P(u) = \sum_{v \in U^s} c_P(u, v) \times \zeta(v) \quad (1)$$

where U^s are all the authors who have published the papers of topic s except u himself, $c_P(u, v)$ is the number of instances of P that connects u and v, and $\zeta(v)$ is the prior probability that v publishes papers of topic s, characterizing v's inherent tendency to follow topic s. In contrast, each $f_P(u)$ characterizes the u's tendency to follow topic s along each meta path, which is caused by the influence along the information network. Thereby, all the features corresponding to different meta paths can be defined by Equation 1. Specifically,

[2]http://arnetminer.org

[3]In fact, all feature scores of our models depend on the topic. Without loss of clarity, we omit s in the relevant notations for brief representation.

$f_{SI}, f_{CI}, f_{CCI}, f_{CCD}$ and f_{VT} can be defined in this way. $\zeta(v)$ is defined as the ratio of v's publications of topic s to all of his/her publications. To precisely quantify an author's recent research interest, a time window should be used. By our statistics on DBLP dataset, we found that most authors concern on a specific research topic for three years or so. Thus, we use a time window of three years for relevant computations of the models. That is, $\zeta(v) = \frac{n_s(v)}{n(v)}$ where $n_s(v)$ is the number of v's publications belonging to topic s and $n(v)$ is the number of all v's publications, all publications are in recent three years. Hence, $\zeta(v)$ characterizes v's inherent tendency to publish a paper of topic s.

Another important model feature is *topic similarity* [1], which quantifies the similarity of topic distributions between two authors.

C. Prediction Model

In fact, predicting whether an author will follow a certain topic or not can be viewed as a binary classification. Thus we can try some effective classification models.

Logistic Regression Model is one of the most widely used binary classifiers, hence we firstly try multiple logistic regression (MLR for short) model to predict author's topic following behavior in HIN. More formally, our MLR model is given in Equation 2. We will refer to our model as *MPT* (meta path + tendency). MPT uses 6 explanatory variables listed in Table I to calculate $\pi(u)$, which is the probability author u follows others to publish papers of a given topic and is formulated as

$$\begin{aligned} logit[\pi(u)] = \alpha + \beta_1 f_{SI} + \beta_2 f_{CI} \\ +\beta_3 f_{CCI} + \beta_4 f_{CCD} + \beta_5 f_{VT} + \beta_6 f_{TS} \end{aligned} \quad (2)$$

Support Vector Machine [9] (SVM for short) is also an effective classification model to predict category variables. Thus we can select SVM model to classify each sample into one of the two classes indicating whether a tested author will follow the topic or not. In our experiments, besides MPT we also evaluate SVM's performance on predicting topic following by importing all the 6 features of Table I into the model.

D. Model Sampling

Dataset Description At first, we brief the dataset we used in the evaluation experiments. In order to construct the HIN, we integrated DBLP publication dataset and Arnetminer [7] citation dataset. Without loss of generality, from the dataset we filtered out the papers published up to year 2011 in seven research categories[4] that are related to *database*, *data mining*, *World Wide Web* and etc. Thus, the HIN for model evaluation contains 193,194 authors, 186,952 papers, 557,916 coauthor relationships and 4,344,955 citation relationships.

Topic Identification To collect samples for model training and testing, we have to identify the topic for a publication that is a preliminary step. At first, without loss of generality we selected 25 (T=25) popular topics, e.g., *Social Network Mining, Query Processing, Rich Media* and etc., from some prevalent conferences such as KDD, ICDM, WWW and etc. As in [1], by referring to track or subject classification for the publications in these conferences we can get some publication samples with identified research topics. Based on these samples, we trained a SVM classifier to identify the rest publications' topics in our dataset. The classifying features of SVM classifier were extracted from the titles of the training samples, i.e., title keywords.

Bias Reducing After identifying topics, we collected publications in year $[2004, 2008]$ as the training data and the publications in year 2009 as the test data. We trained the model for each topic individually since the model's parameters are topic sensitive. Suppose now we need to generate samples for a certain topic s, we only considered those authors who published at least 3 papers in one year as the *valid* samples. Then, all valid authors were collected for each year t in $[2004, 2008]$. Concretely, each pair $< u, t >$ ($2004 \leq t \leq 2008$ and u is a valid sample) can be regarded as one training pair sample for topic s. Now, for each pair sample $< u, t >$, we need to assign 0 or 1 to the binary response variable Y. We set Y=1 if author u publishes at least one paper of topic s or other topics *closely related* to s during a time window $[t, t+2]$, otherwise Y=0. We selected a three-year time window because that it generally takes one or two years (or even more) for an author to follow a certain topic. Moreover, it also takes time for a topic to be diffused to more authors especially for a new topic. At last, we have 13,734 training samples and 2,788 test samples for each topic.

[4]http://academic.research.microsoft.com

III. Evaluation

A. Model Training

We first justify the effectiveness of explanatory variables which can reported through training MPT model. For this sake, we begin with the results for a concrete topic, i.e., *Social Network Mining*. Table II lists the results of parameter estimation when we force all explanatory variables into MPT. All parameters are estimated by maximum likelihood method. From the table, we can see that in MPT, all the explanatory variables can explain the response variable ($Sig.$ <0.05) except for f_{SI}. Hence f_{SI} should be excluded from MPT model for this topic when predicting. Furthermore, β_5 has the largest $Wald$ value in all β_is, implying that f_{VT} is the most significant feature to characterize an author's topic following behavior for topic *Social Network Mining*.

Then, we built MPT model for all 25 topics and recorded the significance of each feature in each topic model. We found that, in general the effectiveness of each feature may vary on different topics. We used $Sig.$ <0.05 as the criteria to import a feature into the model of predicting a given topic. Then we counted the number of each feature to be imported. The results show that, f_{CCI} and f_{TS} are imported by all topics' models. It implies that *co-citing* and *topic similarity* are most significant to impact topic following behavior through different topics.

TABLE II. MPT's parameter estimation. *S.E.* is standard error of parameters, *Wald* and *Sig.* are Wald Chi-square and P-value that test the null hypothesis of coefficient.

feature	para.	value	S.E.	Wald	Sig.
$const.$	α	-1.349	0.043	965.92	0.00
f_{SI}	β_1	0.589	0.343	2.947	**0.086**
f_{CI}	β_2	3.183	0.715	19.82	0.00
f_{CCI}	β_3	13.227	1.111	141.79	0.00
f_{CCD}	β_4	3.732	0.868	18.47	0.00
f_{VT}	β_5	5.139	0.343	**224.58**	0.00
f_{TS}	β_6	2.550	0.281	82.37	0.00

B. Baselines

The first baseline (denoted as *LR1*) was proposed by [2]. The second baseline (denoted as *LR2*) was proposed in [1]. They were both designed for the same problem of predicting topic following behavior in homogenous coauthor networks. The third baseline (denoted as as *MPC* (meta path + count)) is a complicated model using meta paths was proposed [4]. But it only considered the number of path instances between two authors. MPC can also be formulated as a logistic regression model

$$logit[\pi(u)] = \alpha + \sum \beta_i f'_P(u) \quad (3)$$

where P is a meta path type and can be SI, CI, CCI, CCD and VT, and $f'_P(u)$ is quantified by $c_P(u, v)$ as defined in Equation 1. All these three baselines belong to logistic regression family.

In fact, topic following behavior can also been predicted by *Collaborative Filtering* (CF for short) model if we regard a topic as an item. An author's rating on a topic can be set as 1 if s/he has followed the topic, otherwise as 0. Specifically, given an author u, we adopt a user-based CF model as the last baseline, to predict u's rating on a certain topic s, i.e., r^s_u, which is formalized as

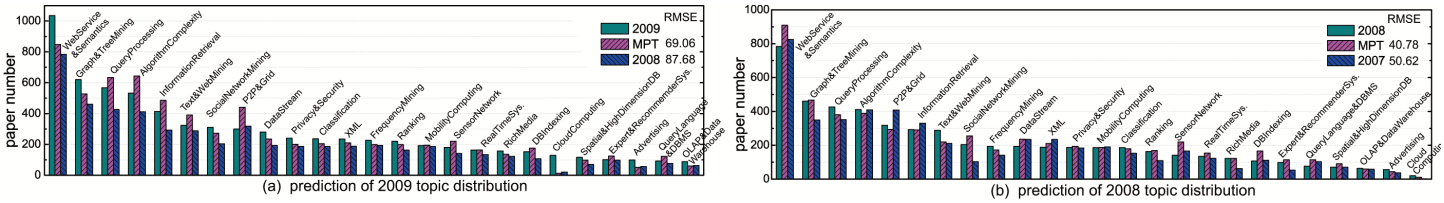

Fig. 3. Comparisons of topic distribution show that MPT prediction is more closer to the ground truth than the naive solution.

$$r_u^s = \sum_{v \in Ngb(u)} \frac{sim(u,v)}{\sum_{v \in Ngb(u)} sim(u,v)} \times r_v^s \quad (4)$$

where $Ngb(u)$ represents u's neighbors who are linked to u along different meta paths in HIN. $sim(u,v)$ is the similarity between u and v that is also quantified by $c_P(u,v)$ as defined in Equation 1. u is predicted to follow s if $r_u^s \geq 0.5$. Obviously, compared to MPT and SVM, this CF model does not consider the feature of topic similarity, i.e., f_{TS}.

C. Prediction Performance

Micro-level Prediction On the micro-level, we use the competitors *not only to predict whether an author follows a new topic, but also to predict following behavior on an old topic of which s/he has published papers*. We use $sensitivity$, $specificity$, $precision$, $accuracy$ and F_β to measure model's prediction performance.

TABLE III. MODEL'S PREDICTION PERFORMANCE ON *Social Network Mining*, $\beta = 1.5$ IN F_β METRIC.

metric	MPT	SVM	MPC	LR2	LR1	CF
recall/sens.	74.52%	**75.13%**	67.64%	74.22%	67.64%	43.78%
F_β	67.03%	**67.49%**	61.05%	64.54%	57.14%	49.40%
accuracy	69.05%	69.29%	64.60%	64.42%	55.85%	**70.19%**
precision	54.67%	54.92%	50.07%	49.90%	42.34%	**61.16%**
specificity	66.04%	66.09%	62.92%	59.03%	49.36%	**84.71%**

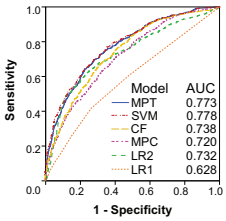

Fig. 4. The ROC curves show that MPT and SVM perform best in prediction.

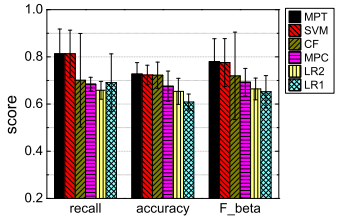

Fig. 5. Performance comparisons on all 25 topics show that our models beat the baselines evidently.

Table III gives the prediction performance of MPT and SVM compared with the competitors against test samples in year 2009, w.r.t. topic *Social Network Mining*. We can see that SVM gets the best recall and F_β to which MPT's performance nearly equals. Moreover, MPT and SVM get a good accuracy very close to CF's. Although CF owns the best specificity and precision, it gets the worst recall and F_β indicating that many real positive samples are mistakenly predicted as negative by CF. It is because that most neighbors did not follow the topic before and topic similarity is ignored in CF prediction, implying CF is not applicable to this scenario. Moreover, we can refer to AUC (area under ROC curve [10]) to learn the performance of binary classification. Therefore, we depicted the AUCs of all models in Fig. 4. Except for LR1, the rest models' AUCs are beyond 0.7 ($AUC > 0.7$ generally implies good prediction performance). MPT's AUC (0.773) almost equals to SVM's AUC (0.778) and they both beat the rest competitors. All these results suggest the performance superiority of MPT and SVM over its competitors implying that our proposed models are practically valuable in real applications. The improvement of our models over MPC shows that our measures which take into account the tendency of neighbors to follow a certain topic are crucial for a more accurate prediction. The improvement of our models over LR1 and LR2 shows that meta-path-based features we extracted from HIN are very effective for modeling topic following. In order to extensively justify our models' performance, we further give the average scores and the standard deviations of *recall*, *accuracy* and F_β on all 25 topics. As shown in Fig. 5, we found that MPT and SVM can consistently beat the competitors cross all the tested topics, especially in *recall* and F_β.

Macro-level Prediction Next, we show the prediction results of topic distribution over all papers in HIN by MPT model. Given an author u, for each topic, we can calculate a $\pi(u)$, i.e, the probability that u will publish papers of this topic according to Equation 2. Thus, by normalizing each $\pi(u)$ over all topics, we get u's probability distribution over all topics. Suppose the author u will totally publish n papers in the next year. Then, we can sample n times from the probability distribution independently. We used the result as his/her predicted publications over different topics. By summarizing all authors' paper numbers, we got the prediction result of topic distribution over all papers of next year. For comparison, we also show the result of a macro-level prediction solution which directly uses the topic distribution of the previous year as the prediction. The comparison results on topic distribution prediction of year 2009 and year 2008 are respectively shown in Fig. 3(a) and (b). It is clear that, our MPT-based predication results are much closer to the ground truth for both two years. For example, the root-mean-square-error (*RMSE*) of MPT prediction on paper number to the ground truth of 2009 is 69.06 which is smaller than the baseline solution (87.68).

IV. CONCLUSION

Driven by some real applications, in this paper we aim to model author's topic following behavior, i.e., publishing papers of a certain research topic in heterogenous information networks. For this purpose, we evaluate the similarity between different authors by seeking the different types of meta paths in the network. Based on the informative features that are extracted and quantified from meta paths, we build a multiple logistic regression and SVM model. Extensive experiments on micro-level and macro-level predictions show that our models can consistently achieve evident superiority over the state-of-the-art competitors.

REFERENCES

[1] D. Yang, Y. Xiao, B. Xu, H. Tong, W. Wang, and S. Huang, "Which topic will you follow?" in *Proc. of ECML-PKDD*, 2012.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proc. of SIGKDD*, 2008.

[3] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415 - 445, 2001.

[4] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Proc. of ASONAM*, 2011.

[5] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen? relationship prediction in heterogeneous information networks," in *Proc. of WSDM*, 2012.

[6] W. Shen, J. Han, and J. Wang, "A probabilistic model for linking named entities in web text with heterogeneous information networks," *Proc. of SIGMOD*, 2014.

[7] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proc. of SIGKDD*, 2008.

[8] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," in *Proc. of VLDB*, 2011.

[9] D. Harris, C. J. C., K. Linda, S. A. J., and V. Vapnik, "Support vector regression machines," *NIPS*, pp. 155 - 161, 1996.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, 3^{rd} ed.* Morgan Kaufmann, 2006.