Infuencee Oriented Topic Prediction: Investigating the Effect of Influence on the Author

Murat Yukselen
yukselen@ceng.metu.edu.tr
Middle East Technical University (METU), Computer Eng. Dept.
Ankara, Turkey

Alev Mutlu
alev.mutlu@kocaeli.edu.tr
Kocaeli University, Computer Eng. Dept.
Kocaeli, Turkey

Pinar Karagoz
karagoz@ceng.metu.edu.tr
Middle East Technical University (METU), Computer Eng. Dept.
Ankara, Turkey

ABSTRACT

In this paper, we study the problem of topic adoption prediction for an author within a social academic network. The previous efforts on the problem use topic similarity and topic adoption of co-authors. We model the problem with an influence detection point of view, and propose that the influence on the author is an important factor. Hence, we define a novel influencee prediction based feature. To this aim, in this work, an algorithm is proposed to calculate the influence propagated towards the author. The effect of this feature is explored together with and in comparison to other features used in the literature for the problem. The experiments conducted on Arnet Miner data set show that accumulated influence on author is effective for predicting topic adoption.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Web searching and information discovery**; **Social recommendation**; • **Computing methodologies** → **Machine learning approaches**.

KEYWORDS

information flow, social networks, link prediction

ACM Reference Format:
Murat Yukselen, Alev Mutlu, and Pinar Karagoz. 2019. Infuencee Oriented Topic Prediction: Investigating the Effect of Influence on the Author. In *Proceedings of ACM Conference (Conference'19).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social network analysis is concerned with analysing the structure of the network and behaviour of individuals forming the network [23]. Although early studies in social network analysis focused on building descriptive models, with increasing amount of social network data, the research direction moved to building predictive models [8]. Such predictive models can be used in a variety of domains such as link prediction for friend recommendation [4, 5], influence detection for advertisement domain [20], and community detection for urban safety domain [15].

In this study, we work on *academic social networks* (also called *scientific collaboration networks* [26]) with a predictive point of view. In academic social networks, nodes represent entities such as authors and papers and edges represent relations such as authoring a paper and co-authorship. There are various studies in the literature on academic social networks aiming to predict collaboration patterns [10, 16, 18]. We focus on the problem of topic adoption, and propose a method to predict topic adoption of an author. More specifically, given an author and a new topic for the author, we aim to predict whether the author will publish a work on the given topic in the next time slot. Yang et al. studied this problem in [26], and propose to use two features, *topic similarity*, and *social influence*, in a regression based model. Social influence denotes the co-authors of the author who already adopted the given topic, and topic similarity denotes the similarity between the given topic and the topics that are already adopted by the author.

In this work, we argue that the amount of influence accumulated on the author for the given topic can affect the topic adoption. To this aim, we propose a new feature, *influencee score*, and propose an algorithm to compute this feature. The proposed algorithm is inspired from the influence detection method given in [24], in which the amount of influence going out from an author to other authors within social stream is computed. For computing influencee score, we invert the flow of the stream and find the amount of influence *towards the author* from other authors through social stream in a time line. The effect of this new feature for topic adoption prediction is analyzed within a multiple logistic regression model together with and in comparison to the features given in [26]. Additionally, we compare the performance against the baseline model of [26], as well. The experiments reveal that the proposed feature improves the prediction accuracy.

The contributions of this work can be summarized as follows:

- For topic adoption in scientific collaboration networks, we propose a new feature, *influencee score* (F_{InfSc}).
- For computing the *influencee score* of an author, an algorithm is proposed.
- The effect of *influencee score* for topic adoption prediction is analyzed within a multiple regression model.

The rest of the paper is organized as follows. In Section 2, related studies are summarized. The background studies, which are adopted in this work, are described in Section 3. Section 4 includes the detailed description of the proposed feature and the algorithm for its computation. We present the experiments and results in Section

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'19, June 2019,
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

5. Finally, the overview of the work and future directions are given in Section 6.

2 RELATED WORK

In this section, we summarize the related studies in the literature on influence detection and social network analysis on scientific collaboration networks.

In a social network, action correlations of the agents are studied and defined as a result of social influence, homophily or confounding (environment) effects [1]. In [17], homophily and influence are studied within various sociological aspects, from relationship types to different sized communities. Network of ties, connections between individuals are affected heavily from homophily and in turn open to receive more influence. Combined effects of social influence and homophily are studied in large scale networks such Wikipedia and LiveJournal in [7]. Social influence and homophily effects are investigated through randomization tests on first-order effects, leaving second-order effects such as community and structural similarity as a future work [13]. The problem of distinguishing social influence and homophily is studied in [2]. In [1], social influence is studied to be more identifiable among other correlations.

In [18], collaboration patterns of authors in scientific papers are studied through statistical network features. It is reported that coauthor networks tend to include simple collaboration patterns. The study in [10] extends the scientific collaboration network structure with citation information in order to analyze topic modeling and evolution. In [16], the authors further use the text content as well as the network structure. They aim to track the popularity of the events and discover the evolution of the events over time as event diffusion on Twitter and DBLP. In [11], individual collaboration networks are studied in order to predict the evolution of collaboration. The study focuses on social collaboration network under computer science field on a 25 year time-window both at community level and individual level. In [24], the authors analyze scientific collaboration network in order to predict topic adoption. The topic similarity and co-author topic adoption are reported as features effective on topic adoption prediction.

Information diffusion [21] is studied on various types of social networks such as blogspace [9]. Through user behavior modeling on features such as neighborhood, topic and recipient, in [6], communication flow predictability is studied on MySpace network. In [14], the authors studied information diffusion on blogosphere data without incorporating post contents. It is reported that information cascades mainly as a tree. Feature implementations of the this work is also compatible with this observation. In [19], retweeting on Twitter is studied as an information diffusion problem for behavior modeling. The authors use conditional random fields with features such as content influence, network influence and temporal decay. Within scientific collaboration network including author-topic interaction, the work in [3] studies group and community growth and evolution. In [26], the authors aim to detect most influential authors in the academic social network and propose a social stream based solution.

Our work shares several common properties with previous studies on academic social networks. In addition to co-authorship network, we use textual content, as in [16], in limited to paper title and abstract. The most similar one among such studies is the one in [26], challenging the same problem of topic adoption prediction. However, the main difference of our work is that a novel feature is proposed. In order to realize this new feature, we incorporate ideas from the literature on information diffusion and influence detection. More specifically, the proposed algorithm for computing our new feature has its roots from the solution given in [24], however the structure of the social stream is changed, and hence the proposed algorithm has considerable differences. Since we have adopted several basics from the works in [26] and [24], in Section 3, we present further details for these two studies.

3 BACKGROUND

This study is motivated by approaches presented in [24, 26] to predict topic adoption of an author. In the subsequent sections we explain these studies in detail.

3.1 Efficient Influence Querying

In [24], authors propose a method to query influencers in dynamic social networks in a context-sensitive and time-aware manner. Authors assume that social stream is generated by propagating contents, represented as a bag of keywords, between users, and user a_j is influential on user a_k if there is a significant content flow from a_j to a_k. To calculate the influence score of a_j on a_k, authors propose a set of concepts as summarized below:

- Valid flow ($\mathcal{F}$): Flow of a keyword, K_i, from a_j to a_k is an ordered set of nodes, $a_j, b_1, \ldots, b_r, a_k$, such that a_j is the initiator of the keyword, there is a directed edge between b_i, b_{i+1}, for all i = 1 ... r - 1 and every node transmits the keyword to its neighbor after receiving the keyword.
- Flow duration (δt): This metric indicates time elapsed between initiation of a keyword K_i by a_j and its transmission to a_k.
- Decayed flow weight: δt indicates latency and large value of δt for K_i may indicate decay in its significance in the flow F. To incorporate this assumption into the model each flow is assigned with a weight given by $2^{-(\lambda \delta t)}$ where λ is a decay factor.
- Aggregate flow path: Aggregate flow path for keyword K_i at time t_c along a particular path, P, is the sum of weights of all valid distinct flows of the keyword along the same path. Two flows are considered distinct if the flows are initiated at different times.
- Aggregate pairwise flow: Aggregate pairwise flow between a_j and a_k for keyword K_i at time t_c is the summation of the aggregate flow paths on every path from a_j to a_k.
- Atomic influence value: Atomic influence value of a_j on a_k is the summation of all aggregate pairwise flow score of each keyword originated in a_j and transmitted to a_k.

Authors implement the above mentioned concepts on a data structure called *Flow Path Tree*, $\mathcal{T}$. In $\mathcal{T}$, root is *null* node, a path from root a leaf corresponds to a valid flow, and each node is associated with a weight corresponding to flow weight from root the that node. Algorithm 1 outlines the influencer score generation process.

Authors investigated performance of the proposed method on an academic social network. In their setting, each node corresponds to

Algorithm 1 Original UpdateFlowPaths algorithm

```
function UPDATEFLOWPATHS($a_i, G(t), \mathcal{S}, \mathcal{T}$)    ▷ $a_i$: Originating Node, $G(t)$: Network, $\mathcal{S}$: Social Stream, $\mathcal{T}$: Flow-Path Tree
  Receive the next message containing keyword $K$ in social stream $\mathcal{S}$ originating at node $a_i$
  Create singleton node $a_i$ in tree $\mathcal{T}$ as child root of node if it does not already exists
  $C \leftarrow a_i$                    ▷ Set of candidate paths for expansion
  Update weight of singleton path containing only node $a_i$ in tree $\mathcal{T}$ by 1
  while $C$ is not empty do
    Delete the first path $P$ from $C$ and denote the first node of $P$ by $a_j$
    for each $a_k \notin P$ in $V(t)$ with an incoming edge to $a_j$ do
      if prefix of path $a_k \bigoplus P$ exists in $\mathcal{T}$ and $a_k$ has propagated keyword $K$ prior to $a_j$ then
        if the complete path $a_k \bigoplus P$ exists in $\mathcal{T}$ then
          Increment weight of last node of path $a_k \bigoplus P$ by 1 in $\mathcal{T}$
        else
          Create last node of $P$ as child for prefix of path $a_k \bigoplus P$ in $\mathcal{T}$ with weight as 1
        end if
        Add $a_k \bigoplus P$ to $C$
      end if
    end for
  end while
end function
```

an author and edges between nodes are placed based on co-author relationship. Keyword set is constructed by extracting uni-, bi-, and tri-grams of words appearing in the title and the abstract of the papers.

3.2 Topic Adoption Prediction for Authors

In [26], Yang et al. investigate the topic-following behavior of researchers and propose a topic-following model to predict topic of the next publication of a researcher. Authors claim that social influence and homophily are driving factors of topic-following for a researcher. In case of scientific collaboration network, social influence corresponds to tending to adopt a topic that is most widely studied among researcher's co-authors and homophily corresponds to similarity of scientific publications. Authors also argue in their study that assigning a weight for these factors is a difficult task and build multiple regression model, with social influence score and homophily score as independent variables, to predict topic adaption. The multiple regression model is given in 1 where F_{SI} and F_{TS} are, respectively, special influence score and homophily score of an author.

$$logit[\pi(x)] = \alpha + \beta_1 F_{SI} + \beta_2 F_{TS} \tag{1}$$

Social influence is calculated according to 2 where $F_{SI}(u, s, t)$ indicates the probability of researcher u will follow topic s in year t. In 2, $N'(u)$ indicates co-authors of u who published on topic s before u, $w(e_{u,v})$ is the weight of the edge between u and v and $f(v, s, t-1)$ indicates the influence from u's neighbor v in $t-1$.

$$F_{SI}(u, s, t) = \sum_{v \in N'(u)} \frac{w(e_{u,v})}{\sum_{v \in N'(u)} w(e_{u,v})} \times f(v, s, t-1) \tag{2}$$

Homophily score of an author, u, is calculated according (3). Topic similarity of u's and his/her co-authors' publications, where u' is aggregated paper counts of u and $U^s_{<t}$ is the aggregated paper counts up to time t. Topic similarity of two authors, u and v, is calculated based on cosine similarity given in (4) where u and v are vectors and v_i indicates number of publications by v in the i^{th} topic.

$$F_{TS}(u, s, t) = sim(u', U^s_{<t}) \tag{3}$$

$$sim(u, v) = cosine(u, v) = \frac{u \cdot v}{\| u \| \| v \|} \tag{4}$$

To evaluate their model, authors consider publications in three-years intervals. To predict topic of a paper published in year t, papers published in years $[t-3, t-1]$ are considered for feature calculations and observation is made in year $[t, t+2]$ interval.

4 PROPOSED METHOD: INCORPORATING INFLUENCE FACTOR IN TOPIC PREDICTION

In this study we propose a new feature, influence score, and present a model that incorporates it together with social influence, and topic similarity. Effect of social influence and topic similarity in topic adoption are discussed in [26] and are implemented in a similar fashion in this study.

Incorporating influencee score in topic adoption is motivated by [24]. However, ideas presented in [24] can not be applied directly as [24] provides mechanisms to capture influencer score, i.e. $I(u, *, keywords(s), t)$, while we are interested in influencee score which is formulated in Equation (5). In this equation, u represents an author, s represents a topic and t represent time. Function $I(\cdot)$ calculates the accumulated influence on user u for topic s, represented by a set of keywords *keywords(s)*, at time t.

$$F_{InfluenceeScore}(u, s, t) = I(*, u, keywords(s), t) \tag{5}$$

Although the data structure to represent scientific collaboration network presented in [24], namely *Flow Path Tree* - $\mathcal{T}$, forms basis for influencee score calculation, it can not be directly adopted in this study. In $\mathcal{T}$, influencers appear under nodes representing keywords and ranked influencers appear 2 levels down from the virtual root node and for efficiency issues nodes at deeper levels are pruned. In this study, we are interested in influencee score and nodes representing influencees appear at deeper levels of the tree, ideally at leaves. Hence, such a pruning mechanism avoids representation of influencees. In order to overcome this limitation, the flow path tree is held in reverse order and the modified algorithm consumes social stream backwards. Figure 1 illustrates the modified data structure where directed arrows between authors (w_y, w_u) and (w_v, w_u) denote influence propagation aligned with time as it becomes more recent near to influencee node w_u under keyword node k_1. Author nodes have influence weights and timestamp for their last update. In that sense author w_u will not have any weight since we are interested in how much influence it receives.

When an influence emerges, influencer node receives weight for the timestamp. The update incorporates decay in a way that when a timestamp changes, here we are receiving older messages and timestamp decreases, weight is decayed according to time delta and new weight is added. Delta of the timestamps between nodes is shown as distance between author nodes in Figure 1. At step 1 for $Stream_{t-1}$, author v and y publish a paper at $t-1$ and because of their author graph G_{t-1} it is a valid influence. At step 2 for $Stream_{t-2}$, it is understood that author y has influenced before, so node w_y's weight is decayed and it is displaced to $t-2$.

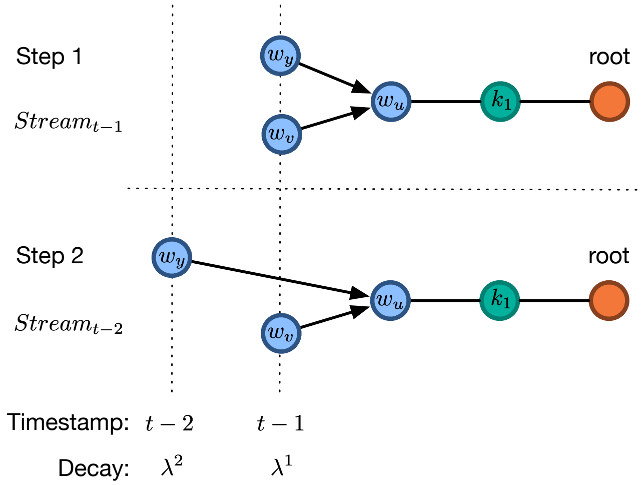


Figure 1: Example flowpath tree

Algorithm 2 Influencee score f_{InfSc}

```
function FeatureInfSc($u, s, t$)    ▷ $u$: author, $s$: topic, $t$: time
  $S_{reversed}$ ⟵ filter the social stream $S$ up to time $t$ with keywords for $s$
  $\tau$ ⟵ a flow-path-tree for author $u$ with all $u$'s neighbors are expanded with weight 0
  $score$ ⟵ 0
  for all $p$ in $S_{reversed}$ do    ▷ $p$: paper as message
    for all $k$ in $p.filteredKeywords$ do
      for all $v$ in $p.authors$ do
        updateFlowPathTree($\tau$, $k$, $v$, $t_p$)    ▷ $t_p$: time of $p$
      end for
    end for
  end for
  return $aggregatedScore(\tau, t)$ ▷ $Influence(*, author_u, *, t)$ as in [24]
end function
```

In Algorithm 2, τ tree is initialized layer by layer as virtual root node, keyword nodes, $author_u$ with zero weight, $author_u$'s friends $author_v$ with zero weight. Reversed stream of papers are then used to construct τ that holds information specifically to answer how much influence $author_u$ received. Updating the flow paths is described in algorithm 3.

Algorithm 3 UpdateFlowPaths function

```
function updateFlowPathTree($\tau$, $k$, $v$, $t$)
  Starting from keyword node $k$, traverse all open
  for all author $node_w$ as direct friend of influencee do
    if $node_w$ is $v$ then
      Increment weight incorporating decaying and return
    end if
    for all child $node_x$ of $node_w$ do
      $dfsPathUpdate(node_x, v, t)$
    end for
  end for
end function
function dfsPathUpdate($node_x, v, t$)
  if $t > t_{node_x}$ then return
  else if $node_x$ is $v$ and $t$ is $t_{node_x}$ then
    increment weight incorporating decay and return
  else if $v$ is friend of $node_x$ then
    append $node_v$ to $node_x$ with decayed weight and return
  end if
  for all already open $node_y$ of $node_x$ do    ▷ these friends already conveyed information to influencee
    recurse with $dfsPathUpdate(node_y, v, t)$
  end for
end function
```

Influence is calculated as an accumulation of all paths from keyword nodes to leaves where path score is summed by decayed weight with respect to author node timestamp. The effect of decay is similar to Katz measure [12]. This aggregation process is given in Algorithm 4.

Algorithm 4 Aggregate Influencee score from τ

```
function aggregatedScore($\tau, t$)    ▷ $\tau$: flow-path tree, $t$: time
  $score$ ⟵ 0
  for all $node_k$ in $\tau$ do
    for all path $p$ starting from $node_k$ do
      for all $node_a$ in $p$ do
        add $weight_{node_a} * decay(t, t_{node_a})$ to $score$
      end for
    end for
  end for
  return $score$
end function
function decay($t_1, t_2$) return $2^{(-decayFactor*(t_1-t_2))}$
end function
```

5 EXPERIMENTS

5.1 Data Sets and Experimental Setting

To evaluate performance of the proposed topic adoption model a set of experiments is conducted on data retrieved from Arnet Miner [25] and Microsoft Academic Graph (MAG) [22].

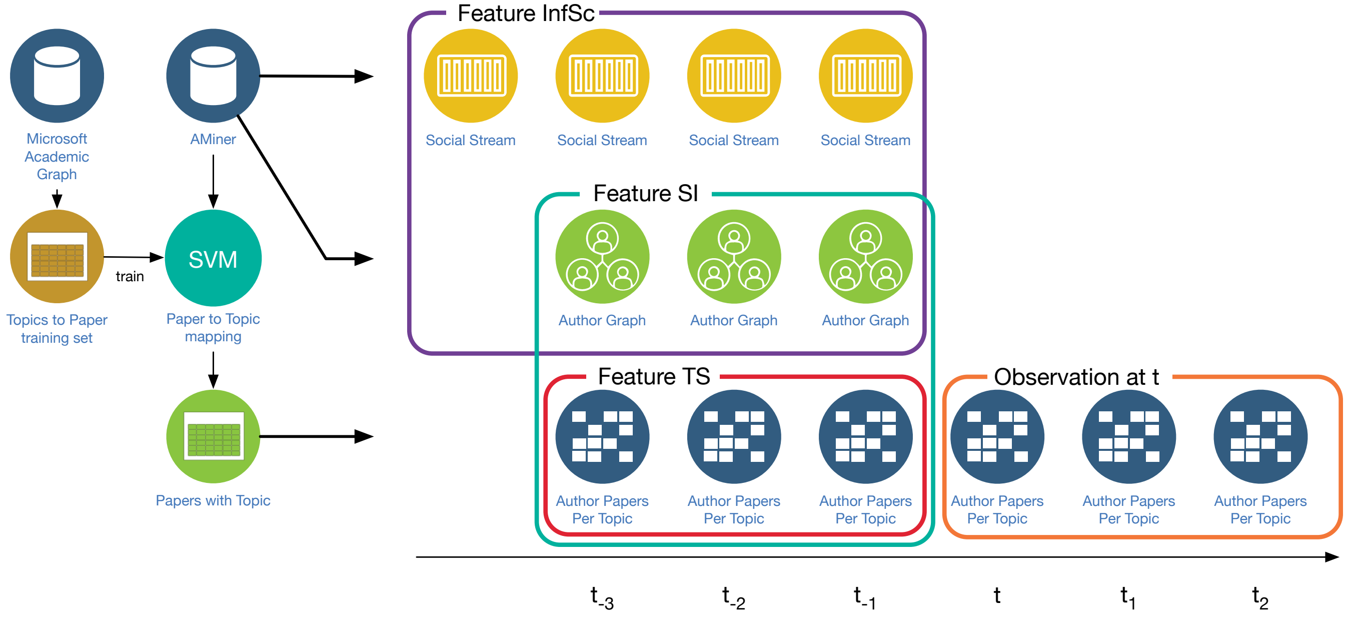


Figure 2: Data flow of the features

MAG database is used retrieve papers to build a SVM to predict topic of a paper. To this aim, we retrieved top 10 papers belonging to topics of algorithm and complexity, classification, information retrieval, privacy and security, and text and web mining. We retrieved uni-, bi-, and trigrams from titles and abstracts of the papers and used these keywords to train SVM.

From Arnet Miner database we retreived papers published between 2001 and 2011 and assigned their topic using our SVM model. In Table 1 we list the number of papers retrieved from Arnet Miner [25] and give yearly tagged papers in their respective columns.

Although papers have topics, in our model network is based on authors and keywords. For this purpose, for each topic we build a keyword dictionary that contains top 30 terms of the topic. To query the stream for a specific topic we use the topic's keyword dictionary. The statistics of the filtered stream is given in Table 2.

Figure 2 illustrates our experimental setting. The leftmost part visualizes the data preparation step, data gathering and SVM training. The right part of the Figure 2 visualizes model training and prediction steps. For an author, say u, that has not published any paper on topic s, his/her stream for the past three years is queried with $keyword(s)$ which is made up of top 30 keywords on the topic. Influence Score feature needs yearly social streams and author graphs in computation. Similarly, Feature Social Influence makes use of Author Graphs and yearly Author Paper Counts per Topic. Feature Topic Similarity is simply calculated based on the cosine similarity of user u with respect to all users on topic s up to time t. Observation result is a simple lookup in 3 consecutive years whether u has a publication on that topic or not.

In order to analyze the effect of the features, various models are trained with Multiple Logistic Regression having different set of features. Each model name is set such that it expresses the set of features used. In that respect, *mlr* as a short name was used in [26] as a model with two features: F_{SI} and F_{TS}. This work focused on adding a third feature as $F_{InfluencesScore}$ (F_{InfSc}). The experiments are conducted with a balanced number of samples for training and test, as given in Table 3. The numbers of positive and negative samples are balanced as well.

Table 1: Yearly paper counts per topic

Year	Total	Alg. C.	Class.	IR	P.& Sec.	W.& T. M.
2001	56783	19464	1678	3990	2448	5833
2002	62029	20262	2096	4458	2903	6358
2003	54874	18184	1900	3913	2528	5672
2004	61169	19635	2315	4402	3279	6034
2005	98203	31166	5036	6575	5591	8539
2006	121189	41547	6246	7991	5896	9894
2007	116194	36312	6496	7913	5778	10448
2008	128681	38931	7878	8754	6029	11124
2009	169575	50779	11545	10381	6967	13764
2010	139357	42951	8446	9060	5653	12555
2011	131262	41350	7553	8556	5078	12355

5.2 Experimental Results

For the selected topics, the parameters of each model are given in Tables 4, 5, 6, 7, and 8 respectively.

In the model parameter tables, Wald and significance values help to interpret the feature contributions such that positive Wald value is considered worthy feature for the model at hand. Similarly parameters with significance value smaller than 0.05 are also useful. In light of this information, model parameters can be interpreted

Table 2: Social Stream Paper Counts per Year for Influencee Score Calculation

Year	count
2001	35447
2002	48473
2003	46550
2004	52027
2005	87755
2006	108749
2007	106364
2008	121537
2009	162249
2010	133040
2011	124158

Table 3: Training and test sample counts per topic

	Alg. Complex.	Class.	Inf. Retrvl.	Prv. & Sec.	Web & Text Mining
train	4043	4090	4115	4174	4153
test	1991	2067	2031	2034	2077

Table 4: Model parameters of Alg. and Complex. Topic

Model	Feature	Par.	Value	Std.Err.	Wald	sig.
F_{CE}	**intercept**	α	0.0427	0.0042	10.0701	0
	F_{CE}	β_1	0.7610	0.0014	532.3658	0
$F_{SI} + F_{TS}$	**intercept**	α	0.0232	0.0040	5.6945	6.2524e-09
	F_{SI}	β_1	-0.0098	0.0078	-1.2601	0.1038
	F_{TS}	β_2	3.7651	0.0074	506.5743	0
$F_{SI} + F_{TS} + F_{InfSc}$	**intercept**	α	-1.3639	0.0111	-121.9350	0
	F_{SI}	β_1	-0.0170	0.0242	-0.7022	0.2412
	F_{TS}	β_2	3.8134	0.0240	158.7485	0
	F_{InfSc}	β_3	0.0853	0.0109	7.8076	3.6637e-15
F_{TS}	**intercept**	α	0.0158	0.0029	5.3932	3.4766e-08
	F_{TS}	β_1	3.8065	0.0055	688.4928	0
$F_{SI} + F_{InfSc}$	**intercept**	α	-0.0411	0.0110	-3.7387	9.3618e-05
	F_{SI}	β_1	0.0127	0.0308	0.4128	0.3398
	F_{InfSc}	β_2	0.1062	0.0145	7.3054	1.6220e-13
$F_{TS} + F_{InfSc}$	**intercept**	α	-1.3688	0.0081	-167.9981	0
	F_{TS}	β_1	3.8420	0.0179	214.6297	0
	F_{InfSc}	β_2	0.0440	0.0069	6.3233	1.3570e-10

Table 5: Model parameters of Classification Topic

Model	Feature	Par.	Value	Std.Err.	Wald	sig.
F_{CE}	**intercept**	α	-2.0091	0.0029	-685.3135	0
	F_{CE}	β_1	1.0696	0.0022	465.7863	0
$F_{SI} + F_{TS}$	**intercept**	α	-2.5683	0.0046	-553.9986	0
	F_{SI}	β_1	0.0879	0.0184	4.7673	9.3854e-07
	F_{TS}	β_2	3.6437	0.0125	291.2114	0
$F_{SI} + F_{TS} + F_{InfSc}$	**intercept**	α	-1.3941	0.0157	-88.3621	0
	F_{SI}	β_1	0.0737	0.0669	1.1013	0.1353
	F_{TS}	β_2	3.1679	0.0394	80.3719	0
	F_{InfSc}	β_3	0.0002	7.7684e-05	2.6408	0.0041
F_{TS}	**intercept**	α	-2.5617	0.0034	-744.2074	0
	F_{TS}	β_1	3.6426	0.0093	390.2773	0
$F_{SI} + F_{InfSc}$	**intercept**	α	-0.0097	0.0100	-0.9700	0.1660
	F_{SI}	β_1	-0.0777	0.0676	-1.1488	0.1253
	F_{InfSc}	β_2	0.0012	8.0652e-05	16.0311	0
$F_{TS} + F_{InfSc}$	**intercept**	α	-1.5236	0.0116	-130.9340	0
	F_{TS}	β_1	3.3502	0.0291	114.9896	0
	F_{InfSc}	β_2	0.0004	5.5377e-05	7.3600	1.0191e-13

Table 6: Model parameters of Information Retrieval Topic

Model	Feature	Par.	Value	Std.Err.	Wald	sig.
F_{CE}	**intercept**	α	-1.4699	0.0038	-377.0467	0
	F_{CE}	β_1	0.9033	0.0021	419.0485	0
$F_{SI} + F_{TS}$	**intercept**	α	-1.5351	0.0064	-237.1857	0
	F_{SI}	β_1	-0.1394	0.0226	-6.1706	3.4505e-10
	F_{TS}	β_2	3.0390	0.0177	170.9132	0
$F_{SI} + F_{TS} + F_{InfSc}$	**intercept**	α	-1.2779	0.0173	-73.5049	0
	F_{SI}	β_1	-0.1334	0.0647	-2.0618	0.0196
	F_{TS}	β_2	2.9658	0.0471	62.8816	0
	F_{InfSc}	β_3	0.3001	0.0131	22.8290	0
F_{TS}	**intercept**	α	-1.5915	0.0047	-331.7382	0
	F_{TS}	β_1	3.1183	0.0132	234.7923	0
$F_{SI} + F_{InfSc}$	**intercept**	α	-0.0509	0.0104	-4.8885	5.2544e-07
	F_{SI}	β_1	-0.1832	0.0652	-2.8073	0.0025
	F_{InfSc}	β_2	0.3852	0.0136	28.1625	0
$F_{TS} + F_{InfSc}$	**intercept**	α	-1.3271	0.0127	-103.7753	0
	F_{TS}	β_1	2.9698	0.0346	85.7041	0
	F_{InfSc}	β_2	0.3566	0.0083	42.8263	0

Table 7: Model parameters of Privacy and Security Topic

Model	Feature	Par.	Value	Std.Err.	Wald	sig.
F_{CE}	**intercept**	α	-1.9830	0.0032	-610.0604	0
	F_{CE}	β_1	1.0433	0.0019	526.6821	0
$F_{SI} + F_{TS}$	**intercept**	α	-2.5836	0.0053	-486.2342	0
	F_{SI}	β_1	-0.0357	0.0168	-2.1268	0.0167
	F_{TS}	β_2	4.5032	0.0153	294.2476	0
$F_{SI} + F_{TS} + F_{InfSc}$	**intercept**	α	-1.7229	0.0164	-104.9034	0
	F_{SI}	β_1	0.1734	0.0540	3.2111	0.0006
	F_{TS}	β_2	4.0029	0.0436	91.7938	0
	F_{InfSc}	β_3	1.7738e-05	1.2257e-06	14.4709	0
F_{TS}	**intercept**	α	-2.5612	0.0039	-648.7085	0
	F_{TS}	β_1	4.4228	0.0114	385.7505	0
$F_{SI} + F_{InfSc}$	**intercept**	α	4.3715e-10	0.0106	4.1060e-08	0.4999
	F_{SI}	β_1	1.0555e-09	0.0591	1.7842e-08	0.4999
	F_{InfSc}	β_2	2.3117e-05	1.3952e-06	16.5688	0
$F_{TS} + F_{InfSc}$	**intercept**	α	-1.3728	0.0123	-111.5988	0
	F_{TS}	β_1	3.1942	0.0328	97.1137	0
	F_{InfSc}	β_2	1.3715e-05	7.5793e-07	18.0959	0

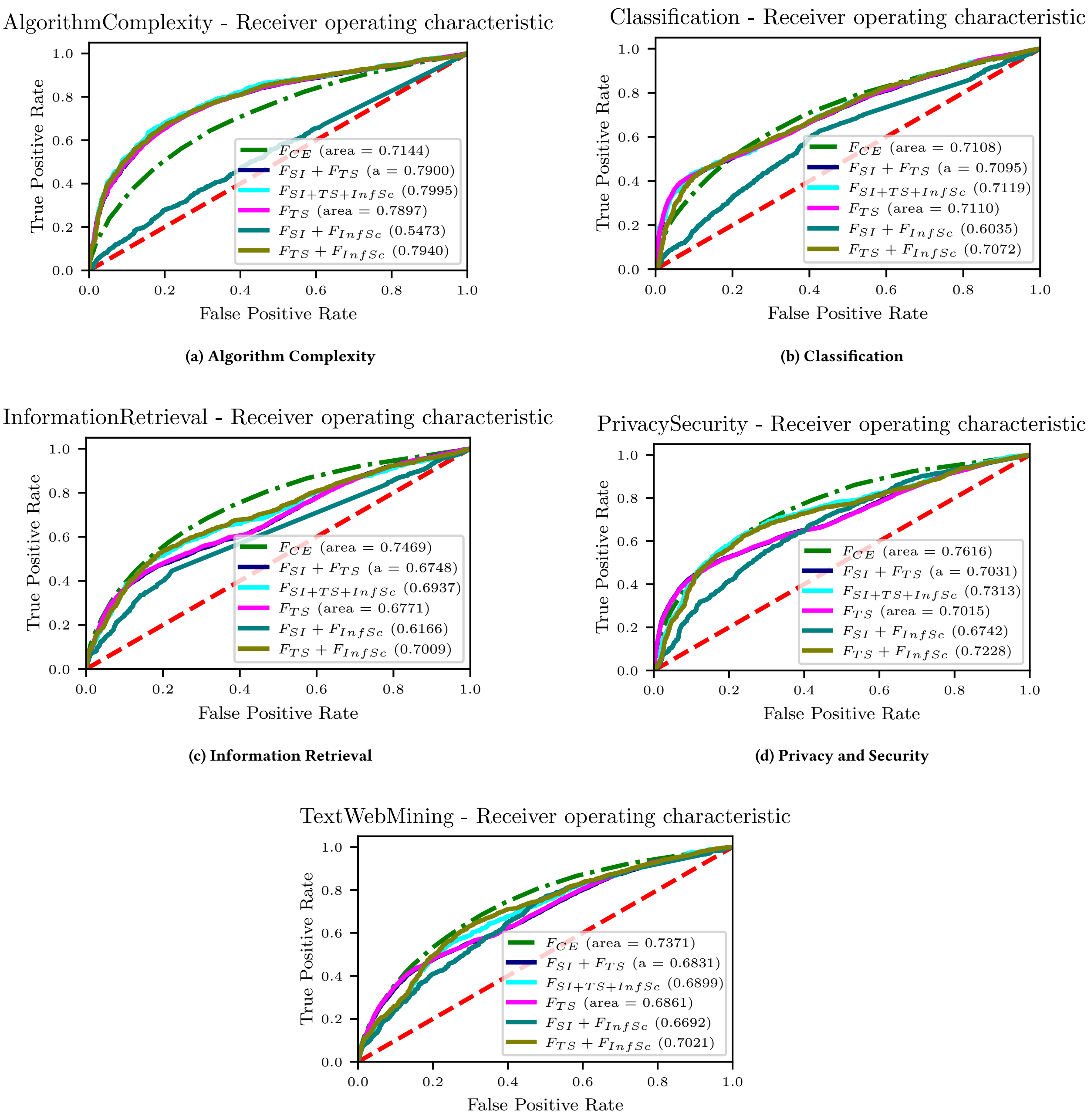
as follows: F_{TS} is the most effective feature and F_{InfSc} follows in second place.

Accuracy of the models per topic are given in Tables 9, 10, 11, 12, and 13 respectively. In terms of F_β scores, proposed "$F_{SI} + F_{TS} + F_{InfSc}$" model and the feature F_{InfSc} performed better in 4 out of 5 topics selected. $\beta = 1.1$ is used similar to [26] favoring recall performance. Although Receiving Operational Characteristic (ROC) [27] value is the highest in Algorithm and Complexity topic (Figure 3(a)), F_β score evaluates our model as subpar.

ROC curves per topic are shown in Figure 3. As seen in the results *Coauthor effect* feature performs better for Privacy & Security topic, Text & Web Mining and Information Retrieval topics. It is simply calculated as logarithm of one plus the number of already active friends on the topic. Although this feature is performing very good with respect to its simplicity, it is not taking into account the time aspect of author interests and does not incorporate any decay and further links in the social network. Proposed model including the feature *Influencee Score* (F_{InfSc}) performs better in all ROC curves.

6 CONCLUSION

In this paper, we investigate the effect of accumulated influence for topic adoption prediction in a scientific collaboration network. We use the term *influencee score* to denote the influence accumulated on an author for a given topic. We hypothesize that the influence



(a) Algorithm Complexity

(b) Classification

(c) Information Retrieval

(d) Privacy and Security

(e) Text and Web Mining

Figure 3: Receiving Operational Characteristic for all topics

accumulated on an influencee for a topic may be effective on topic adaption. The influencee score calculation method is inspired from influencer detection work in [24]. We modify the influencer score calculation such that the social stream is played backwards in order to accumulate the influence coming from various sources. We incorporate the influencee score as a feature within the framework presented in [26].

The experiments conducted on Arnet Miner data set show that the proposed feature, F_{InfSc}, improves the prediction performance, especially recall value, when used together with the features presented in [24]. The prediction performance is better when F_{InfSc}

Table 8: Model parameters of Text and Web Mining Topic

Model	Feature	Par.	Value	Std.Err.	Wald	sig.
F_{CE}	**intercept**	α	-1.3166	0.0040	-324.8517	0
	F_{CE}	β_1	0.8476	0.0021	390.1226	0
$F_{SI} + F_{TS}$	**intercept**	α	-1.4135	0.0063	-221.7997	0
	F_{SI}	β_1	-0.1682	0.0226	-7.4396	5.2069e-14
	F_{TS}	β_2	3.0144	0.0171	175.7761	0
$F_{SI} + F_{TS} + F_{InfSc}$	**intercept**	α	-1.2157	0.0160	-75.7983	0
	F_{SI}	β_1	-0.3031	0.0597	-5.0707	2.0679e-07
	F_{TS}	β_2	2.9363	0.0427	68.6542	0
	F_{InfSc}	β_3	4.5301e-05	8.2357e-07	55.0053	0
F_{TS}	**intercept**	α	-1.4237	0.0047	-300.9280	0
	F_{TS}	β_1	3.0240	0.0128	235.7353	0
$F_{SI} + F_{InfSc}$	**intercept**	α	-9.0498e-09	0.0098	-9.1708e-07	0.4999
	F_{SI}	β_1	-2.9751e-09	0.0599	-4.9587e-08	0.4999
	F_{InfSc}	β_2	5.7770e-05	8.4983e-07	67.9780	0
$F_{TS} + F_{InfSc}$	**intercept**	α	-1.2412	0.0120	-102.8170	0
	F_{TS}	β_1	2.9433	0.0324	90.5676	0
	F_{InfSc}	β_2	4.2191e-05	6.7121e-07	62.8586	0

Table 9: Model performance of Algorithms and Complex.

model	recall	F_β	acc.	prec.	spec.
F_{CE}	1	0.8845	0.7761	0.7761	0
$F_{SI} + F_{TS}$	1	0.8943	0.7929	0.7929	0
$F_{SI} + F_{TS} + F_{InfSc}$	0.7136	0.7372	0.7368	0.7680	0.7071
F_{TS}	1	**0.8949**	0.7940	0.7940	0
$F_{SI} + F_{InfSc}$	0.2235	0.3091	0.5299	0.5759	0.5188
$F_{TS} + F_{InfSc}$	0.6981	0.7273	0.7297	0.7662	0.6962

Table 10: Model performance of Classification

model	recall	F_β	acc.	prec.	spec.
F_{CE}	0.4805	0.5042	0.7177	0.5362	0.7850
$F_{SI} + F_{TS}$	0.3019	0.4212	0.7629	0.8071	0.7571
$F_{SI} + F_{TS} + F_{InfSc}$	0.7265	**0.6695**	0.6279	0.6114	0.6529
F_{TS}	0.3052	0.4247	0.7653	0.8074	0.7598
$F_{SI} + F_{InfSc}$	0.4475	0.5084	0.5755	0.6087	0.5558
$F_{TS} + F_{InfSc}$	0.7111	0.6631	0.6302	0.6131	0.6541

Table 11: Model performance of Information Retrieval

model	recall	F_β	acc.	prec.	spec.
F_{CE}	0.6845	0.6442	0.6850	0.6015	0.7579
$F_{SI} + F_{TS}$	0.6124	0.5724	0.5953	0.5305	0.6608
$F_{SI} + F_{TS} + F_{InfSc}$	0.7898	**0.6902**	0.6120	0.5988	0.6416
F_{TS}	0.6158	0.5709	0.5929	0.5245	0.6633
$F_{SI} + F_{InfSc}$	0.3813	0.4742	0.5992	0.6724	0.5703
$F_{TS} + F_{InfSc}$	0.7962	**0.6927**	0.6154	0.5986	0.6534

Table 12: Model performance of Privacy and Security

model	recall	F_β	acc.	prec.	spec.
F_{CE}	0.6180	0.5751	0.7263	0.5304	0.8287
$F_{SI} + F_{TS}$	0.4116	0.5017	0.7581	0.6827	0.7754
$F_{SI} + F_{TS} + F_{InfSc}$	0.8193	**0.7007**	0.6219	0.5963	0.6834
F_{TS}	0.3983	0.4921	0.7570	0.6882	0.7721
$F_{SI} + F_{InfSc}$	1	0.6859	0.4970	0.4970	0
$F_{TS} + F_{InfSc}$	0.8030	0.6967	0.6200	0.6006	0.6642

Table 13: Model performance of Text and Web Mining

model	recall	F_β	acc.	prec.	spec.
F_{CE}	0.7433	0.6628	0.6645	0.5861	0.7581
$F_{SI} + F_{TS}$	0.6644	0.6057	0.6014	0.5473	0.6666
$F_{SI} + F_{TS} + F_{InfSc}$	0.8259	**0.7001**	0.6186	0.5912	0.6875
F_{TS}	0.6725	0.6105	0.6037	0.5493	0.6708
$F_{SI} + F_{InfSc}$	0.8953	0.6940	0.5793	0.5456	0.7249
$F_{TS} + F_{InfSc}$	0.8395	**0.7095**	0.6245	0.5975	0.6961

is used together with F_{TS} only. Another interesting observation is that, on the contrary to results reported in [26], F_{CE} provides a good performance for most of the topics especially in terms of specificity.

As a future work, further feature combinations can be explored to improve the prediction performance. Another interesting enhancement direction is extending the scientific collaboration network with additional elements such as publication venues.

ACKNOWLEDGMENTS

This work is partially supported by TUBITAK with grant number 117E566.

REFERENCES

[1] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and Correlation in Social Networks. *SIGKDD* (2008), 7–15.

[2] S. Aral, L. Muchnik, and A. Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.

[3] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th {International Conference on Knowledge Discovery and Data mining}* (2006), 44–54.

[4] Hakan Bagci and Pinar Karagoz. 2016. Context-Aware Friend Recommendation for Location Based Social Networks Using Random Walk. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW’16 Companion).* 531–536.

[5] Chien Chin Chen, Shun-Yuan Shih, and Meng Lee. 2016. Who should you follow? Combining learning to rank with social influence for informative friend recommendation. *Decision Support Systems* 90 (2016), 33 – 45.

[6] Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. 2007. Contextual Prediction of Communication Flow in Social Networks. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007* (2007), 57–65.

[7] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. 2008. Feedback Effects between Similarity and Social Influence in Online Communities. *SIGKDD* (2008), 1–14.

[8] Pedro Domingos. 2005. Mining social networks for viral marketing. *IEEE Intelligent Systems* 20, 1 (2005), 80–82.

[9] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. *Proceedings of the 13th conference on World Wide Web - WWW ’04* (2004), 491.

[10] Qi He, Bi Chen, and C Lee Giles. 2009. Detecting Topic Evolution in Scientific Literature : How Can Citations Help ? *Cikm* (2009), 957–966.

[11] Jian Huang, Ziming Zhuang, Jia Li, and C. Lee Giles. 2008. Collaboration Over Time: Characterizing and Modeling Network Evolution. *Wdsm* (2008), 107–116.

[12] Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 1 (1953), 39–43.

[13] Timothy La Fond and Jennifer Neville. 2010. Randomization tests for distinguishing social influence and homophily effects. *Proceedings of the 19th international conference on World wide web - WWW ’10* (2010), 601.

[14] Jure Leskovec, M. McGlohon, Christos Faloutsos, N. Glance, and M. Hurst. 2014. Information Propagation and Network Evolution on the Web. *Ml.Cmu.Edu* January 2009 (2014), 1–25.

[15] Keping Li and Shanshan Wang. 2018. A network accident causation model for monitoring railway safety. *Safety Science* 109 (2018), 398 – 402. https://doi.org/10.1016/j.ssci.2018.06.008

[16] Cindy Xide Lin, Qiaozhu Mei, Ba Zhao, and Jiawei Han. 2010. PET : A Statistical Model for Popular Events Tracking in Social Communities. *KDD10* (2010), 929–938.

[17] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Homophily in Social Networks. *Annu. Rev. Sociol.* 27 (2001), 415–444.

[18] M. E. J. Newman. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* 101, Supplement 1 (2004), 5200–5205.

[19] Huan Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. 2011. Retweet modeling using conditional random fields. *Proceedings - IEEE International Conference on Data Mining, ICDM* (2011), 336–343.

[20] Sebastian A. Rios, Felipe Aguilera, J. David Nunez-Gonzalez, and Manuel Grana. 2019. Semantically enhanced network analysis for influencer identification in online social networks. *Neurocomputing* 326-327 (2019), 71 – 81.

[21] Everett M. Rogers. 2003. *Diffusion of Innovations* (5th ed.). Free Press, New York, NY.

[22] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*. ACM, 243–246.

[23] Jaideep Srivastava. 2008. Data mining for social network analysis. In *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*. IEEE, xxxiii–xxxiv.

[24] Karthik Subbian, Charu C. Aggarwal, and Jaideep Srivastava. 2016. Querying and Tracking Influencers in Social Streams. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16* (2016), 493–502.

[25] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner : Extraction and Mining of Academic Social Networks. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), 990–998.

[26] Deqing Yang, Yanghua Xiao, Bo Xu, Hanghang Tong, Wei Wang, and Sheng Huang. 2012. Which topic will you follow? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7524 LNAI, PART 2 (2012), 597–612.

[27] Marh H. Zweig and Gregory Campbell. 1993. STATISTICA - ROC curve. *Clinical Chemistry* 39, 4 (1993), 561–577.