Collaborative Topic Regression for Predicting Topic-Based Social Influence

Asso Hamzehei[1,3] Shanqing Jiang [1] Danai Koutra[2]
Raymond Wong[1]
Fang Chen[1,3]

[1] University of New South Wales, Sydney, Australia
a.hamzehei@unsw.edu.au
wong@cse.unsw.edu.au
shanqing.jiang@student.unsw.edu.au
[2] University of Michigan, MI, USA
dkoutra@umich.edu
[3] Data61-CSIRO, Sydney, Australia
fang.chen@data61.csiro.au

Technical Report
UNSW-CSE-TR-201610
July 2016

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Social science studies have acknowledged that the social influence of individuals is not identical. Social networks structure and shared text can reveal immense information about users, their interests, and *topic-based* influence. Although some studies have considered measuring user influence, less has been on measuring and estimating topic-based user influence. In this paper, we propose an approach that incorporates both network structure and user-generated content for topic-based influence measurement and prediction. We predict topic-based individual influence on unobserved topics, based on observed influence of users on the topics in which they have shown interest by posting about them in a social network. A collaborative topic-based social influence model is proposed to learn user and topic latent spaces for estimating each user's social influence on an unobserved topic. We perform experimental analysis on Twitter data and show that our model outperforms benchmarks on accuracy, recall, and precision for predicting topic-based user influence.

1 Introduction

Although social influence has been an area of interest for researchers in sociology and more recently in computer science, still there is no agreement on its definition. A very early definition for influential people is "individuals who were likely to influence other persons in their immediate environment" [1]. Social influence has either been studied to identify influential users (opinion leaders or authorities), topical or topic-based influential users [2].

Social science studies, e.g. [3], have acknowledged the fact that the social influence of individuals is not identical. Katz [1] introduced three main factors that are related to an individual's social influence such as: Who one is, what one knows, and whom one knows. The individual's social influence can be much more easily observed on social media while it is confirmed that the social influence factors are similar in social networks to those in the real society [4, 5]. For example, Eirinaki et al [6] introduced two factors (popularity and activity) as factors related to social influence on Online Social Networks (OSN).

One of the main measures studied for influence is information diffusion which measures how important a user is in spreading information in the network. This is equivalent to identify central and hub nodes in the network [7, 8]. Opinion leaders and discussion starters also have been studied as a measure of social influence [9]. A user's position in the network [7], content [10], and activities [11] have been also studied as influence measures. Another aspect of studied influence has been the scale of affected users by a post on social network or intensity of emotional and cognitive impact [12].

According to [13], influential users have different influences on different topics and a very influential user is not necessarily influential on all topics. It is indicated in [14] that topic-based influence measures are more effective and functional than the global ones. One of the differences of topic-related influence studies to network structure analysis is that it takes the posts' (e.g., tweets) content into account. When we consider user influence on topics, no longer the whole network needs to be analyzed, which improves the performance of measures.

However, there are drawbacks and shortcomings in the topic-based influence studies. In most of the existing works, they have aimed at making influential user detection more effective in retrieving the top N users only. Less effort is dedicated in discriminating influential from non-influential users. Also, approaches that uses supervised learning (e.g., SVM) suffer from their dependency on labeled data, which is extremely expensive to prepare for the immense data of social networks. Another considerable issue in these studies is their approach evaluation. This is a difficult task as influence is subjective. More importantly, prediction of user influence is remained as a problem to address in the state-of the-art.

Topic-based user influence identification and prediction are important challenges and the focuses of this paper. This task is significantly important for different applications such as marketing, election campaigns, or recruiting employees for a company. In this work, we predict topic-based user influence on unobserved topics based on observed influence of users on the topics in which they have shown their interests by posting in social networks. We propose a Social Influence Collaborative Topic Regression (SICTR) approach to learn user, topic, and social factor latent spaces for estimating user social influence on an

unobserved topic in social network. Our approach represents users with their topic interests and their social influence on each observed topic.

In more detail, our contributions are:

- We propose a novel topic-based influence prediction approach, to integrate the user-topic relationships, topic content information, and social connections between users into the same principled model. We adopt the collaborative topic regression (CTR) model [15], which has been successfully applied to article recommendation, to combine both user-topic and topic content information for influence prediction.
- Instead of considering user-to-user influence and global user influence, the proposed model considers individuals influence and interests in a topic, which gives the capability of predicting one's influence on a new topic.
- The usefulness of considering content of topics and co-occurrence in the user-topic matrix are confirmed by our experiments on topic-based user influence prediction: topic-aware methods show better performance over other approaches that just consider a generic item and neglect its characteristics (content-base, e.g. LDA, and item-based, e.g. collaborative filtering methods).
- Finally we have prepared a unique dataset from real-world social networks for testing and evaluating the proposed approach that contains all the social media related metadata.

The remainder of this paper is organized as follows. We first discuss existing approaches for topic-based influence analysis in Section 2. We then present the background in Section 3. Next, we define the research problem, and then propose our approach and algorithms in Section 4. We describe our dataset and discuss the results in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

One of the main approaches to study user influence in social networks has been through network structure as well as user's position and connectivity in the network. The traditional centrality measures such as closeness and betweenness are measured for users, to discover how well connected a user is to the rest of users in the network and whether a user is acting as a hub [16]. The major adopted algorithms for network structure based influence measurement include PageRank [17] and HITS [18]. Numerous works have applied PageRank algorithm variations on social network graph to rank user influence according to the network structure. An example of PageRank algorithm variations is the work by Kwak et al [19], in which they ranked users by applying PageRank on follower/following graph in Twitter (along with number of followers and number of retweets). The network structure is relatively static compared to the activities of users in social networks. Some studies have included the social network related meta data (in case of Twitter, the meta data are retweets, mentions, and likes) [8].

Topic-based Influence. Following the influence studies (overall user influence) on social networks, less studies have shed light on topic-based influence.

More recently, topic-based influence studies have combined content of user posts with link-based metrics. Haveliwala [17] proposed a topic-sensitive extension of PageRank to rank query results in regards to the query topics. The idea of topic-sensitive PageRank was later used and adjusted for social networks such as Twitter for ranking topic-based user influence. Topical authorities were also studied in [11] by Pal et al. They proposed a Gaussian-based ranking to rank users efficiently. They used probabilistic clustering to filter feature space outliers and showed that mentions and topical signals are more important features in ranking authorities. Xiao et al [20] aimed at detecting topic related influential users by looking at hashtag user communities where hashtags are pre-identified from news keywords. They proposed RetweetRank an MentionRank as content-based and authority-based influential users. Similarly, [10] worked on detecting topical authorities with the assumption that retweeting propagates topical authority. Montangero and Furini [21] also measured Twitter topic-based user influence where they identify topics by hashtags. Although hashtags can reveal the tweet's topic correctly, over 80% of tweets do not have hashtags. These results are neglecting the majority of tweets and can mislead a topic-based user influence, as 4 out of 5 of her tweets are not considered for measuring her influence. In [22], they estimated Twitter user influence for topics of conversations based on PageRank. For that purpose they build a topic information exchange graph to take the information diffusion and degree of information shared into account for user influence estimation. They manually considered seven topic categories and later assign each tweet to those categories through an n-gram model. However, their approach is unable to identify topics in the lower level of the main categories. For example, if someone is detected as influential in the sports category we do not know which sport the influence belongs to. In [23], they offered TwitterRank, a PageRank extension, that measures user influence by calculating topical similarities of users and their network connections. For topic identification, they used the unsupervised text categorization technique, LDA, by aggregating all tweets of a user into a document. Although this approach is presented as topic-sensitive, this approach cannot discriminate the user influence for the topics. In [24] they proposed another extension of PageRank, and unlike [23], it does not need predefined topics for topic-based user influence. In [25], their topic-based influence framework considers retweet frequency and link strength. The link strength is estimated by poisson regression-based latent variable model on user's frequency of retweeting each other. In a recent work by Katsimpras et al [26], they proposed a supervised random walk algorithm for topic sensitive user ranking. As it is obvious from the algorithm name, it needs labeled data which is not very practical in many cases specially with the volume of social networks.

It is worth mentioning that similar works exist that are only after the identification of global influencers instead of influencers for specific topics. An example of such works is [27] where they extended the Linear Threshold Model and Independent Cascade Model to be topic-aware, the topics are still obtained based on the network structure, while totally ignoring the valuable content information.

Collaborative Topic Regression. Information between users and between items is considered valuable to improve recommendation performance. Wang et al [15] proposed Collaborative Topic Regression (CTR) that utilizes user and item information into topic modeling based Collaborative Filtering (CF) models to further improve recommendation performance. CTR is ex-

plained in details in Section 3.2. CTR is further extended in some studies. For example, [28,29] proposed two models (i.e., CTR-SMF and CTR-SMF2) to incorporate user social network into CTR to further improve item recommendation performance. Wang et al [30] proposed a model to incorporate item social relationship into CTR to further improve tag recommendation performance in social tagging systems.

In contrast to other works, in this paper we take one step further in topic-based influence measurement. We propose an approach that measures topic-based user influence and adopt CTR to predict user's influence on an unobserved topic.

3 Background

Next, we give preliminaries for Probabilistic Topic Modeling and Collaborative Topic Regression.

3.1 Probabilistic Topic Modeling

Given a set of documents denoted by $D = [d_1, \ldots, d_i]$, Topic Modeling generates a set of t topics denoted by $\mathcal{T} = [t_1, \cdots, t_t]$. Each topic is related to a weighted representation over m words denoted by $t_t = [w_1 \cdots w_m]$, where w_t is the weight representing the contribution of word m to topic t. Probabilistic topic modeling , such as Latent Dirichlet Allocation (LDA), represents a low dimensional space of corpus by detecting a set of latent topics. The basic idea of Probabilistic Topic Modeling is having a Z hidden variable for each word's co-occurrence in the collection of documents. Z can range among k topics where each topic is a distribution over a fixed vocabulary. Given a corpus, a document may contain multiple topics and the words are assumed to be generated by those topics. A probabilistic topic model can be generated over a process as follows:

1. Obtain a distribution over topics to generate a document (in LDA this distribution is drawn from a Dirichlet distribution with a corpus-specific hyperparameter α)
2. Then for each word to be generated;
   (a) Assign topics by drawing upon the document-specific distribution over topics
   (b) Finally, generate a word from distribution of topics over words in dictionary, which means words of each document come from a mixture of topics.

We aim to use probabilistic topic modeling to represent items as a set of topics and also detect social network users interest by applying topic modeling on their timelines.

3.2 Matrix Factorization (MF) and Collaborative Topic Regression (CTR)

CF analyzes the relationships between users and their associations with items by relying on historical user behavior (e.g., movies rating), without requirement of

explicit user profiles. A basic approach of CF is neighborhood-based methods which analyze the relationship between items or users. For the item-based approach, the rating of a user on an item j is estimated based on her ratings on similar items, while user-based approach estimates item j's rating by looking at the rating behavior of other users with similar interests. Another CF approach is known as latent factor model, (e.g., matrix factorization [31]), which estimates a rating by utilizing both user and item patterns. It factorizes user-item matrix into a user-specific matrix and a item-specific matrix. The objective function of a matrix factorization model can be formulated as follows,

$$\mathcal{L} = \sum_{i,j} (U_i^T V_j - r_{i,j})^2 + \lambda_U \|U\|_F + \lambda_V \|V\|_F \tag{3.1}$$

in which the first term is the difference between observed and prediction and the rest are regularization terms.

CTR [15] is proposed on top of the matrix factorization and utilizes probabilistic topic modeling. It assumes that items (documents such as news and movie reviews) are generated by a topic model which represents as topic latent vector v_j. In CTR, users are represented by topic interests. Similar to matrix factorization, CTR computes latent parameter of users u_i and items v_j. Latent variable ϵ_j captures the differences between topics for a user based on the users ratings on items. Equation 3.2 draw the ϵ_j as:

$$\epsilon_j \sim N(0, \lambda_v^{-1} I_k) \tag{3.2}$$

where N is probability density function of the Gaussian distribution with mean zero and variance equal to λ_v^{-1} and I is the indicator function that is equal to 1 if user i rated item j and equal to 0 otherwise.

CTR assumes that item latent vector v_j is close to topic proportion θ_j so that $v_j = (\theta_j + \epsilon_j)$ and draw it as:

$$v_j \sim N(\theta_j, \lambda_v^{-1} I_k) \tag{3.3}$$

The generative process of our SICTR model is inspired by CTR [15].

4 Topic-based Social Influence Prediction

4.1 Problem Definition

Assume $G(\mathbb{v}, e)$ denotes a social network graph, where users are the vertex set $\mathbb{V} = \{\mathbb{v}_i\}_{i=1}^m$ and users relationships are the edges set of E. Assume that users publish a set of texts $D = [d_1, d_2, \ldots, d_i]$, and talk about different topics $\mathcal{T} = [t_1, t_2, \ldots, t_t]$. Each user text (post) d_i holds one or more topics and receives engagement from other users by replying, liking, or re-publishing it. The engagement of other users in a post can reveal the influence of that particular post among its audience.

We denote $A = \{a_{ii'}\}$ as the $n \times n$ matrix which shows the social ties among users in the social network G. For the pair of users i and i', $a_{ii'} \in [0, 1]$ shows the weight of the relationship between users u_i and $u_{i'}$, which we treat as the influence of user i on user i' (the higher the value of $a_{ii'}$, the higher the corresponding influence). The matrix A is not symmetric, as the influence of

user i on user i' is not necessarily equal to influence of user i' on user i. We also assume that user post is visible to all users in G.

Quantifying the topic-based influence of each user based on social ties and other users' engagement in activities, we can identify the influence of user i on topic t, represented as F_{it}. Then we have matrix $F = [F_{it}]_{I \times t}$ that represents influence of all the users in all identified topics.

Given a list of users, topics, and social influence of each user on those topics, we are interested in predicting an unknown value in $F_{it} = [F_{1t}, F_{2t}, \ldots, F_{It}]^{\mathcal{T}}$; the social influence of user i on a new topic. Specifically, we aim to estimate user influence on an unobserved topic, based on observed influence of users on the topics in which they have showed their interest by posting in social networks.

We expect that an influential user in topic t can be influential on a similar topic $t + 1$. Assuming that there are patterns among users with similar topic-based influences, the prediction can be performed with CF algorithms. However, the content-based methods use only the content information for recommendation. For example, if we want to predict influence for topic t_t , we can use the influence from the nearest neighbor in $\mathcal{T}$, where $\mathcal{T}$ is the set of topics, based on the topics content similarity. We can also treat each topic as a label and use multi-label methods to train classifiers based on content information. Co-occurrence based methods use only the user-topic matrix F for prediction. For instance, if two topics t_t and t_k occur simultaneously for many users, and t_t is associated with u_i, we should also expect similar influence of t_k to u_i . Both content-based methods and co-occurrence based approaches neglect useful information. As a result, they cannot achieve satisfactory performance in social influence prediction.

4.2 Our Approach

To measure and predict social influence on unobserved topics in a social network, we propose SICTR (standing for Social Influence-based Collaborative Topic Regression), which predicts topic-based influence. In a nutshell, our model performs a two-part representation of the users: (i) latent feature representation of the users according to the social network, and their connections to other users, and (ii) latent feature representation of the users based on the topics they are active in. Our method extends CTR, a well-known method that combines CF with topic modeling, to fit a model that uses the latent topic space to explain both the observed ratings and the observed words. We propose SICTR on top of CTR to predict topic-based user influence in social networks. SICTR computes the latent parameter of users u_i and topics v_t.

Social network-based representation. We want to derive a k-dimensional feature U from the social network g to represent users. Let $U \in R^k$ be latent user metric with column vector U_i for user-specific latent feature vector. We have user and factor feature vectors after placing zero-mean spherical Gaussian prior on them as follows:

$$p(U \mid \sigma_U^2) = \prod_{i=1}^{m} N(U_i \mid 0, \sigma_U^2 I) \tag{4.1}$$

where $N(x \mid \mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2, and I_k is a k-dimensional identity matrix.

Table 4.1: Key notations

Symbol	Description
$U, \|U\|_n$	the set and the number of users, respectively
V	the set of topics
u	user latent vector
v	topic latent vector
t	a topic
k	number of latent dimensions
F_{it}	influence of user i in t
$F_f(i,t)$	follower strength influence measure for user i in t
$F_a(i,t)$	activity influence measure for user i in t
$F_e(i,t)$	engagement influence measure for user i in t
$F_c(i,t)$	centrality influence of measure for user i in t
α	offset term
β	topic bias parameter
λ_u	regularization parameter for u
λ_v	regularization parameter for v
θ_t	k-dimensional topic distribution for t
ϕ_k	word distribution for topic k
ω_{tn}	n^{th} word of document in topic t
z_{tn}	the topic for the n^{th} word in topic t
N_d	number of words in document d
ϵ_t	topic t's latent offset
c_{it}	precision parameter for F_{it}
I_k	k-dimensional identity matrix

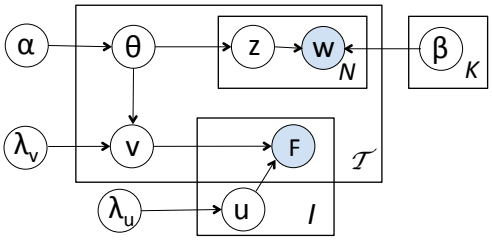


Figure 4.1: The graphical model for SICTR.

Topic-based representation. In our model, SICTR, items are topics in the collection of text that users post in a social network, and users are represented by their topic interests. SICTR predicts user influence on a topic according to similarity of items and other users' influence on similar topics. We identify the topics by applying probabilistic topic modeling, LDA, on all the user-generated text. Each topic contains a set of posts with all their related information and metadata, such as; content, replies, and republishing. For each tuple of $(user_i, topic_t)$, we measure the influence of user i on topic t as details of influence measurement shown in Algorithm 1 and Section 4.4. That way, SICTR generates a topic latent space and a user latent space.

An important part of SICTR is generating topic latent vector $v_t = (\theta_t + \epsilon_t)$, where ϵ_t captures users interest in topic t and it assumes item latent vector v_t is close to topic proportion θ_t. The expectation of F_{it} is a linear function of θ_t, $E[F_{it}|u_i, \theta_t, \epsilon_t] = u_i^T(\theta_t + \epsilon_t)$.

Moreover, from SICTR we know that item latent vector v_t is close to topic proportion θ_t and it generates item latent vector as $v_t = (\theta_t + \epsilon_t)$ where $\epsilon_t \sim N(0, \lambda_v^{-1} I_k)$ is equivalent to $v_t \sim N(\theta_t, \lambda_v^{-1} I_k)$ then

$$P(V \mid \sigma_v^2) \sim N(\theta_t, \lambda_v^{-1} I_k) \tag{4.2}$$

where V is the set of topics and $\lambda_v = \sigma_F^2/\sigma_V^2$

Taking LDA into account, through Bayesian inference we have:

$$\begin{aligned} p(U, V \mid F, \sigma_F^2, \sigma_U^2, \sigma_V^2) \\ \propto p(F \mid U, V, \sigma_F^2) \\ \times p(U \mid \sigma_U^2) p(V \mid \sigma_V^2) \end{aligned} \tag{4.3}$$

Figure 4.1 shows the graphical model of SICTR. Algorithm 2 describes the generative process of SICTR.

4.3 Learning the Parameters of SICTR

We use an EM-style algorithm to learn the parameters [28]. Maximization of the posterior is equivalent to maximizing the complete log-likelihood of U, V, $\theta_{1:t}$, F given U, V, and β.

$$L = \frac{-\lambda_u}{2}\sum_i u_i^T u_i - \frac{-\lambda_v}{2}\sum_t (v_t - \theta)^T (v_t - \theta_t) + \sum_t \sum_n log(\sum_k \theta_{tk}\beta_{k,w_{tn}}) - \sum_{it} \frac{c_{it}}{2}(F_{it} - u_i^T v_t)^2 \tag{4.4}$$

where $\lambda_u = \sigma_F^2/\sigma_U^2$, $\lambda_v = \sigma_F^2/\sigma_V^2$ and Dirichlet prior (α) is set to 1. We optimize this function using gradient ascent by iteratively optimizing the CF on social network variables u_i, v_t and topic proportions θ_t. For u_i, v_t, maximization follows similar to matrix factorization. Given a current estimate of θ_t, taking the gradient of L with respect to u_i, v_t and setting it to zero helps to find u_i, v_t in terms of U, V, C,F, λ_v, λ_u. Solving the corresponding equations will lead to the following update equations:

$$u_i \leftarrow (VC_iV^T + \lambda_u I_k)^{-1}(VC_iF_i) \tag{4.5}$$

$$v_t \leftarrow (UC_tU^T + \lambda_v I_k)^{-1}(UC_tF_t + \lambda_v \theta_t) \tag{4.6}$$

where C_i is diagonal matrix with c_{it}; $t = 1 \cdots J$ as its diagonal elements and $F_i = (F_{it})_{t=1}$ for user i. For each topic t, C_t and F_t are similarly defined. Note that c_{it} is precision parameter for influence matrix F_{it}. Equation 4.6 shows how topic proportions θ_t affect the topic latent vector v_t, where λ_v balances this effect. Given U and V, we can learn the topic proportions θ_t . We define $q(z_{tn} = k) = \phi_{tnk}$ and then we separate the topics that contain θt and apply Jensen's inequality:

$$\begin{aligned} L(\theta_t) &\geq -\frac{\lambda_v}{2}(v_t - \theta)^T(v_t - \theta_t) \\ &+ \sum_n \sum_k \phi_{tnk}(log\theta_{tk}\beta_{k,w_{tn}} - log\phi_{tnk}) \\ &= L(\theta_t, \phi_t) \end{aligned} \tag{4.7}$$

The optimal ϕ_{ink} satisfies $\phi_{tnk} \propto \theta_{tk}\beta_{k,wtn}$. Note that we cannot optimize θ_t analytically, so we use projection gradient approaches to optimize $\theta_{1:t}$ and other parameters U, V, and $\phi_{1:t}$. After we estimate U, V, and ϕ, we can optimize for β,

$$\beta_{kw} \propto \sum_t \sum_n \phi_{tnk} 1\left[w_{tn} = w\right] \tag{4.8}$$

4.4 Influence Measurement

We define social influence in a social network as importance of a user in the social network graph, user's activities, and involvement of others in the user's posts. Social influence can be analyzed through different modalities network structure and user's position in the network, scale of a user's post diffusion in the network, a user's activities and engagement in the social network, and message content that a user broadcast in the network [32].

Algorithm 1 Influence Measurement

Input: List of topics, collection of user posts for each topic, interaction graphs, number of friends of each user.
Output: Matrix of user influence on each topic.

for topic in *topics* **do**
  for user in *users* **do**
    $F_f(i) \leftarrow \#friends$
    $F_a(i,t) \leftarrow \sum_{d_i \in D_t} \delta(d_i)$
    $F_e(i,t) \leftarrow \sum_{d_i \in D_t} (\delta(r_i) + \sum_{m \in m_i} \delta(m))$
    $F_c(i,t) = P(u_i, G(D_t))$
    $F^*_{it} \leftarrow$
    aggregation of $F_f(i,t), F_a(i,t), F_e(i,t), F_c(i,t)$
Return matrix of user influence on topics

From the network structure, we identify influence related attributes, such as user friends and centrality of user in the social network. From the content of broadcasted text, we can identify one or more topics, thus, the influence of that user on different aspects. For instance, in Twitter, a post can contain user mentions, receive replies, and get retweeted by other users. All this information can reveal social influence of a user.

Let denote $D = \{D_t\}_{t=1}^{\mathcal{T}}$ as the set of collected texts, where D_t is texts related to topic t where there are $\mathcal{T}$ topics. Each text d_i contains a set of attributes as $(u_i, dt_i, c_i, r_i, m_i, f_i)$ where u_i is the author of the text, dt_i is the text timestamp, c_i is the text, r_i is the list of users republished the text, m_i is the list of mentions for that text, and f_i is the number of followers of the text author.

We define the following dimensions for measuring social influence of a user on a topic as following:

Follower strength: This measure depicts the number of friends a user has in the network. This value is constant across all topics for a user and is independent of topics. It shows the strength of social ties of a user. Although the number of social connections can be an indicative of influence, it does not carry information on any specific topic. The following influence measures are more topic-specific.

Activity: This measure captures topic-related activities of a user. $F_a(i,t)$ denotes influence of user i in terms of activities related to topic t and we define it as:

$$F_a(i,t) = \sum_{d_i \in D_t} \delta(d_i) \tag{4.9}$$

where $\delta(d_i)$ is 1 if d_i belongs to texts set for topic t and is 0 otherwise. It intuitively measures the volume of topic t-related activities of user i.

Engagement: This is an important indicator of a user's topic specific influence in a social network, since it takes other users' feedback on user i's activities into account. We define it as

$$F_e(i,t) = \sum_{d_i \in D_t} (\delta(r_i) + \sum_{m \in m_i} \delta(m)) \tag{4.10}$$

where $\delta(r_i)$ is the number of times d_i is republished by other users and $\delta(m)$ is the number of mentions or replies of d_i.

Network centrality: Centrality of a user is another indicator of her influence in a social network. PageRank was introduced first for ranking webpages for search engines, and can be used here to calculate topic specific centrality of users in the social graph. To that end, we perform PageRank on the induced graph of interactions on a specific topic t. The interaction graph is a better representative of the topical relevance of two users rather than friendship graph [33]. We denote it as:

$$F_c(i,t) = P(u_i, G(D_t)) \tag{4.11}$$

where $G(D_t)$ is a graph corresponding to users over documents set D_t for topic t. $P(u_i, G(D_t))$ indicates the PageRank score of user i in the graph $G(D_t)$. In this work, we reconstruct the interaction graph, (e.g., retweet and mention graphs from Twitter), to measure topic specific centrality of users by PageRank.

Aggregating Influence Scores: The four influence measures described above F_f, F_a, F_e, F_c will be aggregated to form a single influence score F^* for user i in topic t. For the first attempt, we averaged the measures which gives every measure the same share in the overall influence score. For the future works, we investigate other methods for aggregating the measures.

Algorithm 2 SICTR generative process

1: Run Algorithm 1 to measure users influence for each identified topic
2: For each user i from the user's collection of posts, draw user latent vector of her topics of interest $u_i \sim N(0, \lambda_u^{-1} I_k)$
3: **for** topic in t **do**
4: Draw topic proportions $\theta_t \sim Dirichlet(\alpha)$
5: Draw topic latent offset $\epsilon_t \sim N(0, \lambda_v^{-1} I_k)$ and set the topic latent vector as $v_t = \epsilon_t + \theta_t(d)$
6: **for** word in w_{tn} **do**
7: Draw topic assignment $z_{tn} \sim Mult(\theta)$
8: Draw word $w_{tn} \sim Mult(\beta_{z_{tn}})$
9: For each user-topic pair (i,t), draw the rate of social influence of user i for topic t, F_{it}

4.5 User Influence Prediction on a New Topic

Having the parameters learned, our SICTR model can be used for in-matrix and out-matrix prediction. In-matrix prediction refers to predicting a user influence on a topic where influence rates of some other users are available for it. Out-matrix prediction refers to predicting user influence of on a topic that no influence data is available for (totally new topic).

For a given observed set of documents, D, prediction of influence of a user on a topic can be predicted as the expected value of:

$$E[F_{it}|D] \approx E[u_i|D]^T \left(E[\theta_t|D] + E[\epsilon_t]\right) \tag{4.12}$$

In-matrix can be predicted by using point estimate of u_i, θ_t and ϵ_t to approximate their expectations (recall $v_t = \theta_t + \epsilon_t$),

Table 4.2: A sample from the influence matrix *before* aggregating the 4 measures of influence of user i in topic t.

User	*Topic1*	*Topic2*	*Topic3*	*Topic4*
vnfrombucharest	[0, 0, 0, 0]	[0, 0, 0, 0]	[0.1, 0.02, 0.51, 0.008]	[0, 0, 0, 0]
CharlieDataMine	[0.12, 0.01, 0.34, 0.16]	[0, 0, 0, 0]	[0.1, 0.014, 0.511, 0.163]	[0.28, 0.198, 0.38, 0.16]
sepehr125	[0, 0, 0, 0]	[0, 0, 0, 0]	[0, 0, 0, 0]	[0, 0, 0, 0]
sDataManagement	[0.12, 0.02, 0.35, 0.67]	[0.05, 0.04, 0.15, 0.67]	[0.6, 0.042, 0.512, 0.674]	[0, 0, 0, 0]
yisongyue	[0, 0, 0, 0]	[0, 0, 0, 0]	[0, 0, 0, 0]	[0.07, 0.05, 0.27, 0.08]

Table 4.3: A sample from the influence matrix *after* aggregating the 4 measures of influence of user i in topic t.

User	*Topic1*	*Topic2*	*Topic3*	*Topic4*
vnfrombucharest	0	0	0.159	0
CharlieDataMine	0.157	0	0.197	0.254
sepehr125	0	0	0	0
sDataManagement	0.29	0.227	0.457	0
yisongyue	0	0	0	0.117

$$F_{it}^{*} \approx (u_i^{*})^T(\theta_t^{*} + \epsilon_t^{*}) = (u_i^{*})^T(v_t^{*}) \tag{4.13}$$

For out-matrix prediction, as a topic is new and no historical influence is available from users then $E[\epsilon_t] = 0$. As a result, the we can predict a user's influence on a new unobserved topic as:

$$F_{it}^{*} \approx (u_i^{*})^T\theta_t^{*} \tag{4.14}$$

Putting everything together, we obtain our proposed SICTR method, which is given in Algorithm 2.

5 Results and Experiments

In this section, we discuss the the details of conducted experiments. It includes the data, the influence measurement, and topic-based user influence prediction by our proposed method.

5.1 Dataset

To validate our proposed method, we collected a unique dataset from Twitter using the Twitter Search API. We targeted the Machine Learning domain and identified 500 users that have mentioned machine learning as a keyword in their profile description. To choose the users, we selected a set of machine learning users as seeds and crawled data for their friends and friends of friends for other machine leaning-related users. For the prepared list of users, we gathered their timeline tweets which for most of the users covers their tweets for the last 5 years. For each tweet, we also, collected the related meta-data such as the list of users who have replied to each tweet (mention list) and the list of users who have retweetd each tweet (retweet list). The final dataset contains 101,363 tweets with their related metadata, mention lists, and retweet lists[1].

[1]Our prepared dataset can be downloaded from https://drive.google.com/*** (anonymized for blind review).

Table 5.1: A sample of topics and their 5 top influencers measured by our proposed topic-based influence measurement system.

Deep Learning		*Text Mining*		*Programming Languages*		*Artificial Intelligence*		*Recommender Systems*	
Screen Name	*F*	*Screen Name*	*F*	*Screen Name*	*F*	*Screen Name*	*F*	*Screen Name*	*F*
kdnuggets	0.63	randal_olson	0.62	analyticbridge	0.70	analyticbridge	0.57	xamat	0.70
analyticbridge	0.49	analyticbridge	0.55	randal_olson	0.49	ML_toparticles	0.55	analyticbridge	0.54
deeplearning4j	0.33	jmgomez	0.53	DataScienceCtrl	0.41	DataScienceCtrl	0.37	kdnuggets	0.40
KirkDBorne	0.31	IBMbigdata	0.51	BernardMarr	0.35	IBMbigdata	0.24	jmgomez	0.36
DataScienceCtrl	0.31	kdnuggets	0.49	eddelbuettel	0.34	kdnuggets	0.21	KirkDBorne	0.32

Table 5.2: Table 5.1 continued- a sample of topics and their 5 top influencers measured by our proposed topic-based influence measurement system.

NLP-BigData		*Neural Networks*		*Social Networks*		*R and Stats*		*BigData-Hadoop*	
Screen Name	*F*	*Screen Name*	*F*	*Screen Name*	*F*	*Screen Name*	*F*	*Screen Name*	*F*
jmgomez	0.51	kdnuggets	0.93	analyticbridge	0.58	kdnuggets	0.62	analyticbridge	0.69
randal_olson	0.48	KirkDBorne	0.43	kdnuggets	0.56	analyticbridge	0.54	mapr	0.60
analyticbridge	0.45	smolix	0.34	mjcavaretta	0.47	randal_olson	0.47	BernardMarr	0.58
stanfordnlp	0.43	mapr	0.32	CharlieDataMine	0.44	DataScienceCtrl	0.36	odbmsorg	0.56
bigdata	0.36	mjcavaretta	0.30	jure	0.35	paulblaser	0.36	infochimps	0.53

5.2 Evaluation

Our experiments contain two main tasks: (i) user influence measurement on the identified topics from the tweet corpus and (ii) prediction of user influence on an unobserved topic through our proposed SICTR. For the earlier task we evaluate the measured user influence through expert opinion and user citations on the topics that the user has published in scientific conferences and journals. We collected publications through Google scholar for validation. For evaluating SICTR, we report recall, precision, and accuracy. Due to the uncertainty of the meaning of zero influence, recall will be the main performance measure while we still report precision and recall. Zero influence mean either the user has not been influential in a topic or her activities are not represented in our dataset. However, we have reported all three measures in the results. For experimental analysis, we split the dataset into 80% train and 20% test datasets.

If we present each user by topics that are estimated as influential from SICTR, recall corresponds to the number of topics that the user i is predicted as influential over the total number of topics that the user i recognized as influential:

$$recall = \frac{\#\ of\ topics\ the\ user\ is\ predicted\ as\ influential}{\#\ of\ topics\ the\ user\ is\ influential\ in} \tag{5.1}$$

5.3 Topic-based Influence Measurement

Next, we proceed with identifying topics from the collection of all tweets and then measuring influence. The number of topics generated by LDA can affect the quality of features that will be used in SICTR. We determined a number of topics through cross validation that we could receive higher recall in user influence prediction. In our proposed approach, we perform probabilistic topic modeling in two different rounds. The first round is for identifying the topics in the tweets dataset and the second round is to create topics latent space to take topics similarity into account for influence prediction.

The user tweets gathered from their timelines, belong to the identified topics with a probability. We set the probability threshold to 0.1 to consider whether

a tweet belongs to a topic. Each tweet is mapped to at least one topic. Now that for each topic we have a collection of related tweets with their mention and retweet lists, we can measure user influence for them. In Section 4.4, we defined influence based on 4 measures; follower strength, activity, engagement, and network centrality. Follower strength will be taken from the number of users follow the user i on Twitter. Activity represents the number of tweets user i has in topic t. Engagement is the sum of number of mentions and retweets for all of user i's tweets in topic t. For measuring network centrality, we build the retweet graph for each topic separately from the corresponding retweet list and measure centrality of that user node through PageRank algorithm. Table 4.2 shows a small sample of the 4 calculated measures of topic-user influence. The zero scores mean that user i did not have any tweet for that corresponding topic. The non-zero scores are normalized to lie in the range of [0,1] and higher score means higher influence for that topic. The measured influence scores are aggregated and a sample of aggregated scores is shown in Table 4.3.

Tables 5.1 and 5.2 show the top 5 influencers for selected topics. The sample of topics presented in the tables contain machine learning topics such as Neural Networks, Deep Learning, Big Data, Social Networks, Text Mining, NLP, and more specific topics such as Hadoop. For the task of validation of the influence results, there is no standard method in the literature to validate the algorithm output. One of the reasons we have chosen the machine learning and data science community on Twitter as our community of study was the wide availability of experts in the domain that allows us to verify the identified influential users through our algorithm. We manually verify the top topic-based influential users through expert opinions, their Twitter, and Google scholar accounts. For example, for the topic NLP, Stanford NLP group appeared in the top 5 influential accounts on Twitter. For "Recommender Systems" topic, Xavier Amatriain, who is known for his works on recommender systems, received a high influence score. Also, for the topic "Neural Networks", Alex Smola was in the top 5 influencers who have extensively published on neural network topic. In the topic "Social Networks", Jure Leskovec, who is well-known in the social networks community, was among the top influencers.

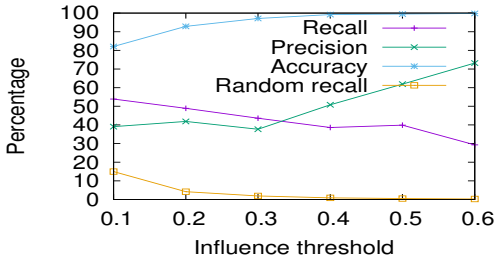


Figure 5.1: Effect of different influence thresholds on SICTR for in-matrix prediction. The number of topics and factors are set to 200 and 50, respectively

5.4 SICTR Prediction Results and Comparisons

In this section, we present the SICTR in-matrix prediction results and compare them with CF, a content-based model that just uses LDA (we call this content-

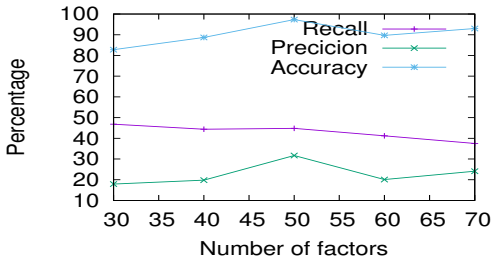


Figure 5.2: Effect of different number of factors on SICTR performance for in-matrix prediction. The number of topics is set to 200, and the influence threshold is set to 0.3. The expected random prediction recall for this figure is 1.9%

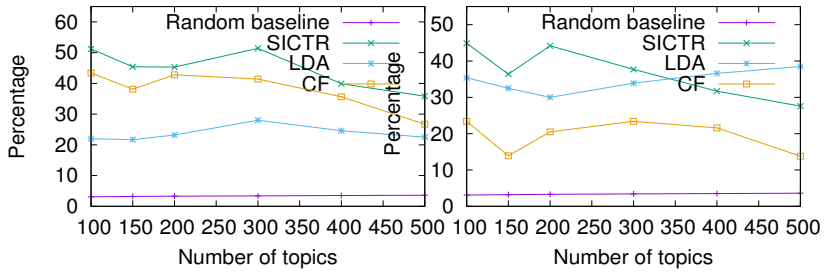
based model as LDA in the rest of the paper), and a random baseline. The random baseline is a random model, where a user randomly is predicted as influential for a topic. The random model predicts a user as influential using the probability of appearance of influential users in the user-topic influence matrix. For the CF model per-user and per-topic latent vectors are fixed with the influence values in the matrix. While, the LDA model behaves as a content-only model. For LDA model, the per-user latent vector is fixed to influence entries in the matrix and per-topic latent vector is only based on the words of the topics. As we mentioned before, in-matrix prediction considers influence prediction for the topics that already exist in the data and influence of at least one user is available for them.

We split the data into train and test datasets. For measuring performance, we use 5-fold-cross-validation. We make sure every topic appears in all the folds so that each topic appears both in the train and test data.

SICTR experimental settings. We evaluate SICTR performance against CF, LDA, and the random baseline. In the models, the a and b are tuning parameters for the parameters c_{it} and d_{it} in Equation 4.4. The parameter λ_v is the precision parameter to balance the diverging of topic latent vector from the topic proportion θ_t. Similar to CTR, we increase λ_v to penalty v_t diverging from θ_t. If we set $v_t = \theta_t$ for per-topic latent vector, SICTR behaves as a content-based method and the topic vector θ_t will be just based on the words in the topic t.

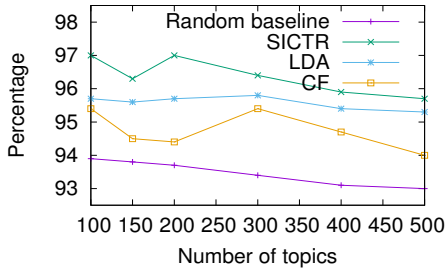
To find the best value of the parameters, we use grid search. For comparisons of the three models, the parameters are set to $k = \frac{number\ of\ topics}{4}$, $\lambda_v = 0.01$, $\lambda_u = 0.01$, a= 1, $b = 0.01$. Moreover, we analyze the SICTR performance for different values of λ_v, number of topics, and influence threshold.

Effect of sparsity on influence prediction. The matrix of topic-based user influence is a sparse matrix. The matrix sparsity can increase with varying influence threshold. By increasing the influence threshold, fewer users will be identified as influential and as a result, the influence matrix becomes more sparse. Also, by increasing the number of topics, the topic-based influence matrix becomes more sparse. Figure 5.1 shows the effect of sparsity on prediction performance of SICTR and Figure 5.3a shows the effect of number of topics on recall. In both figures, when the matrix becomes more sparse, recall decreases. However, according to Figure 5.1, accuracy and precision for SICTR increase



(a) Recall

(b) Precision



(c) Accuracy

Figure 5.3: Recall, Precision, and Accuracy comparisons between SICTR, CF, LDA, and Random baseline. The x-axis shows experiments for different number of topics. The influence threshold is set to 0.3

when the matrix becomes more sparse.

Effect of influence threshold on SICTR performance. The topic-based user influence is in the range from zero to one. Zero indicates no influence over that topic and one means the user has the highest influence for all four influence measures. As we discussed earlier, higher influence threshold results in less number of users being identified as influential, and, consequently, makes the influence matrix more sparse. In Figure 5.1, we show how different thresholds of influence affect the influence prediction performance. The figure shows that when influence threshold is increased, our model predicts influential users more precisely. However, due to sparsity, the recall measure decreases.

Effect of Lambda. For this experiment, we fix the number of topics to 200, the influence threshold to 0.3, the number of factors to 50, and the remaining parameters as explained in experiment settings. We vary $\lambda_v \in [0.01, 0.1, 1, 10, 100, 100]$ to evaluate the SICTR performance for different λ_v values. As λ_v is the precision parameter to balance the divergence of the topic latent vector from topic proportion θ_t and is specific for SICTR, its change does not affect CF and LDA performance. The results show that SICTR outperforms the CF by 14% in the best case scenario with $\lambda_v = 100$ and beats the CF in all experiments. This suggests that taking topic features into account improves the influence prediction performance. Also, it indicates that a user's influence on a new unobserved topic can be estimated by similarity of the new topic with the previously observed topic-based user influences.

Effect of the number of latent factors. Figure 5.2 shows the effect of

the number of latent factors on SICTR prediction performance. To analyze this effect, we run SICTR with fixing the number of topics to 200 and the influence threshold to 0.3. It can be seen that when the rate of number of latent factors to the number of topics increases, then the recall measure decreases. The best performance is achieved when the ratio is $\frac{1}{4}$.

Effect of number of topics on SICTR, CF, and LDA. We evaluate SICTR overall performance against CF, and LDA. Figure 5.3 shows the performance of the three different models in terms of recall, precision, and accuracy. In our dataset, the best performance is achieved for 300 topics. SICTR outperforms CF and LDA models in recall measure on all tested number of topics. Figure 5.3b shows that by increasing the number of topics the content of topics (LDA) have more impact on precision of influence prediction than matrix factorization and even SICTR. The reason is that by increasing the number of topics, some major topics will be divided into smaller ones and the tweets of the topics will also be divided into smaller groups. Eventually, the user influence on the major topic will be scattered among the smaller topics which makes it harder for SICTR to predict them. Figure 5.3c shows the prediction accuracy for both topic-based influential and non-influential users where SICTR continues to outperform CF and LDA. Furthermore, the results show that our model achieves 97% accuracy in predicting user influence on a new topic for both true positive and true negative prediction.

6 Conclusions

In this study, we presented a collaborative topic regression model, SICTR, to predict topic-based user influence in social networks. We identified topics from user posts on social networks, measured each user's influence on each topic, based on the influence definition we proposed, and used SICTR to predict user influence for unobserved topics. Our study's main contributions include:

- proposing a method to measure topic-based influence for social network users
- proposing a topic-based user influence prediction approach in social network that incorporates both network structure and user-generated content for topic-based influence measurement and prediction,
- opening a new discussion for user influence prediction in social networks that has not been explored in the literature.

We tested our topic-based influence measurement system and influence prediction model (SICTR) using a unique dataset that we collected from Twitter, which we will make available online.

In future work, we are interested to measure topic-based user influence over time, and study how influence changes over time and improves the topic-based influence prediction. We will also investigate other methods for combining influence measures.

Bibliography

[1] E. Katz, "The two-step flow of communication: An up-to-date report on an hypothesis," *Public opinion quarterly*, vol. 21, no. 1, pp. 61–78, 1957.

[2] F. Riquelme, "Measuring user influence on twitter: A survey," *arXiv:1508.07951*, 2015.

[3] E. Katz and P. F. Lazarsfeld, *Personal Influence, The part played by people in the flow of mass communications.* Transaction Publishers, 1955.

[4] B. Libai, R. Bolton, M. S. Bügel, K. De Ruyter, O. Götz, H. Risselada, and A. T. Stephen, "Customer-to-customer interactions: broadening the scope of word of mouth research," *JSR*, vol. 13, no. 3, pp. 267–282, 2010.

[5] D. Eccleston and L. Griseri, "How does web 2.0 stretch traditional influencing patterns," *IJMR*, vol. 50, no. 5, pp. 591–161, 2008.

[6] M. Eirinaki, S. P. S. Monga, and S. Sundaram, "Identification of influential social networkers," *IJWBC*, vol. 8, no. 2, pp. 136–158, 2012.

[7] X. Jin and Y. Wang, "Research on social network structure and public opinions dissemination of micro-blog based on complex network analysis," *JNW*, vol. 8, no. 7, pp. 1543–1550, 2013.

[8] B. Hajian and T. White, "Modelling influence in a social network: Metrics and evaluation," in *PASSAT*, 2011, pp. 497–500.

[9] L. B. Jabeur, L. Tamine, and M. Boughanem, "Active microbloggers: identifying influencers, leaders and discussers in microblogging networks," in *SPIRE*, 2012, pp. 111–117.

[10] J. Hu, Y. Fang, and A. Godavarthy, "Topical authority propagation on microblogs," in *CIKM*, 2013, pp. 1901–1904.

[11] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *WSDM*, 2011, pp. 45–54.

[12] A. R. McNeill and P. Briggs, "Understanding twitter influence in the health domain: A social-psychological contribution," in *WWW*. ACM, 2014, pp. 673–678.

[13] F. Probst, D.-K. L. Grosswiele, and D.-K. R. Pfleger, "Who will lead and who will follow: Identifying influential users in online social networks," *BISE*, vol. 5, no. 3, pp. 179–193, 2013.

[14] M. Kardara, G. Papadakis, A. Papaoikonomou, K. Tserpes, and T. Varvarigou, "Large-scale evaluation framework for local influence theories in twitter," *Information Processing Management*, vol. 51, no. 1, pp. 226–252, 2015.

[15] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *KDD*, 2011, pp. 448–456.

[16] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *PKDD*, 2011, pp. 18–33.

[17] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW*, 2002, pp. 517–526.

[18] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *JACM*, vol. 46, no. 5, pp. 604–632, 1999.

[19] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010, pp. 591–600.

[20] F. Xiao, T. Noro, and T. Tokuda, "Finding news-topic oriented influential twitter users based on topic related hashtag community detection," *JWE*, vol. 13, no. 5-6, pp. 405–429, 2014.

[21] M. Montangero and M. Furini, "Trank: ranking twitter users according to specific topics," in *CCNC*, 2015, pp. 767–772.

[22] M. Cataldi and M.-A. Aufaure, "The 10 million follower fallacy: audience size does not prove domain-influence on twitter," *KAIS*, vol. 44, no. 3, pp. 559–580, 2015.

[23] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *WSDM*, 2010, pp. 261–270.

[24] J. Sung, S. Moon, and J.-G. Lee, *The Influence in Twitter: Are They Really Influenced?*, 2013, pp. 95–105.

[25] X. Liu, H. Shen, F. Ma, and W. Liang, "Topical influential user analysis with relationship strength estimation in twitter," in *ICDM*, 2014, pp. 1012–1019.

[26] G. Katsimpras, D. Vogiatzis, and G. Paliouras, "Determining influential users with supervised random walks," in *WWW*, 2015, pp. 787–792.

[27] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," in *ICDM*, 2012, pp. 81–90.

[28] S. Purushotham, Y. Liu, and C.-C. J. Kuo, "Collaborative topic regression with social matrix factorization for recommendation systems," in *ICML*, 2012.

[29] C. Chen, X. Zheng, Y. Wang, F. Hong, and Z. Lin, "Context-aware collaborative topic regression with social matrix factorization for recommender systems," in *AAAI*, 2014.

[30] H. Wang, B. Chen, and W.-J. Li, "Collaborative topic regression with social regularization for tag recommendation." in *IJCAI*, 2013.

[31] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[32] V. R. Embar, I. Bhattacharya, V. Pandit, and R. Vaculin, "Online topic-based social influence analysis for the wimbledon championships," in *KDD*, 2015, pp. 1759–1768.

[33] M. J. Welch, U. Schonfeld, D. He, and J. Cho, "Topical semantics of twitter links," in *WSDM*, 2011, pp. 327–336.