TSIM: Topic-based Social Influence Measurement for Social Networks

Asso Hamzehei[1,3]   **Shanqing Jiang**[1]   **Danai Koutra**[2]   **Raymond K. Wong**[1]

Fang Chen[1,3]

[1] School of Computer Science and Engineering
University of New South Wales,
Sydney, Australia
Email: `assoh,wong@cse.unsw.edu.au,`
`shanqing.jiang@student.unsw.edu.au`

[2] University of Michigan
Ann Arbor, Michigan, USA,
Email: `dkoutra@umich.edu`
[3] Data61, CSIRO, Sydney, Australia
Email: `fang.chen@data61.csiro.au`

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 network structure, user-generated content for topic-based influence measurement, and user's interactions in the network. We perform experimental analysis on Twitter data and show that our proposed approach can effectively measure topic-based user influence.

Keywords: Topic-based social influence, Social networks analysis, influence measurement

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" (Katz 1957). Social influence has either been studied to identify influential users (opinion leaders or authorities), topical or topic-based influential users (Riquelme 2015).

Social science studies, e.g.(Katz & Lazarsfeld 1955), have acknowledged the fact that the social influence of individuals is not identical. Katz (Katz 1957) 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 (Libai et al. 2010, Eccleston & Griseri 2008). For example, Eirinaki et al (Eirinaki et al. 2012) 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 (Jin & Wang 2013, Hajian & White 2011). Opinion leaders and discussion starters also have been studied as a measure of social influence (Jabeur et al. 2012). A user's position in the network (Jin & Wang 2013), content (Hu et al. 2013), and activities (Pal & Counts 2011) 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 (McNeill & Briggs 2014).

According to (Probst et al. 2013), influential users have different influences on different topics and a very influential user is not necessarily influential on all topics. It is indicated in (Kardara et al. 2015) 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 measurement and identification 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 measure topic-based user influence on observed topics in which they have shown their interests by posting in social networks. Our approach, called TSIM (Topic-based Social Influence Measurement), incorporates network structure, user generated contents, users history of activities, and network users engagement in user's activity. 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 measurement approach to integrate the user-topic relationships, topic content information, and social connections between users into the same principled model.

- 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.

- 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 (Romero et al. 2011). The major adopted algorithms for network structure based influence measurement include PageRank (Haveliwala 2002) and HITS (Kleinberg 1999). 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 (Kwak

et al. 2010), 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) (Hajian & White 2011).

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 (Haveliwala 2002) 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 (Pal & Counts 2011) 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. In (Kong & Feng 2011), Kong et al intended to identify and rank users that are posting quality tweets. They defined a topic-based high quality tweet with the author's topic-specific influence, topic related author's behavior. They applied their proposed metric on graph of following and retweets. Xiao et al (Xiao et al. 2014) 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, (Hu et al. 2013) worked on detecting topical authorities with the assumption that retweeting propagates topical authority. Montangero and Furini (Montangero & Furini 2015) 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 (Cataldi & Aufaure 2015), 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 (Weng et al. 2010), 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

Table 1: Key notations

Symbol	Description
t	a topic
$F_{i,j}$	influence of user u_i in t_j
$F_f(i,j)$	follower strength influence measure for user i in t_j
$F_a(i,j)$	activity influence measure for user u_i in t_j
$F_e(i,j)$	engagement influence measure for user u_i in t_j
$F_c(i,j)$	centrality influence of measure for user u_i in t_j
X_i	rank of user u_i calculated through PageRank
I^{op}	operation of identifying authorities
L	asymmetric adjacency matrix representing directed edges
$P_{i1,i2}$	probability of user u_{i1} engage with user u_{i2}
$D_{out}(i)$	number of user that points to user u_i in graph G
N_d	number of words in document d

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 (Sung et al. 2013) they proposed another extension of PageRank, and unlike (Weng et al. 2010), it does not need predefined topics for topic-based user influence. In (Cano et al. 2014) a PageRank-based user influence rank algorithm introduced that the user links have weights based on their topics of interest similarities. In (Liu et al. 2014), 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 (Welch et al. 2011), they found out that topical relevance is better detectable through the retweet link rather than following links. They used two variations of PageRank algorithm to on retweet and following graphs for that purpose. In a recent work by Katsimpras et al (Katsimpras et al. 2015), 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 (Barbieri et al. 2012) 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.

3 Background

Next, we give preliminaries for Probabilistic Topic Modeling and Pagerank.

3.1 Probabilistic Topic Modeling

Given a set of documents denoted by $D = [d_1, \ldots, d_q]$, Topic Modeling generates a set of t topics denoted by $\mathcal{T} = [t_1, \cdots, t_j]$. Each topic is related to a weighted representation over m words denoted by $t_j = [w_1 \cdots w_m]$, where w_j is the weight representing the contribution of word w_m to topic t_j. Probabilistic topic modeling , such as Latent Dirichlet Al-

location (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 j 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 PageRank

PageRank is a webpages ranking algorithm that calculate rank X_i for vertex v_i based on the rank of other vertices in the graph that point to vertex v_i. Assume $G(\mathbb{V}, E)$ denotes a directed graph, where the set $\mathbb{V}$ of vertices consists of i users and users relationships are the edges set E. Considering u_i as a user equal to vertex v_i in the graph G, the directed edge $(i1, i2)$ exists if user u_{i1} is connected to user u_{i2}. The directed vertices of the graph G contained in the asymmetric adjacency matrix $L = (L_{i1,i2})$, where $L_{i1,i2} = 1$ if $u_{i1} \rightarrow u_{i2}$ and $L_{i1,i2} = 0$ otherwise. Out-degree $D_{out}(i)$ is the number of users that points to user u_i.

$$X_i = \sum_{(j,i) \in E} D_{out}(j)^{-1} X_i \qquad (1)$$

The above equation is a recursive function that gives any vertex points to vertex v_i, a fraction of the rank inversley.

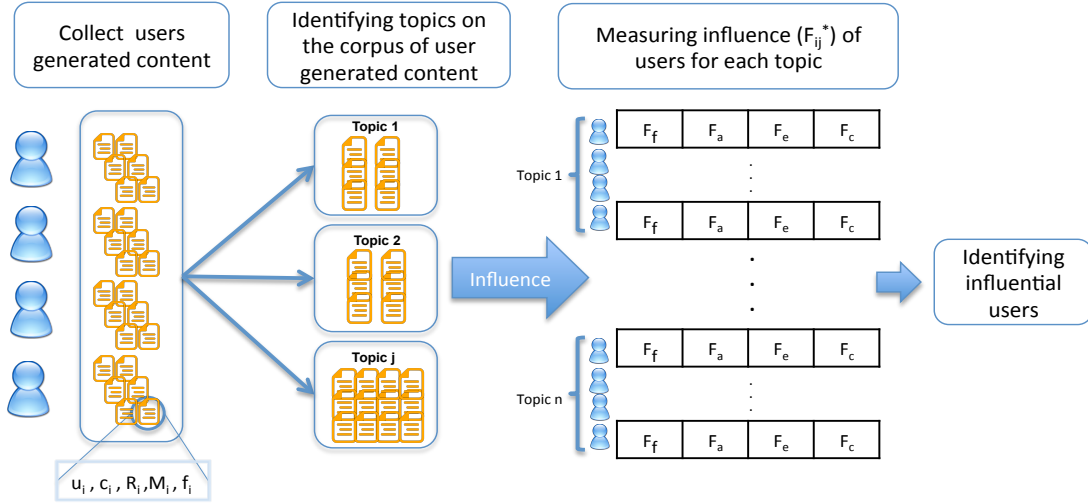Figure 1: TSIM WorkFlow.

In PageRank, each out-going link from v_i is weighted by $1/o_i$, thus every node has the same total out-going weights. Each node has a total of one vote. PageRank uses an idea that a "good" node should connect to or be pointed to by other "good" nodes. However, instead of mutual reinforcement, it adopts a web surfing model based on a Markov process in determining the scores:

$$x = I^{op}(x) \qquad (2)$$

where the I^{op} is an authority that is pointed to by many hubs and the I^{op} operation is defined to be

$$I^{op}(.) = L^T D_{out}^{-1} \equiv P^T. \qquad (3)$$

This amounts to rescale the adjacency matrix L such that each row is sum-to-one. Thus, $P = (P_{i1,i2})$ is a stochastic matrix, since $\sum_v P_{i1,i2} = 1, P_{i1,i2} \geq 0$. $P_{i1,i2}$ represents the probability of a web surfer making a transition from webpage v_{i1} to v_{i2} . Starting from any webpage v_i, a surfer goes to any one of the hyperlinked webpages with equal probability $1/o_i$.

At any moment, millions of people are using the social networks. PageRank assumes the users follow the random surfing model in viewing and engaging with the rest of network. They will reach the equilibrium (stationary) distribution under general conditions. If a node has a high probability in the equilibrium distribution, that means more nodes will point to that node. Therefore, the equilibrium distribution of users in social network is a measure of a node's importance, which is the authority score in PageRank. The equilibrium distribution x is determined by

$$P^T x = \lambda x \qquad (4)$$

and x satisfies $\sum_k x(k) = 1$. One can obtain the solution iteratively. Note that $\lambda = 1$ if the Markov process has an equilibrium distribution x. PageRank models two types of random jumps on the Internet.

(i) Link-tracking jump: a user often follows other users in the network by simply clicking on them; this is modeled by $L^T D_{out}^{-1}$.

(ii) Link-interrupt jump: a user sometimes observe to engage with a user that they are not already connected to each other. PageRank models such link-interrupt jump with a simple uniform distribution $(1 - \alpha)/n$. The full stochastic matrix of transition probability is

$$P^T = I^{op}(.) = \alpha L^T D_{out}^{-1} + (1 - \alpha)(1/n)ee^T \qquad (5)$$

where $\alpha = 0.8 \sim 0.9$. Here $e = (1, 1, , 1)^T$; thus ee^T is a matrix of all 1's (Arasu et al. 2002).

4 Topic-based Social Influence Measurement

4.1 Problem Definition

Assume $G(\mathbb{v}, E)$ denotes a social network graph, where users are the vertex set $\mathbb{V}$ and users relationships are the edges set of E. Assume that users publish a set of texts $D = [d_1, d_2, \ldots, d_q]$, and talk about different topics $\mathcal{T} = [t_1, t_2, \ldots, t_j]$. Each user text (post) d_q holds one or more topics and receives engagement from other users by replying, liking, or republishing 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 social networks, we can identify the influence of user i on topic j, represented as F_{ij}. Then we have matrix $F = [F_{ij}]_{i \times j}$ that represents influence of all the users in all identified topics.

Table 2: A sample from the influence matrix *before* aggregating the 4 influence measures of user u_i in topic t_j.

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 3: A sample from the influence matrix *after* aggregating the 4 influence measures of user u_i in topic t_j.

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

4.2 Our Approach

To measure social influence on an observed topic in a social network, we propose TSIM, Topic-based Social Influence Measure, which measures topic-based individuals influence in social networks. In a nutshell, our model contains two main phases:

- Identifying topics on social networks according to users generated contents, and

- measuring individuals influence for the detected topics.

Figure 1, shows our approach's work flow. First, we collect user-generated content from the social media. For each user-generated content, we collect related information such as list of users that have re-published the content (R_i) and list of users that have engaged in that content (M_i) as well as metadata connected to the content. 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 $(user_i, topic_j)$, we measure the influence of user u_i on topic t_j as F_{ij}^* which comprises of four measures F_f, F_a, F_e, F_c. The details of influence measurement shown in Algorithm 1 and Section 4.3. The measure F_{ij}^* identifies user influence for the identified topics and users can be ranked according to their F_{ij}^* score for each topic.

4.3 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(Embar et al. 2015).

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

---

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.

1: **for** topic in *topics* **do**
2:     **for** user in *users* **do**
3:         $F_f(i) \leftarrow \#friends$
4:         $F_a(i,j) \leftarrow \sum_{d_i \in D_t} \delta(d_i)$
5:         $F_e(i,j) \leftarrow \sum_{d_i \in D_t} (\delta(R_i) + \delta(M_i))$
6:         $F_c(i,j) = PR(u_i, G(D_t))$
7:         $F_{ij}^* \leftarrow$
8:         aggregation of $F_f(i,j), F_a(i,j), F_e(i,j), F_c(i,j)$
9: Return matrix of user influence on topics

---

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_t as the set of collected texts related to topic t_j from the set of topics $\mathcal{T}$. Each text d_i contains a set of attributes as $(u_i, c_i, R_i, M_i, f_i)$ where u_i is the author of the text, 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 scale: 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.

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

Table 4: 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	
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
analyticbridge	0.49	analyticbridge	0.55	randal_olson	0.49	ML_toparticles	0.55
deeplearning4j	0.33	jmgomez	0.53	DataScienceCtrl	0.41	DataScienceCtrl	0.37
KirkDBorne	0.31	IBMbigdata	0.51	BernardMarr	0.35	IBMbigdata	0.24
DataScienceCtrl	0.31	kdnuggets	0.49	eddelbuettel	0.34	kdnuggets	0.21

Table 5: Table 4 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	
Screen Name	F	Screen Name	F	Screen Name	F	Screen Name	F
jmgomez	0.51	kdnuggets	0.93	analyticbridge	0.58	kdnuggets	0.62
randal_olson	0.48	KirkDBorne	0.43	kdnuggets	0.56	analyticbridge	0.54
analyticbridge	0.45	smolix	0.34	mjcavaretta	0.47	randal_olson	0.47
stanfordnlp	0.43	mapr	0.32	CharlieDataMine	0.44	DataScienceCtrl	0.36
bigdata	0.36	mjcavaretta	0.30	jure	0.35	paulblaser	0.36

Table 6: Table 5 continued- a sample of topics and their 5 top influencers measured by our proposed topic-based influence measurement system.

| Recommender Systems		BigData-Hadoop		Database		Visualization	
Screen Name	F	Screen Name	F	Screen Name	F	Screen Name	F
xamat	0.70	analyticbridge	0.69	analyticbridge	0.58	analyticbridge	0.90
analyticbridge	0.54	mapr	0.60	randal_olson	0.55	DataScienceCtrl	0.45
kdnuggets	0.40	BernardMarr	0.58	IBMbigdata	0.36	hmason	0.42
jmgomez	0.36	odbmsorg	0.56	OracleAnalytics	0.32	KirkDBorne	0.35
KirkDBorne	0.32	infochimps	0.53	MarkLogic	0.30	paulblaser	0.31
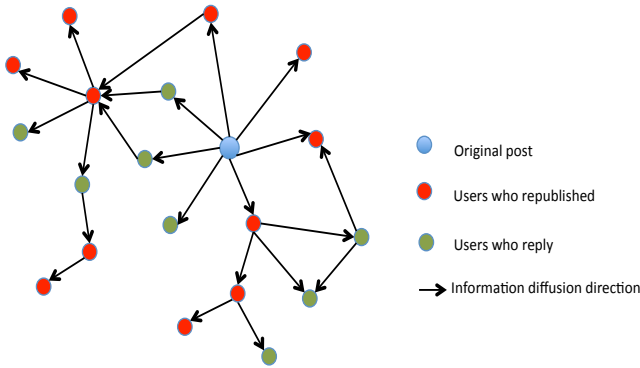
Figure 2: Retweet and Mention Graph.

$$F_a(i,j) = \sum_{d_i \in D_t} \delta(d_i) \qquad (6)$$

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_j-related activities of user u_i.

Topic-based Attractiveness: This measure indicate how other users are attracted to user i's post. It takes other users' feedback on user u_i's activities into account. We define it as

$$F_e(i,j) = \sum_{d_i \in D_j} (\delta(R_i) + \delta(M_i)) \qquad (7)$$

where $\delta(R_i)$ is the number of times d_i is repub-

lished by other users and $\delta(M_i)$ 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. Figure 2 shows an interaction graph of users on a post generated by user 1. To that end, we perform PageRank on the induced graph of interactions on a specific topic t_j. The interaction graph is a better representative of the topical relevance of two users rather than friendship graph (Welch et al. 2011). We denote it as:

$$F_c(i,j) = PR(u_i, G(D_t)) \qquad (8)$$

where $G(D_t)$ is a graph corresponding to users over documents set D_t for topic t. $PR(u_i, G(D_t))$ indicates the PageRank score of user u_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 u_i in topic t_j. 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.

5 Results and Experiments

In this section, we discuss the the details of conducted experiments. It includes the data and the influence measurement performed 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 core 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 among 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 retweeted each tweet (retweet list). The final dataset contains 101,363 tweets with their related metadata, mention lists, and retweet lists. The network that is built on retweet list contains 301870 nodes.

5.2 Evaluation

Our experiments contains a main task of user influence measurement on the identified topics from the tweet corpus.

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. The community of study is intentionally chosen as researchers then we are able to cross-validate our results through the users influence in research community measured by topics of their publications and citations.

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 TSIM. 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 for identifying the topics in the tweets dataset.

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 4.3, 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 u_i on Twitter. Activity represents the number of tweets user u_i

has in topic t_j. Engagement is the sum of number of mentions and retweets for all of user u_i's tweets in topic t_j. 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 2 shows a small sample of the 4 calculated measures of topic-user influence. The zero scores mean that user u_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 3.

Tables 4, 5, and 6 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, Database, Visualization, 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.

5.4 Implications and Applications

This section describes the real-world implication and applications of our model. Identifying topic-based influential users is similar to the problem of finding experts and authorities. Spotting the elite group of users for topics can improve available systems such as search engines. The query result for both contents and users can be returned and ranked using the score provided by our system.

One of the main applications of this work is in Marketing. Marketing campaigns can be implemented through the influential users in the related topic to have more productive and cost effective campaign. Influential users act as hubs in the network and have a central position in the network in terms of information diffusion, also they attract and engage more users into their conversations.

Our model is able to detect the new and surprising topics. This capability gives the strength to our model that works in real-world and detect new topics and related influential users. As a result, there wouldn't be a need for manually defining the topics

and consequently the recent and new topic would not be missed. TSIM, also can be applied to detect topics at what period get viral and who are influential in those topic in different period of time.

6   Conclusions

In this study, we have presented a Topic-based Social Influence Measurement, TSIM, to measure topic-based user influence in social networks. We have identified topics from user posts on social networks, and measured each user's influence on each topic. TSIM is then used to calculate user influence for the observed topics. Our main contributions include:

- the proposal of a effective method to measure topic-based influence for social network users

- opening a new discussion for user influence prediction in social networks that has not been explored in the literature.

Finally, we have tested TSIM using a unique dataset that we collected from Twitter, which we are making it available online.

In future work, we are interested to measure topic-based user influence over time, and study how influence changes over time. Prediction of user influence on unobserved topics is also currently under our investigation. We will also investigate other methods to combine influence measures.

References

Arasu, A., Novak, J., Tomkins, A. & Tomlin, J. (2002), Pagerank computation and the structure of the web: Experiments and algorithms, *in* 'WWW', pp. 107–117.

Barbieri, N., Bonchi, F. & Manco, G. (2012), Topic-aware social influence propagation models, *in* 'ICDM', pp. 81–90.

Cano, A. E., Mazumdar, S. & Ciravegna, F. (2014), 'Social influence analysis in microblogging platforms–a topic-sensitive based approach', *SWJ* **5**(5), 357–372.

Cataldi, M. & Aufaure, M.-A. (2015), 'The 10 million follower fallacy: audience size does not prove domain-influence on twitter', *KAIS* **44**(3), 559–580.

Eccleston, D. & Griseri, L. (2008), 'How does web 2.0 stretch traditional influencing patterns', *IJMR* **50**(5), 591–161.

Eirinaki, M., Monga, S. P. S. & Sundaram, S. (2012), 'Identification of influential social networkers', *IJWBC* **8**(2), 136–158.

Embar, V. R., Bhattacharya, I., Pandit, V. & Vaculin, R. (2015), Online topic-based social influence analysis for the wimbledon championships, *in* 'KDD', pp. 1759–1768.

Hajian, B. & White, T. (2011), Modelling influence in a social network: Metrics and evaluation, *in* 'PASSAT', pp. 497–500.

Haveliwala, T. H. (2002), Topic-sensitive pagerank, *in* 'WWW', pp. 517–526.

Hu, J., Fang, Y. & Godavarthy, A. (2013), Topical authority propagation on microblogs, *in* 'CIKM', pp. 1901–1904.

Jabeur, L. B., Tamine, L. & Boughanem, M. (2012), Active microbloggers: identifying influencers, leaders and discussers in microblogging networks, *in* 'SPIRE', pp. 111–117.

Jin, X. & Wang, Y. (2013), 'Research on social network structure and public opinions dissemination of micro-blog based on complex network analysis', *JNW* **8**(7), 1543–1550.

Kardara, M., Papadakis, G., Papaoikonomou, A., Tserpes, K. & Varvarigou, T. (2015), 'Large-scale evaluation framework for local influence theories in twitter', *Information Processing Management* **51**(1), 226–252.

Katsimpras, G., Vogiatzis, D. & Paliouras, G. (2015), Determining influential users with supervised random walks, *in* 'WWW', pp. 787–792.

Katz, E. (1957), 'The two-step flow of communication: An up-to-date report on an hypothesis', *Public opinion quarterly* **21**(1), 61–78.

Katz, E. & Lazarsfeld, P. F. (1955), *Personal Influence, The part played by people in the flow of mass communications*, Transaction Publishers.

Kleinberg, J. M. (1999), 'Authoritative sources in a hyperlinked environment', *JACM* **46**(5), 604–632.

Kong, S. & Feng, L. (2011), A tweet-centric approach for topic-specific author ranking in micro-blog, *in* 'ADMA', pp. 138–151.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010), What is twitter, a social network or a news media?, *in* 'WWW', pp. 591–600.

Libai, B., Bolton, R., Bügel, M. S., De Ruyter, K., Götz, O., Risselada, H. & Stephen, A. T. (2010), 'Customer-to-customer interactions: broadening the scope of word of mouth research', *JSR* **13**(3), 267–282.

Liu, X., Shen, H., Ma, F. & Liang, W. (2014), Topical influential user analysis with relationship strength estimation in twitter, *in* 'ICDM', pp. 1012–1019.

McNeill, A. R. & Briggs, P. (2014), Understanding twitter influence in the health domain: A social-psychological contribution, *in* 'WWW', ACM, pp. 673–678.

Montangero, M. & Furini, M. (2015), Trank: ranking twitter users according to specific topics, *in* 'CCNC', pp. 767–772.

Pal, A. & Counts, S. (2011), Identifying topical authorities in microblogs, *in* 'WSDM', pp. 45–54.

Probst, F., Grosswiele, D.-K. L. & Pfleger, D.-K. R. (2013), 'Who will lead and who will follow: Identifying influential users in online social networks', *BISE* **5**(3), 179–193.

Riquelme, F. (2015), 'Measuring user influence on twitter: A survey', *arXiv:1508.07951* .

Romero, D. M., Galuba, W., Asur, S. & Huberman, B. A. (2011), Influence and passivity in social media, *in* 'PKDD', pp. 18–33.

Sung, J., Moon, S. & Lee, J.-G. (2013), *The Influence in Twitter: Are They Really Influenced?*, pp. 95–105.

Welch, M. J., Schonfeld, U., He, D. & Cho, J. (2011), Topical semantics of twitter links, *in* 'WSDM', pp. 327–336.

Weng, J., Lim, E.-P., Jiang, J. & He, Q. (2010), Twitterrank: Finding topic-sensitive influential twitterers, *in* 'WSDM', pp. 261–270.

Xiao, F., Noro, T. & Tokuda, T. (2014), 'Finding news-topic oriented influential twitter users based on topic related hashtag community detection', *JWE* **13**(5-6), 405–429.