1-2021

Discovering hidden topical hubs and authorities across multiple online social networks

Ka Wei, Roy LEE
Singapore Management University, roylee@smu.edu.sg

Tuan-Anh HOANG

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

Discovering Hidden Topical Hubs and Authorities Across Multiple Online Social Networks

Roy Ka-Wei Lee, Tuan-Anh Hoang, and Ee-Peng Lim

Abstract—Finding influential users in online social networks (OSNs) is an important problem with many possible useful applications. Many methods have been proposed to identify influential users in OSNs. PageRank and HITs are two well known examples that determine influential users through link analysis. In recent years, new models that consider both content and social network links have been developed. The Hub and Authority Topic (HAT) model is one that extends HITS to identify topic-specific hubs and authorities by jointly learning hubs, authorities, and topical interests from users' relationship and textual content. However, many of the previous works are confined to identifying influential users within a single OSN. These models, when applied to multiple OSNs, could not learn influential users under a common set of topics nor address platform preferences. In this paper, we therefore propose the MPHAT model, an extension of HAT, to jointly model the topic-specific hub users, authority users, their topical interests and platform preferences. We evaluate MPHAT against several existing state-of-the-art methods in three tasks: (i) modeling of topics, (ii) platform choice prediction, and (iii) link recommendation. Based on our extensive experiments in multiple OSNs settings using synthetic datasets and real-world datasets from Twitter and Instagram, we show that MPHAT is comparable to state-of-the-art topic models in learning topics but outperforms the state-of-the-art models in platform prediction and link recommendation tasks. We also empirically demonstrate the ability of MPHAT to determine influential users within and across multiple OSNs.

Index Terms—Hub, authority, topic model, online social networks

1 Introduction

Online social networks (OSNs), such as Facebook, Twitter and Instagram, have grown phenomenally in recent years. It was reported that as of August 2017, Facebook has over 2 billion monthly active users, while Instagram and Twitter have over 700 million and 300 million monthly active user accounts respectively [1]. The vast amount of content and social data generated by these platforms has made them important resources for marketing campaigns such as diffusion of advertising messages and promotion of new products. Identifying influential users in OSNs is therefore critical to these marketing applications.

Many research works have proposed methods to identify influential users in OSNs. For example, there are works that determine users' social influence by network centrality measures [2], [3], [4], [5]. Other works adapted HITS [6] and PageRank [7] algorithms which were originally proposed to determine hub and authority web pages through analyzing the link structure of a web graph to identify influential users in OSNs [8], [9], [10]. Nevertheless, these existing works are either not topic specific or confined to identifying influential users within a single OSN.

Topic and platform specificities are important when analyzing the hub and authority users as they provide more insights about users and reveal in which OSN platforms they are influential. To illustrate the usefulness of topic specificity, consider an example of two users, u_1 and u_2, sharing similar ego network structures. HITS will assign u_1 and u_2 similar authority and hub scores. However, if u_1 is a popular food content contributor who is followed by many food-loving users, while u_2 is a prominent politician followed by many users interested in politics, it is more appropriate to infer that u_1 and u_2 are authority users on food-related and political topics respectively. Platform specificity is also important in identifying influential users across multiple OSNs. Suppose a user u_3 posts much food content and is followed by many food-loving users in an OSN p_1 but is less active in another OSN p_2, i.e., u_3 contributes less content and forms fewer relationships in p_2. While u_3 is regarded as an authority user on food-related topics, her authority on this topic is found in OSN p_1 but not p_2.

The understanding of users' hub and authority specific to topics and platforms is vital to many important applications such as social recommendation, viral marketing, and social sensing. For example, one can address the cold-start problem in social recommendation [11] by recommending a user u on a platform p to follow other users on p who are highly authoritative on topics that u is interested in. Similarly in marketing, companies may enhance the effectiveness of campaigns by hiring hub or authority users in topics related to the campaigns to disseminate the campaign messages so as to maximize their reach [12]. To sense social trends, one could also follow a few selected users who are hubs or authorities across different topics and on different platforms [13].

- *R.K.-W. Lee and E.-P. Lim are with the School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902. E-mail: {roylee.2013, eplim}@smu.edu.sg.*
- *T.-A. Hoang is with the L3S Research Center, Leibniz University of Hanover, Hannover 30167, Germany. E-mail: hoang@l3s.de.*

Manuscript received 5 Sept. 2018; revised 21 May 2019; accepted 6 June 2019. Date of publication 14 June 2019; date of current version 7 Dec. 2020. (Corresponding author: Roy Ka-Wei Lee.)
Recommended for acceptance by K. Selcuk Candan.
Digital Object Identifier no. 10.1109/TKDE.2019.2922962

1.1 Research Objectives

In this work, we aim to model topic and platform specific hub and authority users across multiple OSNs. A simple approach for this could be to first apply the existing topic-specific hub and authority models (e.g., HAT [14]) on multiple OSNs separately, followed by comparing the list of top topical authority and hub users identified in the different OSNs. However, the topics separately learned from different platforms may not be comparable and thus making it is hard to compare users' hub and autority across the OSNs.

Another possible approach that overcomes the incomparability of latent topics learned from the different OSNs is to combine the multiple OSN datasets and learn the users' topics, hub and authority scores from the single combined dataset. For example, we can first combine a user u's generated posts and links from OSNs p_1 and p_2 into a single combined dataset and then apply HAT on the combined dataset. The problem with this approach is that it assumes u has identical hub and authority scores in the two platforms. In other words, if u is an authority on food-related topic in platform p_1, she is also the food authority in platform p_2. In the real-world context, this assumption does not hold as u might be more popular in one platform than the other.

We thus propose Multiple Platforms Hub and Authority Topic (MPHAT) model to learn users' topical interests, platform preferences, topic-specific hub and authority scores simultaneously. We first develop a process for generating both users' content and links in the context of multiple OSNs. We then describe the parameters learning steps of MPHAT. To evaluate MPHAT, we conduct experiments on both real-world and synthetic datasets. For experiments on real-world dataset, we evaluate MPHAT on (i) its effectiveness in learning topics from user generated content, (ii) its ability to predict the platform choice of users' publish post, and (iii) its ability to recommend platform-specific topical influential users through link prediction in a multiple OSNs setting. On synthetic datasets, we further evaluate MPHAT's ability to recover platform-specific topical hub and authority users provided by the ground truth.

1.2 Contributions

Our main contributions in this work consist of the following.

- We propose a topic-based model, Multiple Platforms Hub and Authority Topic (MPHAT) model, which to the best of our knowledge, is the first model that jointly learns user topics, platform preferences, hub and authority users across multiple online social networks.
- We apply the MPHAT model on real-world datasets and demonstrate that (a) MPHAT is comparable to state-of-the-art topic models in learning topics from user generated content, and (b) MPHAT outperforms other models in user link recommendation tasks for both single and multiple platform settings. Empirically, we also applied MPHAT to identify topic-specific hubs and authorities across Instagram and Twitter.
- We also conduct experiments on synthetic datasets to verify the effectiveness of MPHAT in identifying platform-specific topical hubs and authorities under different dataset parameter settings.

1.3 Paper Outline

The rest of this paper is organized as follows: We first discuss the related works in Section 2. We then present our proposed MPHAT model in Section 3. Sections 4 and 5 present the experimental evaluations that we have conducted on real-world and synthetic datasets respectively. The empirical study on the real-world data using MPHAT model will also be discussed in Section 4. Finally, we conclude the paper and discuss the future works in Section 6.

2 Related Works

In this section, we review prior works that are closely related to ours. These works can be broadly categorized into three categories: (i) identifying globally influential users, (ii) finding topic-specific influential users, and (iii) analyzing users' behaviors and topical interests across multiple OSNs.

2.1 Topic Oblivious Influential Users

There are many studies on identifying topic oblivious influential users in OSNs. Most of the existing works find influential users based on analyzing user relationships only [4], [5], [15], [16], [17], behaviors only [8], [18], [19], [20], [21], or both [2], [22], [23], [24], [25], [26], [27], [28].

User Relationships. Many previous works apply network centrality measures to identify influential users [4], [5], [16]. Kayes et al. [4] aggregated network centrality measures such as *degree* [29], *betweenness* [30], *closeness* [29] and *eigenvector* [31] to measure and identify influential bloggers. There are also works which extended HITS algorithm [6] to find influential users in OSNs. Romero et al. [8] proposed the influence-passivity (I-P) algorithm to measure Twitter users' influence and passivity from their retweet activities. Gayo-Avello [15] applied HITS on Twitter follow links to identify and differentiate influential users from spammers. Shahriari and Jalili [17] modified the HITS and PageRank [7] algorithms to analyze and rank users in signed OSNs. Unlike these works, our paper extends HITS to identify topic-specific hub and authority users across multiple OSNs.

User Behaviors. Besides user relationships, user behaviors, e.g., *retweet* and *mention* in Twitter, can also be used to determine influential users in OSNs. Khrabrov and Cybenko [18] adapted PageRank [7] algorithm to Twitter *mention* behavior to identify influential Twitter users. Silva et al. [20] employed a similar approach to find and recommend influential users based on other users' *retweet* activities. Aral and Walker [19] conducted a randomized experiment on Facebook to identify influential and susceptible users based on users' product sharing and adoption behaviors.

User Relationship and Behavior. Some studies have also identified influential users by analyzing both user relationships and behaviors. Agarwal et al. [22] proposed a model that utilizes the page-linking behaviors to measure the influence of bloggers. Ghosh and Lerman [23] applied centrality measures on Digg users' friendship and voting behavior to identify influential users. Cha et al. [2] evaluated the influence of Twitter users using *follower*, *mentions* and *retweets* counts. Other works have also analyzed both user ego

networks and tweet activities to find influential users in Twitter [24], [25], [26], [27], [28].

2.2 Topic Specific Influential Users

Existing works that identified topic-specific influential users in OSNs can be further categorized into two subgroups: (a) *independent models* which model topics and user influence in separate steps [3], [9], [10], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], and (b) *joint models* that jointly model users' topical interests and influence [14], [43], [44], [45].

Independent Model. Many existing works have the modeling and extraction of topics as the first and separate step in the identification of topic-specific influential users. In the simplest manner, the topics of user-generated content are first determined by performing keyword matching with a topical lexicon [3], [32], [33], [34], [36], [37], [40]. For example, in a study to identify topical authorities in Twitter, Pal and Counts [32] first extracted tweets covering three topics: "oil spill", "world cup" and "iphone" using simple substring matching before applying models to determine the topical authorities from the users' *retweet* behavior. Oro et al. [40] proposed *social media authoritative user (SocialAU)* model which includes a three-layer network (i.e., *user-item-lexicon*) for finding authority and hub users of a pre-defined selected topic by extending the TOPHITS, a model proposed by Kolde et al. [46] to analyze a semantic graph that combines anchor text with the hyperlink structure of the web.

Instead of pre-defined topics, some studies use topic modeling such as *Latent Dirichlet Allocation* (LDA) [47] in the first step [10], [35], [38], [39], [41], [42]. For example, Weng et al. [10] first applied LDA to learn the latent topics from users' tweets before applying a PageRank-like model called TwitterRank to measure topic-specific influence of Twitter users. Huang et al. [41] also applied LDA in a similar manner before applying their graph partitioning model to find influential users in Twitter. Hoang and Lim [39] learned the latent topics using Twitter−LDA [48], a model which extends LDA to short-text messages, before analyzing the virality and susceptibility of Twitter users.

Joint Modeling. There are relatively very few works that jointly model user topical interests and influence altogether. Liu et al. [43] proposed two-step model which consists of a generative model to learn the direct influence between users and a topic-level influence propagation method to mine the indirect and global influence. The generative models the generation of a user posts, which is assumed to be either influence by his or her friends who have the same interests or generate depending on his her interests. Bi et al. [44] introduced FLDA, a Bernoulli-Multinomial mixture model which models the users' topic-specific influence and content-independent popularity. Barbieri et al. [45] proposed the WTFW model, which models topical and social relationships of users. The model learns the authoritative and susceptible users for each topic, and it considers a topic-specific susceptible user as one who is interested in the topic (e.g., posting topic-related content), and a topic-specific authority user as one who is followed by many topic-specific susceptible users. Our previous work [14] extended HITS and proposed the Hub and Authority Topical model (HAT) which jointly models the users' topical interests, hub and authority scores simultaneously. HAT considers a topic-specific hub user as one who is not only interested in a topic but also follows many authority users of that topic. Conversely, the model considers an authority on a topic as one who is followed by many hub users of the topic. Unlike the existing works which are confined to finding topic-specific influential users in a single OSN, our proposed model, MPHAT, is able to identify topic-specific hubs and authorities across multiple OSNs by jointly learning the users' topical interests, platform preferences, topic-specific hubs and authorities scores from users' relationships and textual content in multiple OSNs.

2.3 Modeling Multiple Social Networks

Modeling multiple OSNs has been a much studied research direction. The following covers two relevant research topics that have been studied.

User Identity Linkage. As many users utilize multiple OSNs to exchange content and build social networks, identifying user accounts from different OSNs belonging to the same users becomes an important research problem, also known as the *user identity linkage problem*. Shu et al. [49] provided a comprehensive survey of user identity linkage methods. The survey groups all user identity linkage methods into two categories: (a) analyzing user behaviors and (b) learning user topical interests. In this paper, we assume that user identity linkage has been performed as we focus on users who already declare their accounts across OSNs. It is also possible to apply user identity linkage on the user accounts across OSNs, before applying the model to identify topic-specific influential users across these OSNs.

Modeling user Topical Interests. There are also works that apply topic models on multiple social media platforms. Guo et al. proposed a model that considers social-relationship among users for topic modeling and applied their model on Sina Weibo and Twitter datasets [50]. Cho et al. designed a model that incorporates users' social interactions and attributes for topic modeling and applied their model on six social media platforms [51]. Many of these works, however, do not link the users across platforms but perform the topic analysis on each platform independently. Lee et al. [52] addressed this gap by proposing MultiPlatform−LDA, which jointly models the user topical interests and platform preferences across multiple OSNs.

3 Proposed Model

In this section, we describe our proposed Multiple Platforms Hub and Authority Topic (MPHAT) model in detail. We begin by introducing the key elements of the model and their notations. Next, we present the principles behind designing the model and its generative process. We then present an algorithm for learning the models parameters and a data sub-sampling strategy to reduce the computational cost.

3.1 Notations and Preliminaries

Our main notations are summarized in Table 1. We use $\mathcal{U}$ to denote the set of users, and use U and V to denote the sets of followers and followees of all users in $\mathcal{U}$ respectively. For each user $u \in \mathcal{U}$, we denote her posts by S_u. Here, we adopt the bag-of-word representation for each post: that is, each post is represented as a multi-set of words, and the word order is

TABLE 1
Notations

Symbol	Description
$\mathcal{U}/U/V$	Sets of users, followers, and followees
$\mathcal{W}$	Vocabulary of words in users' content, and $\lvert\mathcal{W}\rvert = W$
S_u	Sets of posts by user u
$N_{u,s}$	Sets of words in post s_u
$w_{u,s,n}$	n-th word of the s-th post by user u
K	Number of topics
τ_k	Word distribution of topic k
X_u	Topic vector of user u
$\eta_{u,k}$	Platform preference vector of user u for topic k
$p_{u,s}$	Platform of s-th post of user u
H_u	Topic-specific hub vector of user u
A_v	Topic-specific authority vector of user v
$r_{u,v,p}$	Relationship between u and v in platform p = 1 if u follows v in platform p, = 0 otherwise
γ	Dirichlet priors of τ_k
$\alpha, \beta, \sigma, \delta$	Prior shape of $X_{u,k}$, $\eta_{u,k,p}$, A_v, and H_u respectively
κ, ϕ	Prior scale of $X_{u,k}$ and $\eta_{u,k,p}$ respectively

ignored. The number of words of the s-th post of user u is then denoted by $N_{u,s}$, while the n-th word of the s-th post is denoted by $w_{u,s,n}$. Lastly, we denote the word vocabulary by W.

In this work, we adopt a topic modeling approach for modeling users' interests, platform preferences, hubs and authorities specific to each topic. Our proposed model, *MPHAT*, consists of the following model elements.

Topic. A topic is a semantically coherent theme of words found in the user posts. Formally, a topic is represented by a multinomial distribution over W (unique) words. For example, a topic about traveling would have high probabilities for words such as *trip*, and *flight*, but low probabilities for other words. Another topic about food would have high probabilities for words such as *coffee* and *sandwich* but low properties for other non food related words.

Topical Interest. This refers to the a user's interests for a specific topic. Formally we assign to every user u a topical interest vector $X_u = (X_{u,1}, \ldots, X_{u,K})$ where K is the number of topics and $X_{u,k} \in (0, +\infty)$ for $k = 1, \ldots, K$.

Platform Preference. For a specific topic k, a user may prefer to share content or connect to other users for topic k in a specific platform that she participates in. We model this user's topical platform preference by assigning to every user u a topic-specific platform preference vector, $\eta_{u,k} = (\eta_{u,k,1}, \ldots, \eta_{u,k,P})$, where P is the number of platforms.

Topic-Specific Authority. This refers to the authority of a user for a topic. A topic-specific authority user is one who attracts connections from others for the topic she is well known for. We thus assign to every user $v \in V$ a topic-specific authority vector $A_v = (A_{v,1}, \ldots, A_{v,K})$ where K is the number of topics and $A_{v,k} \in (0, +\infty)$ for $k = 1, \ldots, K$.

Topic-Specific Hub. This refers to users with connections to many other users for specific topics. We assign to every user $u \in U$, a topic-specific hub vector $H_u = (H_{u,1}, \ldots, H_{u,K})$ where K is again the number of topics and $H_{u,k} \in (0, +\infty)$ for $k = 1, \ldots, K$.

3.2 Model Design Principles

Our MPHAT model is designed to generate users' posts and following links based on their topical interests, platform preferences, hubs, and authorities in a multiple OSN environment. We employ topic modeling approaches for generating user posts from topics. We also use a factorization approach to generate the following links from topic-specific platform preferences, hubs and authorities.

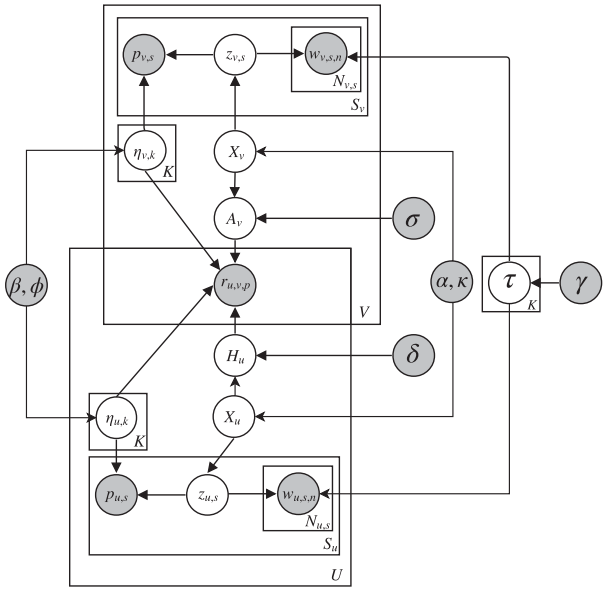


Fig. 1. Plate diagram of MPHAT model.

The notable point in our model is in the explicit and direct modeling of the relationships among topical interests, platform preferences, hubs and authorities. In MPHAT, user topical interests and platform preferences not only determine post content and which platform the content will be shared, but also play important roles in determining hubs and authorities. The relationships are however not deterministic, but probabilistic in nature. We postulate that it is necessary for a user to be interested in a topic before she becomes an authority or hub for that topic. However, users having strong interest in a topic may not be authoritative or hub for that topic. Moreover, different topical hub or authority users can be found on different platforms. MPHAT model therefore learns for each user the numerical scores of her topic-specific hub and authority, as well as her platform preferences. Also, unlike the existing models that return scores normalized across users, topics, or platforms, MPHAT aims at learning users' explicit, unnormalized scores, which can be used directly or normalized when required.

3.3 Generative Process

We depict the plate diagram of the MPHAT model in Fig. 1 and summarize in Algorithm 1 its generative process. Recall that the number of topics K is given and suppose that there are P OSN platforms. We denote the word distribution of topic k by τ_k and assume that it is sampled from a given Dirichlet prior with parameter γ. User posts and following links are then generated as follows.

Generating Topic Interest Vectors. For each user u (also the user v in the plate diagram), the k-th dimension of her topical interest vector, $X_{u,k}$, is sampled from the Gamma distribution with shape α and scale κ. Gamma distribution is chosen over Gaussian because we want the values of topical interests to be positive values.

Generating Topic-Specific Platform Preference Vectors. We follow a similar approach to generate user's topic-specific platform preference vector. For every user u and every topic k, the p dimension of u's platform preference vector specific to topic k $\eta_{u,k,p}$ is sampled from the Gamma distribution with shape β and scale ϕ.

Algorithm 1. Generative Process for MPHAT Model

⊡ "Generating topics"
for each topic k **do**
  sample the topic's word distribution $\tau_k \sim Dir(\gamma)$
end for
⊡ "Generating user topical interests and topic-specific platform preferences"
for each user u **do**
  for topic $k \in \{1, \ldots, K\}$ **do**
    sample u's interest in topic k: $X_{u,k} \sim \Gamma(\alpha, \kappa)$
    for platform $p \in \{1, \ldots, P\}$ **do**
      sample u' preference for platform p on topic k: $\eta_{u,k,p} \sim \Gamma(\beta, \phi)$
    end for
  end for
end for
⊡ "Generating user topic-specific authorities and hubs"
for each topic k **do**
  for each user $v \in V$ **do**
    sample v's authority on topic K: $A_{v,k} \sim \Gamma(\sigma, \frac{X_{v,k}}{\sigma})$
  end for
  for each user $u \in U$ **do**
    sample u's hub on topic k: $H_{u,k} \sim \Gamma(\delta, \frac{X_{u,k}}{\delta})$
  end for
end for
⊡ "Generating posts"
for each user u **do**
  for each post s **do**
    sample topic $z_{u,s} \sim Multi(\theta_u)$ where $\theta_u = \mathsf{s}(X_u)$
    for each word slot n **do**
      sample the word $w_{v,s,n} \sim Multi(\tau_{z_{v,s}})$
    end for
    sample platform $p_{u,s} \sim Multi(\Omega_{uz_{u,s}})$ where $\Omega_{uz_{u,s}} = \mathsf{s}(\eta_{u,z_{u,s}})$
  end for
end for
⊡ "Generating following relationship"
for each pair of source user u and target user v **do**
  sample the relationship: $r_{u,v,p} \sim Bernoulli(f(H_u^{pT} A_v^p, \lambda))$
end for

Generating Posts. To generate the s-th post of user u, the post's topic $z_{u,s}$ is first sampled from the multinomial distribution with parameter $\theta_u = \mathsf{s}(X_u)$. Here $\mathsf{s}(X)$ is the Softmax function[1] that converts an arbitrary vector to a probabilistic vector of the same dimension size. Similar to other previous works on modeling user content in social networks [48], we assume that each post has only one topic as it contains limited amount of text. The post's content is then generated by sampling its words. Each word $w_{u,s,n}$ is sampled from the word distribution of the chosen topic, i.e., $\tau_{z_{u,s}}$, independently from the other words. Lastly, the platform on which the post is shared is sampled from the multinomial distribution $\Omega_{u,z_{u,s}} = \mathsf{s}(\eta_{u,z_u})$.

Generating Topic-Specific Hub and Authority Vectors. MPHAT incorporates two main ideas in generating user topic-specific hubs and authorities vectors. First, MPHAT models the users' topic-specific hub and authority values as positive numeric values. Second, MPHAT probabilistically relates these hub and authority values to user topical interests. Hence, we propose to model a user's topic-specific hub and authority scores using Gamma distributions whose means are the user's interest for the topics. Specifically, the topic-specific authority score of user $v \in V$ for topic k, $A_{v,k}$, is sampled from the Gamma distribution with shape σ and scale $\frac{X_{v,k}}{\sigma}$. Similarly, the topic-specific hub score of user $u \in U$ for topic k, $H_{u,k}$, is sampled from the Gamma distribution with shape δ and scale $\frac{X_{u,k}}{\delta}$. Due to the property of Gamma distributions[2], both $A_{v,k}$ and $H_{u,k}$ share the same expectation $X_{u,k}$.

Generating Links. We use $r_{u,v,p}$ to denote the relationship between u and v on platform p: $r_{u,v,p} = 1$ if u follows v on p, and = 0 otherwise. To generate $r_{u,v,p}$, we first derive the platform-specific authority vector of v on platform p, A_v^p, by taking the element-wise product of A_v and vector $\mathsf{s}(\eta_{v,1,p}, \ldots, \eta_{v,K,p})$. Similarly, the platform-specific hub vector of u on platform p, H_u^p, is defined by the element-wise product of H_u and vector $\mathsf{s}(\eta_{u,1,p}, \ldots, \eta_{u,K,p})$. Finally, we sample $r_{u,v,p}$ from the Bernoulli distribution with mean $f(H_u^{pT} A_v^p, \lambda)$. Here $H_u^{pT} A_v^p$ is the dot product of H_u^p and A_v^p; and f is the function to scale it to [0, 1) and is defined below.

$$f(x, \lambda) = 2\left(\frac{1}{e^{-\lambda x} + 1} - \frac{1}{2}\right), \tag{1}$$

where $\lambda \in (0, 1)$ is an engineering parameter.

The likelihood of forming a following link from u to v is therefore factorized into u's topic-specific hub scores, v's topic-specific authority scores and their platform preferences. The likelihood is high when these scores and platform preferences correlate (i.e., u has high hub in topics that v has high authority, and both of them have high preference for the same platform), and is low otherwise.

3.4 Model Learning

Given the prior γ, and the parameters α, β, δ, σ, ϕ, κ, and λ, we learn the other parameters in MPHAT model using maximum likelihood approach. In other words, we solve the following optimization problem.

$$\{X^*, \eta^*, A^*, H^*, Z^*, \tau^*\} = \arg.\max_{X,\eta,A,H,Z,\tau} L(\mathcal{D}|\Psi). \tag{2}$$

In Equation (2), $\Psi = \{X, \eta, A, H, Z, \tau, \alpha, \beta, \delta, \sigma, \phi, \kappa, \lambda, \gamma\}$ where X represents for the set of X_u for all users $\{u\}$. η, A and H are similarly defined. Z represents for the bag of topics of all posts, while τ represents for the set of all topic word distributions $\{\tau_k\}$. Lastly, $L(\mathcal{D}|\Psi)$ is the likelihood function of the observed data $\mathcal{D}$ (i.e., posts and following links) given the value of all the parameters.

Similar to LDA-based models, the problem in Equation (2) is intractable [47]. We therefore make use of Gibbs-EM

1. https://en.wikipedia.org/wiki/Softmax_function

2. https://en.wikipedia.org/wiki/Gamma_distribution

method [53] for learning in MPHAT model. Specifically, we first randomly initialize X, η, A, H, and τ. We then iteratively perform the following steps until reaching a convergence or exceeding a given number of iterations.

- To sample Z while fixing X, η, A, H, and τ. The topic $z_{u,s}$ is sampled according to the following equation.

$$P(z_{u,s} = k|\theta_u, \tau) \propto \theta_{u,k} \times \prod_{n=1}^{N_{u,s}} \tau_{k,w_{u,s,n}}, \quad (3)$$

  where, again, $\theta_u = \mathsf{s}(X_u)$
- To optimize X, η, A, H, and τ while keeping Z unchanged. In this step, we make use of the alternating gradient descent method [54]. That is, we iteratively optimize X, η, A, H, or τ while fixing all the others.

3.5 Parallelization

As suggested by Equation (3), the sampling of a post's topic is independent from that of all the other posts. Hence, we can use multiple child processes, each corresponding to a small set of users, to sample the topics for the users' posts simultaneously. Also, in the alternating steps for optimizing X, we can parallelize the computation as the optimization of a user's topic interest vector is independent of that of all other users' topic interest vectors. Similarly, we can parallelize the alternating optimization of A, H, η, and τ.

In our implementation, in sampling Z, we build a process pool, and submit a process for sampling topic for posts of $\frac{1}{N}$ of the users where N is the pool's size. In the ideal case, we can reduce the running time of sampling Z to N times. Similarly, we use process pool to reduce the running time in the alternating optimization steps.

3.6 Data Sub-Sampling

Like previous factorization and mixed membership models, the MPHAT model considers both link and non-link relationships of all pair of users. This makes the overall complexity of the MPHAT model to be $O(N_u^2)$ where N_u is the number of users, which is not practical for large scale social networks. We therefore choose to use a data sub-sampling method to reduce the computational cost. To do that, for each user u, we keep all u's out links (i.e., the links where u follows other users) and $m\%$ of its out non-links (i.e., the no-links where u does not follows some other users). These $m\%$ non-links are selected from the followees of u's followees (i.e., the 2-hops non-existent links). This selection strategy retains only a subset of relationships that carry strong signal of users' hub and authority values, while filtering out the remaining data that may contain noise.

4 Experiments on Real-World Dataset

Ideally, we should evaluate MPHAT by comparing the authority and hub users identified by the model with ground truth authority and hub users. However, it is difficult to have find ground truth in real-world datasets. For such datasets, we evaluate MPHAT against some baseline methods on three tasks: (i) modeling of topics, (ii) users' platform choice prediction, and (iii) link recommendation. We first introduce the real-world datasets which we have collected for our model evaluation. Next, we describe the experiments conducted and report the results. Finally, we present several empirical findings on the topics, hub and authority users learned by the MPHAT model.

4.1 Dataset

Our model evaluation requires multiple datasets that allow us to observe user topical interests and preferences. Furthermore, as we are interested in studying authorities and hubs across online social networks, we require some users to have accounts on multiple OSNs. Public datasets that satisfy the above requirements are not available. Thus, we specially collect two datasets from two popular social networking platforms that fulfill our requirements, namely Twitter, a short-text microblogging site, and Instagram, a photo-sharing social media site. Both Twitter and Instagram support directed relationships among users, which reflect the preferences of users towards *following* other authority users. Furthermore, the hub and authority users in the two platforms may differ with respect to different topics.

For Twitter data, we collected a set of Singapore-based Twitter users who declared Singapore locations in their user profiles. These users were identified by an iterative snowball sampling process starting from a small seed set of well known Singapore Twitter users followed by traversing the follow links to other Singapore Twitter users until the sampling iteration did not get any more new users. From these users, we obtain a subset of users who are active, i.e., have more than 50 directed links, and posted at least 40 tweets between October and December 2016. Subsequently, we retrieve the posts of these *active* Twitter users. A similar approach is used to retrieve the data of active Instagram users who have more than 50 directed links and posted at least 10 posts between October and December 2016.

To identify users having accounts on both Twitter and Instagram among the above active Twitter users, we obtain a subset of users who mention their Instagram accounts in their Twitter bio descriptions. If a mentioned Instagram account is active and do not exist in our subset of active Instagram users, we retrieve the posts and links of that account and add it to our Instagram user set. A similar approach is used to retrieve users who have mentioned their Twitter accounts in their Instagram bio descriptions. Table 2 shows the statistics about the collected datasets. In total, we gathered 5,633 Instagram users and 5,401 Twitter users. Among the gathered users, 932 pairs of Twitter and Instagram user accounts are owned by the same users, i.e., these users have active accounts on the two OSNs.

4.2 Experiment Setup

We evaluate MPHAT model in three tasks, namely, (i) topic modeling, (ii) platform choice prediction, and (iii) link recommendation. The first task focuses on comparing the topics learned by MPHAT with those learned by other baseline models. The second task applies MPHAT to predict users' platform choices as they publish posts. Finally, the last task applies MPHAT to the prediction of missing links in OSNs. Note that three evaluation tasks will be conducted in the multiple OSNs setting. For example, in the first task,

TABLE 2
Statistics for Instagram and Twitter Datasets

	Instagram	Twitter
Total users	5,633 (932)	5,401 (932)
Total links	342,719 (22,529)	276,299 (25,379)
Avg Links/user	60 (24)	51 (27)
Max followers	803 (217)	2,048 (421)
Max followings	672 (147)	991 (172)
Min followers	5 (5)	5 (5)
Min following	5 (5)	5 (5)
Total posts	636,593 (121,856)	944,035 (143,317)
Max posts/user	200 (200)	200 (200)
Min posts/user	10 (40)	40 (40)
Avg posts/user	113 (130)	174 (153)

Numbers in () refer to counts that involve users with accounts on both platforms and the links among these accounts only.

we not only model the topics in individual platforms (i.e., Twitter and Instagram separately) but also topics across both OSNs. In the second task, we predict the platform choices of users who have accounts on multiple OSNs. Finally in the last task, we train MPHAT with user relationships from multiple OSNs and predict links to users in individual platforms.

4.2.1 Baselines

For topic modeling, we compare MPHAT with HAT [14], LDA [47] and TW_LDA [48]. HAT is a generative model which jointly models the latent topics, users' topical interests, hub and authority scores from user posts and relationships found at one OSN only. LDA and TW_LDA are two popular topic models for text documents and Twitter content respectively.

For platform choice prediction, we compare MPHAT with MultiPlatform-LDA (MultiLDA) [52] and TW_LDA. MultiLDA learns the user's platform preferences from their posts. Although TW_LDA does not model platform choices, we could infer the posts' platform based on the popular platform choice for the topics learned using TW_LDA.

For link recommendation, we compare MPHAT against several baselines: HAT, HITS, WTFW, and common user interests learned by LDA and TW_LDA. The intuition for interest-based baselines is that user who share common interests are likely to follow each other due to homophily [55]. WTFW models the topic-specific and social relationships among users, while HITS returns the authority and hub scores of users based on the relationship network structure.

4.2.2 Parameter Setting

In our experiments, the parameter setting of LDA, TW_LDA, WTFW, and HAT methods are set to the default values as recommend in their origin. HITS method is parameter free. For MPHAT methods, we found that the Gibbs-EM algorithm converges around after 200 alternating iterations, each iteration includes 10 gradient descent steps. Topics' prior is set to a symmetric Dirichlet distribution with $\gamma = 0.001$ as widely used in previous works. Both shape α and scale κ of the Gamma prior of users' topical interest X_{uk} are set to 2 for all users u and all topics k. This setting makes X_{uk}'s mean and standard deviation close to 4 and 3 respectively. That means X_{uk} deviates moderately with respect to its mean, hence, $\mathsf{s}(X_{uk})$ is moderately but not extremely skewed toward any topic. This is reasonable as we expect that it is very less likely that users totally focus on some single topic. Similarly, both shape β and scale ϕ of the Gamma prior of users' platform preference η_{ukp} are set to 2 as we do not expect users, who have account on multiple platforms, to totally focus on some single platform. Also, the shapes σ and δ of Gamma priors of users' authority and hub are set to 2. This makes the means of users' authority A_{uk} and hub $H_{u,k}$ close to their topical interest $X_{u,k}$. The scaling parameter λ is set to 0.01 through empirical evaluation on a list values.

4.2.3 Evaluation Metrics

Topic Modeling Evaluation. For evaluation on topic modeling, we compute the likelihood of the training set and perplexity of the test set when MPHAT and the baselines are applied to the OSN datasets. The model with higher likelihood and lower perplexity is considered superior in this task.

Platform Choice Prediction Evaluation. For evaluation on platform choice prediction, we get the models to predict users' platform choices given the content of the test posts. The platform choice of a test post is predicted by MPHAT by first assigning the posts topic using the trained MPHAT, and then selecting the most probable platform for the assigned post topic where the most probable platform is determined by the users topic-specific platform preference distribution.

For TW_LDA which does not model platform choices, we generate the predicted platform choice of a given test post by first assigning the particular posts topic using the trained TW_LDA, and then returning the most popular platform choice for the assigned topic according to the training set

Finally, we compute accuracy to measure the accuracy of platform choice prediction. *Accuracy* for platform choice prediction is defined as:

$$Accuracy = \frac{\#\text{posts with platform correctly predicted}}{\#\text{posts in all platforms}}.$$

Link Recommendation Evaluation. For evaluation on link recommendation, we first define the link recommendation task as recommending new links to user in a given OSN platform, i.e., we want to recommend users other users to follow in a specific OSN platform. Thus, given a user u, we first rank her predicted *following* and *non-following* of a specific OSN in the test set by some link scores. Then, we recommend u other users v who are in the specific platform and are higher on the link scores.

For MPHAT, the link score, $score_{\mathsf{MPHAT}}(u, v, p)$ that user u would follow user v is measured by the likelihood that $r_{u,v,p} = 1$ as computed based on the two users' hub, authority, and platform preference as described in Section 3.3 on *Generating links*. Similarly, for HAT, the score, $score_{\mathsf{HAT}}(u, v)$, is the likelihood that u follows v as computed based on the two users' hub and authority learnt by HAT.

For HITS, the score is measured by taking the product of u's hub (h_u) and v's authority (a_v):

$$score_{\mathsf{HITS}}(u, v) = h_u \cdot a_v. \quad (4)$$

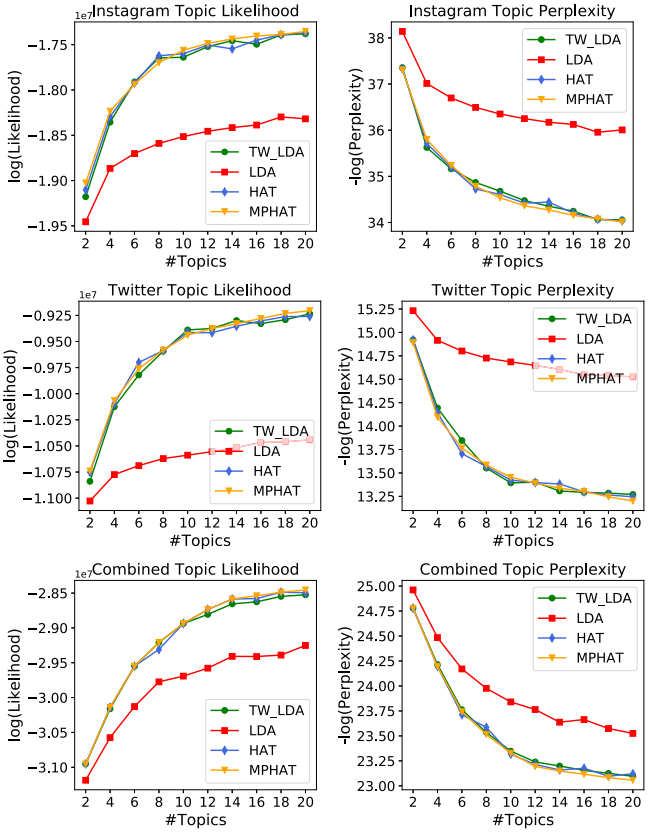


Fig. 2. Likelihood and perplexity of topics modeled in Instagram, Twitter, and combined datasets.

For LDA, the score is measured by taking the inner product of the topical interests θ_u and θ_v:

$$score_{\text{LDA}} = \sum_{k=1}^{K} \theta_{u,k} \cdot \theta_{v,k}. \quad (5)$$

The same way is also applied to measure links' scores in TW_LDA. Lastly, for WTFW, we directly use the link scores returned by the model.

Finally, we use *precision at top k* and Mean Reciprocal Rank (MRR) [56] to measure the accuracy of link recommendation. *Precision at top k* is defined as:

$$Prec_k = \frac{\sum_{u \in u_k} |L_u \cap L'_{u,k}|}{k \cdot |u_k|},$$

where u_k is a set of users with at least k positive links, L_u and $L'_{u,k}$ are the set of u's positive links and set of top k predicted links for u.

4.2.4 Training and Test Datasets

We generate three pairs of training and test datasets which will be used in our experiments: (i) *Instagram*, (ii) *Twitter* and (iii) *combined* datasets.

For *Instagram* datasets, we randomly select 80 percent of Instagram posts and links from each user who have an account on Instagram to form the training set and use the remaining posts and links as the test set. A similar process is applied to generate the *Twitter* training and test dataset. The *Instagram* and *Twitter* datasets are used to conduct single platform link recommendation experiments.

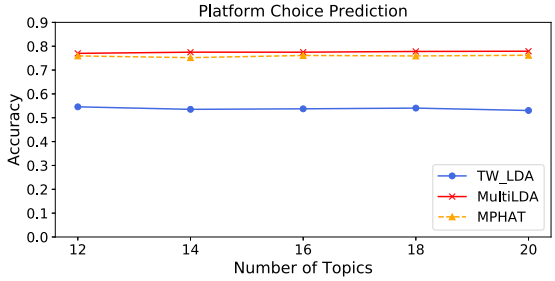


Fig. 3. Accuracy of platform choice prediction at various number of topics.

For the *combined* datasets, we randomly select 80 percent of platform-specific posts and links from each user to form the training set and use the remaining posts and links as the test set. When combining the two platforms, the users who have accounts on both Twitter and Instagram will be unified into a single user identity. The *combined* datasets are used to conduct multiple OSNs setting experiments in the three evaluation tasks, i.e., topic modeling, platform choice prediction and link recommendation.

4.3 Evaluation on Topic Modeling

We evaluate the topic modeling of MPHAT and the baselines on three datasets mentioned in Section 4.2.4. Fig. 2 shows the likelihood and perplexity achieved by MPHAT, HAT, LDA and TW_LDA. As expected, the larger the number of topics, the higher likelihood and lower perplexity are archived by all models. The quantum of improvement, however, reduces as the number of topics increases.

Fig. 2 also shows that MPHAT outperforms LDA, and is comparable to HAT and TW_LDA in the topic modeling task. This result supports the insights from previous work which suggested that standard LDA does not work well for short social media text as both Instagram photo captions and Twitter tweets are much shorter than normal documents [48]. A possible explanation for the similar results achieved by MPHAT, HAT and TW_LDA can be due to the three models assuming that each post has only one topic.

Interestingly, we also observe that MPHAT, HAT and TW_LDA have outperformed LDA more in Twitter than Instagram. A possible explanation can again be attributed the different length of the post in different OSNs; Twitter tweets as shorter with a 140 character limit, while Instagram photo captions are longer with no limitation in length imposed.

4.4 Evaluation on Platform Choice Prediction

We next evaluate MPHAT and and the baselines in a platform choice prediction task using the *combined* dataset. The task predicts the platforms to be used for posts from users with accounts on both Instagram and Twitter. Fig. 3 shows the *accuracy* of MPHAT, MultiLDA and TW_LDA for each platform with number of topics varying from 12 to 20. We observe that MPHAT and MultiLDA outperforms TW_LDA by about 35 percent in this prediction task. The figure also shows that the prediction results do not change significantly for different number of topics.

We also observe that MultiLDA outperforms MPHAT by a very small margin. A possible reason for this observation could be due to the noise introduced by the user

TABLE 3
Multiple Platform Instagram and Twitter Link Recommendations

Method	P@1	P@2	P@3	P@4	P@5	MRR
Instagram						
LDA	0.017	0.017	0.018	0.019	0.020	0.065
TW_LDA	0.015	0.017	0.017	0.017	0.018	0.059
HITS	0.069	0.065	0.057	0.051	0.050	0.135
WTFW	0.086	0.070	0.058	0.052	0.048	0.141
HAT	0.087	0.078	0.073	0.067	0.064	0.160
MPHAT	**0.114**	**0.104**	**0.097**	**0.090**	**0.086**	**0.200**
Twitter						
LDA	0.020	0.019	0.019	0.018	0.017	0.067
TW_LDA	0.017	0.017	0.018	0.019	0.019	0.067
HITS	0.100	0.094	0.084	0.078	0.076	0.203
WTFW	0.152	0.125	0.109	0.100	0.093	0.261
HAT	0.196	0.163	0.144	0.129	0.117	0.305
MPHAT	**0.226**	**0.182**	**0.156**	**0.141**	**0.130**	**0.337**

TABLE 4
Stratified Instagram and Twitter Link Recommendations

Method	P@1	P@2	P@3	P@4	P@5	MRR
Instagram						
HAT (All)	0.087	0.078	0.073	0.067	0.064	0.160
MPHAT (All))	0.114	0.104	0.097	0.090	0.086	0.200
%Improvement	31%	31%	32%	33%	33%	25%
HAT (MP)	0.032	0.035	0.037	0.038	0.040	0.096
MPHAT (MP)	0.047	0.065	0.066	0.066	0.063	0.152
%Improvement	43%	84%	77%	72%	57%	59%
Twitter						
HAT (All)	0.196	0.163	0.144	0.129	0.117	0.305
MPHAT (All))	0.226	0.182	0.156	0.141	0.130	0.337
%Improvement	15%	11%	8%	9%	11%	10%
HAT (MP)	0.050	0.057	0.056	0.055	0.052	0.126
MPHAT (MP)	0.073	0.075	0.070	0.062	0.059	0.161
%Improvement	46%	29%	24%	12%	12%	28%

relationships; MultiLDA learns the users' platform preference from their posts, while MPHAT considers both users' posts and relationships when learning the users' platform preference. Some users, albeit few, might form a lot of relationships in Twitter but they seldom tweet, and this could lead MPHAT to infer that the user has stronger preference in Twitter.

4.5 Evaluation on Link Recommendation

In link recommendation experiments, we consider all links in test datasets as positive instances, and in principle, all the non-existent links as negative instances. Nevertheless, due to the sparsity of OSNs, the number of possible non-links is enormous. Thus, we limit the negative instances to all the nodes which are 2-hops away from the source node of each positive link, which is about 100 times the number of positive instances. The evaluation on link recommendation are conducted in two settings: (i) *multiple platforms* and (ii) *single platform* link recommendation.

In the multiple platforms link recommendation setting, we train MPHAT and the baseline models on the *combined* training dataset and perform link recommendation in individual OSNs separately using the *combined* test dataset. This experiment aims to evaluate the models when recommending links in multiple OSNs. To further analyze the model effectiveness, we will present the recommendation results involving (a) all types of links and (b) links among users who have accounts on both platforms (i.e., MP Links) using the *combined* test dataset.

In single platform link recommendation setting, the models are trained on a single OSN training dataset, say *Instagram* training dataset, and the link recommendation is performed on the same single OSN test dataset, i.e., *Instagram* test dataset. The purpose of this experiment setting is to evaluate MPHAT's ability in single platform link recommendation compared with other single platform methods even though MPHAT is designed for multiple OSN setting.

4.5.1 Multiple Platforms Link Recommendation

Table 3 shows the multiple platforms link recommendation results for Instagram and Twitter. Note that for MPHAT and the topic-specific baselines, i.e., HAT, WTFW, LDA and TW_LDA, the number of topics learned is set to 18 as beyond which, the quantum of improvement on topic likelihood and perplexity are significantly reduced (see Section 4.3).

We observe that MPHAT outperforms all baselines in both *precision at top k* and MRR for both Instagram and Twitter. When measured by MRR, MPHAT significantly outperforms HITS by more than 50 and 60 percent in Instagram and Twitter respectively. This suggests that the topical context is important in link recommendation. MPHAT also improves the MRR of the common user interests baselines by more than two-fold. This also suggests the importance of network information in link recommendation.

Considering both OSNs, MPHAT and HAT also outperform WTFW by more than 10 percent in MRR. Interestingly, this demonstrates the importance of hub when modeling topical links; WTFW models susceptibility as users who are interested in a particular topic, while MPHAT and HAT model hub as users who are not only interested in a topic but follow users who are also authority users in that topic. Finally, when measured by MRR, MPHAT outperforms HAT by more than 25 and 10 percent in Instagram and Twitter respectively. This demonstrates MPHAT's superiority over HAT in recommending links in multiple OSNs setting.

Table 4 shows the results for links among users who have accounts on both platforms. We observe that MPHAT has significant improvement over HAT for both *all links* and *MP links*. In particular, MPHAT observes 25 percent improvement by MRR over HAT for *all links* recommendation in both Instagram and Twitter. This could be attributed to MPHAT model design, which considers the users' platform preferences. For example, when user u, who has accounts on both Instagram and Twitter, is an authority for a specific topic k, HAT will recommend other Instagram and Twitter users who are hub for topic k to follow u. However, suppose u is actually more active in Instagram. She is more likely to be an authority for topic k in Instagram only. MPHAT, which models u's platform preferences, would instead recommend only Instagram users who are hub for topic k to follow u.

We also note that the MRRs for Instagram and Twitter *MP links* are lower than *all links* recommendation for both models. We examined the learned model parameters and

TABLE 5
Single Platform Instagram and Twitter Link Recommendations

Method	P@1	P@2	P@3	P@4	P@5	MRR
Instagram						
LDA	0.018	0.019	0.019	0.019	0.019	0.062
TW_LDA	0.020	0.018	0.017	0.017	0.017	0.059
HITS	0.078	0.070	0.063	0.057	0.054	0.145
WTFW	0.099	0.082	0.071	0.064	0.059	0.167
HAT	0.103	0.092	0.086	0.081	0.078	0.182
MPHAT	**0.123**	**0.113**	**0.106**	**0.100**	**0.097**	**0.211**
Twitter						
LDA	0.017	0.017	0.018	0.019	0.019	0.067
TW_LDA	0.024	0.025	0.025	0.024	0.023	0.080
HITS	0.055	0.066	0.064	0.064	0.065	0.169
WTFW	0.169	0.146	0.132	0.123	0.115	0.296
HAT	0.220	0.166	0.144	0.130	0.120	0.319
MPHAT	**0.220**	**0.182**	**0.159**	**0.146**	**0.135**	**0.335**

found that most of the users who have accounts on both OSNs are topical authorities but not strong hubs. On average, 48.91 percent of the top 100 authority users across the 18 topics are users on both OSNs. Conversely, only 19.91 percent of the top 100 hub users across the 18 topics are users on both OSNs. These characteristics of the users on both OSNs make it harder to recommend *MP links* to these users because most of them they are authorities and they have less propensity to follow other authorities.

4.5.2 Single Platform Link Recommendation

Table 5 shows the single platform link recommendation results for Instagram and Twitter. Note that for topic-specific models, the number of topics learned in the training phase is set to 8 and 10 for Instagram and Twitter respectively.

Similar to link recommendation in multiple platform setting, we observe that MPHAT outperforms all baselines measured by *both precision at top k* and MRR for both Instagram and Twitter. This shows that MPHAT can also perform well in single platform link recommendation.

We also note that MPHAT outperforms HAT by a small margin. A possible reason could be due to the learning of the models. HAT learns the users' topical interests using projection gradient descent and this constraint might result in trapping the learned parameter in a local optimal. MPHAT, on the other hand, does not have this constraint as it generates the users' topical interests by applying Softmax function on the learned unconstrained user latent factor.

Interestingly, we also observe that the MRR of single platform link recommendation is higher for most models than that of multiple platform link recommendation. A possible explanation could be the additional noise introduced when we combined the Instagram and Twitter datasets to form the *combined* dataset. For example, when recommending Instagram links in the test dataset, we train the models using the Twitter and Instagram links in the *combined* training dataset. The additional Twitter links might be noise in modeling influence of Instagram users, thus making the Instagram link recommendation task more difficult for multiple platforms. The effect of this additional cross-platform noise is further discussed in an empirical analysis in Section 4.6.2.

4.6 Empirical Analysis

In this section, we first examine the topic-specific platform preferences of users learned by the MPHAT model. Next, we empirically compare the authority and hub users learned by HITS, HAT, and MPHAT. Note that the analysis is conducted on the *combined* dataset.

4.6.1 Topic-Specific Platform Preferences

Other than the users' topical interests, authorities and hubs, MPHAT also learns the topical platform preferences of users on multiple platforms. Here, we showcase the platform preference of users on Instagram and Twitter. Fig. 4 shows the distributions of platform preferences of users with accounts on multiple OSNs for four selected sample topics, namely, "sports", "current affairs", "beauty" and "gourmet".

Generally, we observe that most of the users have 0.5 platform preference across multiple topics. This is due to the users not publishing any posts related to those topics; in such situation the default platform preferences becomes $1/P$ when P is the number of platforms. More interestingly, we also observe that the distribution of platform preferences differs across the four topics. This observation supports previous research work [52] that suggests that users have different platform preferences for different topics. For example, for the topics on "sports" and "current affairs", the right-leaning bar charts of users' platform preference for Twitter suggest that the users on multiple platforms prefer to generate their "sports" and "current affairs" content in Twitter, and also link to other Twitter users who have displayed interests on the two topics.

The study on users' topical platform preference also has implications for users' topical authority and hub values. Suppose that "sports" is a popular topic on Twitter and a user, u, who has accounts on both Twitter and Instagram, is identified as a "sports" authority, it is likely that u also has a stronger platform preference for Twitter on "sports" topic.

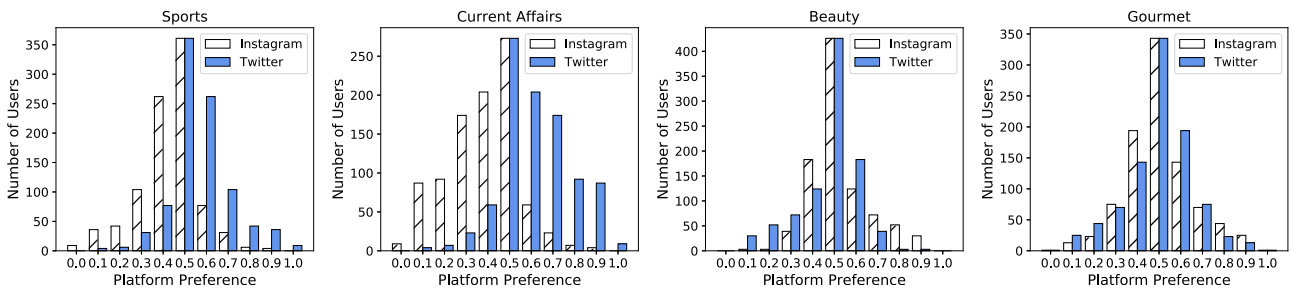


Fig. 4. Distributions of platform preferences for *sports*, *current affairs*, *beauty*, and *gourmet* topics.

TABLE 6
A Sample of Authority and Hub Users in *Combined* Dataset Learned by HITS, HAT, and MPHAT

Topic	Top 10 Keywords	Top 5 Authority Users	Top 5 Hub Users
		HITS	
-	-	^{C}@xiaxue, ^{T}@blxcknicotine, ^{C}@naomineo_ (lifestyle blogger), ^{C}@benjaminkheng, ^{C}@toshrock (celebrity)	^{T}@blxcknicotine, ^{C}@naomineo_ (lifestyle blogger), ^{C}@benjaminkheng, ^{C}@flyirene (celebrity), ^{T}@herbertrsim (businessman)
		HAT	
Beauty	beauty, makeup, skincare, treatment, clozette, collection, lip, foundation, facial, lipstick	^{I}@sephorasg, ^{I}@laneigesg (cosmetics brand), ^{I}@benefitcosmeticssg (lifestyle blogger), ^{I}@beautifulbuns_sg (fashion magazine), ^{I}@thewowoshop (cosmetics ecommerce)	^{I}@sephorasg, ^{I}@etudehousesingapore, ^{I}@laneigesg (cosmetics brand), ^{I}@benefitcosmeticssg (lifestyle blogger), ^{I}@a_must_shop (cosmetic ecommerce)
Sports	game, team, united, arsenal, manutd, league, fans, football, goal, footy_jokes	^{T}@lfc, ^{T}@arsenal (football club), ^{T}@ufc (sports news media), ^{T}@futballtweets, ^{T}@empireofthekop (sports blogger)	^{T}@redsports, ^{T}@empireofthekop, ^{T}@futballtweets, ^{T}@coutinhoflair, ^{T}@theredcardtv (sport blogger)
Current Affairs	business, marketing, digital, trump, tech, ai, data, china, fintech, startup	^{C}@stcom, ^{T}@channelnewsasia (news media), ^{T}@mrbrown (satire blogger), ^{T}@eskimon (businessman), ^{T}@govsingapore (government)	^{T}@wtfsg (satire blogger), ^{T}@eskimon, ^{T}@herbertrsim, ^{T}@alansoon (business), ^{T}@robinhicks_ (editor)
		MPHAT	
Beauty	beauty, makeup, skincare, treatment, natural, facial, oil, lip, foundation, clozette	^{C}@jamietyj, ^{C}@bongqiuqiu, ^{C}@bellywellyjelly, ^{C}@Xiaxue, ^{C}@xchubbykitty (lifestyle blogger)	^{I}@ilrpsg (skin-care brand), ^{C}@william82sg, ^{C}@JoannaLHS, ^{I}@makeupforeversg, ^{I}@benefitcosmeticssg (lifestyle blogger)
Sports	arsenal, game, manutd, team, league, football, united, goal, mufc, liverpool	^{C}@stcom, ^{T}@channelnewsasia (news media), ^{T}@lfc , ^{T}@arsenal (football club), ^{T}@redsports (sport blogger)	^{T}@alb_s_fc (football club), ^{T}@redsports, ^{T}@footbalifact, ^{T}@futballtweets (sport blogger), ^{T}@theutdreview (fan group)
Current Affairs	business, marketing, digital, trump, tech, ai, data, china, fintech, startup	^{C}@stcom (news media), ^{T}@eskimon (businessman), ^{T}@mrbrown (satire blogger), ^{C}@papsingapore (political party), ^{T}@govsingapore (government)	^{C}@pinkdotsg (social group), ^{T}@alansoon, ^{C}@skinnylatte, ^{T}@mrscotteddy, ^{C}@mediumshawn (businessman)

^{I}@, ^{T}@ and ^{C}@ denotes Instagram, Twitter, and multiple OSNs users, respectively.

This is because most of the sports-loving users and hubs who follow u are likely to be from Twitter. Note that u may also have other Instagram followers. However, these Instagram followers may not contribute much in determining the u's authority in "sports" topic because majority of the "sports" topical hubs that link to u are in Twitter. Another empirical example on topical platform preference's effects on topical authority and hub is discussed in Section 4.6.2.

4.6.2 Hub and Authority Users

Table 6 shows samples of the authority and hub users learned by HITS, HAT and MPHAT. HITS basically determines the authority and hub users strictly by the network structures. Thus, the top authority and hub users identified by HITS are popular Twitter and Instagram users with many followers. On the other hand, MPHAT and HAT are able to identify authority and hub users for specific topics. For example, for the "sports" topic, MPHAT was able to identify popular football clubs and news media and a sports blogger as top authority users. These users often post sports-related content and are followed by many users interested in sports. Similarly, the top sports topic hub users identified by MPHAT are also sports bloggers and fan group who have followed the sports topic authority users. Similar observations are made in HAT.

Interestingly, we also observe the topic-specific authority and hub users identified by MPHAT different from those that are identified by HAT. Particularly for the topic on "beauty", MPHAT have identified popular lifestyle bloggers who have accounts on both Instagram and Twitter as authority users, while HAT identified cosmetics brands and lifestyle bloggers who only have Instagram accounts as authorities. A possible reason for the difference could be the additional cross-platform noise in modeling influence of users with accounts on multiple OSNs, which we have briefly discussed in Section 4.5.2.

To investigate this further, we first examine the top 100 hub users for the topic on "beauty" and found that they all

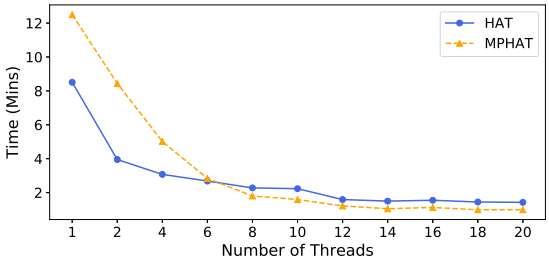


Fig. 5. Running time of HAT and MPHAT with various number of processes.

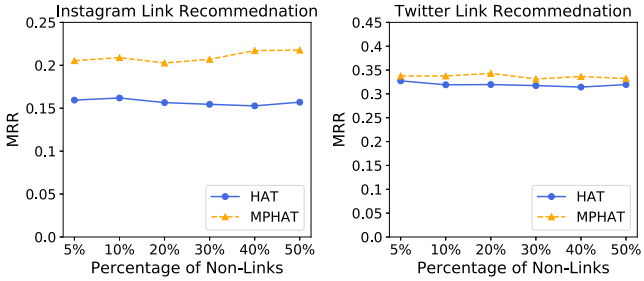


Fig. 6. MRR for Instagram and Twitter link recommendation with various percentage of non-link sampled.

have accounts on Instagram. This suggests that "beauty" is a popular topic in Instagram and the authority users followed by these hub users should also have an account in Instagram. Many of these top 100 hub users follow the top 5 "beauty" authority users identified by HAT and MPHAT.

However, HAT has given lower authority scores to the users who have accounts on multiple OSNs because they are also followed by other non-hub users in Twitter, i.e., noise from the links in other OSNs are introduced in HAT's modeling of the users' topical authority. MPHAT mitigates these noise by considering the topical platform preferences of users on multiple OSNs when learning their topical authority and hub scores from the users' links in multiple platforms. We examined the topical platform preferences of the top 5 "beauty" topic authority users identified by MPHAT and found that these authority users have an average 0.62 platform preferences score for Instagram, i.e., they have a stronger preference for the Instagram platform on the "beauty" topic. MPHAT weighs the "beauty" topical authority scores of these users by their platform preferences for Instagram, and reduces the effect of the noise among the links from Twitter.

4.7 Efficiency of Parallel Implementation

We now examine the efficiency of the parallel implementation of the learning algorithm in MPHAT as presented in Section 3.5. Fig. 5 shows the running time of a full iteration of the algorithm when the number of parallel processes is varied from 1 to 20. The figure clearly shows that, as we expected, the running time drop dramatically when the number of parallel processes starts increasing. This shows the efficacy of our parallel implementation. It is also expected that the running time does not decrease significantly after a certain number of processes due to trade off between the actual computing time and the additional time spent for managing the process pool.

4.8 Data Sub-Sampling Analysis

In Section 3.6, we discussed a data sub-sampling method used to reduce computation cost of HAT and MPHAT. We now empirically examine the effect on link recommendation of the data sub-sampling method. Note that the experiments are conducted in the *multiple platforms* link recommendation setting described in Section 4.5.

Fig. 6 shows the HAT and MPHAT's MRR for Instagram and Twitter link recommendation with various percentage of non-link sampled. The link recommendation results are observed to be consistent even when we increase the percentage of non-links sampled for training as the data sub-sampling process is not random but bias to more informative non-links (i.e., followees of users' followees). Thus, the additional less informative non-links would not improve the link recommendation performance significantly.

5 Experiments on Synthetic Datasets

We now evaluate the accuracy of MPHAT in recovering topics and users' authority and hub specific to topics. We also examine how MPHAT behaves in different data settings. To do so, we need the access to the ground truth value of those variables, which is however not available in any real dataset. We therefore address this shortcoming by generating synthetic datasets for conducting the experimental evaluations.

5.1 Synthetic Data Generation

We employ the following steps to generate a dataset with N users on P platforms, whose posts covering K topics and using a vocabulary with W words.

Generating users' Topical Interest. Given K topics, for each user u, we randomly choose 10 percent of topics to be ones that u is interested in. That is, the topical interest vector of u, X^g_u, is randomly generated such that the distribution $Softmax(X^g_u)$ (i.e., applying Softmax function on X^g_u) mostly skews on u's interested topics. Also, $\{X^g_u\}_u$ are also normalized across users such that: if users u and v are interested topics k_i and k_j respectively then X^g_{u,k_i} is similarly as large as X^g_{v,k_j}, and they are both much larger than other X^g_{w,k_l} for users w not interested in topics k_l. This normalization does not affect $Softmax(X^g_u)$ but creates clear and distinctive users' topical interest for more accurate comparison among models.

Generating users' Platform Preference. Given P platforms, as suggested by observations from real datasets used in the Section 4, we randomly choose a large subset of users, says 70 percent, to have accounts on only a single platform, and the remaining users have accounts on all P platforms. For each user u having account on only a single platform, says p, her platform preference vector $\omega^g_{u,k}$ is generated with $Softmax(\omega^g_{u,k})$ totally focused on the p-th element for any topic k. Otherwise, u has accounts on multiple platforms and $\omega^g_{u,k}$ is defined to have either (i) $Softmax(\omega_{u,k})$ return uniform distribution of platforms u has accounts on, or (ii) $Softmax(\omega^g_{u,k})$ return a distribution that skews 90 percent on a certain platform. We generate two synthetic datasets with all the users on multiple platforms either adopting uniform or skewed platform preference distributions. The two

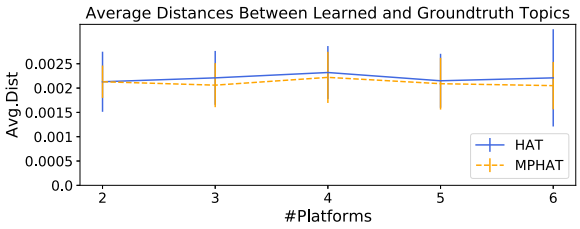


Fig. 7. Distances between learned and ground truth topics on *uniform* datasets.

synthetic datasets help to evaluate the models more comprehensively.

Generating users' Hub and Authority. For each topic k, we randomly choose a small proportion, says q, of users interested in k (refer to the previous step for generating users' topical interest) to be authority users of topic k. Similarly, q of users interested in k are randomly chosen to be hub users of topic k. As q becomes sufficiently larger, the users who are both authority and hub will increase. If v is among the authority users of k, her authority score $A^g_{v,k}$ is set to $X^g_{v,k}$ plus a small perturbation μ, ($\mu > 0$). Otherwise, $A^g_{v,k}$ is set to be much smaller than $X^g_{v,k}$. Similarly, the hub score $H^g_{u,k}$ of user u on topic k is set in the same way. As users' topical interest $X^g_{*,k}$'s are normalized, A^g_{v,k_i} is similarly as large as A^g_{u,k_j} if v and u are authoritative on topics k_i and k_j respectively. The same observations are held for users' hub scores. These result in a clear separation between authority (or hub) users and non-authority (or non-hub) users in the synthetic datasets. Such a separation helps to evaluate the models more accurately.

Generating Topics' Word Distribution. Given W words in vocabulary, for each of K topics, its word distribution is randomly generated such that the distribution skews on 10 percent of the words. Again, this skewness is to create clear and distinctive topics.

Generating the Posts and Relationships. For each user u, we generate a random number between T_{min} and T_{max} of posts, and for each u's post a random number between L_{min} and L_{max} of words. The posts' topic, words and following links are generated similar to the generative process described in Section 3.3.

5.2 Experiment Setup

We generate the synthetic datasets with $N = 1000$ users, $K = 10$ topics, and the number platforms P is varied from 2 to 6. We set the authority and hub perturbation μ to 0.1, and set T_{min} and T_{max} for post generation to be 100 and 200 respectively. For each topic k, 10 percent of the users who are interested in topic k are also randomly selected as the topical hub and authorities. For each setting, we generate two datasets: the *skewed* dataset and the *uniform* dataset. In the skewed dataset, users show platform preference to generate posts and relationships (i.e., their platform preference distribution are skewed) while the uniform dataset has users not having any platform preference.

In this section, we compare MPHAT against HAT only, the best performing baseline method on our real datasets. For training MPHAT and HAT models, we adopt the parameter settings described in Section 4.2.2.

We first evaluate the two models in recovering the topics by comparing their learned topics with ground truth topics. That

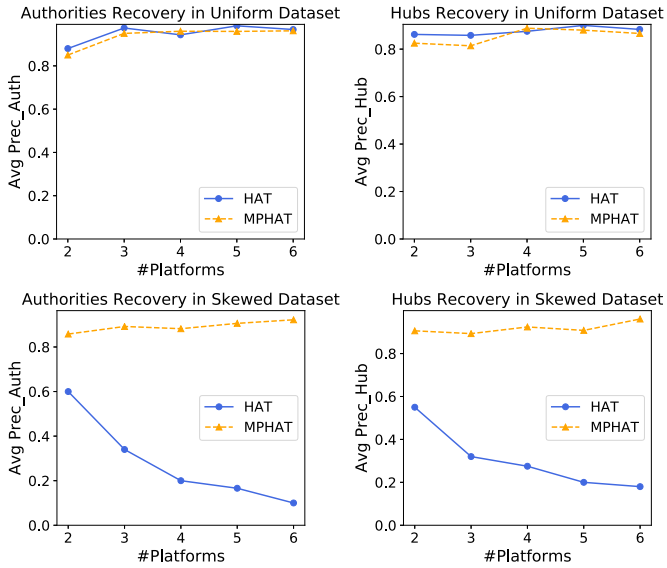


Fig. 8. $Prec_{Auth}@10\%$ and $Prec_{Hub}@10\%$ on *uniform* datasets.

is, we find the best matching between K learned topics and K ground truth topics such that the total distance of the K pairs is minimized. Here the distance between a pair of topics is the euclidean distance between their word distributions. A small total distance between matching topics is desired.

To evaluate MPHAT and HAT's accuracy in identifying topical hub and authority users, we rank users by the model computed hub and authority scores for each topic, and compare the top 10 percent users in the ranked lists with the ground truth topical hub and authority users. We measure the model's precision by $Prec_{Auth}@10\% = \frac{|\tau_p \cap \tau_g|}{|\tau_p|}$ for each topic k, where τ_p is the set of top 10 percent authorities predicted by the model and τ_g is the set of authorities in the ground truth. The precision in recovering ground truth topical hubs, $Prec_{Hub}@10\%$, is defined in a similar manner.

5.3 Performance in Ground Truth Topic Recovery

Fig. 7 shows the mean and standard deviation of euclidean distance over K groud truth topics and their matching topics learned byHAT and MPHAT on *uniform* datasets. We observed that the mean distance is consistently small across different number of platforms: the mean stays below 0.0025 for both MPHAT and HAT. This suggests that both models have learned the topics well, which is important for identifying the ground truth topic-specific hubs and authorities. Similar observations are made when applied HAT and MPHAT on the *skewed* datasets, implying the effectiveness and robustness of the models in recovering topics.

5.4 Performance in Ground Truth Hubs and Authorities Recovery

Fig. 8 shows the accuracy of MPHAT and HAT in recovering ground truth hubs and authorities - as measured by $Prec_{Auth}@10\%$ and $Prec_{Hub}@10\%$ - on *uniform* and *skewed* datasets involving 2 to 6 platforms. From the figure, we observe that both MPHAT and HAT performs well in identifying topical hub and authority users in the *uniform* dataset, while MPHAT outperforms HAT in the *skewed* dataset. The results are reasonable as HAT is designed to identify topical

hubs and authorities in a single platform setting, and is thus able to perform well in the *uniform* dataset. It however yields poor results for *skewed* dataset; as the number of platform increases, HAT's performance deteriorates further. On the other hand, MPHAT performs very well in both data settings. MPHAT learns the users' topical platform preference and is thus able to perform well in identifying the topical hubs an authorities in both synthetic datasets. We have also varied q% in {10%, 20%, . . . , 50% } and generated synthetic datasets and conducted the same experiments with each value of q. These experiments returned results that are consistently similar to ones shown in Fig. 8. These results conclude that MPHAT significantly outperforms HAT in recovering ground truth hubs and authorities.

6 Conclusion

In this paper, we have proposed a novel generative model called Multiple Platform Hub and Authority Topic (MPHAT) model, which jointly models user's topic-specific hubs, authorities, interests and platform preferences. We evaluated MPHAT using synthetic and real-world datasets and benchmarked against the state-of-the-art. Our experiments on Twitter and Instagram datasets show that our proposed MPHAT outperforms LDA and achieves comparable results as TW_LDA in topic modeling. On platform prediction, MPHAT outperforms the TW_LDA baseline method and is able to predict which platform a user would publish his or her posts with reasonable accuracy. On link recommendation, MPHAT outperforms the baseline methods in MRR by at least 10 percent. We have empirically shown that MPHAT is able to identify hub and authority users within and across Twitter and Instagram for different topics. Our experiments on synthetic datasets also show that our proposed model outperforms baseline method in identifying hub and authority users in multiple OSNs setting. For future works, we would like extend our model to include non-topical relationship among users. Currently, our model assumes that all links among users are topical although a user may follow each other for social reasons (e.g., they are friends).

Acknowledgments

This work is supported by the National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, and National Research Foundation (NRF).

References

[1] "Most famous social network sites worldwide as of August 2017, ranked by number of active users (in millions)," Statista, Hamburg, Germany, 2017. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 65–74.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 65–74.

[4] I. Kayes, X. Qian, J. Skvoretz, and A. Iamnitchi, "How influential are you: Detecting influential bloggers in a blogging community," in *Proc. Int. Conf. Social Informat.*, 2012, pp. 29–42.

[5] N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi, "Identifying high betweenness centrality nodes in large social networks," *Social Netw. Anal. Mining*, vol. 3, no. 4, pp. 899–914, 2013.

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

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[8] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 18–33.

[9] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 807–816.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 261–270.

[11] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 393–402.

[12] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018.

[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.

[14] R. K.-W. Lee, T.-A. Hoang, and E.-P. Lim, "Discovering hidden topical hubs and authorities in online social networks," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 378–386.

[15] D. Gayo-Avello, "Nepotistic relationships in twitter and their impact on rank prestige algorithms," *Inf. Process. Manag.*, vol. 49, no. 6, pp. 1250–1280, 2013.

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

[17] M. Shahriari and M. Jalili, "Ranking nodes in signed social networks," *Social Netw. Anal. Mining*, vol. 4, no. 1, 2014, Art. no. 172.

[18] A. Khrabrov and G. Cybenko, "Discovering influence in communication networks using dynamic graph analysis," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, 2010, pp. 288–294.

[19] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Sci.*, vol. 337, no. 6092, pp. 337–341, 2012.

[20] A. Silva, S. Guimarães, W. Meira Jr, and M. Zaki, "Profilerank: Finding relevant content and influential users based on information diffusion," in *Proc. 7th Workshop Social Netw. Mining Anal.*, 2013, Art. no. 2.

[21] B. Sun and V. T. Ng, "Identifying influential users by their postings in social networks," in *Ubiquitous social media analysis*. Berlin, Germany: Springer, 2013.

[22] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. Int. Conf. Web Search Data Mining*, 2008, pp. 207–218.

[23] R. Ghosh and K. Lerman, "Predicting influential users in online social networks," in *Proc. Workshop Social Netw. Mining Anal.*, 2010.

[24] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa, "Turank: Twitter user ranking based on user-tweet graph analysis," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2010, pp. 240–253.

[25] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.

[26] I. Anger and C. Kittl, "Measuring influence on twitter," in *Proc. 11th Int. Conf. Knowl. Manage. Knowl. Technol.*, 2011, Art. no. 31.

[27] L. B. Jabeur, L. Tamine, and M. Boughanem, "Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks," in *Proc. 19th Int. Conf. String Process. Inf. Retrieval*, 2012, pp. 111–117.

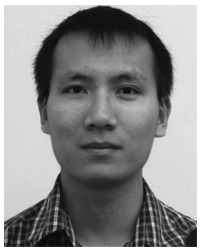
[28] X. Li, S. Cheng, W. Chen, and F. Jiang, "Novel user influence measurement based on user interaction in microblog," in *Proc. IEEE/ACM Int. Conf. Advances Social Netw. Anal. Mining*, 2013, pp. 615–619.

[29] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Netw.*, vol. 1, no. 3, pp. 215–239, 1978.

[30] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[31] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *J. Math. Sociology*, vol. 2, no. 1, pp. 113–120, 1972.
[32] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 45–54.
[33] J. Hu, Y. Fang, and A. Godavarthy, "Topical authority propagation on microblogs," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1901–1904.
[34] X. Liu, H. Shen, F. Ma, and W. Liang, "Topical influential user analysis with relationship strength estimation in twitter," in *Proc. IEEE Int. Conf. Data Mining Workshop*, 2014, pp. 1012–1019.
[35] G. Katsimpras, D. Vogiatzis, and G. Paliouras, "Determining influential users with supervised random walks," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 787–792.
[36] R. K.-W. Lee and E. P. LIM, "Measuring user influence, susceptibility and cynicalness in sentiment diffusion," in *Proc. Eur. Conf. Inf. Retrieval*, 2015, pp. 411–422.
[37] M. Montangero and M. Furini, "Trank: Ranking twitter users according to specific topics," in *Proc. Consum. Commun. Netw. Conf.*, 2015, pp. 411–422.
[38] A. Aleahmad, P. Karisani, M. Rahgozar, and F. Oroumchian, "Olfinder: Finding opinion leaders in online social networks," *J. Inf. Sci.*, vol. 42, no. 5, pp. 659–674, 2016.
[39] T.-A. Hoang and E.-P. Lim, "Microblogging content propagation modeling using topic-specific behavioral factors," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2407–2422, Sep. 2016.
[40] E. Oro, C. Pizzuti, N. Procopio, and M. Ruffolo, "Detecting topic authoritative social media users: A multilayer network approach," *Trans. Multimedia*, vol. 20, no. 5, pp. 1195–1208, 2018.
[41] M. Huang, G. Zou, B. Zhang, Y. Gan, S. Jiang, and K. Jiang, "Identifying influential individuals in microblogging networks using graph partitioning," *Expert Syst. Appl.*, vol. 102, pp. 70–82, 2018.
[42] Z. Z. Alp and Ş. G. Öğüdücü, "Identifying topical influencers on twitter based on user behavior and network topology," *Knowledge-Based Syst.*, vol. 141, pp. 211–221, 2018.
[43] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 199–208.
[44] B. Bi, Y. Tian, Y. Sismanis, A. Balmin, and J. Cho, "Scalable topic-specific influence analysis on microblogs," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 513–522.
[45] N. Barbieri, F. Bonchi, and G. Manco, "Who to follow and why: Link prediction with explanations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1266–1275.
[46] T. G. Kolda, B. W. Bader, and J. P. Kenny, "Higher-order web link analysis using multilinear algebra," in *Proc. 5th IEEE Int. Conf. Data Mining*, 2005, pp. 8.
[47] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, 2003.
[48] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retrieval*, 2011, pp. 338–349.
[49] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *SIGKDD Explorations*, vol. 18, no. 2, pp. 5–17, 2016.
[50] W. Guo, S. Wu, L. Wang, and T. Tan, "Social-relational topic model for social networks," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1731–1734.
[51] Y.-S. Cho, G. Ver Steeg, E. Ferrara, and A. Galstyan, "Latent space model for multi-modal social data," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 447–458.
[52] R. K.-W. Lee, T.-A. Hoang, and E.-P. Lim, "On analyzing user topic-specific platform preferences across multiple social media sites," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1351–1359.
[53] J. A. Bilmes, et al., "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *Int. Comput. Sci. Institute*, vol. 4, no. 510, 1998, pp. 126.
[54] N. Boyd, G. Schiebinger, and B. Recht, "The alternating descent conditional gradient method for sparse inverse problems," *J. Optimization*, vol. 27, no. 2, pp. 616–639, 2017.
[55] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociology*, vol. 27, no. 1, 2001, pp. 415–444.
[56] E. M. Voorhees and L. Buckland, "Overview of the TREC 2003 question answering track," in *Proc. Text REtrieval Conf.*, 2003, pp. 54–68.

Roy Ka-Wei Lee is a assistant professor of computer science at University of Saskatchewan. He received the PhD degree in information systems from the Singapore Management University, in 2018. He is a research scientist at the Living Analytics Research Centre. His research lies in the intersection of data mining, machine learning, and social computing. In particular, he is interested in studying user behaviors and information diffusion across multiple social networks.



Tuan-Anh Hoang received the BSc degree from the Hanoi University of Science, in 2006, and the PhD degree from Singapore Management University, in 2015. His research interests include social and information networks analysis and mining, and user profiling and personalization. He is particularly interested in modeling user behavioral factors in information propagation and social interactions.

Ee-Peng Lim is a professor of information systems at Singapore Management University. His research interests include social network and web mining, information integration, and digital libraries. He has published more than 300 refereed journal and conference papers in these areas. He is currently the faculty director of the Living Analytics Research Center which focuses on urban and social analytics. He is a member of the Singapore's Social Science Research Council and also serves as the Steering Committee Chair of Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD).