# Identifying Influencers in Thai Internet Forum based on Topic-oriented Gravity Model

Jirateep Tantisuwankul
Massive Information and Knowledge Engineering Laboratory, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
jirateep.t@ku.th

Bundit Manaskasemsak
Massive Information and Knowledge Engineering Laboratory, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
un@mikelab.net

Arnon Rungsawang
Massive Information and Knowledge Engineering Laboratory, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
arnon@mikelab.net

## ABSTRACT

The task of identifying influencers provides a lot of benefits for various practical applications such as recommendation systems, viral marketing, and information monitoring. This issue can traditionally be solved via a network structure with several proposed graph algorithms. However, most of them employ a global computation with much time-consuming; some consider only undirected and unweighted networks which may be inconsistent with the nature of data. Inspired by the law of gravity in Physics, we present the Topic-oriented Gravity Model (TopicGM) that investigates a directed and weighted network incorporating users' topical aspects. The key concept is that an individual is first represented as a textual content he created or read. Afterwards, TopicGM simply adopts a topic modeling, i.e., the Hierarchical Dirichlet Process (HDP), to classify topics over those contents. A topical network is then constructed where nodes represent individuals and an edge connects two individuals in the direction from the poster to the reader with a topical confidence weight. Finally, we apply the gravity formula to calculate influence scores and rank individuals. The experimental results, conducted on real-world data gathered from Pantip.com (famous Thai web forum), show that our approach outperforms many state-of-the-art baselines by accurately identifying influencers within the top of rankings.

## CCS CONCEPTS

• **Human-centered computing**; • **Empirical studies in collaborative and social computing**; • **Information systems**; • **Web log analysis**;

## KEYWORDS

influencer identification, topic model, gravity model, social network, viral marketing


*CSAI 2020, December 11–13, 2020, Zhuhai, China*

ACM ISBN 978-1-4503-8843-6/20/12…$15.00
https://doi.org/10.1145/3445815.3445859

**ACM Reference Format:**
Jirateep Tantisuwankul, Bundit Manaskasemsak, and Arnon Rungsawang. 2020. Identifying Influencers in Thai Internet Forum based on Topic-oriented Gravity Model. In *2020 4th International Conference on Computer Science and Artificial Intelligence (CSAI 2020), December 11–13, 2020, Zhuhai, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3445815.3445859

## 1 INTRODUCTION

Nowadays, the rapid growth of various social media applications has played an important communication role in our daily lives [1, 2]. Many social media platforms such as Facebook and Twitter allow millions of users to exchange and share their information at different levels with different influences [3]. An influencer is a person who can create a chain reaction of influence that takes a word-of-mouth approach and spreads information to reach abroad audiences [4, 5]. This kind of user has an ability to control the dissemination of both positive and negative things [6, 7] and expand the scope of information transmission [8]. In marketing, many brands seek to find potential influencers with the ability to spread their product listings and persuade other users to increase their prospects for future customers.

The task of identifying influencers is not straightforward to explore, especially in a large and complex network like real-world communities [9, 10]. Many studies have been proposed to assess how much an individual can influence others; however, most of them such as centrality approaches based on degree [11], K-shell [12], and betweenness [13, 14], consider only undirected and unweighted networks and finally give rough ranking results. Multiple individuals with different influential abilities in the same range are placed in the same rank.

Recent studies [3, 15, 16] have adopted the concept of the law of gravity in finding influencers and have proven to be more effective than traditional methods. However, the approaches are still simulated on an undirected and unweighted network which may be inconsistent with the behavior of online users on some media platforms. A user action by clicking a like or replying to comments on other people's tweet or forum posts can indicate the interest of that user and the model should represent this implicit influence arising in the direction of those posters to the user, for instance.

Moreover, topical aspects are another important issue and evidently affect the influence and adoption [17, 18]. In fact, people have their own interests and are more likely to be influenced by their friends with the similar interests. A celebrated singer is more likely to have an impact on behaviors of his/her fans rather than those of

non-fans, for instance. Therefore, in this work, we also take into account topical aspects in order to accurately simulate information dissemination and identify influencers.

To summarize, our contributions are as follows:

- We concentrate on interesting topics of online users. Here, we analyzed the content those users posted and then employed the Hierarchical Dirichlet Process (HDP) [19] to calculate the topic distribution.
- Based on the content reading activity obtained from the clickstream log, we defined a social graph corresponding to each topic. The graphs are directed and weighted, where a node represents a person and an edge expresses an implicit influence from poster to reader associated with the topic distribution.
- We proposed the TopicGM algorithm that adopts the law of gravity to assess an individual's influence score to finally get the first $k$ influencers for a specific topic.
- We conducted experiments on a real-world data from Pantip.com; the results show that our approach outperforms all state-of-the-art baselines by accurately identifying influencers within the top of rankings.

The remainder of this paper is organized as follows. Section 2 reviews background and research studies related to ours. Section 3 describes the Pantip dataset that we used in this study. Section 4 details the proposed TopicGM algorithm. Section 5 reports performance evaluation comparing with several state-of-the-art baselines. Finally, Section 6 concludes the paper.

## 2 BACKGROUND AND RELATED WORK

This section presents a literature review on the issue of social influence. First, we mention a number of traditional studies. We then briefly review the law of gravity and the use in influencer identification. Finally, related work about influence analysis based on topical aspects will be covered.

### 2.1 Traditional Influencer Identification

Several approaches have been introduced to simulate the mechanism of social influence. Most of them are explored on undirected and unweighted networks. We can classify them into 3 categories: neighborhood-based, path-based, and random walk-based methods.

First, the neighborhood-based method focuses on the number of connected neighbors of the user being considered and is based on the assumption that the greater the number, the more influential the user is. Degree centrality is the simplest method that directly sets the number of neighbors as an influence score [11]. H-index has been proposed for the domain of scientific research to rank researchers [20]. The method defines the value $HI(u)$ as the maximum integer $n$ to a researcher $u$ with the condition that there exist at least $u$'s neighbors whose degrees are not less than $n$. K-shell decomposition is another approach applied for identifying the most influential spreaders in complex networks [12]. The method starts with removing nodes with only one link and assigning them to the group 1-shell, and then recursively assigns nodes with 2 links to 2-shell. This process continues until no nodes remain, and the influencers are selected from shells with greater $k$ values.

Second, the path-based method focuses on the number of shortest paths in the network through the node being considered. That is, betweenness centrality measures how much a given node is in-between others [14], so it is used to capture the influence relative to the flow of information between others in a network. The betweenness centrality of a node $u$, $BC(u)$, is defined by $\sum_{v \neq u \neq w} \frac{\sigma_{v,w}(u)}{\sigma_{v,w}}$ where $\sigma_{v,w}$ is the total number of shortest paths from node $v$ to node $w$ and $\sigma_{v,w}(u)$ is the number of those paths that pass through $u$. Closeness centrality [21] is another way of detecting nodes that are able to spread information through a network; it measures the average farness (inverse distance) of a node to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes. Hence, the closeness centrality of a node $u$, $CC(u)$, is defined by $\frac{N-1}{\sum_{v \neq u} d_{u,v}}$ where $N$ is the number of nodes in the graph and $d_{u,v}$ is the distance between nodes $u$ and $v$.

Last, the random walk-based method focuses on measuring how accessible the rest of the network is from a given start node; this implies the ability to propagate information of that node. The well-known algorithms include HITS [22] and PageRank [23]. Both are originally introduced for ranking authoritative web pages in the Web. The former assigns hub and authority scores to each web page and relies on the assumption that a good hub represents a page that points to many other authority pages, while a good authority represents a page that is linked by many good hub pages. For the latter algorithm, PageRank considers the authority score only and is simply based on the assumption that a good authority page is linked by many other good authorities. Moreover, in epidemiology, SIR is a random walk-based model proposed to approximate the number of infected people from a contagious illness in a closed population over time [24]. The algorithm starts by leaving one of the nodes in the network, named a seed $u$, in the infected state ($\mathcal{I}$) and the others in the susceptible state ($\mathcal{S}$). The infected node can spread to susceptible neighbors with probability $\beta$, then the currently infected node is set in the recovered state ($\mathcal{R}$) with probability $\lambda$. However, nodes in $\mathcal{R}$ cannot spread or be infected again. The algorithm repeats until no more infected nodes in the network so that the score of $u$ is defined as the number of nodes existing in $\mathcal{R}$.

### 2.2 Gravity-based Approaches

In Physics, the law of gravity was introduced by Sir Issac Newton. It describes the gravitational force $F$ between two objects with mass $m_1$ and $m_2$ being inversely proportional to the square distance $d$ between them, as shown in Eq. 1). The law of gravity is applied in various sciences to describe and predict the behavior of gravitational interactions.

$$F \propto \frac{m_1 m_2}{d^2} \tag{1}$$

Many studies have also adopted this concept to identify influencers [3, 15, 16] by letting an influence score $S_{u,v}$ between users $u$ and $v$ be a force of gravity, degree centrality $DC$ of the individual be the mass, and the smallest number of hops $d_{u,v}$ between them be the distance. Then, $S_{u,v}$ is defined as:

$$S_{u,v} = \frac{DC_u DC_v}{d_{u,v}^2} \tag{2}$$

Thus, the gravity model computes $GM(u)$ by the sum of influences between $u$ and his neighbors in the network:

$$GM(u) = \sum_{u \neq v} s_{u,v} \tag{3}$$

In addition, Ma et al. [16] presented *LG*, an extension of *GM*, by using the k-shell centrality *KS* instead of degree centrality. The authors also limited the distance between the considered user and his neighbors only along the path within the truncation radius $r$, in order to decrease the complexity of iteration runs, defined in Eq. 4). Similarly, Li et al. [15] proposed the local gravity model, *LGM*, using the degree centrality with the limited distance, defined in Eq. 5). Niu et al. [3] proposed logarithmic gravity model, *LogGM*, by assigning the logarithm of the smallest number of hops between two users to the limited distance instead, defined in Eq. 6).

$$LG(u) = \sum_{d_{u,v} < r, u \neq v} \frac{KS_u KS_v}{d_{u,v}^2}, \tag{4}$$

$$LGM(u) = \sum_{d_{u,v} < r, u \neq v} \frac{DC_u DC_v}{d_{u,v}^2}, \tag{5}$$

$$LogGM(u) = \sum_{d_{u,v} < r, u \neq v} \frac{DC_u DC_v}{\ln\left(d_{u,v} + e - 1\right)} \tag{6}$$

As can be seen, all approaches define an undirected network. The influence impact between two individuals is thus symmetrical, but it is not the case in the real world.

### 2.3 Topic-Constrained Influence Analysis

Most studies on influence analysis tend to determine user relationships based on friendships or followerships only. However, in reality, media content has a great effect on online user behavior. People usually read, click, like, or even adopt it if they are interested. A retweet may imply that Twitter users agree with the poster and wish their followers to read it; commentators respond to a forum owner if they wish to collaborate with that forum topic, for instance. Therefore, topical aspects are important factors that should be taken into account. Some studies have taken this concept and proven that topical aspects significantly enhance the performance of influencer identification [17, 18].

## 3 DATASET

Internet Forum (a.k.a., webboard) is a website where people being interested in the same topic can exchange and discuss their idea. It allows users to create threads for both announcing news or events and asking and answering doubts. An Internet forum is a tree-like structure, i.e., forums can contain subforums. Each subforum contains different topics or discussion areas, called threads. A thread can be responsive by other users who wish to engage with content. To keep in peace, some Internet forum allows only members to post messages.

Pantip.com, a popular Thai-language website and discussion forum, was founded in 1997 by Wanchat Padungrat, a former electronics engineer. "Pantip" in Thai means a thousand (Pan) of tricks (tip). Pantip has a wide variety of topics that attract a large number of users. In particular, people who are interested in on-going trends or current social events can easily get up-to-date information from here. The forums are separated into 38 different main topics and categorized with more than 15,000 tags. In addition, there are special corners for interesting threads: Pantip Now, Pantip Pick, and Pantip Trend.

Thanks to Pantip's engineering team for collaborating in this study, we got the clickstream data gathered from January 1 until June 30, 2019. Within this log, we focus only on behavior when users click to read threads so that we have 2.1 billion raw transactions in total, or an average of 11.5 million reading clicks per day. Each transaction is recorded in a JSON format consisting of several fields: thread_id, comment_id, and reply_id, representing a specific thread, a thread comment, and a response to comment, respectively. Since Pantip's platform allows both members and non-members to read content but only members can post the content, the member_id field is therefore used to identify those registered users and the trans_cookies field is used to distinguish between the rest of the users.

In addition to the transaction log to answer our questions about who read on which thread (or comment or response to comment), we also would like to know who created it on what topic and how it will be interesting to readers, in order to identify influencers. To achieve this requirement, we retrieve thread description which Pantip has organized into a hierarchical structure. The top level contains thread data such as id, created_time, title, content, owner_id, and voting_score. The second and last levels correspond to the thread's comments and responses to that comment (if any); each of which has the same data fields except the title.

Finally, we preprocess all the data by mapping the clickstream log and the thread description to produce reading paths. A path is represented as a quintuple attribute: reader_id (determined by either member_id or trans_cookies), thread_id, comment_id, reply_id, and owner_id (determined by member_id). That is, who reads which thread, which comment, and which response to comment of whom. Moreover, we simply assume that all readers will be influenced once by the same thread owner so that the duplicate path is the same. We eventually got 1.8 billion records left for the experiments. Last but not least, the textual content and votes for the threads will be analyzed later, as we will discuss in the next section.

## 4 TOPIC-ORIENTED GRAVITY MODEL

As mentioned, most of the previous studies represent user behavior and relationships as an undirected and unweighted network. However, in real scenarios, when people are interested in someone's posted content, they tend to take certain actions such as liking, requesting a friend, or even following him/her. These actions indicate an influence impact that is passed from the content owner to the reader, which can be measured in weight by the number of those actions or by estimating how interested both persons are on the same topic. Here, we focus on the action of reading since it happens in most cases more than others.

In this section, we present our Topic-oriented Gravity Model (TopicGM) which considers both the direction of influence and the topic-based reading activity as weight. In the followings, we first explain how textual content can be categorized into different topics. We will then detail our influence network construction and influence score computation.

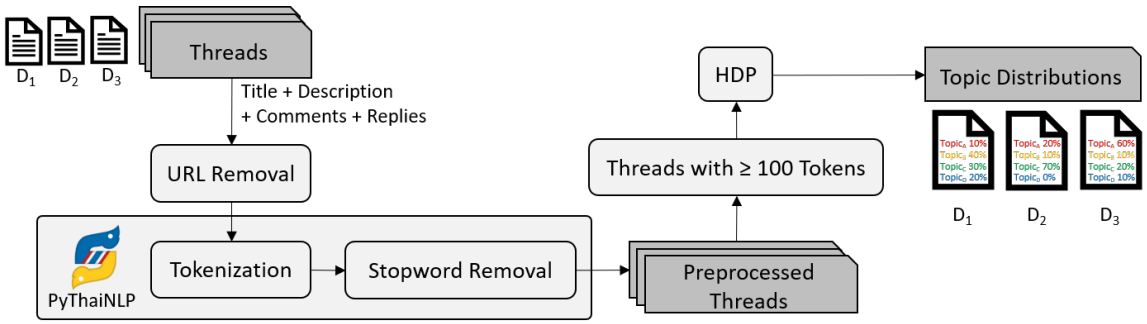


**Figure 1: Procedures for modeling thread topics.**

## 4.1 Topic Model

Our first effort is to estimate how interested an individual user is over topics based on textual content he/she associates with. Figure 1 shows an overview of steps in our topic model. First, we represented a thread entity as a concatenation of title, content, and all helpful comments and responses with at least two votes. Embedded URLs were clearly removed. Then, we used the open-source PyThaiNLP library [25] to tokenize Thai text and remove all stopwords. Since texts that are too short are usually difficult to identify the correct topic, we thus ignored thread entities with fewer than 100 words, so we had 668,345 thread entities. Finally, we exploited the existing topic modeling algorithm, i.e., HDP [19], to compute topic distributions. The algorithm finds the appropriate number of topics and assigns each thread entity into the topics. In our experiments, HDP reported 11 meaningful topics along with each topic's probability for each thread entity. This value indicates how much a thread entity is related to a particular topic; however, we will assume it is not related if the value is less than 0.3.

## 4.2 Influence Network Construction

We hypothesize that people will not be influenced if the topic of the content they read does not match their interests. Thus, we can consider the activities (i.e., reading transactions) of users to construct separate influence networks based on the topic they are involved in so that we can identify influencers for a particular topic.

Given a specified topic $t$, the influence network is defined as a directed and weighted graph $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t, \mathcal{W}^t)$ consisting of sets of individuals, influence directions, and influence weights, respectively. Hence, an influence effect $e^t_{u,v} \in \mathcal{E}^t$ directed from the thread owner $u \in \mathcal{V}^t$ to the reader $v \in \mathcal{V}^t$ will be assigned with a weight $\omega^t_{u,v} \in \mathcal{W}^t$, defined as:

$$\omega^t_{u,v} = \sum_{\tau_{u,v} \in T} \mathcal{P}(t|\tau_{u,v}), \tag{7}$$

where $T$ is a data collection containing all reading transactions, $\tau_{u,v}$ denotes a $v$'s reading activity on the $u$'s thread content, and $\mathcal{P}(t|\tau_{u,v})$ is the probability that the $u$'s content involved by $v$ is related to the topic t.

## 4.3 Influence Score Computation

Our main objective is to identify influencers for a specific topic by applying the law of gravity in analysis. The original concept defines a gravitational force as symmetry; however, this is not the case for the effects of influence. In this study, we hypothesize that individuals who are defined as influencers depend on three factors: the ability to spread information, other readers' ability to adopt it, and the distance to any reader.

Similar to degree centrality, the ability to spread information of a content owner, in our case, is determined by the number of his readers. The difference is that we take into account the topical aspects for the content they read. Hence, given a specified topic $t$, we let a weighted out-degree centrality $WDC^t_{out(u)}$ be the owner $u$'s ability to spread his information out, defined as:

$$WDC^t_{out(u)} = \sum_{\forall v: e_{u,v} \in \mathcal{E}^t} \omega^t_{u,v} \tag{8}$$

Likewise, the ability of a reader is determined by the number of owners of the content he interacts with. Let $WDC^t_{in(v)}$ be the reader $v$'s ability to receive information, defined as:

$$WDC^t_{in(v)} = \sum_{\forall u: e_{u,v} \in \mathcal{E}^t} \omega^t_{u,v} \tag{9}$$

For the distance factor, we defined it as how far away any reader is from the content owner. In our case, the distance $d^t_{u,v}$ is the number of hops along the shortest path from the source content owner $u$ to the reader $v$ in $\mathcal{G}^t$.

Finally, our Topic-oriented Gravity Model presents a new formula to calculate an influence score for a given content owner $u$ based on the topic $t$, defined as:

$$TopicGM(u;t) = \sum_{d^t_{u,v} < r, u \neq v} \frac{WDC^t_{out(u)} \times WDC^t_{in(v)}}{\left(d^t_{u,v}\right)^2} \tag{10}$$

Inspired by [15, 16], the authors recommended limiting the distance within three hops as this will be less effective for influence if the reader is too far away from the content owner. We therefore set $r = 3$ for all experiments. After influence scores of all users $u \in \mathcal{V}^t$ are calculated, we sort them from highest to lowest to get the ranked list $\mathcal{L}^t$ of influencers.

# 5 PERFORMANCE EVALUATION

We conducted the experiments on Pantip.com data (i.e., clickstream log and thread content description) which were available on January 1 until June 30, 2019 as described in Section 3. In the followings, we

**Table 1: The details of 11 selected topics provided by the HDP algorithm.**

| Topic | #nodes (users) | #edges | Edges per node | #threads | Description |
|---|---|---|---|---|---|
| 1 | 581,784 | 1,315,349 | 2.26 | 20,559 | AOA (K-POP band) |
| 2 | 141,657 | 158,021 | 1.12 | 205 | News broadcaster |
| 3 | 33,591 | 35,439 | 1.06 | 894 | Animation recommendation |
| 4 | 14,350 | 14,878 | 1.04 | 415 | Following famous people |
| 5 | 1,041,014 | 2,223,079 | 2.14 | 23,334 | Traveling trip reviews |
| 6 | 161,493 | 212,125 | 1.31 | 3,879 | English text including in threads |
| 7 | 3,525,431 | 10,517,704 | 2.98 | 61,728 | Trivial conversation |
| 8 | 1,881,698 | 4,463,281 | 2.37 | 45,782 | Horoscope and entertainment |
| 9 | 1,901,192 | 4,166,654 | 2.19 | 20,771 | Love and tips |
| 10 | 1,012,453 | 180,672 | 1.78 | 9,749 | Money and investment |
| 11 | 1,353,633 | 2,637,235 | 1.95 | 15,172 | Love, politics, and religion |

will mention to the evaluation metrics and the compared baselines, and then report the performance comparisons.

## 5.1 Evaluation Metrics

We report performance in terms of measuring ranking accuracy. First, the *precision@k* metric captures the correctness in identifying influencers within the top of ranking. It is defined as the fraction of the first $k$ ranked individuals that are truly influencers. Second, the *NDCG@k* metric indicates the quality of ranking. It returns higher value if potential influencers are ranked closer to the top. Formally, $NDCG@k = \frac{DCG@k}{IDCG@k}$ for $DCG@k = \sum_{i=1}^{k} \frac{2^{l_i}-1}{\log_2(i+1)}$, where $l_i$ is the true label of item at rank $i$ (i.e., 1 if it is an influencer, otherwise 0) and $IDCG@k$ is the $DCG$ for ideal ranking where all $l_i$'s are 1.

## 5.2 Compared Baselines

We employed 9 state-of-the-art methods: gravity model (GM) [15], degree centrality (DC) [11], weighted k-shell centrality (WKS) [26], betweenness centrality (BC) [14], closeness centrality (CC) [21], H-index (HI) [20], HITS [22], PageRank (PR) [23], and SIR [24], to achieve the problem of identifying influencers. However, some of these original versions were proposed to operate on undirected and/or unweighted network. To deal with fair comparisons, we therefore optimize them for our directed and weighted network, detailed as follows.

We implement an extended version of DC, named weighted degree centrality (WDC), with a weighted edge. Let $\psi_{u,v}$ be the number of $u$'s threads that $v$ has read. Then, the $u$'s centrality score, $WDC(u)$, is defined by $\sum_{\forall v: e_{u,v} \in \mathcal{E}} \psi_{u,v}$. Similarly, we also present WBC and WCC, extensions of BC and CC, by defining a shortest path $P_{v_1 \to v_n} = v_1, v_2, \ldots, v_n \in \mathcal{V}^n$ with the distance $d_{v_1,v_n} = \sum_{i=1}^{n-1} (\psi_{v_i,v_{i+1}})^{-1}$. For HITS and PageRank, since both estimate authorities based on in-coming references by others but contrary to the influence that considers out-going relationships with others, we thus defined a score for $u$ in the reverse-edge manner as $\frac{\psi_{u,v}}{\sum_{\forall w: e_{w,v} \in \mathcal{E}} \psi_{w,v}}$. In SIR, we have modified an edge weight as $\frac{\psi_{u,v}}{\sum_{\forall w: e_{u,w} \in \mathcal{E}} \psi_{u,w}}$. Moreover, we have included topical aspects in all baselines. Each baseline works separately on a network $\mathcal{G}^t$ according to each topic $t$ as detailed in Section 4.1.

## 5.3 Experimental Results

We first report the results of our topic model. As mentioned, we used the HDP algorithm to categorize a set of Pantip threads. After some investigation, we selected clusters with 100 to 100,000 threads since clusters that are too small and too large would give too specific and too broad topics, respectively. Table 1 provides details of the selected 11 topics.

In the followings, we will report a comparison of influencers ranked by different algorithms. The major challenge is that we do not have the goal standard. Thus, we recommend using the Borda method [27] to vote answers for all algorithms. The Borda count is a family of single-winner election methods in which voters rank candidates in order of preference. In our case, a candidate (i.e., influencer) $c$ will get $1/\rho_{ij}$ points as he was ranked at the $i^{th}$ position of the ranked list $\mathcal{L}_j$ introduced by the $j^{th}$ algorithm. Hence, the $c$'s total score is the sum of points received by all the algorithms. With these scores of all candidates, we will sort them from the highest to the lowest to get the influencer ranking standard.

Tables 2 and 3 report the average *precision@k* and *NDCG@k* scores over 11 topics of 18 influencer ranking algorithms. As can be seen, the algorithms applying the concept of gravitation (i.e., TopicGM and GM) can identify more potential influencers at the top of the rankings. Moreover, the algorithms that incorporate the topical aspects give higher scores than they do not in most cases. Last, the TopicGM algorithm can more accurately identify the top 100 influencers rather than all the baselines.

After thoroughly examining the top 100 influencers ranked by TopicGM, we found some potential influencers that cannot be identified by traditional GM, indicating that the topical aspects have indeed a significant impact.

# 6 CONCLUSION

This work presents the TopicGM algorithm, aiming to identify potential influencers in the online media platform. The algorithm has applied the concept of the gravitation and incorporated topical aspects in the analysis. Experiments were conducted on Thai Internet forum (i.e., Pantip.com) dataset. The results have proven that

**Table 2: Average *precision@k* of different influencer ranking algorithms.**

| | Topic-oriented | | | | | | | | | Non Topic-oriented Algorithms | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **k** | **Topic GM** | **Topic WDC** | **Topic WKS** | **Topic WBC** | **Topic WCC** | **Topic HI** | **Topic HITS** | **Topic PR** | **Topic SIR** | **GM** | **WDC** | **WKS** | **WBC** | **WCC** | **HI** | **HITS** | **PR** | **SIR** |
| 10 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.97 | **1.00** | 0.94 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.97 | 0.97 | 1.00 | 0.94 |
| 20 | 0.99 | 0.89 | 0.72 | 0.66 | 0.84 | 0.52 | 0.43 | 0.96 | 0.37 | **1.00** | 0.84 | 0.42 | 0.83 | 0.84 | 0.38 | 0.38 | 0.93 | 0.35 |
| 30 | **0.90** | 0.78 | 0.53 | 0.54 | 0.74 | 0.37 | 0.28 | 0.88 | 0.23 | 0.73 | 0.70 | 0.26 | 0.70 | 0.70 | 0.23 | 0.23 | 0.77 | 0.22 |
| 40 | **0.78** | 0.68 | 0.44 | 0.46 | 0.66 | 0.31 | 0.20 | 0.74 | 0.17 | 0.60 | 0.52 | 0.19 | 0.52 | 0.52 | 0.17 | 0.17 | 0.58 | 0.16 |
| 50 | **0.67** | 0.59 | 0.37 | 0.40 | 0.58 | 0.27 | 0.17 | 0.65 | 0.14 | 0.49 | 0.45 | 0.16 | 0.44 | 0.45 | 0.13 | 0.13 | 0.50 | 0.12 |
| 60 | **0.60** | 0.52 | 0.32 | 0.35 | 0.50 | 0.25 | 0.14 | 0.58 | 0.11 | 0.42 | 0.39 | 0.13 | 0.38 | 0.39 | 0.11 | 0.11 | 0.42 | 0.10 |
| 70 | **0.53** | 0.46 | 0.29 | 0.31 | 0.45 | 0.22 | 0.12 | 0.51 | 0.10 | 0.37 | 0.33 | 0.11 | 0.33 | 0.33 | 0.09 | 0.09 | 0.36 | 0.09 |
| 80 | **0.47** | 0.42 | 0.27 | 0.28 | 0.41 | 0.20 | 0.10 | 0.45 | 0.08 | 0.33 | 0.29 | 0.10 | 0.29 | 0.29 | 0.08 | 0.08 | 0.32 | 0.07 |
| 90 | **0.43** | 0.38 | 0.25 | 0.25 | 0.38 | 0.18 | 0.09 | 0.41 | 0.07 | 0.30 | 0.26 | 0.09 | 0.26 | 0.26 | 0.07 | 0.07 | 0.29 | 0.07 |
| 100 | **0.39** | 0.34 | 0.24 | 0.23 | 0.35 | 0.17 | 0.08 | 0.38 | 0.07 | 0.27 | 0.24 | 0.09 | 0.24 | 0.24 | 0.06 | 0.07 | 0.27 | 0.06 |

**Table 3: Average *NDCG@k* of different influencer ranking algorithms.**

| | Topic-oriented Algorithms | | | | | | | | | Non Topic-oriented Algorithms | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **k** | **Topic GM** | **Topic WDC** | **Topic WKS** | **Topic WBC** | **Topic WCC** | **Topic HI** | **Topic HITS** | **Topic PR** | **Topic SIR** | **GM** | **WDC** | **WKS** | **WBC** | **WCC** | **HI** | **HITS** | **PR** | **SIR** |
| 10 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.98 | **1.00** | 0.97 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.98 | 0.98 | **1.00** | 0.97 |
| 20 | **1.00** | 0.93 | 0.82 | 0.77 | 0.90 | 0.67 | 0.61 | 0.98 | 0.57 | **1.00** | 0.90 | 0.60 | 0.90 | 0.89 | 0.57 | 0.58 | 0.96 | 0.56 |
| 30 | **0.94** | 0.85 | 0.67 | 0.67 | 0.82 | 0.54 | 0.47 | 0.92 | 0.43 | 0.83 | 0.79 | 0.46 | 0.79 | 0.79 | 0.43 | 0.44 | 0.85 | 0.42 |
| 40 | **0.85** | 0.77 | 0.59 | 0.60 | 0.75 | 0.47 | 0.38 | 0.82 | 0.35 | 0.73 | 0.66 | 0.38 | 0.66 | 0.66 | 0.36 | 0.36 | 0.71 | 0.34 |
| 50 | **0.77** | 0.70 | 0.52 | 0.54 | 0.69 | 0.42 | 0.34 | 0.75 | 0.30 | 0.64 | 0.59 | 0.33 | 0.59 | 0.59 | 0.30 | 0.31 | 0.64 | 0.29 |
| 60 | **0.71** | 0.64 | 0.47 | 0.48 | 0.62 | 0.39 | 0.30 | 0.69 | 0.27 | 0.57 | 0.54 | 0.29 | 0.53 | 0.53 | 0.27 | 0.27 | 0.57 | 0.26 |
| 70 | **0.65** | 0.59 | 0.43 | 0.45 | 0.58 | 0.36 | 0.27 | 0.63 | 0.24 | 0.52 | 0.48 | 0.26 | 0.48 | 0.48 | 0.24 | 0.24 | 0.52 | 0.23 |
| 80 | **0.60** | 0.55 | 0.40 | 0.41 | 0.54 | 0.33 | 0.24 | 0.58 | 0.22 | 0.48 | 0.44 | 0.24 | 0.44 | 0.44 | 0.22 | 0.22 | 0.47 | 0.21 |
| 90 | **0.56** | 0.51 | 0.38 | 0.38 | 0.51 | 0.30 | 0.23 | 0.54 | 0.20 | 0.45 | 0.41 | 0.22 | 0.41 | 0.41 | 0.20 | 0.20 | 0.44 | 0.19 |
| 100 | **0.53** | 0.48 | 0.36 | 0.36 | 0.48 | 0.29 | 0.21 | 0.51 | 0.19 | 0.42 | 0.39 | 0.21 | 0.38 | 0.38 | 0.19 | 0.19 | 0.41 | 0.18 |

our efforts outperform several state-of-the-art baselines. In additional, the topical aspects have a significant impact on identifying influencers.

For the future work, we plan to extend more experiments on other social media platforms such as Facebook and Twitter. TopicGM could be further enhanced by other techniques for the definition of mass and distance in the gravity model. The topic modeling is one of the important parts in TopicGM that can be improved. Moreover, it is possible to include other types of interactions such as emotion clicking, liking, or sharing, in the study.

## REFERENCES


[1] Frank Bauer and Joseph T. Lizier. 2012. Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach. EPL 99, 6 (October 2012), 68007.

[2] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. In Proceedings of the 28th International Conference on International Conference on Machine Learning. 561–568.

[3] Jianwei Niu, Haifeng Yang, and Lei Wang. 2017. Logarithmic gravity centrality for identifying influential spreaders in dynamic large-scale social networks. In Proceedings of the 2017 IEEE International Conference on Communications. 1–6.

[4] Zeynep Z. Alp and Şule G. Öğüdücü. 2019. Influence factorization for identifying authorities in Twitter. Knowl.-Based Syst. 163 (January 2019), 944–954.

[5] Naohiro Matsumura, Hikaru Yamamoto, and Daisuke Tomozawa. 2008. Finding influencers and consumer insights in the blogosphere. In Proceedings of the 2nd International Conference on Weblogs and Social Media. 76–83.

[6] Javier Borge-Holthoefer and Yamir Moreno. 2012. Absence of influential spreaders in rumor dynamics. Phys. Rev. E 85 (February 2012), 026116.

[7] Andrea Clementi, Angelo Monti, Francesco Pasquale, and Riccardo Silvestri. 2011. Information spreading in stationary Markovian evolving graphs. IEEE Trans. Parallel Distrib. Syst. 22, 9 (September 2011), 1425–1432.

[8] Seth A. Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 33–41.

[9] Duan-Bing Chen, Rui Xiao, An Zeng, and Yi-Cheng Zhang. 2014. Path diversity improves the identification of influential spreaders. EPL 104, 6 (January 2014), 68006.

[10] Muhammad U. Ilyas and Hayder Radh. 2011. Identifying influential nodes in online social networks using principal component centrality. In Proceedings of the 2011 IEEE International Conference on Communications. 1–5.

[11] Phillip Bonacich. 1972. Factoring and weighting approaches to status scores and clique identification. J. Math Sociol. 2, 1, 113–120.

[12] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. 2010. Identification of influential spreaders in complex networks. Nat. Phys. 6 (August 2010), 888–893.

[13] Ulrik Brandes. 2008. On variants of shortest-path betweenness centrality and their generic computation. Soc. Networks 30, 2 (May 2008), 136–145.

[14] Linton C. Freeman. 1977. A set of measures of centrality based on betweenness. Sociometry 40, 1 (Mar 1977), 35–41.

[15] Zhe Li, Tao Ren, Xiaoqi Ma, Simiao Liu, Yixin Zhang, and Tao Zhou. 2019. Identifying influential spreaders by gravity model. Sci. Rep. 9 (June 2019), 8387.

[16] Ling-ling Ma, Chuang Ma, Hai-Feng Zhang, and Bing-Hong Wang. 2016. Identifying influential spreaders in complex networks based on gravity formula. Physica A 451 (June 2016), 205–212.

[17] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Topic-aware social influence propagation models. Knowl. Inf. Syst. 37, 3 (April 2013), 555–584.

[18] Bundit Manaskasemsak, Rattana Phuangpanya, and Arnon Rungsawang. 2017. Topic-constrained influence maximization in social networks. In Proceedings of the 3rd International Conference on Communication and Information Processing. 405–410.
[19] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing clusters among related groups: Hierarchical Dirichlet processes. In Proceedings of the 17th International Conference on Neural Information Processing Systems. 1385–1392.
[20] Linyuan Lü, Tao Zhou, Qian-Ming Zhang, and H. Eugene Stanley. 2016. The H-index of a network node and its relation to degree and coreness. Nat. Commun. 7 (January 2016), 10168.
[21] Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. Soc. Networks 1, 3 (July 2002), 215–239.
[22] Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604–632.
[23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.
[24] Herbert W. Hethcote. 2000. The mathematics of infectious diseases. SIAM Rev. 42, 4 (December 2000), 599–653.
[25] PyThaiNLP: Thai Natural Language Processing in Python. Retrieved September 17, 2020 from http://doi.org/10.5281/zenodo.3519354
[26] Antonios Garas, Frank Schweitzer, and Shlomo Havlin. 2012. A k-shell decomposition method for weighted networks. New J. Phys. 14, 2 (August 2012), 353–358.
[27] Peter Emerson. 2012. The original Borda count and partial voting. Soc. Choice Welfare 40, 2 (October 2011), 353–358.