A Social Sensing Model for Event Detection and User Influence Discovering in Social Media Data Streams

Lei-lei Shi, Lu Liu, Member, IEEE,
Yan Wu, Liang Jiang, John Panneerselvam, Roy Crole

Abstract—Online Social Networks (OSNs) have emerged as a major platform for sharing information through social relationships and are one of the major sources of big data. Social networks can even accommodate sharing of live streaming data among the connected users. However, social information on social networks is often locally exploited rather than capturing the changes in the entire network over time. Obtaining user's influence statistics is limited only in their local vicinity, which may not facilitate capturing the changes in the user and post influences across the entire network, thereby resulting in lower accuracy whilst measuring user's topical influence. Moreover, low-influence users always exist in the network publishing low-quality posts. With the objectives of accurately capturing highly influential users and posts, this paper proposes a novel dynamic social sensing model named DPRank (Dynamic PageRank) model to evaluate the dynamic topical influence of the users of social information on social networks during the social information evolution. We deploy our proposed model to real-world Twitter datasets, which demonstrates the effectiveness of our proposed model against notable existing methods whilst identifying the true influence of users and posts in a dynamically evolving social network.

Index Terms- Social Sensing; Dynamic Post influence; Dynamic User influence; PageRank

I. INTRODUCTION

The emergence of online social network (OSN) services such as Facebook, Twitter, Instagram and Google+ etc., is exerting positive impacts on data-centric applications by the way of exploiting the social elements of OSNs [1-4]. Most of today's data-centric applications utilize various kinds of sensors to continuously collect massive amounts of data. Moreover, social media big data shared via social networks has become an indispensable part of our lives [5-7] as OSNs are a rich source of shared messages between users [43-48]. Thus, information propagation has become a fundamental component of OSNs with their increasing popularity [8, 49-51]. In general, businesses use information propagation in social networks to sell their goods to the entire social network. In fact, comments and opinions of products from the OSN users often have an influential impact on product sales. Generally, user interests propagate through the network, such that users tend to follow other peer users with common or similar interest [52-54]. It is common that a given user's influence on a similar user tends to change the opinion and behaviours of the latter [9-11, 55].

- *Lei-lei Shi, Yan Wu, and Liang Jiang are with the School of Computer Science and Telecommunication Engineering, Jiangsu University, China.*
- *Lu Liu and Roy Crole are with the School of Informatics, University of Leicester, UK. Corresponding author: Lu Liu, Email: l.liu@leicester.ac.uk*
- *John Panneerselvam is with the School of Electronics, Computing, and Mathematics, University of Derby, UK.*

Identification of influential users in social networks has been the focus of recent researches in the OSN domain [8-14]. A ranking algorithm called PageRank has been heavily deployed in this context [15, 36, 41]. However, these studies only focus on building social networks according to the frequency and number of users' posts being published, replied or retweeted without the publication time of the posts into account, which is an important characteristic whilst analysing the influence of posts. It is worthy of note that posts published, replied or retweeted earlier characterize better probabilities of being retweeted or replied to some degree. Moreover, users with lower influence always exist in OSNs who tend to publish many low-quality posts for attracting people's attention and guiding the wrong facet in social networks. Selecting high-influence users from a collection of users can largely improve the efficiency and accuracy of user influence-based computing services.

With the intention of achieving the aforementioned objective, this paper proposes a dynamic social sensing model, named DPRank, for analysing social information on social networks to identify the high-influence users and further to characterize the real dynamic influence of the users. Specifically, first, an influential user automatic filtering-based HITS method is used to create a high-quality training dataset. Based on this, a new post influence detection model, named Dynamic Network Structure (DNS) model, is created to comprehensively identify the real dynamic influence of posts. The proposed model can efficiently explore the real influence of a post on different topics. Then a method of calculating the topic influence contribution of each user who publishes high-quality posts in social networks is presented by exploiting the user influences and the influence of different topics in social networks. Finally, an improved PageRank algorithm is applied to extract the dynamic topical influence of users in social networks during event diffusion and evolution.

The main contributions of this paper are listed as follows:

1) A high influence user automatic filtering method [43] is introduced to create a small high-quality training data set by selecting high-influence users from a collection of users [28]. Then, an improved method is designed to accurately identify the real influence of posts over time. Specifically, we start from specific posts in the network topology and postulate that the connection property and surrounding influence of posts should not be neglected, which has been the case with most of the traditional methods [8-15]. Based on this, a new post dynamic influence detection method named Dynamic Network Structure (DNS) [16] model is introduced to comprehensively assess the dynamic influence of posts. With the proposed DNS model, which considers local posts network structure and a time-varying factor of each post, the real influence of posts on different topics in social networks can be effectively identified during event diffusion and evolution.
2) This paper introduces an improved PageRank algorithm-based model [17] to discover the topic influence the contribution of each user under different topics who publishes or replies or retweets posts in social networks. Specifically, the number of publishes or replies or retweets is used to calculate the effective distance between two users and their corresponding influences on each other. Then the degree of influence between users is utilized to form a new dynamic network of users, based on both the quality of the users and their corresponding influence in the community of different interests. Finally, the newly formed user networks are exploited to extract a dynamic ranking of the real importance of users under various topics based on the PageRank algorithm.
3) Real world Twitter datasets are used in the experimentation to validate the effectiveness of our proposed model, against the existing state-of-the-art methods of identifying the influential users dynamically under different topics during event diffusion and evolution.

The remainder of this paper is structured as follows: Section II presents a review of related works of user influence discovery in social networks. Section III introduces our dynamic user topic influence extraction model. Section IV discusses the experiments and results and Section V concludes this paper along with outlining our future research directions.

II. RELATED WORK

Owing to the recent developments and increasing popularity of social networks, research on identifying the user influence has gained considerable attention. Identifying user influence mainly depends on the diffusion ability, topic popularity, individual characteristics and network structure [37-38].

Lee et al. [12] simulated the information dissemination in a concerned network based on twitter datasets to obtain the influence of each user by calculating the number of effective customers with different topics. Besides Bakshy et al. [13] constructed the spread of URL cascade tree according to the links existing in the Twitter dataset by spreading the topics among seed users to measure the influence of each seed user. Aggarwal et al. [14] proposed a stochastic information flow model in order to find representative authority users on Twitter based on diffusion ability with different topics. However, these researchers have not given enough emphasis to users' individual characteristics, interest community influence and time factors. Such factors aid effective capturing of users' community influence and dynamic influence in microblogging network during event diffusion and evolution. Meanwhile, users frequently publish or reply to posts whose potential influence should include these dynamic factors.

Measuring user influence heavily depends on individual characteristics and community network structure. For example, Liu et al. [18] proposed a generative graph model to detect the influence of users with various interests in heterogeneous networks. Besides Cha et al. [19] exploited the number of followers of users for discovering a given user's influence in the Twitter networks. Pal et al. [20] considered the number of individual posts, replies, the number of forwards, followers, respectively to calculate the forwarding influence and the spread influence of users on Twitter. Moreover, a few other methods rank users through analyzing the user behaviours [21-22, 39, 41], including copying, replying, retweeting and so forth. A few other existing methods [23-24] pay attention to other characteristics like tweet quality and utilized different centrality measures such as Eigen-vector centrality. However, not all such methods have considered the factors of posts publish time and topic popularity into account, which can provide inferences on users' dynamic influence on various topics in microblogging networks during event diffusion and evolution.

In an attempt of addressing such shortcomings, Chai et al. transformed attribute measures into four categories such as activity, reputation, quality and centrality [9], thereby resolving the issues of topics popularity. Moreover, Wang et al. postulated that the influential users in each hot community should be detected for propagating information across the whole social network [25], which can largely improve the accuracy of influence identification. Barbieri et al. postulated that cascading is a local phenomenon [26]. Besides, Xiao et al. transformed the attribute measures into the three different classes, in order to discover influential users in networks [27]. With the objectives of overcoming the aforementioned drawbacks in the existing works of detecting user influences in social networks, this paper considers the interested community factor into account along with the changing dynamism of both users' interest and event topics over time whilst enabling information propagation. To this end, our proposed model finds the dynamic influence of influential users during event evolution, in particular, our proposed model targets the global users as opposed to local users. Our proposed model stands out from the existing methods, as our proposed model discovers the dynamic influence of the influential users with popular topics in dynamic microblogging network during event diffusion and evolution.

III. A DPRANK MEASURE

The proposed DPRank model consists of three main components. An influential user filtering-based HITS (Hyperlink-Induced Topic Search) method creates a small high-quality training dataset by selecting high-influence users from a collection of users. A DNS model is used to discover the real dynamic influence of posts,

which forms the basis for discovering the real dynamic influence of users. An improved PageRank method is used for evaluating the dynamic influence of users with various event topics over time.

A. Extracting High-influence Users

In the original HITS algorithm [28], a link is used to represent the hyperlinks between Web pages. However, in our proposed method, the link represents an operational relationship between a user and a post such as publishing or retweeting or replying. Further, the improved HITS algorithm creates a directed link from users to their posts, with regards to the corresponding individual user's operations [1-2, 43-44].

B. Dynamic Network Structure model

Most often, the global influence of posts may not be applicable, but propagating through [16] the complex networks can help us choose a high-quality post from a network of neighbour posts along with obtaining its neighbour posts' information.

1) Local Structure Model (LSM)

The proposed model images a microblogging network with m posts and utilizes an $m*m$ matrix to represent the link relationship of posts. Then, we introduce the measure $p_{ij,}$ which implies that post i characterizes a p_{ij} percentage of spreading hot events information to post j. $p_{ij} = 0,$ implies that post i and post j have no linking relationship.

$$P=\begin{bmatrix} p_{11}, p_{12}, \ldots, p_{1m} \\ p_{21}, p_{32}, \ldots, p_{3m} \\ \ldots\ldots\ldots\ldots \\ p_{m1}, p_{m2}, \ldots, p_{mm} \end{bmatrix} \quad (1)$$

In order to obtain the spreading probability p_{ij} with post i and post j, we apply the Independent Cascade Model (IC) [32] which is the most widely used information propagation method and it is often used to simulate information propagation in social networks. Specifically, in an IC model, an inactive neighbour node can be activated by linking it with all the active nodes in accordance with the activation probability between them. The higher the activation probability, the easier it is for the nodes to be activated. The information dissemination process of the IC model is shown in Figure 1.

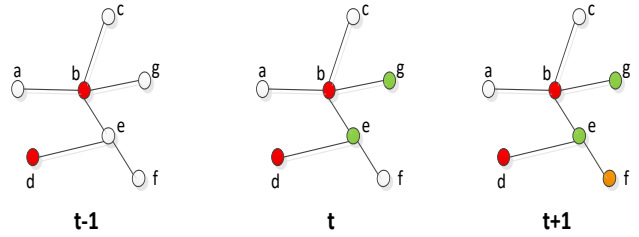

Figure 1. Information dissemination process of IC model

In Figure 1, there are two active nodes b and d at time t-1, and these two nodes intend to activate their neighbour nodes with a certain activation probability. The nodes a, c, e, g can be activated by the node b while the node e can be activated by the node d. However, b and d characterize only one opportunity to activate their inactive neighbour nodes; if this activation fails, they cannot attempt to activate again. Moreover, as shown in Figure 1, b and d have a common neighbour node e, and they can activate e at the same time, but the node e can only be activated by one of them. It means if the node e is already activated by b, there is no need for node d to activate it. In the next timeframe t, the nodes e and g which are successfully activated are added to the active node set. Finally, the above process continues until no new nodes remain to be activated.

As we know, in the IC model, the activation probability is generated randomly. However, the activation probability of users is related to the social relationships among users and topics in the process of hot events propagation in a microblogging network, and users often characterize different activation probability for different topics. Therefore, a topic popularity ranking [33] based information propagation model is proposed which calculates the post activation probability P_{ij}^{t} for specific topics to simulate the information propagation in a more realistic way.

The activation probability P_{ij}^{t} is influenced by the following factors. First, it is closely related to the social connections between posts; more connection times imply a more intimate relationship between posts and have a higher activation probability. Therefore, user intimacy can be used to represent the degree of intimacy between posts.

Definition 1: User intimacy, $C_{u,v}$, defines the frequency of the connection between users u and v. It can be obtained by the ratio of the connection time between u and v to the connection time between u and other users, as shown in equation (2).

$$C_{u,v}=\frac{R_{u,v}}{\sum_{i=1}^{n} R_{u,V_i}+\sum_{i=1}^{n} R_{v,V_i}},\left(u,v,V_i \in V\right) \quad (2)$$

where, R_{u,V_i} denotes the connection time between users u and V_i, $R_{u,v}$ denotes the connection time between users u and v, V denotes all the users, and n denotes the size of V.

In addition, P_{ij}^{t} is also influenced by users' topic popularity, such that user topic with higher popularity enables easy and quicker information propagation. Therefore, topic popularity can affect the activation probability P_{ij}^{t} between two given users.

Definition 2: Topic popularity, $TP_{u,v}^{t}$, denotes the popularity degree of a topic between two given users. The topic popularity $TP_{u,v}^{t}$ can be calculated using the following formula.

$$TP^{t}_{u,v} = \frac{Hub^{t}{}_{u,v}}{Hub^{t_i}{}_{\max} + Hub^{t_i}{}_{\min}}, (u, v \in V, t_i \in t) \quad (3)$$

where, $Hub^{t}{}_{u,v}$ denotes the hub value of a user in topic t, $Hub^{t_i}{}_{\max}$ denotes the maximum hub value of a user's topics, and $Hub^{t_i}{}_{\min}$ denotes the minimum hub value of a given user's topics.

In summary, the activation probability P^{t}_{ij} is influenced by the user intimacy $C_{u,v}$ and the topic popularity $TP^{t}_{u,v}$, so that the activation probability of post i to j for a specific topic t is calculated as follows.

$$P^{t}_{ij} = C_{u,v} \times TP^{t}_{u,v} \quad \left(P^{t}_{ij} \in [0,1]\right) \quad (4)$$

Before considering the real influence of a post in the network, it is essential to identify the posts those are linked to this post and to divide them into respective classes according to their ability of event diffusion. As we know, any given post in the network should have one or more directly linked posts, such that the connected posts are classified as neighbour posts in class 2, class 3 through to class T etc., where T denotes the vicinity of the neighbour nodes of the interested node.

As we can see from Fig. 2,

To start with, the post we want to analyze (post 1 in Figure 2) is denoted as class 1.

Then, the posts which have a direct link with class 1 (post 2, 3, 4, 5) is denoted as class 2.

Next, the posts which have a direct link with class 2 posts (post 6, 7, 8, 9, 10 in Figure 2) is denoted as class 3 for each post of class 2. The posts that have been classified before are ignored when encountered later (e.g., post 5 in Figure 2).

Finally, we define the higher-class numbers from previous class posts until reaching class T.

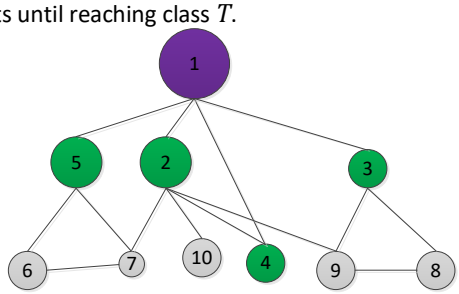

Figure 2. An example of defining neighbour classes.

Now, we obtain the numbers of links between two adjacent classes, denoted as n_{ij}. Then, the total real influence through links between class G posts to class H posts is defined as in equation 5.

$$L_{GH} = \sum_{i} \sum_{j} f_{i,j} \quad (5)$$

where, i ranges over all the posts in class G, j ranges over all the posts in class H. G and H usually have fewer posts than T.

Now, the posts are assigned in a $T*T$ matrix N

$$N = \begin{bmatrix} L_{11}, L_{12}, X, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ L_{21}, L_{22}, L_{23}, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ X, L_{32}, L_{33}, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X, X, X, \ldots\ldots, L_{(T-2)(T-2)}, L_{(T-2)(T-1)}, X \\ X, X, X, \ldots, L_{(T-1)(T-2)}, L_{(T-1)(T-1)}, L_{(T-1)T} \\ X, X, X, \ldots\ldots\ldots\ldots\ldots, X, L_{T(T-1)}, L_{TT} \end{bmatrix} \quad (6)$$

X in matrix N means that there are no connections between these two classes, thus X is always equal to zero. When we view a sub-network, we look at different classes with different emphasis. A controlled e_{GH} should be defined to present our perspective and show our emphasis. For example, as shown in Fig. 2, the links existing between class 1 and class 2 posts would be more important than those between class 2 and class 3, thus e_{12} characterize larger significance than e_{23}. Moreover, in order to be normalized, the value of e is within the range of 0-1.

2) Dynamic Network Structure model (DNS)

In order to obtain suitable values e_{GH} to present our concerns about various classes, an automatic variable scoring method [16, 33] is essential for ranking e_{GH}. Consequently, suitable values of e_{GH} can be automatically achieved.

Thus, we use the variable scoring method [16, 33] in 2-dimensional space to automatically select a set of suitable variables e_{GH}.

$$E = \begin{bmatrix} L_{11} * E_{11}, L_{12} * E_{12}, X, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ L_{21} * E_{21}, L_{22} * E_{22}, L_{23} * E_{23}, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ X, L_{32} * E_{32}, L_{33} * E_{33}, \ldots\ldots\ldots\ldots\ldots, X, X, X \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X, X, X, \ldots\ldots, L_{(T-2)(T-2)} * E_{(T-2)(T-2)}, L_{(T-2)(T-1)} * E_{(T-2)(T-1)}, X \\ X, X, X, \ldots, L_{(T-1)(T-2)} * E_{(T-1)(T-2)}, L_{(T-1)(T-1)} * E_{(T-1)(T-1)}, L_{(T-1)T} * E_{(T-1)T} \\ X, X, X, \ldots\ldots\ldots\ldots\ldots, X, L_{T(T-1)} * E_{T(T-1)}, L_{TT} * E_{TT} \end{bmatrix} \quad (7)$$

X is equal to 0.

Now, a Dynamic Network Structure (DNS) model is generated for all the users and posts as shown in equation 8.

$$DNS = \sum_{i=1}^{T} L_{ii} * e_{ii} \quad (8)$$

3) Identification of dynamic influence of posts

In Figure 2, DNS, with a post's link property and influence, holds true. Hence, the neighbouring post with highest DNS value is chosen as the dynamic influence of each post.

C. The improved PageRank algorithm

A new method based on the PageRank algorithm [15, 29-31] to analyse the weighted microblogging network is introduced in this section according to the posts published or retweeted by the users.

PageRank provides a more complex model for defining the importance of posts. Thus, we can adopt PageRank to detect the high influence users in the microblogging network [31].

Let U_i be a user ranging over all users in social networks. The PageRank model is denoted as shown in formula (9):

$$PR(U_i) = (1-d) + d\left(\frac{PR(U_1)}{C(U_1)} + \frac{PR(U_2)}{C(U_2)} \ldots + \frac{PR(U_n)}{C(U_n)}\right) \quad (9)$$

where, $PR(U_A)$ is the PageRank value of user U_A, $PR(U_i)$ is the PageRank value of the user of U_i which links to user U_A,$C(U_i)$ is the number of retweets or reply links from user U_i and d is set between 0 and 1.

1) Evaluate users' community influence

Generally, earlier posts get replied to, or retweeted, to generate new posts. When an important post is published, many retweets or replies would quickly follow in a relatively shorter time. Fig. 3 illustrates the proposed post retweeting network model. The nodes in Fig. 3 represent the retweeting posts, where post a linking to post b means that the post a is a retweet or a reply to post b.

The method of building the post network [31] is described as follows. The post network G_p is defined as $Gp = (P, C, Wc)$, where P is the set of posts, C is the set of edges (retweet or reply relationships between posts), and Wc is the set of weights, $Wc_{(ab)}$ associated with each edge links a pair of posts (p_a, p_b), which is defined as follows:

$$Wc_{ab} = \frac{\sigma_{ab}}{y_a - y_b + 1} \tag{10}$$

where, σ_{ab} = 1 only if post p_a retweets or replies post p_b, otherwise σ_{ab} = 0. y_a is the time of publishing of post p_a. y_b is the time of publishing of post p_b.

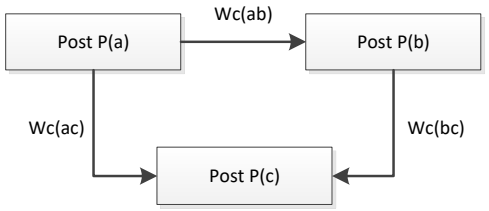

Figure 3. Decomposition graph of retweeting network model

A weighted number of retweets c is used to present a post's real influence:

$$c_a = \sum_{b=1}^{n} Wc_{ab} \tag{11}$$

Users usually receive different scores in each event when propagated according to their influences, and ranked based on their hub score. The process of calculating the value of the n^{th} user obtained in a post with the weighted number of retweets or replies of that post is shown in equation 12.

$$S_{n^{th}a} = \frac{(1 - r) * r^{n-1}}{1 - r^n} * c_a \tag{12}$$

where, $S_{n^{th}a}$ depicts the score of n^{th} user obtained for the post P_a, n^{th} defines the rank of a user for this post, n represents the total number of users of the post and c_a denotes the number of retweets for the corresponding post. In this paper, r is set to 0.7. Thus, the influence of the first user is above 30%.

The total value S_i of user i is defined as the sum of all values the user has obtained from each post.

2) Weighed user community network

However, the weighted number of all retweets can only represent the influence of the high-quality posts published or retweeted or replied by users, but not the real influence of users in the whole network, since the networks are always local. However, the methods based on networks, focus merely on the topological structure and ignore the characteristics of each user's topic influence on different topics or events. Hence it is necessary to combine the network topology with the users' topic influence by changing the network to indicate the topic characteristics of each user with the topological structure.

The weight between users in the microblogging network presents the frequency of their retweets or replies. The frequency of interactions can be represented by the effective distance between two users, which can be defined as:

$$d_{ij} = 1 - log(f_{ij}/F_i) \tag{13}$$

where, d_{ij} denotes the effective distance between two users i and j, f_{ij} denotes their interaction time, and F_i is the total number of interactions of user i. Equation 13 projects the proportion of interaction (f_{ij}/F_i, the value set to [0, 1]) to effective distance, the value set to $[0, inf]$, in this situation user j with higher proportion will have a shorter distance to the user i.

The importance degree of the relationship between users is calculated based on their effective distance. The more important the relationship between each user is, the closer the users are, and the more important the retweet relationship should be. With the scores of influence for different topics of users and effective distance between users, their relationship degree is computed using equation 14.

$$W_{ij} = \frac{k * S_i * S_j}{d^2_{ij}} \tag{14}$$

The rebuilt microblogging network C is denoted as $C = (V, E, W)$, where V is the set of users, E is the set of links (retweet or reply relationship between users), and W is the set of weights W_{ij} associated with each link connecting a pair of users (v_i, v_j), which is defined by equation 14.

IV. EXPERIMENTS

In this section, we use two popular influence models and their improved versions as our baseline methods, namely HITS-based methods [2, 28, 42] and PageRank-based methods [15, 40], validate the performance of our proposed model.

A. Dataset

The dataset is collected from Twitter [34], which is composed of 1,500,000 posts and 36,052 users. As discussed above, to reduce the impact of the bump phenomenon, we only included those users who published or retweeted or replied to posts in our dataset.

B. Baseline Approaches

The effectiveness of the proposed DPRank method is validated by comparing our proposed model against HITS +TS-LDA [44] and PageRank+HEE [43].

C. Evaluation Methods

The overlap coefficient [35] is defined as the size of the intersection divided by the minimum of the size of the two sets.

$$overlap(X,Y)=\frac{|X\cap Y|}{\min(|X|,|Y|)} \tag{15}$$

In this paper, the overlap coefficient is mainly used in the two sets of the most influential nodes obtained from the ranking algorithm and the Independent Cascade Model (IC) [32]. The higher the overlap is, the more reliable the ranking algorithm.

Meanwhile, the following evaluation criteria are also used to verify the proposed model: Recall and Precision. They are defined as follows:

The Recall rate indicates the proportion of detecting influential users being relevant, as shown below:

$$\text{Recall}=\frac{|influentials_r\cap influentials_t|}{|influentials_t|} \tag{16}$$

where $influentials_r$ represents a result set of influential users discovered by the proposed model, $influentials_t$ represents a true set of influential users discovered by HITS, PageRank and the proposed model.

The precision rate indicates the proportion of the relevant influential users in the real set of influential users, as shown below:

$$\text{Precision}=\frac{|influentials_r\cap influentials_t|}{|influentials_r|} \tag{17}$$

D. Experiment results

1) Recall and precision: Fig. 4 and Fig. 5 depict the recall and precision comparisons of the three methods. The performance of our proposed model is compared with HITS+TS-LDA and PageRank+HEE methods. It is evident, on the basis of our adopted evaluation criteria, that the effect of the proposed method based on the DNS model is better than two benchmarks. This performance improvement is due to the three basic characteristics of the users and posts: the ordinary users, the low-quality posts and the changing microblogging network structure. The compared methods ignore to exploit the above characteristics to detect relevant and accurate influential users dynamically from far too many low-quality posts and low-influence users. The proposed DPRank method based on HITS, DNS and PageRank, takes full advantage of the filtration function of the HITS algorithm, the dynamic influence feature of posts from DNS model and the dynamic influence feature of users from PageRank.

Furthermore, our proposed method exhibits better recall and precision than the other two benchmarks. As the number of detected influential users increases gradually, the recall rate of all the methods also increases because of the increasing number of correct influential users. However, where the number of real influential users is limited, the recall rate should exhibit a steady trend with an increasing number of detected influential users. On the contrary, the precision rate of all the methods decreases as the number of influential users' increases. This is because more inaccurate influential users are detected.

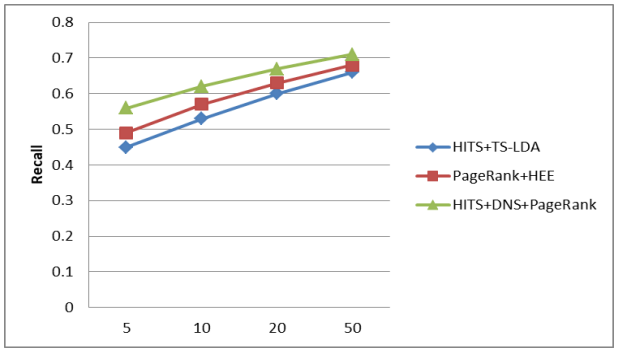

Figure 4. The Recall comparison of three methods

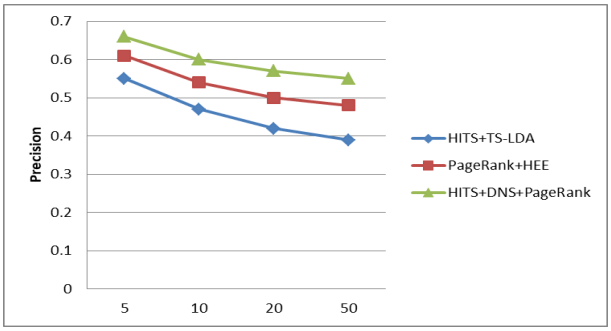

Figure 5. The precision comparison of three methods

2) The proper number of popular user topics: As shown in Fig. 6 and Table I, the proper number of popular topics can be obtained according to the results of Minimum Distance [43,44]. Moreover, high-quality users can be detected efficiently and effectively by our proposed DPRank method according to their hub values. This is because our proposed model effectively filters the low-influence users and non-popular topics according to the improved HITS method. This significantly increases the efficiency of user influence discovery of our proposed model in comparison with the existing user influence measure tactics.

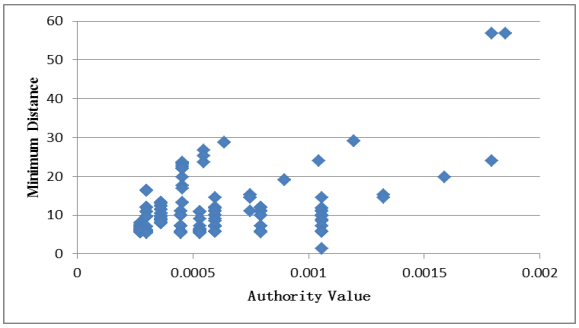

Figure 6. The number of topics from the DPRank method

TABLE I. MINIMUM DISTANCE AND AUTHORITY OF POSTS

Post ID	Authority Value	Minimum Distance
668664904677617664	0.001852832	56.80668975
668942051610873856	0.001792382	56.80668975
668946589063364608	0.001194922	29.12043956
668943987470811136	0.001194922	29.12043956
668946589063364608	0.001194922	29.12043956
668943987470811136	0.001194922	29.12043956
681697568456192001	0.000636248	28.7923601
681693469564383232	0.000636248	28.7923601
681697568456192001	0.000636248	28.7923601
681695337304702976	0.000545355	26.73948391

3) The analysis of initial influential users: As shown in Table III, it can be observed that the degree and hub values of users for topics can distinguish the importance of users under each popular topic initially. Meanwhile, we can also discover the number of influential users for each popular topic from Table IV, by setting a different number of initial influential users. Besides the proper number of influential users under popular topics can also be discovered from Fig. 7. With an increase in the number of initial influential users, the sphere of influence of the proposed model reaches its peak efficiency when the number of initial influential users is 10 and then saturates. Hence, the proper number of popular topics is set to 5, which verifies the correctness of the number of popular topics identified in Table II based on the key posts. The top 10 initial influential spreaders and the popular topics they belong to are shown from Table III and Table VI, which plays a key role in the influence diffusion of a given topic.

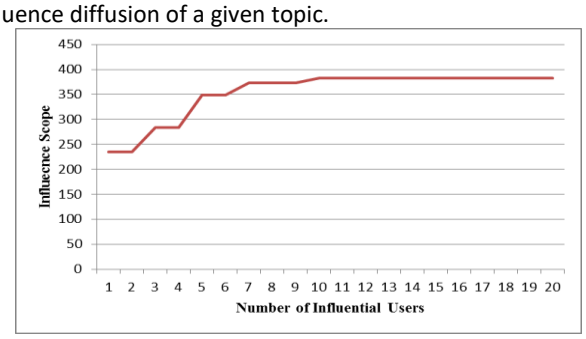

Figure 7. The proper number of influential users under popular topics

TABLE II. THE RESULT OF DIFFERENT PARAMETER

Method	Number of popular topics under the different *k*						
	k=1	*k=2*	*k=4*	*k=8*	*k=10*	*k=15*	*k=20*
TAS-HITS	1	1	3	4	5	5	5

TABLE III. DEGREE AND HUB VALUE OF TOP 10 INFLUENTIAL USERS UNDER TOPICS

User ID	Hub Value	Degree	Topic
339283603	0.003429355	24535	Sport
1679619506	0.003233392	2869	Sport
3693887599	0.003135411	334	Music
933364430	0.002253576	1157	Sport
4068440360	0.00186165	377	Economy
1000421510	0.001665687	1458	Music
2168821905	0.001567705	21973	Life
3254047099	0.001567705	489	Life
2310175028	0.001273761	1778	Music
863205451	0.000979816	44	Emotion

TABLE IV. THE NUMBER OF INFLUENTIAL USERS IN POPULAR TOPICS

Popular Topic	The number of Users
Sport	235
Economy	64
Music	49
Life	25
Emotion	10

4) Sphere of influence: It can be observed from Table V that the final sphere of influence of the proposed DPRank model is better than IC model, thus the influential spreaders discovered by our proposed model can be regarded as the most proper set when compared with the IC model. This is because the activation probability of the users considers the topic popularity and the intimacy between users in our proposed DPRank model.

TABLE V. THE FINAL SPHERE OF INFLUENCE OF INFLUENCE MAXIMIZATION

User ID	IC	DPRank
339283603	237	237
1679619506	174	185
3693887599	168	164
933364430	164	164
1000421510	153	164
4068440360	153	164
1367531	123	123
3254047099	123	123
2310175028	123	123
2168821905	102	123

TABLE VI. TOP 10 INITIAL INFLUENTIAL USERS MINING AND THE POPULAR TOPICS THEY BELONG TO

User ID	Authority Value	Popular Topic
339283603	0.051440325	Sport
1679619506	0.03233392	Sport
3693887599	0.03135411	Music
933364430	0.020282184	Sport
1000421510	0.01303155	Music
4068440360	0.011757792	Life
1367531	0.011757792	Economy
3254047099	0.011659809	Life
2310175028	0.010973935	Music
2168821905	0.010973935	Life

5) Correlation coefficient analysis: Firstly, traditional methods (i.e. HITS method, PageRank method) are used to grade all the users in microblogging networks. Table VII

presents the correlation coefficient of the influential users according to different centrality of our model and the IC models. In fact, a large number of users obtain the same score when using the HITS method and the PageRank method. However, the HITS method and the PageRank method present the same ranking for users whose degree is not large.

TABLE VII. CORRELATION COEFFICIENT COMPARISON OF THE TOP 10 INFLUENTIAL USERS

DataSet	(DPRank,IC)	(PageRank,IC)	(HITS,IC)
Twitter	0.162	0.156	0.152

TABLE VIII. THE INFLUENTIAL USERS COMPARISON OVER TIME, RANKED BY SPHERE OF INFLUENCE

User ID	December 28	March 10	May 12
1	339283603	339283603	339283603
2	1679619506	1679619506	1679619506
3	3693887599	933364430	3693887599
4	933364430	1367531	933364430
5	1000421510	3693887599	1367531
6	4068440360	1000421510	4068440360
7	1367531	2310175028	716307126
8	3254047099	4068440360	468646961
9	2310175028	3254047099	3553368433
10	2168821905	716307126	401127939

6) Accuracy: Table VIII shows changes in the user's sphere of influence ranking from December 28, 2015 to May 12, 2016, for posts with non-static influences. Days shown in Table VIII detected the top 10 influential users ranked by their influence degree, in which the dataset of the second day and the third day show greater fluctuation in the ranks of influential users, in fact, some of the influential users are eliminated later. This is due to the fact that the influence of users always changes over time affected by their published or retweeted posts. Moreover, the DPRrank model can dynamically identify the user's influence and improve the accuracy of identification, since over time important posts will naturally receive more replies and retweets. Therefore, the performance of our proposed DPRank method is much better than the HITS and PageRank methods, the accuracy of identifying the most influential spreader is improved significantly in the DPRank. This also validates the effectiveness of our proposed method.

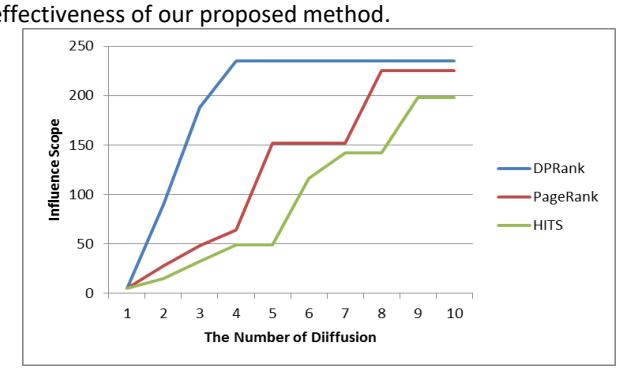

Figure 8. The comparison of dynamic sphere of influence under the "sport" topic

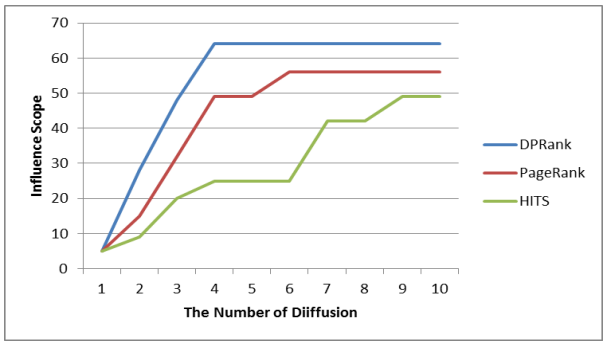

Figure 9. The comparison of dynamic sphere of influence under the "economy" topic

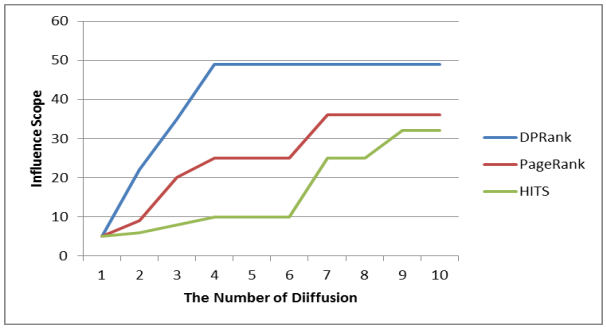

Figure 10. The comparison of dynamic sphere of influence under the "music" topic

7) Dynamic Influence analysis: For influential user analysis, experiments are conducted including HITS, PageRank and DPRank. Top 5 users in the dataset are selected as the influential users set. Then, the sphere of influence is calculated for each topic. Fig. 8, Fig. 9 and Fig. 10 presents the sphere of influence of influential users for top 3 popular topics. The dynamic sphere of influence of each topic with the DPRank method is higher than the other two methods in all Top 3 popular topics. This is because of the fact that our proposed DPRank method can change the Top 5 influential users dynamically during influence diffusion when compared with the PageRank and HITS methods. Thus, it can be concluded that the DNS model can identify the importance of posts dynamically, for enabling an automated discovery of the real influence of users during event evolution.

V. CONCLUSION AND FUTURE WORK

The hot event information on social networks is often local which can change dynamically over time whilst the user influence statistics are computed. Existing methods of influential user detection in social networks have not given enough emphasis to this changing nature of user influences. To this end, this paper proposed an efficient event detection and user influence discovery model to better detect the real influence of a post on hot events information evolution. Meanwhile, we applied our proposed model to real-world Twitter dataset and validated its efficiency against two notable existing methods

namely the PageRank and HITS methods, which demonstrates the effectiveness of our proposed model.

In our future work, we plan to explore the applicability of our proposed model in the field of multiple scholarly topic propagation, to capture varying responses of different users under different topics. Moreover, we plan to develop a prediction framework to predict the scholarly topical influence of influential users by combining all related features of users and posts.

ACKNOWLEDGEMENT

This work was partially supported by the Natural Science Foundation of Jiangsu Province under Grant BK20170069, UK-Jiangsu 20-20 World Class University Initiative programme, and UK-Jiangsu 20-20 Initiative Pump Priming Grant. Lu Liu is the corresponding author.

REFERENCES

[1] Y. Zhou, H. Xu, and L. Lei, "Event Detection Based on Interactive Communication Streams in Social Network," in Eai International Conference on Mobile Multimedia Communications, 2016, pp. 54-57.

[2] X. Sun, Y. Wu, L. Liu, and J. Panneerselvam, "Efficient Event Detection in Social Media Data Streams," in IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015, pp. 1711-1717.

[3] M. Hasan, M. A. Orgun, and R. Schwitter, TwitterNews+: A Framework for Real Time Event Detection from the Twitter Data Stream: Springer International Publishing, 2016.

[4] Q. Gao, F. Abel, G.J. Houben, et al., A comparative study of users' microblogging behavior on Sina Weibo and Twitter, in: User Modeling, Adaptation, and Personalization, Springer, Berlin, Heidelberg, 2012, pp. 88-101.

[5] C. Bigonha, T. N. Cardoso, M. M. Moro, M. A. Gonçalves, and V. A. Almeida, "Sentiment-based influence detection on Twitter," Journal of the Brazilian Computer Society, vol. 18, pp. 169-183, 2012.

[6] A. Aldhaheri and J. Lee, "Event detection on large social media using temporal analysis," in Computing and Communication Workshop and Conference, 2017.

[7] P. Yan, "MapReduce and Semantics Enabled Event Detection using Social Media," Journal of Artificial Intelligence & Soft Computing Research, vol. 7, 2017.

[8] J. del Campo-Ávila, N. Moreno-Vergara, and M. Trella-López, "Bridging the gap between the least and the most influential Twitter users," Procedia Comput. Sci., vol. 19, no. 6, pp. 437–444, 2013.

[9] W. Chai, W. Xu, M. Zuo, and X. Wen, "ACQR: A Novel Framework to Identify and Predict Influential Users in Micro-Blogging," in PACIS, 2013, p. 20.

[10] J. Zhou, Y. Zhang, and J. Cheng, "Preference-based mining of top-K influential nodes in social networks," Future Generation Computer Systems, vol. 31, pp. 40-47, 2014.

[11] C. Gao, D. Wei, Y. Hu, S. Mahadevan, and Y. Deng, "A modified evidential methodology of identifying influential nodes in weighted networks," Physica A: Statistical Mechanics and its Applications, vol. 392, pp. 5490-5500, 2013.

[12] C. Lee, H. Kwak, H. Park, and S. Moon, "Finding influentials based on the temporal order of information adoption in twitter," in International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, Usa, April, 2010, pp. 1137-1138.

[13] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February, 2011, pp. 65-74.

[14] C. C. Aggarwal, A. Khan, and X. Yan, "On Flow Authority Discovery in Social Networks," in Eleventh Siam International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, Usa, 2011, pp. 522-533.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, ODU, Norfolk.,1999.

[16] L. Chen and D. Wang, "An improved acquaintance immunization strategy for complex network," Journal of Theoretical Biology, vol. 385, p. 58, 2015.

[17] K. Zhou, A. Martin, and Q. Pan, "A similarity-based community detection method with multiple prototype representation," Physica A: Statistical Mechanics and its Applications, vol. 438, pp. 519-531, 2015.

[18] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October, 2010, pp. 199-208.

[19] M. Cha, H. Haddi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter," 2010.

[20] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February, 2011, pp. 45-54.

[21] G. Razis and I. Anagnostopoulos, "InfluenceTracker: Rating the impact of a Twitter account," in IFIP International Conference on Artificial Intelligence Applications and Innovations, 2014, pp. 184-195.

[22] B. Sun and V. T. Ng, "Identifying influential users by their postings in social networks," in Ubiquitous Social Media Analysis, ed: Springer, 2013, pp. 128-151.

[23] S. Kong and L. Feng, "A tweet-centric approach for topic-specific author ranking in micro-blog," in International Conference on Advanced Data Mining and Applications, 2011, pp. 138-151.

[24] L. A. Overbey, C. Paribello, and T. Jackson, "Identifying influential twitter users in the 2011 egyptian revolution," in International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2013, pp. 377-385.

[25] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1039-1048.

[26] C. Xiao, Y. Zhang, X. Zeng, and Y. Wu, "Predicting User Influence in Social Media," JNW, vol. 8, pp. 2649-2655, 2013.

[27] Barbieri, Nicola, Bonchi, Francesco, & Manco, Giuseppe. '' Cascade-based community detection''. 2013,pp 33–42 of: Proceedings of the sixth acm international conference on web search and data mining. WSDM '13. New York, NY, USA: ACM.

[28] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," Computer Networks and ISDN Systems, vol. 30, pp. 65-74, 1998.

[29] M. Franceschet, "PageRank: Standing on the shoulders of giants," Communications of the ACM, vol. 54, pp. 92-101, 2011.

[30] G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," Information Processing & Management, vol. 12, pp. 297-312, 1976.

[31] J. Liu, Y. Li, Z. Ruan, G. Fu, X. Chen, R. Sadiq, et al., "A new method to construct co-author networks," Physica A Statistical Mechanics & Its Applications, vol. 419, pp. 29-39, 2015.

[32] D. Kempe, J. Kleinberg, and Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137--146.

[33] Y. Li, C. Jia, and J. Yu, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 438, pp. 321-334, 2015.

[34] Twitter, REST API Resources, 2019, Available at: https://dev.twitter.com

[35] H. F. Inman and E. L. B. Jr, "The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities," Communication in Statistics-Theory and Methods, vol. 18, pp. 3851-3874, 1989.

[36] F. Riquelme and P. González-Cantergiani, "Measuring user influence on Twitter: A survey," Information Processing & Management, vol. 52, pp. 949-975, 2016.

[37] J. Zhou, G. Wu, M. Tu, B. Wang, Y. Zhang, and Y. Yan, "Predicting user influence under the environment of big data," in IEEE International Conference on Cloud Computing and Big Data Analysis, 2017, pp. 133-138.

[38] A. Amalanathan and S. M. Anouncia, "A review on user influence ranking factors in social networks," International Journal of Web Based Communities, vol. 12, p. 74, 2016.

[39] K. Lee, J. Mahmud, J. Chen, M. Zhou, J. Nichols, Who will retweet this? detecting strangers from twitter to retweet information, ACM Trans. Intell. Syst.Technol. (TIST) 6 (3) ,2015.

[40] N. Ohsaka, T. Maehara, K.-i. Kawarabayashi, Efficient pagerank tracking in evolving networks, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 875–884.

[41] G. Katsimpras, D. Vogiatzis, G. Paliouras, Determining influential users with supervised random walks, in: Proceedings of the 24th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2015, pp. 787–792.

[42] I. Karagiannis, A. Arampatzis, P.S. Efraimidis, G. Stamatelatos, Social network analysis of public lists of pois, in: Proceedings of the 19th Panhellenic Conference on Informatics, ACM, 2015, pp. 61–62.

[43] L. L. Shi, L. Liu, Y. Wu, L. Jiang, and J. Hardy, "Event Detection and User Interest Discovering in Social Media Data Streams," IEEE Access, ,2017, 5: 20953-20964.

[44] L. Shi, Y. Wu, L. Liu, X. Sun, and L. Jiang, "Event detection and identification of influential spreaders in social media data streams," Big Data Mining & Analytics, vol. 1, pp. 34-46, 2018.

[45] Y.H. Guo, L. Liu, Y. Wu, J. Hardy, Interest-Aware Content Discovery in Peer-to-Peer Social Networks, ACM Transactions on Internet Technology, 18(3), 2017

[46] L. Liu, N. Antonopoulos, M. H. Zheng, Y. Z. Zhan, Z. J. Ding, A Socio-ecological Model for Advanced Service Discovery in Machine-to-Machine Communication Networks, ACM Transactions on Embedded Computing Systems, Vol 15(2), 2016, pp 38:1–38:26.

[47] L. Jiang, L. L. Shi, L. Liu, B. Yuan, Y. J. Zheng, An Efficient Evolutionary User Interest Community Discovery Model in Dynamic Social Networks for Internet of People, IEEE Internet of Things Journal, in press, 2019.

[48] L. Jiang, L. L. Shi, L. Liu, M. Ali Yousuf, J. J. Yao, User Interest Community Detection on Social Media Using Collaborative Filtering, Wireless Networks, Springer, 2019.

[49] L. L. Shi, L. Liu, Y. Wu, L. Jiang, A. Ayorinde, Event Detection and Multi-source Propagation for Online Social Network Management, Journal of Network and Systems Management, Springer, in press, 2019.

[50] Y. Wu, C. G. Yan, L. Liu, Z. J. Ding and C. J. Jiang, An Adaptive Multilevel Indexing Method for Disaster Service Discovery, IEEE Transactions on Computers, Vol 64(9), September 2015, pp 2447 - 2459.

[51] L. Shi, Y. Wu, L. Liu, X. Sun and L. Jiang, "Event Detection and Key Posts Discovering in Social Media Data Streams," in 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, United Kingdom, 2017 pp. 1046-1052.doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.159.

[52] Y. Lin, X. Wang, F. Hao, Y. Jiang, Y. Wu, G. Min, D. He, S. Zhu, and W. Zhao, "Dynamic Control of Fraud Information Spreading in Mobile Social Networks," IEEE Transactions on Systems, Man, and Cybernetics: Systems, DOI: 10.1109/TSMC.2019.2930908, In Press.

[53] X. Cheng, Y. Wu, G. Min, and A.Y. Zomaya, "Network Function Virtualization in Dynamic Networks: A Stochastic Perspective," IEEE Journal on Selected Areas in Communications, vol. 36, no. 10, pp. 2218-2232, 2018.

[54] W. Miao, G. Min, Y. Wu, H. Huang, Z. Zhao, H. Wang, and C. Luo, "Stochastic Performance Analysis of Network Function Virtualisation in Future Internet," IEEE Journal on Selected Areas in Communications, vol. 37, no. 3, pp. 613-626, 2019.

[55] H. Wang, Y. Wu, G. Min, J. Xu, and P. Tang, "Data-driven Dynamic Resource Scheduling for Network Slicing: A Deep Reinforcement Learning Approach," Information Sciences, vol. 498, pp. 106-116, 2019.

Lei-lei Shi received the B.S. degree from Nantong University, Nantong, China, in 2012, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2015. He is currently working towards the Ph.D. degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include Event Detection, Data Mining, Social Computing and Cloud Computing.

Lu Liu is the Head of School of Informatics and Professor of Informatics at the University of Leicester, UK. Prof. Liu received his Ph.D. degree from University of Surrey in 2007 and M.S. degree from Brunel University in 2003. Prof. Liu's research interests are in the areas of data analytics, service computing, cloud computing, Artificial Intelligence and the Internet of Things. He is a Fellow of British Computer Society (BCS). Liu's IEEE membership year starts from 2007.

Yan Wu received the M.S. degree from Shandong University of Science and Technology, Qingdao, China, in 2009, and the PhD degree from Tongji University, Shanghai, China, in 2014. He is currently a Lecturer with the School of Computer Science and Telecommunication Engineering in Jiangsu University, China. His research interests include formal methods, service-oriented Computing, and Cloud Computing.

Liang Jiang received the B.S. degree from Nanjing University of Posts and Telecommunications, China, in 2007, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2011. He is currently working towards the Ph.D. degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include OSNs, Computer Networks and Network Security.

John Panneerselvam is a Lecturer in Computing at the University of Derby, United Kingdom. John received his PhD in Computing from the University of Derby in 2018 and received the M.Sc. degree in the University of Derby, Derby, U,K, in 2013.. He is an active member of IEEE and British Computer Society, and a HEA fellow. His research interests include cloud computing, fog computing, Internet of Things, big data analytics, opportunistic networking and P2P computing. He has won the best paper award in IEEE International Conference on Data Science and Systems, Exeter, 2018.

Roy Crole is the Deputy Head of the Department of Informatics and Associate Professor of Informatics at the University of Leicester. He received his M.A. degree in 1990, M.Math in 2011 and Ph.D. degrees from the University of Cambridge in 1992; he is a fellow of the HEA and the Cambridge Philosophical Society. Dr. Crole's research is in Categorical Type Theory, the Semantics of Programming Languages, and Communicating Systems.