IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 11, 2020, accepted September 24, 2020, date of publication October 8, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029657

Fuzzy Relative Willingness: Modeling Influence of Exogenous Factors in Driving Information Propagation Through a Social Network

SUMAN KUNDU[1,3], TOMASZ KAJDANOWICZ[1], PRZEMYSLAW KAZIENKO[1], (Senior Member, IEEE), AND NITESH CHAWLA[1,2], (Senior Member, IEEE)

[1]Department of Computational Intelligence, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland
[2]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
[3]Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur, Karwar 342037, India

Corresponding author: Tomasz Kajdanowicz (tomasz.kajdanowicz@pwr.edu.pl)

This work was supported in part by the National Science Centre, Poland Research, under Project 2016/23/B/ST6/01735 and Project 2016/21/D/ST6/02948, and in part by the Statutory Funds of Department of Computational Intelligence.

ABSTRACT A high percentage of information that propagates through a social network is sourced from different exogenous sources. E.g., individuals may form their opinions about products based on their own experience or reading a product review, and then share that with their social network. This sharing then diffuses through the network, evolving as a combination of both network and external effects. Besides, different individuals (nodes in a social network) have different degrees of exposition to their external sources, as well. Modeling this influence of external sources is important in order to understand the diffusion process and predict future content sharing patterns. Recognizing this fusion of intrinsic (network) effect and exogenous (external) effect, this paper develops a novel fuzzy relative willingness (FRW) model. Leveraging a fuzzy set approach provides a way to handle the uncertainties arising within the human concept of willingness. We demonstrate that FRW is able to accurately identify both top-k most content producers and diffusion effect based on external influence. We also demonstrate that the fuzzy set theory provides a compelling framework to model uncertainties pertaining to the influence as well as the susceptibility of individuals for both network and exogenous effects.

INDEX TERMS Social networks, information diffusion, exogenous factors.

I. INTRODUCTION

Quantifying how much a person is willing to accept information from external sources is an interesting problem in the context of information diffusion. Information that propagates in the network can be twofold because of its source. The first kind of information is acquired from one of the user's neighbors in the social network (internal influence), and the other one is brought into the network from outside (exogenous influence). The external source can be anything, starting from legacy books through a more dynamic online news portal, in-person discussion to television, from family relationships to virtual networks like blogs. Each individual has a different degree of exposition to these external sources. In this context, one question arises: 'is understanding persons' willingness to adopt from external a valuable factor for predicting the future content sharing pattern in the whole network?'

Information diffusion has been forefront of the research for quite sometime [1]–[4] with the objective of finding the influential nodes [3], [5]–[17]. Most of the previous work, with very few exceptions [4], [18], only considered peer-influence in information diffusion. However, recently Li *et al.* [19] show that about 50% to 70% of the information cascade is due to the exogenous factors. Therefore, it is essential to understand whether a person is willing to share information from exogenous sources or not in order to explain the information diffusion process better.

The present study is an attempt in the direction of quantifying willingness to adopt from exogenous sources. We define a new problem of identifying nodes' willingness in the

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao Liu.

context of information propagation. The concept of 'willing' is rather perceived than certain. One can be less willing or more willing, hence, binary classification of whether a person is willing or not would not be suitable in such a scenario. The uncertainties involved within the concept of willingness itself needs to be taken care off. Fuzzy sets being a well known mathematical tool for managing uncertainties involved with the imprecise concepts is a natural choice for characterizing the willingness of a node. In this paper, we use a fuzzy set to express the uncertainties and propose a fuzzy relative willingness (FRW) measure, which characterizes the relative willingness of a node to adopt from exogenous factors. In order to validate, extensive simulations have been performed on synthetic data set to check the membership against the prior configurations. We perform basic comparisons with few baseline algorithms, which show that the proposed model better explains the prior configurations than the baseline algorithms. FRW is further used in a newly re-framed perspective on target set selection (seed selection), i.e., in the top-k content producer problem. Comparative results on two real-world networks show that selecting top-k nodes based on proposed FRW measure outperforms other baseline algorithms in predicting the number of externally influenced future content shares. In summary, the contribution of the paper is as follows:

1) We define a new problem of quantifying willingness to adopt and propagate from exogenous influence.
2) We use a simple methodology to express willingness in the framework of a fuzzy set and propose a fuzzy relative willingness (FRW) measure to quantify willingness to adopt.
3) Finally, we use the FRW measure in a newly defined ranking problem, which maximizes the number of externally influenced future content shares in the network. The problem is referred to as the top-k content producer problem.

The rest of the paper is organized as follows. Related literature is discussed in Section II. The proposed problem of quantifying adoption willingness, its mathematical foundations, an algorithm, and complexity analysis are described in Section III. Section IV reports various experiments, corresponding results and discussions, whereas the final conclusions can be found in Section V.

II. RELATED WORK

The work on diffusion of innovation has been started by sociologist [1], [20] and currently being explored by different field of studies including computer science for last few decades. Since the pioneering work of [5], [21] on viral marketing, different diffusion models [1]–[4] and algorithms to find out the influencing individuals [3], [5]–[17] have been developed. It also includes interacting spreading processes in multilayer social networks [22], [23]. During these time, the main objective was to use the information diffusion for viral marketing, diagnosis and controlling epidemic spread and identifying threats among others. The overall aim was to maximize the influence, e.g. [24], [25]. Most of these diffusion models considered only the effect of peer-influence except a few exceptions. The paper [26] model information propagation to identify whether a propagation is peer-driven or authority-driven, whereas [19] and [27] tried to estimate the magnitude of the external influence in the network.

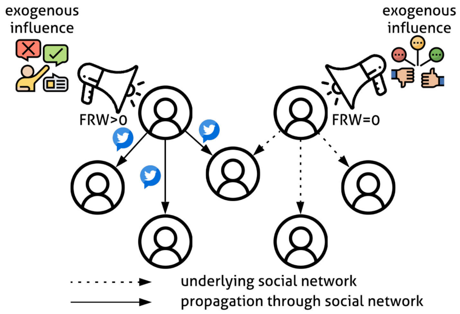


FIGURE 1. The example of how Fuzzy Relative Willingness (FRW) quantifies susceptibility to external factors and the associated initiation of information diffusion in the social network.

Reference [18] modeled information diffusion process incorporating both the peer and exogenous effects; here the exogenous effects were calculated from a time function called event profile. However, another important task of identifying nodes' willingness to accept new information from exogenous sources or willingness to share external information in the network is not attempted. Quantifying such attributes of a person will provide an opportunity to study the aforementioned problems in a different dimension. For example, in viral marketing, we could reduce our search to the more willing individuals only; for blocking a threat we could push the correction measures to the members who wish to accept the changes rather whose influence is higher.

Quantifying willingness of a node to adopt from exogenous influence, as per our best knowledge, is not attempted earlier. The best know problem addressed in this direction is the target set selection which attempt to select nodes based on the influence in the network. Readers may refer [28] for a comprehensive related work on target set selection problem. Different popular algorithms used for this task includes centrality like degree, diffusion degree [11], [29], degree discount [9] based heuristics methods, Prefix excluding Maximum Influence Arborescence (PMIA) [30] and Network Discovery of Influencers using Flows (NDIF) [31] among others. Table 1 list the properties of our proposed FRW based top-k content producer in comparison with the other comparative methods.

III. FUZZY RELATIVE WILLINGNESS

Considering exogenous factors, we are trying to quantify the relative willingness of each node to adopt and propagate a piece of information as compared to other nodes in the network. For a given social network (e.g., friendship network, following-follower network) and content stream (e.g., list of tweets, status, hashtag) shared within the network, we assume that network actors are externally influenced to share content

TABLE 1. Comparison of Our Method to others.

Feature	Degree	Diffusion Degree	NDIF	Degree Discount	PMIA	FRW
Data-Driven	✗	✗	✓	✗	✗	✓
Diffusion Model Dependency	✗	✓	✗	to some extent	✓	✗
Accuracy	✗	✗	✗	✗	✗	✓

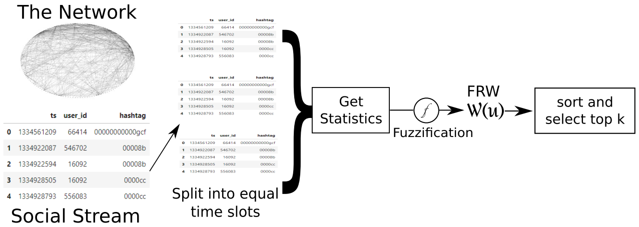


FIGURE 2. Block diagram showing the procedure for quantifying fuzzy relative willingness (FRW) and top-*k* content producers in the network.

if none of their neighbors used the same piece of information anytime before them. In order to calculate the willingness, we split the content stream into equal-length time spans. As we were trying to gather a comparative view of externally influenced nodes, we calculate the global statistics like mean and standard deviation of externally influenced social shares per user per time span. Fuzzy relative willingness score is then calculated based on how much a node's externally influenced social share differs as compared to the global average. A block diagram of the process is shown in Figure 2.

The concept of 'willing' is related to human behavior. Hence, it is imprecise in nature. Uncertainties within it have been expressed in the concept of fuzzy set theory here. A fuzzy transformation function is used to transform a node's social share count into a fuzzy membership value, which depicts its position compared to the global average value and characterizes the willingness. *FRW* is mathematically defined in the following sections, along with the necessary introduction of notation used (Table 2) and complexity analysis.

A. MODEL

Let a social network be represented with a graph $G(V, E)$ where V is the set of nodes, and E is the set of edges. Also, let the stream of all the content shared in the network be $\mathcal{P}$, which is a list of tuples $p(h, u, t)$. Here, h is a content (tweet) or content token (e.g. hash-tag or keyword) shared by a node (user) $u \in V$ at a time-stamp t. We call h as a post in the rest of the paper. The problem is to quantify the relative willingness to adopt $\mathcal{FRW}(u)$ of a node u for a given network G and content stream $\mathcal{P}$.

Definition 1 (Length of Time Slot t_s): The stream $\mathcal{P}$ has been split into equal time-length slots. Let the number of these time slots be $N_t \in \mathbb{N}$. Then, the length of a single time slot is calculated as:

$$t_s = \frac{t_{max} - t_{min}}{N_t} \tag{1}$$

TABLE 2. Explanation of the notation used.

Used symbol	Explanation
G	graph
V	set of nodes
E	set of edges
$\mathcal{P}$	stream of content shares
h	content of a share, e.g. tweet, hash-tag, keyword
u	node (user), $u \in V$
t	timestamp of a share
$p(h, u, t)$	single content share
N_t	number of time slots
$E^j(u)$	set of externally influenced posts in jth time slot for a node u
$\Gamma(u)$	set of u's neighbors in the graph
m	mean number of externally influenced post per user over all time slots
σ	standard deviation of externally influenced post per user over all time slots
a, b	parameters of Fuzzy Relative Willingness
$\mathcal{FRW}(u)$	Fuzzy Relative Willingness to adopt of node u

where t_{max} and t_{min} is the time-stamp of the very last and the very first post in $\mathcal{P}$ respectively.

Definition 2 (Externally Influenced Post): A user's post is said to be externally influenced if it is observed that none of his neighbors shared the same post earlier within the same time slot. Thus, a set $E^j(u)$ of externally influenced posts p for a node $u \in V$ in time slot j is defined as:

$$\begin{aligned} E^j(u) = \{&p(h, u, t^u) \in \mathcal{P} | \nexists q(h, v, t^v) \in \mathcal{P}\ \forall v \in \Gamma(u), \\ &(t^v < t^u); t_{min} + (j-1)t_s < t^u, t^v \leq t_{min} + jt_s\} \end{aligned} \tag{2}$$

Here $0 < j \leq N_t$ is the index of the time slot the observed post $p \in \mathcal{P}$ belongs to. $\Gamma(u)$ represents the set of u's neighbors. We restrict our search for a known neighbor to a particular time slot. This provides a sense of memory window to each node as well as it reduces the running time of the algorithm. Further, we considered only the short-term influence not long-term influence. The assumption of using a particular time slot in the definition is inline with this.

Definition 3 (m): m is the mean number of externally influenced post per user over all time slots. That is,

$$m = \frac{\sum_{j=1}^{N_t} \sum_{u \in V} |E^j(u)|}{|V| * N_t} \tag{3}$$

Definition 4 (σ): σ is the standard deviation of externally influenced post per user over all time slots. That is,

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N_t} \sum_{u \in V} ||E^j(u)| - m|^2}{|V| * N_t}} \tag{4}$$

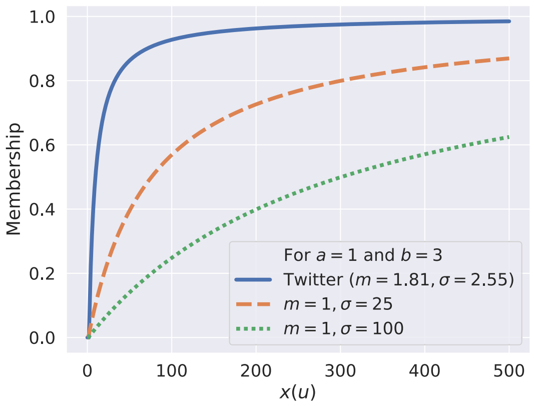


FIGURE 3. Examples of MS large fuzzy transformation function for distinct parameters.

Definition 5 (Fuzzy Relative Willingness): Fuzzy Relative Willingness $\mathcal{FRW}(u)$ of a node u is calculated with the following MS Large fuzzy transformation function [32].

$$\mathcal{FRW}(u) = \begin{cases} 1 - \frac{b \times \sigma}{x(u) - a \times m + b \times \sigma} & \text{when } x(u) > a \times m \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Here, $x(u) = \frac{\sum_{j=1}^{N_t} |E^j(u)|}{N_t}$ is the mean number of externally influenced posts for a particular user u; a and b are user-defined parameters of FRW. The choice of the MS Large transformation function is motivated by the fact that it allows us to assign the users who are more likely to adopt and have many posts with larger $x(u)$ (membership) values. In addition, we wanted a slow growth as the post increases to a certain extent. Figure 3 show the plots of MS Large function for different values of m and σ as a reference. However, one may choose another fuzzy function depending upon the properties they wish to express.

1) PROPERTIES OF THE PARAMETERS a AND b

The parameter a controls the minimum number of posts, after which a person is labeled as a willing person via an FRW membership value > 0. If $a = 0$, then all the nodes having one or more externally influenced posts are assigned with a positive FRW membership, whereas, for $a = 1$, only the nodes with more than the mean number of externally influenced posts are assigned with an FRW membership.

The parameter b is the scaling factor of the standard deviation used in the FRW. The value $b = 0$ will assign each node (qualified based on a) with a membership value of 1.

In other words, the selection of coefficients depends on the definition of anomalies. It is highly likely that no anomaly is present when the concentration is less than the mean, implying that a is 1. The tolerance between 'anomaly' and 'background' is around the mean plus one standard deviation, implying that b is also 1.

Algorithm 1 Willingness Calculation

```
Input: $G(V, E)$, $\mathcal{P}$, $t_s$
$\mathcal{P} \leftarrow Sort(\mathcal{P})$    ▷ sort by timestamp
$k \leftarrow 1$, $start \leftarrow t_{min}$, $end \leftarrow start + t_s$
while $start < t_{max}$ do
    $events \leftarrow \{p(h_t, u, t) \in \mathcal{P} | start \leq t < end\}$
    Initialize: $KT$, $KTC$, $PTC$ with empty dictionaries
    ▷ $KT$ : known tags, $KTC$ : $KT$ count, $PTC$ : total post count
    for all $p(h, t, u) \in events$ do
        if $h \in KT[u]$ then
            $KTC[u] \leftarrow KTC[u] + 1$
        else
            $KT[u] \leftarrow KT[u] \cup \{h\}$
        $PTC[u] \leftarrow PTC[u] + 1$
        for all $v \in \Gamma(u)$ do
            $KT[v] \leftarrow KT[v] \cup \{h\}$
    for all $u \in V$ do
        $|E^k(u)| = PTC[u] - KTC[u]$
    $k \leftarrow k + 1$, $start \leftarrow end$, $end \leftarrow start + t_s$
Calculate $m$ and $\sigma$ using Equations 3 and 4
for all $u \in V$ do
    Calculate $\mathcal{FRW}(u)$ using Equation 5
```

The estimation method as well as sensitivity study over a and b parameters is out of scope in this work. We would like to redirect the reader to works dealing with this problem directly, e.g. [32]

B. ALGORITHM

In order to effectively calculate the FRW values, the main computational challenge is to identify the posts which are externally influenced. The natural way is to search for the source in the network recursively. Although the search is limited to the immediate neighbors, it is time-consuming. We take a top-down approach to reduce execution time. The algorithm to calculate FRW of all the nodes in the network is shown in Algorithm 1. The inputs to the algorithm are the social graph $G(V, E)$, the list of social posts $\mathcal{P}$, and the time span t_s used for identifying the externally influenced posts. We sort $\mathcal{P}$ according to their time-stamps, and the algorithm then linearly traverse each post. In each step, it updates two sets of counters, one is a total number of posts of the user (PTC), and another is a total number of known posts (KTC) of the user (i.e., shared by some neighbor beforehand in the time slot). It then updates the known keywords (KT) of all its neighbors as it is being shared with them from the current post owner.

C. COMPLEXITY

The proposed algorithm runs for each post of the social stream. In each step, two major activities are performed. The former is to check whether the content is prior known by the content owner. The latter is to update all of its

TABLE 3. Synthetic Data set.

Properties/Name	Net200
Network Type	Undirected
Nodes	148
Edges	604
Avg. Degree	8

neighbors' known content list. We manage this list using a hash set so that a typical search operation can be considered as $O(1)$. Updating neighbors' list of known content can be very dependent on the number of neighbors one possesses. The average degree can be considered as a good indicator for typical cases. Finally, we calculate the FRW values for all the nodes in the network. So the complexity of the algorithm is $O(|\mathcal{P}| + \frac{2\times|\mathcal{P}|\times|E|}{|V|} + |V|)$; where $|V|$ is the number of nodes, $|E|$ is the number of edges and $|\mathcal{P}|$ is the number of tokens in the social stream.

IV. EXPERIMENTS AND RESULTS

We conducted various experiments analyze the validity of the proposed FRW model using both synthetically generated and real-world data sets. Accuracy is measured by the fraction of externally influenced nodes identified (or true positive rate): $I_g = \frac{|Support(\mathcal{FRW})\cap T|}{|T|}$, where T returns the set of nodes selected as externally influenced nodes prior to the simulation, and $Support(\mathcal{FRW})$ is the support set of FRW. zIn addition, experiments are performed to see how the accuracy is affected by the change of different diffusion parameters. The simulations were carried out over a synthetically generated network Net200 (Table 3). LDBC DATAGEN [33] is used to generate this data, which is based on Facebook degree distribution.

We evaluate efficacy of FRW to predict the future states on a real-world data set: a Twitter social graph and hash tag data (Table 5). FRW is compared with several other baseline algorithms, where we also considered the effect of cascade size and the parameter of FRW. All software is written in Pythin. SOIL [34] simulation package is utilized for the simulations. All codes will be made publicly available after review.

Let us first describe the simulation processes and its information diffusion parameters.

A. THE DIFFUSION MODEL OF THE SIMULATION

The diffusion models like the independent cascade model or linear threshold model do not consider the influence of external factors. Our study focuses on the external factors and nodes' willingness. So a more comprehensive model is required. We use a simple diffusion model inspired by the tutorial on SOIL [35] for the simulations. In this process, each node $v \in V$ have three parameters expressing external exposure (Λ_v), the probability of infection from external factors (P_e^v) and the probability of infection by the internal connections (P_i^v). Λ_v is a Boolean parameter that determines whether node v is exposed to the external source or not. If $\Lambda_v = 1$ then v will be influenced to propagate the message by the external information with a probability of P_e^v. A node v gets influenced by its neighbor with a probability of P_i^v. The state transition diagram of the diffusion process is shown in Figure 4.

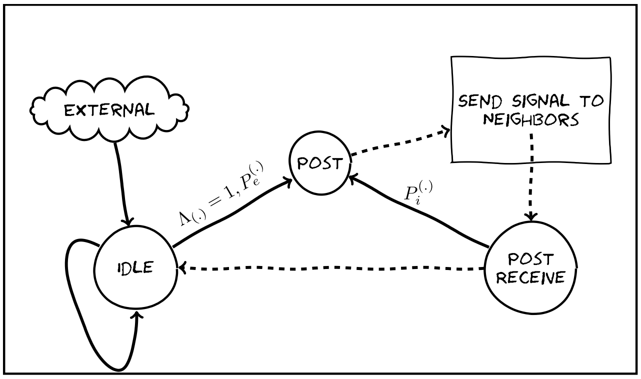


FIGURE 4. State transition diagram of the diffusion process.

B. ACCURACY WITH DIFFUSION PARAMETERS

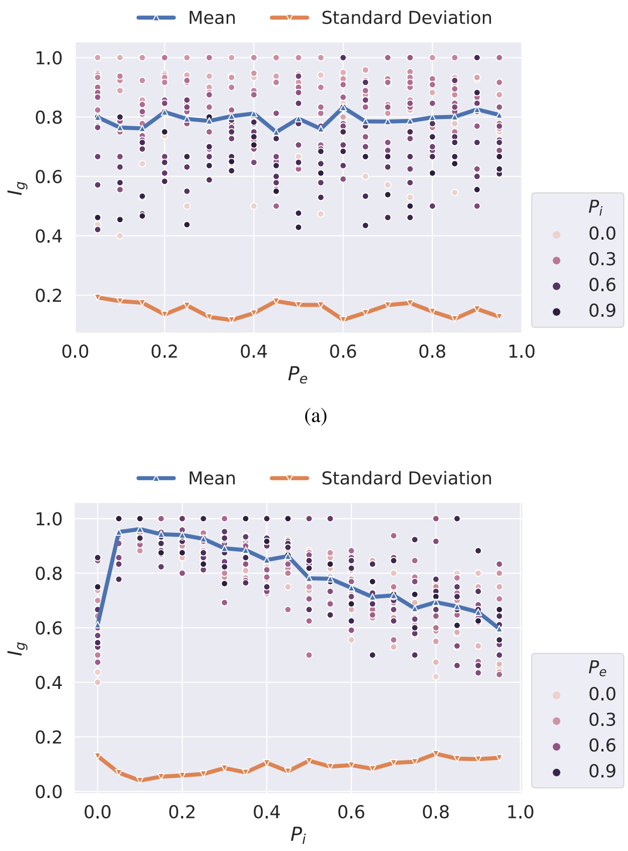
We varied the diffusion parameters in three different combinations viz. (i) keeping both P_e and P_i fixed for all the nodes in the network, (ii) varying P_e while keeping P_i fixed for all nodes and (iii) varying P_i for a fixed value of P_e. The obtained results and details of the simulations are presented below.

1) GLOBALLY FIXED PROBABILITY OF EXTERNAL AND INTERNAL INFLUENCES: P_e, P_i

In this setting, we assign a fix value for P_e^v and P_i^v for all $v \in V$ in a particular simulation. We randomly choose few nodes (about 10%) and assign it with $\Lambda = 1$. In various simulations, parameter P_e has been incremented within interval (0.0, 1.0) and parameter P_i in [0.0, 1.0) with the same increment value of 0.05. Thus, we run 380 simulations on Net200 (Table 3). A single simulation is performed with 2000 time steps (characterizing the time-stamp in real data set) and $t_s = 100$ is considered while calculating FRW values using Equations 2-5.

Figure 5a shows the variation of I_g with the value of P_e. Each dot in the plot corresponds to a single simulation in the network. The hue of the dots indicates the P_i value. It is evident that the mean I_g remains nearly constant at around 0.8, whereas the standard deviation very slowly decreases with rising P_e. We found that the accuracy is greater than 0.9 for 30% simulations and greater than 0.6 for about 88% of the simulations.

Variation of I_g with the change in internal influence probability P_i is shown in Figure 5b. Hue represents the external influence probability in this scatter plot. The blue line indicating the mean I_g clearly shows that it drops with the increase in P_i. At the same time, the standard deviation represented by the orange line reveals a slow increase. Despite the decrease, the mean value of I_g is above or equal to 0.6 for all the simulations. The same is evident from the heat map



(a)

(b)

FIGURE 5. Variation of I_g with (a) P_e and (b) P_i.

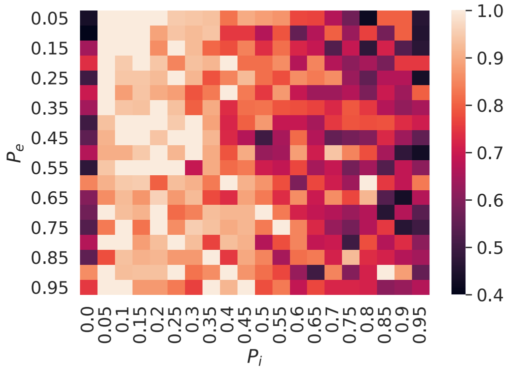


FIGURE 6. Heat map of I_g for P_e and P_i.

in Figure 6. We can see there that the higher accuracy is obtained when P_i is between 0.05 to 0.6 irrespective of the value of P_e.

2) VARIABLE VALUE OF P_e

In this experiment, we randomly selected 10 seed nodes and each of them is assigned with a P_e value taken from (0.0, 1.0] with a step 0.1. We were interested in testing the model behavior with uniformly distributed P_e. Simulations were carried out for the values of P_i between [0.0, 1.0) with step value 0.05. Twenty such simulations are further repeated for 10 different randomly chosen sets of seed nodes. Thus, total 200 simulations have been performed with this setting. The I_g values for 10 experiments are shown using box-whisker in Figure 7. Each label in x-axis corresponds to the results of experiments on one set of randomly chosen 10 seed nodes and the box-whisker graph of that label summarizes the outcome of 20 different simulations (for different values of P_i) as stated above. It is evident that the median value for accuracy is at least 0.7, if considering each experiment separately. The mean and standard deviation of all the 200 simulation together is found to be 0.814 and 0.158, respectively.

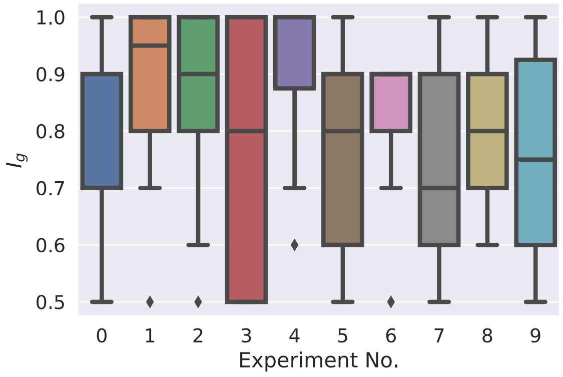


FIGURE 7. Box whisker plot of I_g for different experiments when variable values of P_e are used.

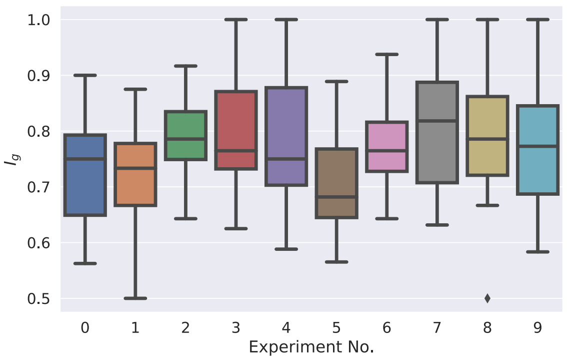


FIGURE 8. Box whisker plot of I_g for different experiments when variable values of P_i are used.

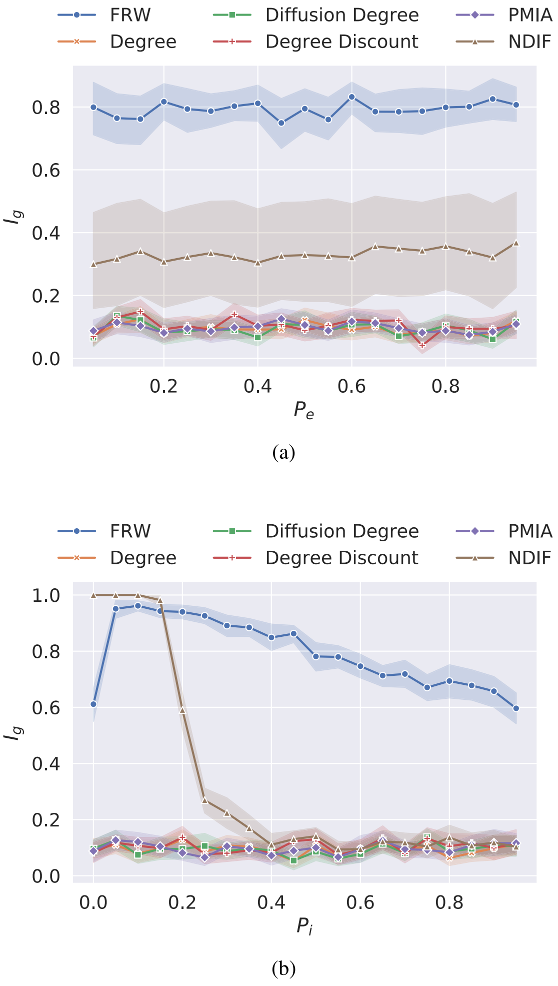
3) VARIABLE VALUE OF P_i

In this experiment we keep a global fixed value for P_e, while we assigned uniform random values to P_i. Similar to Section IV-B.2 different values of P_e is used for different simulations. The same process is repeated for 10 times with different uniform random values of P_i. The results for individual experiments are shown in Figure 8. The mean value of I_g is 0.7647. The minimum and maximum median accuracy are found to be 0.68 for experiment no. 5 and as much as 0.81 for experiment no. 7.

C. COMPARATIVE STUDY

To the best of our knowledge, no attempt was made to measure the willingness of the node to adopt from external sources. Accordingly, no similar algorithm is available in the literature to compare with straightforwardly. Therefore, in our experiments, we choose popular seed selection algorithms as a baseline. There are several algorithms for seed selection in the literature [28]. In our study, we tried to compare our method with well known seed selection algorithms of different approaches. In particular, we used centrality based, path based and content based seed selection methods. The details of these methods are as follows:

- Degree Centrality: One of the classic approaches where top-k influential nodes are selected based on their degree scores. We use $k = |\{v \in V|\Lambda_v = 1\}|$. Predicted true positive is calculated by the cardinality of the intersection of the seed selected by degree measure and the set $\{v \in V|\Lambda_v = 1\}$.
- Diffusion Degree [11], [29]: Diffusion degree includes the diffusion parameters along with the degree measure. Similarly to degree, top k nodes are chosen based on the node ranking upon diffusion degree.
- Degree Discount [9]: Node's centrality is calculated here by discounting edges of already selected seeds from the degree.
- Prefix excluding Maximum Influence Arborescence (PMIA) [30]: This is a path based method. Nodes are selected based on the expected influence of the paths connecting the node with others in the network.
- Network Discovery of Influencers using Flows (NDIF) [31]: This is a content-based algorithm, where content flow is used to identify the flow paths. Then, the algorithm greedily chooses the seeds. This algorithm is the closest algorithm to our proposed methods as it also uses content in order to find seeds. The algorithm takes the network structure and content streams as the input.



(a)

(b)

FIGURE 9. Comparative results for different algorithms. Color patch showing the confidence interval at 95%.

TABLE 4. Precision, Recall and f-Score of different algorithms.

Algorithm	Recall	Precision	f-score
Degree	0.0942631	0.0942631	0.0942631
Degree Discount	0.103773	0.103773	0.103773
Diffusion Degree	0.0938844	0.0938844	0.0938844
NDIF	0.330682	0.330682	0.330682
PMIA	0.0973923	0.0973923	0.0973923
FRW	0.792716	1	0.884374

Comparative results for different algorithms for the globally fixed P_i and P_e are shown in Figure 9. Each data plotted here is the average over different P_i (Figure 9a) and P_e (Figure 9b) values. It is evident from the figures that overall, the proposed method detects the externally influenced nodes with very high accuracy compared to the baseline algorithms. As expected, NDFI performs better compared to the other structure-based algorithms. It is found that for P_i between 0.0 and 0.15, accuracy of NDIF is very high. However, as P_i increases the accuracy falls sharply and from $P_i = 0.4$ it stabilizes around 0.1. On the other hand, the proposed method maintains high accuracy, with the lowest level of 0.6 for $P_i = 0.95$. Table 4 shows the average values of Precision, Recall and f-score for different methods. Note that in the case of proposed FRW method the FP is zero so the Precision is 1.0, and for the other methods FP = FN as we take the top-k nodes where k is equal to the number of desire seeds. Hence in the case of these comparing methods desire nodes (TP + FN) is equal to the retrieved nodes (TP + FP).

The results NDIF are better for smaller values of P_i, and as we increase cascade size from 0 to 4, the performance deteriorates. The results are shown in Figure 10. FRW's performance is also affected by the user-defined parameter a. The effect of different values of a is presented in Figure 11. Reducing a would produce higher accuracy. However, in practical scenarios, reducing a to very low values results can result in higher miss-classification.

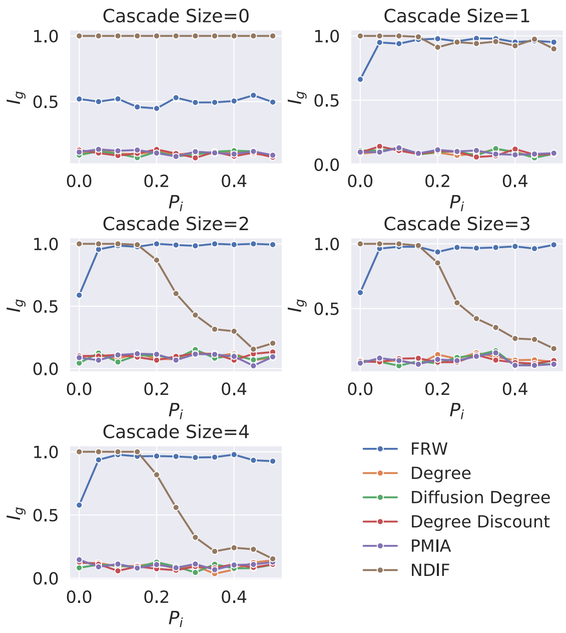


FIGURE 10. I_g **vs** P_i **for different maximum cascade size.**

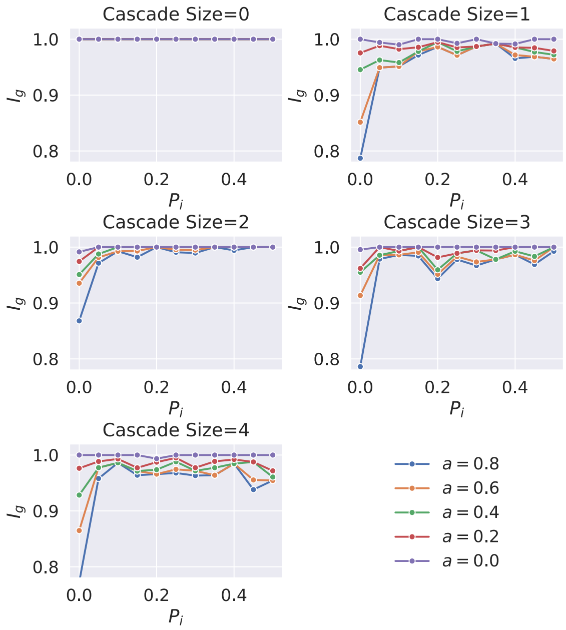


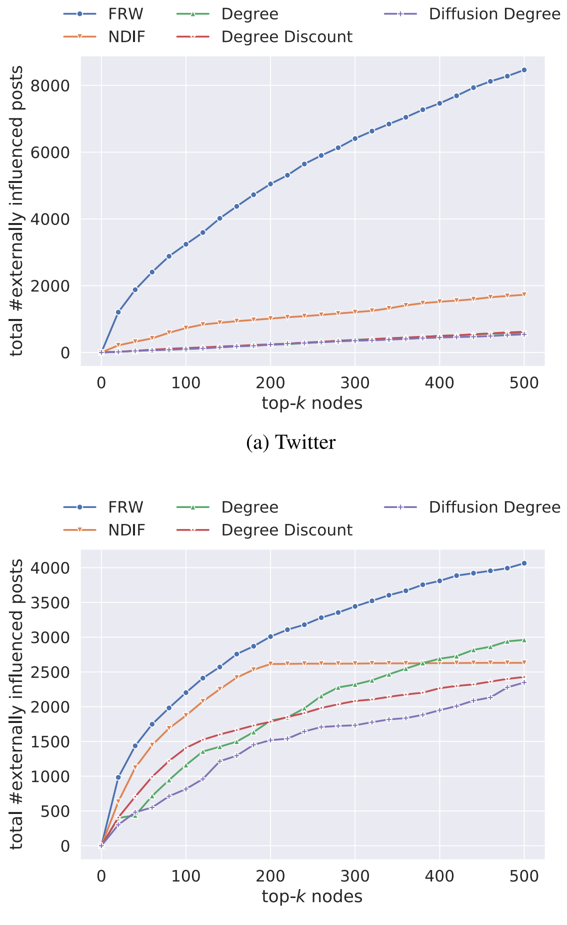
FIGURE 11. **Effect of** a **with maximum cascade size.**

D. PREDICTION OF TOP-*k* A CONTENT PRODUCER

In order to predict the top-*k* number of users who may create higher number of posts influenced by exogenous factors, we formulate the problem as a ranking problem. We split the data into two parts, namely, training set and testing set. With the training data, we selected the top-*k* nodes based on their FRW model predicted score. These top-*k* nodes created the most number of externally influenced posts in the network and intuitively it is expected that these nodes would continue to do so in the future as well. We verified this from the test data set. The experiment is conducted with two real-world social network viz. Twitter [36] and Weibo [37] network. The Twitter network contains reciprocal links of Twitter user. The data was collected between March 24, 2012 and April 25, 2012. The network properties are shown in Table 5. It is an undirected network along with the hashtags used in tweets. Each hashtag has been associated with a time stamp to indicate when it was used. Weibo data, on the other hand, is a following-follower network collected for 2009-2012. In our experiment, we took the content stream for the year 2011. In twitter, first 25 days (about 75% hashtags) were used as training data, and the rest (about 8 days) data was used as testing set. In Weibo, we used 11 months data for identifying the top-*k* content producer and the last month of the year is

TABLE 5. **Real-World Data Set.**

Properties/Name	**Twitter**	**Weibo**
Nodes	475,311	1,776,950
Edges	2,735,341	308,489,739
Avg. Degree	11.5097	173.61
Total #tag Posts	12,054,205	8,454,839
Total Days	33	365
Type	Undirected	Directed



(a) Twitter

(b) Weibo

FIGURE 12. **Comparative results of top-*k* content producer for different real-world data.**

used as a test set, similar to Twitter. Similarly, we get the top-k nodes by the comparing methods and total number of externally influenced posts in the test set is computed. These results are compared and shown in Figure 12. X-axis shows the value of k and the y-axis shows the number of externally influenced posts in the test set by the top-k nodes. It is evident that FRW performs better than the baseline algorithms to a large extent.

V. CONCLUSION

We considered a new problem of estimating users' willingness to adopt information from external sources in the context of information diffusion in the network. Given a social network and stream shared in the network, we provided an algorithm and a measure to identify the willingness of a node, i.e. openness of a node to external influences. The proposed method is called fuzzy relative willingness (FRW). Empirical analysis demonstrates that FRW is able to discover the externally influenced nodes with high accuracy. Comparative analysis with baseline algorithm reveals that FRW is better than the most similar state-of-the art approach for most of the cases.

The FRW generated scores are also used to predict the top-k content producer, that is the top-k nodes who share and spread the external information in the network. This concept was tested on both directed and undirected real-world networks, namely, Twitter (undirected) and Weibo (directed). Experimental studies demonstrate that our algorithm can better predict the number of future posts compared to all other methods.

Quantification of human willingness, i.e. openness, will provide a new research direction in the domain of information diffusion and target set (seed) selection. The presented algorithm is relatively simple, but the idea of representing willingness with fuzzy membership values will provide an opportunity to use other well-known membership functions to suit the data set and the applications they are used for.

REFERENCES

[1] M. Granovetter, "Threshold models of collective behavior," *Amer. J. Sociol.*, vol. 83, no. 6, pp. 1420–1443, May 1978.

[2] D. J. Watts, "A simple model of global cascades on random networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 9, pp. 5766–5771, Apr. 2002.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2003, p. 137.

[4] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis, "Infectious disease modeling of social contagion in networks," *PLOS Comput. Biol.*, vol. 6, no. 11, pp. 1–15, 2010.

[5] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Edmonton, AB, Canada, 2002, pp. 61–70.

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

[7] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen, "Maximizing influence in a competitive social network: A follower's perspective," in *Proc. 9th Int. Conf. Electron. Commerce (ICEC)*, Minneapolis, MN, USA, 2007, pp. 351–360.

[8] R. Narayanam and Y. Narahari, "A shapley value-based approach to discover influential nodes in social networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 8, no. 1, pp. 130–147, Jan. 2011.

[9] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Paris, France, 2009, pp. 199–208.

[10] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," *Data Mining Knowl. Discovery*, vol. 25, no. 3, pp. 545–576, Apr. 2012.

[11] S. K. Pal, S. Kundu, and C. A. Murthy, "Centrality measures, upper bound, and influence maximization in large scale directed social networks," *Fundamenta Informaticae*, vol. 130, no. 3, pp. 317–342, 2014.

[12] J. Jankowski, P. Bródka, P. Kazienko, B. K. Szymanski, R. Michalski, and T. Kajdanowicz, "Balancing speed and coverage by sequential seeding in complex networks," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 891.

[13] J. Jankowski, P. Bródka, R. Michalski, and P. Kazienko, "Seeds buffering for information spreading processes," in *Proc. Int. Conf. Social Informat. (SocInfo)*. Cham, Switzerland: Springer, 2017, pp. 628–641.

[14] J. Jankowski, B. K. Szymanski, P. Kazienko, R. Michalski, and P. Bródka, "Probing limits of information spread with sequential seeding," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 13996.

[15] J. Jankowski, M. Waniek, A. Alshamsi, P. Bródka, and R. Michalski, "Strategic distribution of seeds to support diffusion in complex networks," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0205130.

[16] F. Montes, A. M. Jaramillo, J. D. Meisel, A. Diaz-Guilera, J. A. Valdivia, O. L. Sarmiento, and R. Zarama, "Benchmarking seeding strategies for spreading processes in social networks: An interplay between influencers, topologies and sizes," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 3666.

[17] R. Michalski, J. Jankowski, and P. Brodka, "Effective influence spreading in temporal networks with sequential seeding," *IEEE Access*, vol. 8, pp. 151208–151218, 2020.

[18] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Beijing, China, 2012, pp. 33–41.

[19] J. Li, J. Xiong, and X. Wang, "Measuring the external influence in information diffusion," in *Proc. 16th IEEE Int. Conf. Mobile Data Manage.*, Pittsburgh, PA, USA, vol. 2, Jun. 2015, pp. 92–97.

[20] E. M. Rogers, *Diffusion of Innovations*, 5th ed. New York, NY, USA: Free Press, 2003.

[21] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, 2001, pp. 57–66.

[22] P. Bródka, K. Musial, and J. Jankowski, "Interacting spreading processes in multilayer networks: A systematic review," *IEEE Access*, vol. 8, pp. 10316–10341, 2020.

[23] Q. Wu and S. Chen, "Spreading of two interacting diseases in multiplex networks," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 30, no. 7, Jul. 2020, Art. no. 073115.

[24] Q. Liqing, G. Chunmei, Z. Shuang, T. Xiangbo, and Z. Mingjv, "TSIM: A two-stage selection algorithm for influence maximization in social networks," *IEEE Access*, vol. 8, pp. 12084–12095, 2020.

[25] P. Li, H. Nie, F. Yin, J. Liu, and D. Zhou, "Modeling and estimating user influence in social networks," *IEEE Access*, vol. 8, pp. 21943–21952, 2020.

[26] A. Anagnostopoulos, G. Brova, and E. Terzi, "Peer and authority pressure in information-propagation models," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Athens, Greece: Springer, 2011, pp. 76–91.

[27] M. Piškorec, N. Antulov-Fantulin, I. Miholić, T. Šmuc, and M. Šikić, "Modeling peer and external influence in online social networks: Case of 2013 referendum in croatia," in *Proc. Int. Conf. Complex Netw. Appl.* Lyon, France: Springer, 2017, pp. 1015–1027.

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

[29] S. Kundu, C. A. Murthy, and S. K. Pal, "A new centrality measure for influence maximization in social networks," in *Pattern Recognition and Machine Intelligence* (Lecture Notes in Computer Science), vol. 6744. Moscow, Russia: Springer-Verlag, 2011, pp. 242–247.

[30] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2010, pp. 1029–1038.

[31] K. Subbian, C. Aggarwal, and J. Srivastava, "Content-centric flow mining for influence analysis in social streams," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, San Francisco, CA, USA, 2013, pp. 841–846.

[32] X. Luo and R. Dimitrakopoulos, "Data-driven fuzzy analysis in quantitative mineral resource assessment," *Comput. Geosci.*, vol. 29, no. 1, pp. 3–13, Feb. 2003.

[33] A. Prat. (2015). *DATAGEN: Data Generation for the Social Network Benchmark*. [Online]. Available: http://ldbcouncil.org/blog/datagen-data-generation-social-network-benchmark

[34] J. M. Sánchez, C. A. Iglesias, and J. F. Sánchez-Rada, "Soil: An agent-based social simulator in Python for modelling and simulation of social networks," in *Advances in Practical Applications of Cyber-Physical Multi-Agent Systems: The PAAMS Collection*, Porto, Portugal. Cham, Switzerland: Springer, 2017, pp. 234–245.

[35] J. F. Sánchez-Rada. (2018). *Soil Tutorial*. [Online]. Available: https://github.com/gsi-upm/soil/blob/master/examples/tutorial/soil_tutorial.ipynb

[36] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Sci. Rep.*, vol. 3, no. 1, Dec. 2013, Art. no. 2522.

[37] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, "Social influence locality for modeling retweeting behaviors," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, 2013, pp. 2761–2767.

SUMAN KUNDU received the B.Tech. degree in information technology from the West Bengal University of Technology, Kolkata, India, in 2005, and the M.E. degree in software engineering from Jadavpur University, in 2009.

His Ph.D. Research was with the Center for Soft Computing Research, Indian Statistical Institute, from 2010 to 2015. He visited the Engine Group for the Postdoctoral Research, Wroclaw University of Science and Technology, from June 2018 to April 2019. He has more than six years of industrial software development experience with ZINFI Software Systems Private Ltd., Kolkata. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur. He has published eight articles in social network analysis, granular computing, and soft computing. His research interests includes social network analysis, network data science, soft computing, crowd sourcing, fuzzy and rough set, and granular computing.

TOMASZ KAJDANOWICZ received the M.Sc. and Ph.D. degrees in computer science from the Wroclaw University of Science and Technology, Wroclaw, Poland, in 2008 and 2012, respectively, and the D.Sc. degree from the Institute of Informatics, Polish Academy of Sciences, Warsaw, in 2020. He is currently a Professor with the Department of Computational Intelligence, Wroclaw University of Science and Technology. He has conducted research with Stanford University, USA, The University of Sydney, Australia, and the Technical University of Dortmund. His research interests include social network analysis, machine learning, and representation learning.

PRZEMYSLAW KAZIENKO (Senior Member, IEEE) is currently a Full Professor and a Leader with the European Centre for Data Science (ENGINE), Wroclaw University of Science and Technology, Poland. He has authored over 200 research articles, including 40 in journals with impact factor related to social network analysis, complex networks, spread of influence, emotion recognition, collective classification, machine learning, sentiment analysis, DSS in medicine, finances and telecommunication, knowledge management, collaborative systems, data mining, recommender systems, information retrieval, and data security. He gave 15 keynote/invited talks for international audience. He is on the Board of the Network Science Society. He served as the Co-Chair for over 20 international scientific conferences and workshops. He also serves as a member of the Editorial Board for *Social Network Analysis and Mining*, *Social Informatics*, the *International Journal of Knowledge Society Research*, and the *International Journal of Human Capital Management*.

NITESH CHAWLA (Senior Member, IEEE) started his tenure-track career in Notre Dame, in 2007. He held a Chaired Full Professor position for a period of nine years. He is currently a Frank M. Freimann Professor of computer science and engineering and the Director of the Research Center on Network and Data Sciences (iCeNSA), University of Notre Dame. He is also a Founder of Aunalytics, a data science software and solutions company. He was a recipient of several awards and honors, including the Outstanding Dissertation Award, the NIPS Classification Challenge Award, the IEEE CIS Outstanding Early Career Award, the IBM Watson Faculty Award, the IBM Big Data and Analytics Faculty Award, the National Academy of Engineering New Faculty Fellowship, and the First Source Bank Technology Commercialization Award. He was also recognized from the Rodney Ganey Award and the Michiana 40 Under 40.