An Interaction-Aware Approach for Social Influence Maximization

Diego Alonso, Ariel Monteserin and Luis Berdun

Abstract—**Microblogging networks are considered a great source of social influence. One of its characteristics is their high dynamism. This fact produces that influential users continuously change according with time and topic. Several social networks metrics have been defined to rank influential users. However, these metrics fail to capture the dynamism of microblogging networks. For this reason, we propose an approach based on Credit Distribution model to identify the influential users of a microblogging social network by performing an online analysis of the users' interactions. Moreover, we present a comparison of our approach with well-known metrics used for influencers ranking. The experiments were carried out in Twitter during sport events (football matches) and new product (video games) launchings. The results showed that our approach outperforms the metric-based rankings in terms of the influence spread. This confirms the importance of being updated for identifying influential users.**

Index Terms—**Social Influence Maximization; Social Network Modeling; Influencers Discovering; Viral Marketing**

I. Introduction

In recent years, there has been a growing interest in the analysis of the influence exerted by users and its spread. Due to the success of social networks as Twitter, specialists in areas such as marketing and computer science have been attracted by this topic ([1]–[3]). One of the key problems of the analysis of influence in social networks is the identification of influential users, i.e., users whose opinion impacts with greater force the tastes or actions of other users.

In order to determine who these influential users are, it is necessary to analyze the spread of influence on social networks. This analysis is very useful since it allows understanding how information is disseminated in a social network [4]. Among the applications that make use of this analysis, such as feed ranking and personalized recommendations, one of the most benefited is viral marketing [5], [6]. In brief, viral marketing aims to select a small set of influential users to adopt a product, in order to trigger a chain reaction of adoptions driven by the word-of-mouth effect of social networks [7].

Motivated by this type of marketing, Kempe et al. [8] posed the problem of influence maximization in social networks. This problem seeks to determine k nodes (called seeds) in the network, so that when activated, the propagation of the expected influence is maximized. In order to solve the influence maximization problem, several diffusion models of social networks are used [9]. Besides, the diffusion process has a time dimension, also known as innovation-decision process [10]. The traditional approaches for solving the influence maximization problem do not consider the time in the diffusion process. However, Goyal et al. [11] highlighted that the time is a key factor for the spread of influence between users. In similar fashion, some researchers ([1], [12]) build the propagation graph from real world log data where actions are timestamped. Nevertheless, as far as our knowledge, only static analysis approaches of the influence maximization problem have been proposed. For this reason, taking into account the dynamism of current social networks, there exist a crucial need to analyze the set of users who maximize the spread of influence in real time.

In this context, we present TSeedS (Twitter Seed Set), an approach based on Credit Distribution (CD) model [13] to identify the influential users of a social network in short-time batches in order to allow users to take marketing decisions in real time. Therefore, our research goal consists in demonstrating the importance of analyzing the set of users who maximize the spread of influence in real time by considering the user interactions, both for modeling the network and for detecting influencers. Besides, we aim to compare the ranking of influential users obtained with our approach with rankings based on well-known social network metrics. To evaluate our approach, we monitored the social network Twitter in two different circumstances, related to console game launches and international sport events. Experimental results provide encouraging evidence for the feasibility of the approach and show the importance of dynamically analyzing the social influence in microblogging networks. Moreover, experiments also show that our interaction-aware model outperforms the spread of influence over the spread achieved by well-known ranking metrics.

The rest of the paper is organized as follows. Section II describes a complete background, including important concept definitions, related works, microblogging networks, ranking metrics, and diffusion models. Section III details the proposed approach. Section IV shows the study cases used to evaluate the approach, the experimental results and discussion. Finally, Section V presents the conclusions and future works.

II. Background

In this Section, we first introduce the key concepts to fully understand the context of our research. In particular, we define our perspective of what is considered influence in social networks and we distinguish it from related concepts such as homophily, correlation, environment, among others. Then, we discuss about social networks and its diverse classifications,

Diego Alonso, Ariel Monteserin and Luis Berdún are with the ISISTAN Research Institute (CONICET-UNCPBA), Campus Universitario, Tandil, Bs. As., Argentina e-mail: ariel.monteserin@isistan.unicen.edu.ar

e.g., static networks and dynamic networks. At this point, we talk about microblogging social networks such as Twitter, which are the type of social networks that we use to evaluate our approach. Next, we analyze the current state-of-art on ranking metrics, including graph-based metrics, Twitter-based metrics, among others. In the last subsection, we present different diffusion models and related works that have been proposed for detecting influential users in social networks. Particularly, we describe the CD model, which is the one that we select for our approach.

A. Important Concepts

First, it is important to define what is considered influence in social networks. One of the most popular conceptions indicates that users are being influenced when performing an action that they see it has been performed before by one of their friends. Rashotte in [14] defines social influence as the change in an individual's thoughts, feelings, attitudes, or behaviors that results from the interaction with another individual or a group. However, when a user performs an action, he/she may have other reasons than influence. For example, it might be possible that the user heard about the action outside the social network or maybe the action is too popular in itself. For this reason, we decide to take as a definition of influence a perspective based on the monitoring of the relations of the users in the networks. As described by Bonchi in [4], if we observe a user v performing an action a at a time t, and a user u (which has a relation with v) performs the same action in a short time delay, say $t+\Delta$, then we can think that the action a was spread from v to u. If we observe that this happens frequently for different actions, then we can conclude that user v is exerting influence on user u, and user v becomes an influential user.

Correctly identifying influence in social networks is important. One of the various applications that make use of influence in social networks is viral marketing. This type of marketing aims to produce exponential increases in the number of people who know (acquire) a brand with the less possible effort (cost) [15]. Typically, the marketing staff decides whether or not to market to an individual based on their characteristics (direct marketing). However, this approach leads to suboptimal marketing decisions by not taking into account the effect that members of a market have on each other's purchasing decisions [5]. At this point, viral marketing takes great advantage by using the word-of-mouth effect of social networks [16]. In other words, viral marketing aims to select a small set of influential users to adopt a product, and subsequently trigger a cascade of further adoptions ([7], [17]). Motivated by this, in [8], the authors pose an algorithmic problem for social networks: "If we can try to convince a subset of individuals to acquire a new product or innovation, and the objective is to trigger a cascade of future adoptions, to which set of individuals should we aim?". In formal terms, Kempe et al. defined the previous as the problem of influence maximization.

On the other hand, time is a fundamental factor of the influence analysis. It is proven that a sublogarithmic time is sufficient to propagate a novelty to all the nodes of the network [18], [19]. It is also argued that, the instant nature of these networks influences the speed at which these events unfold [20]. In [17], the authors proposed the time-constrained influence maximization problem. They showed that in many viral marketing applications, it is crucial to consider the spread of influence before a fixed time. Chen et al. [21] and Shi et al. [22] modeled this problem as maximizing the influence spread on a bounded time. Nonetheless, the conventional influence maximization models do not consider the time dimension. Several authors approached this issue by adapting the set of influential users in different intervals of time ([23]–[25]).

B. Microblogging Social Networks

Nowadays, the most popular online social networks have millions of active users. Typically, the vertices in these online social networks are the users and the edges represent friendships or subscriptions. For our experiments we decide to use Twitter social network, which since 2010 is being studied extensively in the contexts of social network analysis, computer science, and sociology [26], [27].

Twitter platform currently generates approximately 65 million tweets (posts on this platform) per day. In particular, Twitter provides microblogging services which allow users to share short texts, videos or images with other users and it is characterized for pointing to simplicity and synthesis. Essentially, users can post and read messages called tweets. Furthermore, Twitter allows users to highlight words by preceding them with a '#' symbol, these words are known as *hashtags*. The use of hashtags allows global discussion of diverse topics by grouping the posts that use the same tag. Moreover, Twitter analyzes the amount of posts with the same *hashtag* in real time and presents a list of them, each of these tags is known as *Trending Topic* (TT). Thus, a user can visualize which are the most important topics being discussed in each moment and, read and comment about them.

Moreover, the platform offers three ways of interaction between the users and the tweets. First, users can comment other user posts. This action is also known as reply. Second, users can like other user posts, which will be added to their liked post list. Third, user posts can be re-posted by other users, this action is called retweet. For the relationships between users, the platform allows one-way and two-way connections. The platform manages these connections as list of *followers* and list of *following*. For example, a user u can connect with a user v, and this means that the user u is subscribing to user v posts. In this case, user v belongs to the list of following of user u, and user u belongs to the list of *followers* of user v.

With all the aforementioned features and services, the Twitter platform became one of the social networks with the highest number of active users. Due to its synthesis and simplicity, Twitter is well-known for the velocity of the real-time propagation of topics and news. That is why, two of the most studied problems in Twitter are the identification of influential users and the analysis of the information spreading [28]. In order to identify the influential users on Twitter, several authors proposed a huge number of techniques and metrics to rank them. In the next subsection, we present many of these metrics.

C. Ranking Metrics

As mentioned before, social networks have immensely grown and identifying influential users has become one of the most important problems to solve. Several authors have addressed this problem and have proposed the use of metrics. Centrality metrics such as *degree*, *betweenness*, *eigenvector*, and *closeness* are the most traditional. The degree of a node (user) refers to its number of adjacent edges [29]. The closeness of a node is based on its distance to all the other nodes, i.e., considers the sum of the shortest paths of a node to all the others [29]. In similar fashion, the betweenness of a user refers to the number of shortest paths in which it is present [29], [30]. In addition, there exist several metrics based on eigenvector, such as PageRank [31] and TrueTop [32]. It is important to note, that centrality measures are different, i.e., are not correlated, despite their conceptual similarities.

However, owing to the high-complex structures of these social networks, in many cases, obtaining centrality metrics is suboptimal or non-practical. Moreover, these type of metrics do not take into account all the available information of the interactions among users, since they are based on graphs and not on the particular services provided by Twitter. At this point, Pal and Counts in [33] introduce a list of metrics based on Twitter features including tweets, interactions, and graph characteristics. For example, Retweet Impact, which is computed as $RI = RT2 * log(RT3)$, considers the number of unique tweets of the user (author) that have been retweeted (*RT2*) with the number of unique users who retweeted the author's tweets (*RT3*).

All these metrics allow us to rank users from social network data at a certain period of time. Analyzing the variation of these metrics in the different periods of time is fundamental for the identification of influential users. Nevertheless, it is important to note that, for real-time applications, these metrics could be really difficult to compute. For a complete survey about metrics of influence on Twitter, see [26].

D. Diffusion Models

In order to determine how influence is spread, which factors to take into account, and with what probability the influence spreads from one node to another, a variety of diffusion models have been proposed over time [34]. For example, in [5], the authors modeled the influence maximization problem using Markov random fields and proposing heuristics for the influential users selection. In particular, authors aim to augment the expected profit compared to the expected profit of no applying any marketing strategy. For this purpose, they distinguished two types of client values. On the one hand, the intrinsic value, which is its value based on the products that he/she usually consumes. On the other hand, the network value, which is higher when he/she positively influences the purchasing probabilities of other users.

Two of the fundamental propagation models are the threshold model and the cascade model. In both models, at each given time (discrete steps), each node is active or inactive and its tendency to activate increases in a monotonous way as the number of active neighbors augments. A node is considered active when it has been influenced or convinced of what it is desired to spread. In addition, these models assign probabilities according to a defined criterion. For instance, in the Linear Threshold model (LT), from the family of the threshold models, each node has attached a random value between 0 and 1 that indicates the percentage of neighbors that must be convinced for its activation. In the Independent Cascade model (IC), from the family of the cascade models, each node has only a single chance to activate a neighbor node with a certain probability. For example, this probability can be assigned taking into account the degree of the node (number of edges relate to it). In similar fashion, the Trivalency model (TR) assigns probabilities to each directed edge over a set of values indicating different types of relation between nodes. For example, the set $\{0.1; 0.01; 0.001\}$ could indicate three levels of influence (high, medium, low). Besides, several authors approached the influence maximization problem by using or adapting these well-known models [17], [35]–[37].

Nonetheless, one of the limitations of these models is that the edge-weighted social graph is assumed as an input to the problem, without addressing the question of how the probabilities are obtained [11]. For this reason, Goyal et al. [13] approached the influence maximization problem considering a data-based perspective. They proposed the Credit Distribution model (CD), which learns how influence flows in a network by directly leveraging available propagation traces. Experimental results showed that this approach outperforms IC and LT models, in terms of predicted seed precision, seed selection quality, run time, and scalability. It is also noteworthy, that this approach is time-aware because it takes the temporal nature of influence into account. For this reasons, we decided to use the CD model for our approach. In Section III, we detail the CD model and explain our approach.

III. TSeedS Approach

In order to identify the influential users of a microblogging social network in short periods of time, we propose a novel approach based on the CD model. Fig. 1 illustrates our approach, which we called TSeedS in reference to Twitter Seed Set. TSeedS intends to avoid some of the limitations of traditional ranking metrics by taking into account the time dimension and the content of the posts. In other words, the influential users of a social network could vary because of the topic and the time when they performed the interactions. Moreover, our approach considers the dynamism of this kind of social networks, which is an important factor for taking viral marketing decisions. At this point, a dynamic social network is a multigraph $G = (V, E)$, where E is a set of edges, and each timestamped edge $(u, v)_t \in E$ represents an interaction (u, v) that occurred at time $t \in N$ [38].

At first, our approach focuses in gathering information of a social network in real time. As mentioned before, we decided to use Twitter for our experiments. However, it is important to note that TSeedS provides support for using a different social network of microblogging. TSeedS subscribes to the streaming API of Twitter in order to listen for network interactions (*tweets*, *retweets*, *replies*) in real time. Particularly, *tweets* are

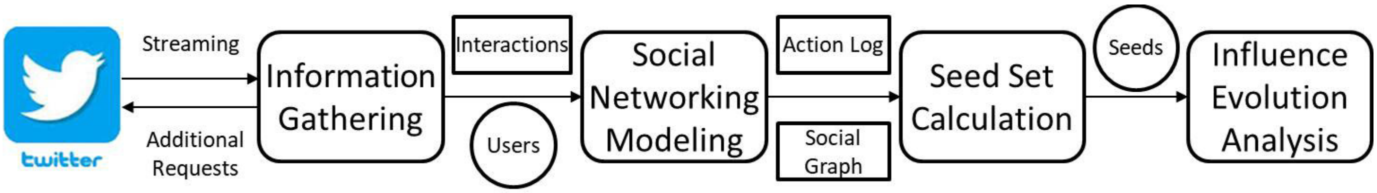

Fig. 1. The four main steps of the TSeedS approach: gathering information, modeling the social network, obtaining influential users (seed set) and analyzing the evolution of influence.

considered original actions, while *retweets* and *replies* are considered propagated actions. It is important to note that, since we are going to work on a topic and not with the totality of publications, i.e. not all the interactions listened are considered valid, the interactions must respect a set of filters previously specified by an expert. To do this, a set of keywords to represent the interest category must be defined in TSeedS. In addition, TSeedS uses a configurable time filter in order to define a window of temporary validity for the actions. For example, considering as a valid interaction the one that its time of original publication does not exceed 60 minutes with respect to the current time. Thus, if TSeedS identifies an interaction (*retweet*, *reply*) which its time of original publication refers to a period greater than the previous 60 minutes, the interaction is discarded.

Once a valid interaction is listened, TSeedS proceeds to establish the relationship between the users involved in a structure called Social Graph and to register the information of the action in a structure called Action Log. The Social Graph (SG) is a representation of a directed graph based on nodes (users) and edges (relationships). If the user involved does not exist in the SG, TSeedS creates a new node representing the user. Furthermore, if the interaction is a propagated action, TSeedS establishes a directed edge from the user who originally performed the action to the one that propagated it. However, our perspective of influence, described in subsection II-A, states that there must be a social relationship between users to consider that influence is spreading. In other words, it is not enough the propagation of an action from one user to the other to relate them in the graph, but it is also necessary to verify that there is a social connection between them. On the Twitter platform, relationships are based on followers and followed, which is similar to a subscription to the user's posts that follows. Therefore, before establishing the edge between the nodes, TSeedS verifies if the user of the interaction is socially linked to the user of the original action.

Nevertheless, due to the fact that the Twitter platform allows users to perform a retweet of a retweet, there may not be a direct link between the interacting user and the original. For this reason, TSeedS analyzes the path established between the user that propagated the action and the original user that posted it, adding the intermediate users that are necessary. For practical purposes, TSeedS analyzes if the user that interacts is socially related to any other user that has recently performed the same retweet. This process is recursively repeated until a direct link is found between a user who interacts and the one who performed the action originally. While incorporating users to the SG, the interaction is also recorded in the Action Log (AL). The AL is a register of all the propagations, in which each entry refers to an action, both the original actions and the propagated ones. Each tuple of the AL is composed by the user id, the action performed id, and the time when the action was performed.

The next stage is to identify the influential users in the generated network. As mentioned before, we chose the CD model to solve this problem [13]. Thus, the problem of influence maximization to be solved under the CD model is reformulated as follows: given a directed graph $SG = (V, E)$, an Action Log AL, and an integer $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that $\sigma_{cd}(S)$ is maximum. σ_{cd} is the expected spread of influence and is computed using the Equation 1, where $K_{S,u}$ represents the total credits assigned to S for influencing the user *u* in all actions. As a result, $\sigma_{cd}(S)$ is the total influence propagated by the users included in the set *S*. For a pair of users *v* and *u*, the average credit given to *v* for influencing *u*, over all actions that *u* performs is denoted by Eq. 2, where A_u is the number of actions performed by user *u* and $\Gamma_{(v,u)}(a)$ is the total credit given to *v* for influencing *u* on action *a*.

$$\sigma_{cd}(S) = \sum_{u \in V} K_{S,u} \tag{1}$$

$$Kv, u = \frac{1}{A_u} \sum_{a \in A} \Gamma_{v,u}(a) \tag{2}$$

$\Gamma_{(v,u)}(a)$ is calculated as described in Eq. 3, where $\gamma_{(w,u)}(a)$ indicates the direct credit given by *u* to a neighbor *w* for action *a* and $N_{in}(u, a)$ is the set of neighbors of *u* which activated on action *a* before. The direct credits are computed by Eq. 4, where $infl(u)$ refers to the user influenciability, i.e. the rate of actions that *u* performs under the influence of at least one of its neighbors; $\tau_{(v,u)}$ is the average time taken for actions to propagate from user *u* to user *v*; and $t(x, a)$ is the time in which user *x* performed action *a*.

$$\Gamma_{(v,u)} = \sum_{w \in N_{in}(u,a)} \Gamma_{v,w}(a).\gamma_{w,u}(a) \tag{3}$$

$$\gamma_{w,u}(a) = \frac{infl(u)}{|N_{in}(u,a)|} \,.\, exp\left(-\frac{t(u,a) - t(v,a)}{\tau_{v,u}}\right) \tag{4}$$

In order to solve the problem of social influence maximization, Goyal et al. [13] developed an algorithm that works under the CD model by scanning the action log AL to learn the influence probabilities in the social network and computing influenciability scores for the users. Then, the seed set is selected under the CD model by using a greedy algorithm with

CELF optimization [39] according to a set of training actions. See [13] for further details on algorithm implementation.

Obtaining the set of influential users (seeds) in short periods of time allows TSeedS to visualize the evolution of the different selected users. It is noteworthy that, each seed is in a particular position of the seed set according to the marginal gain (total credits reached by the user). To summarize, the approach aims to identify the influential users in a specific topic in order to maximize the propagation of the influence and also allows to visualize the evolution of user's influenciability over time.

IV. Experiments

To evaluate our approach, we ran experiments comparing the performance of TSeedS with the performance of other metrics that allow us to ranking influential users in social networks. In particular, we studied two different cases of the real world: sport events and new console games launching. In subsection IV-A, we explain how we generated our datasets and describe them. In addition, we describe the evaluation method by analyzing the metrics that were used for comparison. In subsection IV-B, we show the obtained results and perform a comparative analysis.

A. *Experimental Setup*

As mentioned before, we studied two real-world scenarios in which the application of viral marketing techniques would be of interest. For the sport events and the new console games launching, we listened to the social network Twitter and analyzed the activity by considering different marks of time. In the following subsections we analyze the datasets and the evaluation method.

1) Datasets : To generate the datasets, we used the Twitter Streaming API. This API allows us to establish a direct channel with the global flow of Twitter interactions in real time. On the one hand, we studied the case of televised sport events. In particular, we analyzed four matches of the UEFA Champions League. Two of these matches were simultaneously played on March 14, 2017 (Event Day 1) between 4:45pm and 6:30pm (GMT-3). The other two matches were also simultaneously played in the same hourly range but on March 15, 2017 (Event Day 2). These sport events were of great repercussion since they were play-offs of a tournament of world interest. We will refer to this dataset as Champions.

On the other hand, we studied the case of the launch of new products. In particular, we analyzed the launch of two games of the Playstation 4 (PS4) console. One of the games was the FIFA 18, a very popular football game, that was launched for PS4 on September 29, 2017 at 2:00pm (GMT-3). The other console game was the NieR: Automata, an action role-playing game, that was launched for PS4 on March 10, 2017 at 2:00pm (GMT-3). In Table I, we summarize the main characteristics of the datasets used for the experiments. By comparison, sport events allowed us to analyze our approach during a long-duration event, while new console games launching allowed us to analyze our approach during a short-duration event. Moreover, we covered two types of event: real life events (a football match) and virtual events (a game launching), which is an event that occurred entirely on social media.

TABLE I
Datasets Summary.

Event	#Users	#Tweets	#Retweets	#Interactions
Champions Day 1	122062	89428	112810	202238
Champions Day 2	81991	62763	66008	128771
FIFA 18	160545	73945	162671	236616
NieR: Automata	36974	49206	17989	67195

2) Evaluation Method: For validating our proposal, we compared the true influence spread (TIS) obtained by different seed sets. The TIS measures how many nodes of the network will be activated after the seed nodes are activated. In particular, we compared the evolution of the TIS of the user rankings generated by our approach with the ones based on the following metrics: (a) Closeness Centrality; (b) Retweet Impact; (c) Degree; (d) Betweenness; (e) PageRank. The Closeness Centrality (CC) is the length of the shortest paths from a node i to everyone else. CC is obtained by $CC(i) = 1 - n \ / \ \sum_{i \neq j} D_{i,j}$, where D is the distance matrix of a network with n nodes. If there is no path from i to j, then we assume that $D_{i,j} = n$. Retweet Impact (RI) is obtained as $RT2 * log(RT3)$, where *RT2* is the number of unique *tweets* retweeted by other users, and *RT3* is the number of unique users who retweeted author's tweets. The Degree (DE) of a user refers to its number of adjacent edges. The Betweenness (BT) of a user refers to the number of shortest paths (the ones calculated for CC) in which it is present. The PageRank (PR) is a metric based on the eigenvector centrality measure, which considers the number of connections of a user and the quality of the users who are connected with it.

In order to analyze the evolution of the TIS, we considered the results obtained in one-hour batches. Moreover, it should be noted that, we assumed two hours as the validity time of an interaction. This means that a retweet made more than two hours after the original tweet has been published will be discarded for the influence analysis. For instance, if we want to identify the influential users at 4 p.m. we will consider valid all those tweets that have been made from 2 p.m. onwards and all those *retweets* of the tweets that belong to this range. We defined this assumption as the sliding window of validity of an interaction. We decided to use a two-hour window of validity after running a variability test of the seed sets obtained by our approach. With this analysis, we discovered that the variability of datasets decreased substantially over time. In particular, with a window of five hours of validity, the variability was about 9%, while with the current window the seed sets generated for the different datasets vary between 48% and 53%. In addition, we did not consider validity windows of less than two hours due to the decrease in the number of interactions. It is also important to take into account the domain that is being analyzed, e.g., in the case of events, consider its duration. Therefore, we have decided to consider validity windows of two hours because they improve the variability by holding a representative number of interactions to be updated. It is worth noticing that, the validity window is a parameter that can be modified by an expert according to the domain of application.

B. Experimental Results and Discussion

We analyzed the TIS from two points of view. The first one consisted in analyzing the variation of the TIS over the hours. The second one consisted in analyzing the variation of the TIS in each specific hour varying the size of the seed set. We used a seed set size of 50 users because it is a common number used in the area of influence maximization [13]. Although this number may seem low in relation to the large number of users of the social network, we believe that it is enough since only a few users exert a great influence. In the following, we analyzed the experimental results obtained according to the proposed cases of study. It is important to highlight that, the TIS obtained by our approach is significantly different from the TIS obtained by the other techniques with a p-value < 0.01 in all the ran experiments.

1) Sport Events: Fig. 2 and 3 show the evolution of the TIS considering a seed set size of 50 users for each day of the sport event. Although both events were very similar, analyzing Table I that summarizes the datasets we can see that the event of day two has had a smaller impact than the event of day one. However, both events have sustained a large influence presence almost doubling the amount of retweets over the number of tweets. It is important to note that, the transversal lines in the graphs mark the beginning and the end of the events.

At a general level, the streamed events show a presence of increasing influence but with a peak achieved after the end of the events. In the Fig. 2 and 3, the graphs show that our approach obtains better results than those obtained by metric-based rankings selecting 50 users at all hours. We can observe that the difference in the TIS achieved is very pronounced just in the hours with the greatest presence of influence in the network, i.e. from 7 p.m. to 9 p.m. This analysis can be deepened and related to what is observed in the Fig. 4, 5, 6 and 7.

Fig. 4-7 show the evolution of the TIS as more seeds are selected in two particular moments: during the event at 6 p.m. and post-event at 9 p.m. The graphs in these figures indicate that in the moments after the event, the selection of the first seven to ten users was essential to achieve a greater TIS. In this figure, it is also evident that in hours of higher influence presence, such as the post-event moment (9 p.m.), the difference between our approach and the other techniques is very noticeable and sustained. This may be due to the fact that each new user that is added to the seed set already makes an important contribution of influence to the network. On the other hand, during event moments, in which the presence of influence is relatively minor, a kind of convergence can be identified between 25 users and 35 users, that is, the contribution made by them is no longer as significant between 25/35 users and 50 users. Notice that this reinforces the choice of a seed set size of 50.

The ranking based on RI, which is dynamic because it is based on the interactions performed, has achieved good results with respect to the other techniques analyzed although it has not exceeded the performance of our approach. It should be noted that the RI-based rankings are closer to our approach when there is less influence on the network, since the most influential users are somewhat more obvious or easy to identify due to the lack of network interaction. Similar results were obtained by DE, though this is a static metric, because the graph is built from the users' interaction (an increment in the interaction produces an increment in the degree). This is not the case of the ranking based on CC, since in the moments where the presence of influence is lower it does not select the ideal users. This may be because the users that should be activated are not necessarily the most strongly connected.

TABLE II
BEST TIS VALUES OBTAINED BY EACH APPROACH.

Event	TSeedS	RI	CC	BT	DE	PR
Champ. Day 1	5532	3635	3140	3592	5121	4362
Champ. Day 2	4270	2914	3590	3590	3986	3725
FIFA 18	4483	4462	3725	3725	3814	3610
NieR: Automata	1380	1232	1379	1232	1236	1172

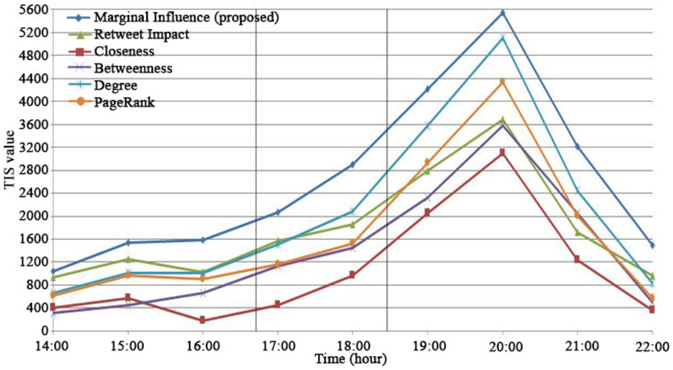

Fig. 2. True influence spread per hours during event 1 considering 50 users.

Table II summarizes the best results obtained by each approach in each event. In summary, our approach achieves better results because it considers the dynamism of the graph, the number of interactions, and the time of propagation of those interactions. While the ranking based on RI considers the dynamism of the graph and the number of interactions, and the ranking based on CC and PR does not consider the interactions and only considers the graph in a static way.

2) New Console Games Launching: While both are games of the same console, analyzing Table I, we can see that the FIFA18 had a greater impact (generating a lot of interactions in a short time) than the one that had the NieR: Automata. We identify that a higher presence of influence was evident in the FIFA18 network, while this did not happen in the NieR: Automata network.

Fig. 8 and 9 show the evolution of the TIS considering a seed set size of 50 users for each console game. In this case, the difference achieved by our approach on the less influential network (NieR: Automata) was not significant. However, the results obtained by our approach were not surpassed by any of the techniques with which it was compared. The biggest

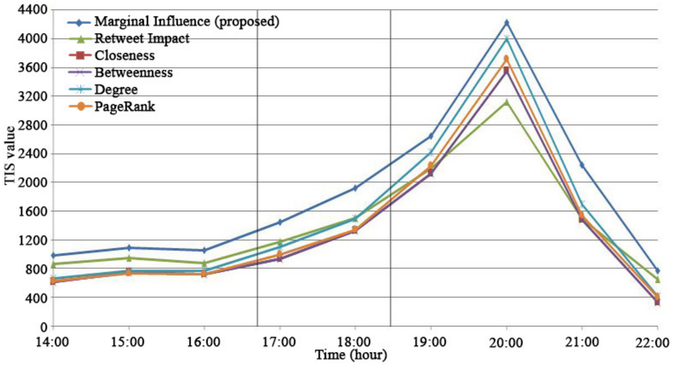

Fig. 3. True influence spread per hours during event 2 considering 50 users.

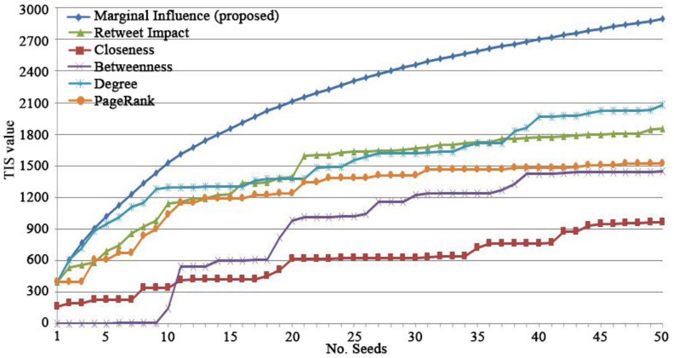

Fig. 4. True influence spread per seeds during event 1 at 6 p.m.

differences for the case of NieR: Automata were achieved with a smaller number of users and when there was a greater presence of influence. That is to say, the convergence (in terms of the number of users who made important contributions of influence) in low-influence networks occurs earlier than in networks with a high presence of influence. Thus, our approach was notoriously superior to the other techniques used when no more than 8 to 10 users were selected. In addition, this analysis corresponds to the NieR: Automata graphs of Fig. 11 and 13, where it is clearly more evident that the convergence of the influence was between 10 users to 20 users.

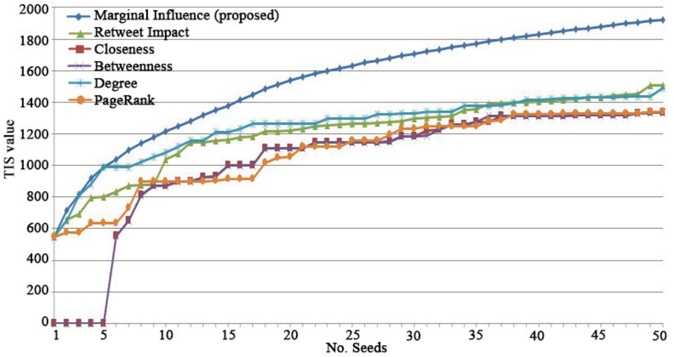

Fig. 5. True influence spread per seeds during event 2 at 6 p.m.

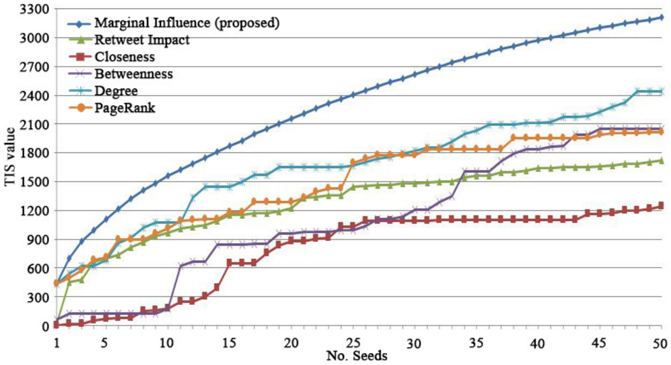

Fig. 6. True influence spread per seeds after event 1 at 9 p.m.

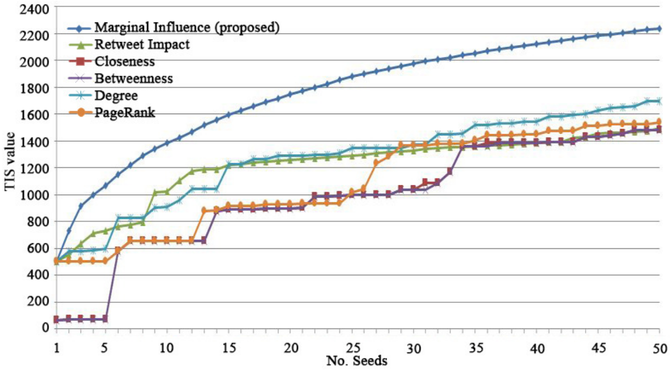

Fig. 7. True influence spread per seeds after event 2 at 9 p.m.

In the case of FIFA18, Fig. 10 and 12 show that our approach had great advantages in the selection of the first seeds when there was a large presence of influence. Analogously to the previously studied cases, our approach was not surpassed by any of the techniques in comparison, even when 50 seeds were considered. This fact can be seen, in Fig. 10, where the convergence of influence, when there is a greater presence of influence (2 p.m.), was with more than 20 users. However, when there was a less presence of influence (6 p.m.), convergence starts at 12 users to 15 users.

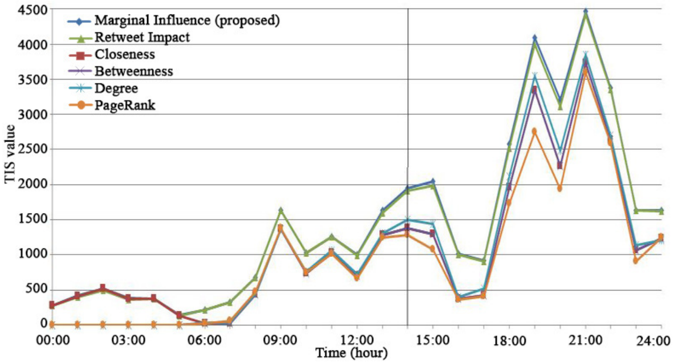

Fig. 8. True influence spread per hours during FIFA18 launching by considering 50 users.

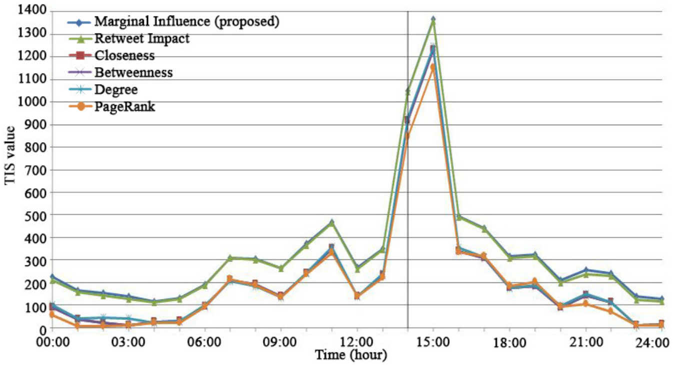

Fig. 9. True influence spread per hours during NieR: Automata launching by considering 50 users.

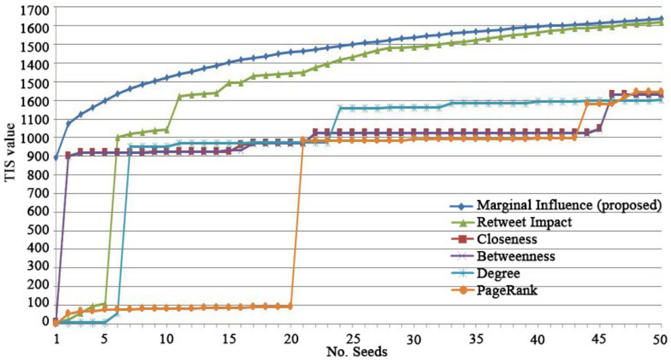

Fig. 12. True influence spread per seeds after FIFA launching at 6 pm.

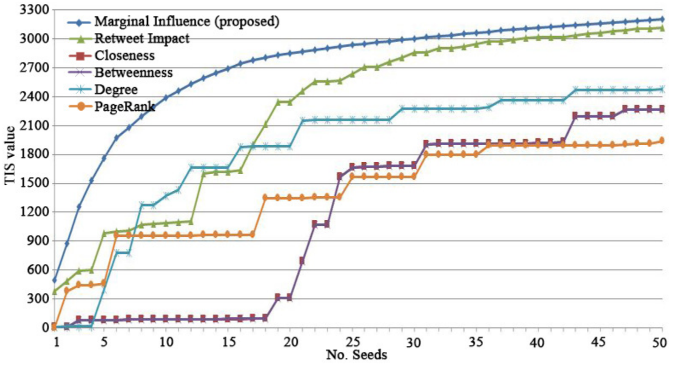

Fig. 10. True influence spread per seeds during FIFA launching at 2 p.m.

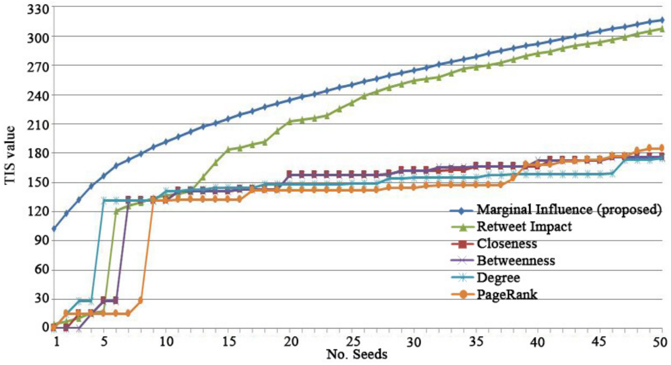

Fig. 13. True influence spread per seeds after NieR: Automata launching at 6 pm.

V. Conclusions and Future Work

In this paper, we present a comparative analysis of techniques to detect influential users in microblogging networks during events with high dynamism. We propose an approach for social network modeling and social influence maximization based on the Credit Distribution model. Moreover, we compare the results obtained by our approach with a set of well-known social network metrics. In short, our approach outperforms the rankings based on the well-known metrics analyzed under any scenario (low or high presence of influence, sustained influence or with marked times). However, our approach makes a greater difference in networks with a high presence of influence. It also shows great advantages for the identification of the most influential users within the influential users, e.g., in our experiments the first 10 users, even in networks with low presence of influence. One of the contributions of the application of our approach is the modeling of the network based on the interactions between users and the identification of intermediate users which allows to make a more realistic influence analysis. Moreover, it is important to remark on the relevance that the proposed approach gives to the time factor, so important for the definition of viral marketing strategies, and the fact that using this approach we could maximize the influence in real-time, which is also vital for viral marketing. Finally, we have shown the importance of computing the influence in different time windows in order to capture the high dynamism of the microblogging social networks.

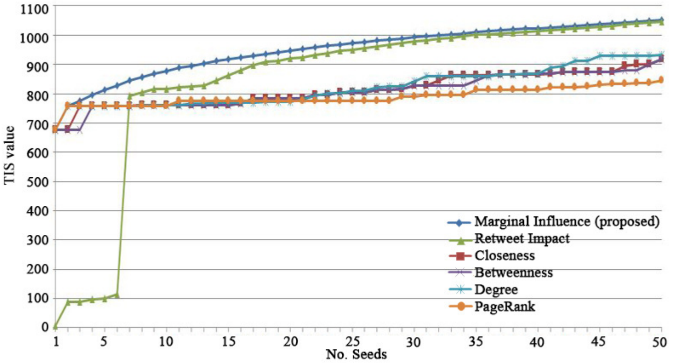

Fig. 11. True influence spread per seeds during NieR: Automata launching at 2 p.m.

Future work will focus on modifying the CD technique to add a budget and a cost function of the users. This way, we could make a more interesting analysis for viral marketing. On the other hand, research lines related to the analysis of the content of interactions could also be analyzed. For example, profiling users and labeling their interactions to perform sentiment analysis or to maximize negative or positive influence. In addition, it would be interesting to identify influential words, phrases, or sentence structures by applying Natural Language Processing techniques to the content of the interactions.

References

[1] M. Gomez-Rodriguez, L. Song, N. Du, H. Zha, and B. Schölkopf, "Influence estimation and maximization in continuous-time diffusion networks," *ACM Trans. Inf. Syst.*, vol. 34, no. 2, pp. 1–33, 2016.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining*, ser. WSDM '11. NY, USA: ACM, 2011, pp. 65–74.

[3] K. Devi and R. Tripathi, "An ltirs model for influence diffusion process," in *2022 14th Int. Conf. on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2022, pp. 285–289.

[4] F. Bonchi, "Influence propagation in social networks: A data mining perspective," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM Int. Conf. on*, vol. 1, Aug 2011, pp. 2–2.

[5] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '02. NY, USA: ACM, 2002, pp. 61–70.

[6] C. Aslay, L. V. Lakshmanan, W. Lu, and X. Xiao, "Influence maximization in online social networks," in *Proceedings of the 11 ACM Int. Conf. on Web Search and Data Mining*, ser. WSDM 18. New York, NY, USA: ACM, 2018, pp. 775–776.

[7] V. Mahajan, E. Muller, and F. M. Bass, "New product diffusion models in marketing: A review and directions for research," *Journal of Marketing*, vol. 54, no. 1, pp. 1–26, 1990.

[8] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '03. NY, USA: ACM, 2003, pp. 137–146.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[10] M. R. Everett, "Diffusion of innovations," *Simon and Schuster*, vol. 12, 2010.

[11] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. of the Third ACM Int. Conf. on Web Search and Data Mining*, ser. WSDM '10. NY, USA: ACM, 2010, pp. 241–250.

[12] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting information diffusion models over social networks for behavioral analysis," in *Proc. of the 2010 European Conf. on Machine Learning and Knowledge Discovery in Databases: Part III*, ser. ECML PKDD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 180–195.

[13] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *PVLDB*, vol. 5, no. 1, pp. 73–84, 2011.

[14] L. Rashotte, "Social influence," *The Blackwell encyclopedia of sociology*, 2007.

[15] C. Long and R. C.-W. Wong, "Viral marketing for dedicated customers," *Inf. Syst.*, vol. 46, pp. 1–23, dec 2014.

[16] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, pp. 211–223, 2001.

[17] B. Liu, G. Cong, Y. Zeng, D. Xu, and Y. M. Chee, "Influence spreading path and its application to the time constrained social influence maximization problem and beyond," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1904–1917, 2014.

[18] B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, pp. 70–75, jun 2012.

[19] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, "Dynamics of information diffusion and its applications on complex networks," *Physics Reports*, vol. 651, pp. 1–34, 2016, dynamics of information diffusion and its applications on complex networks.

[20] P. Beaumont, "The truth about twitter, facebook and the uprisings in the arab world," in *The Guardian*, vol. 25, 2011.

[21] W. Chen, W. Lu, and N. Zhang, "Time-critical influence maximization in social networks with time-delayed diffusion process," in *Proc. of the Twenty-Sixth AAAI Conf. on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, pp. 592–598.

[22] T. Shi, J. Wan, S. Cheng, Z. Cai, Y. Li, and J. Li, "Time-bounded positive influence in social networks," in *2015 Int. Conf. on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, 2015, pp. 134–139.

[23] D. Golovin and A. Krause, "Adaptive submodularity: A new approach to active learning and stochastic optimization," *CoRR*, vol. abs/1003.3967, 2010.

[24] Y. Chen and A. Krause, "Near-optimal batch mode active learning and adaptive submodular optimization," in *Proc. of the 30th Int. Conf. on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conf. Proc., vol. 28. JMLR.org, 2013, pp. 160–168.

[25] S. Vaswani and L. V. S. Lakshmanan, "Adaptive influence maximization in social networks: Why commit when you can adapt?" *ArXiv*, vol. abs/1604.08171, 2016.

[26] F. Riquelme and P. G. Cantergiani, "Measuring user influence on twitter: A survey," *Inf. Process. Manag.*, vol. 52, pp. 949–975, 2016.

[27] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications*, vol. 164, p. 114006, 2021.

[28] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter," in *Proc. of the 20th Int. Conf. on World Wide Web*, ser. WWW '11. NY, USA: ACM, 2011, pp. 695–704.

[29] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, p. 215, 1978.

[30] X. Jin and Y. Wang, "Research on social network structure and public opinions dissemination of micro-blog based on complex network analysis," *J. Networks*, vol. 8, no. 7, pp. 1543–1550, 2013.

[31] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120.

[32] J. Zhang, R. Zhang, J. Sun, Y. Zhang, and C. Zhang, "Truetop: A sybil-resilient system for user influence measurement on twitter," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 2834–2846, 2016.

[33] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining*, ser. WSDM '11. NY, USA: ACM, 2011, pp. 45–54.

[34] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852–1872, 2018.

[35] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks," *2012 IEEE 12th Int. Conf. on Data Mining*, pp. 918–923, 2012.

[36] W. Liang, C. Shen, X. Li, R. Nishide, I. Piumarta, and H. Takada, "Influence maximization in signed social networks with opinion formation," *IEEE Access*, vol. 7, pp. 68 837–68 852, 2019.

[37] G. Xie, Y. Chen, H. Zhang, and Y. Liu, "Mbic: A novel influence propagation model for membership-based influence maximization in social networks," *IEEE Access*, vol. 7, pp. 75 696–75 707, 2019.

[38] M. Lahiri and M. Cebrian, "The genetic algorithm as a general diffusion model for social networks," in *Proc. of the Twenty-Fourth AAAI Conf. on Artificial Intelligence*, ser. AAAI'10. AAAI Press, 2010, pp. 494–499.

[39] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '07. NY, USA: ACM, 2007, pp. 420–429.

Diego G. Alonso is a Teacher Assistant at UNICEN University, Argentina. He received his PhD on Computer Science in 2020. His research interests include influence maximization algorithms, signal processing, computer vision, and natural user interfaces.

Ariel Monteserin is a researcher at ISISTAN Research Institute at CONICET-UNICEN, Argentina. He received his PhD on Computer Science in 2009. He is an Associate professor in the Computer Science Department at Univ. Nac. del Centro de la Pcia. de Bs. As (UNICEN), Tandil, Argentina. His main interests are negotiation among intelligent agents, influence maximization algorithms and smart cities.

Luis Berdun is a researcher at ISISTAN Research Institute (CONICET-UNICEN), Argentina, and a Professor at UNICEN University, Argentina. He received a Master degree in Systems Engineering in 2005, and a PhD degree in Computer Science in 2009. His research interests include intelligent aided software engineering, planning algorithms, and knowledge management.